Methods Notes

Apoorva Lal *

Tuesday 22nd October, 2024 14:14

Contents

*Originally written as notes for Spring 2020 Comprehensive Exam in the methods sequence at Stanford University. Now maintained as a stats/metrics notebook for reference. *Thanks for comments & corrections*: Kyle Butts, William Thistlethwaite, Dawen Liang

1 Brand-name Distributions

Dist Sample Space	Notation	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}[X]$	Rationale	Relations
Bernoulli $\{0,1\}$	$\text{Bern}(p)$	$(1-p)^{1-x}$	$p^x(1-p)^{1-x}$	p	$p(1-p)$	Poisson - Coin flips	$\text{Bern}(1) = \text{Bin}(1, p)$
Binomial $0 \cup \mathbb{Z}^+$	$\text{Bin}(n,p)$	$I_{1-p}(n-x, x+1)$	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$	sum of n binomials	Approx: $\mathbf{np \approx n(1-p) >> 0}$ $\text{Bin}(n,p) \approx \mathcal{N}(np, np(1-p))$ $\mathbf{np >> n(1-p)}$: $\text{Bin}(n,p) \approx \text{Poi}(np)$
Multinomial Nonnegative integer vectors sum to n	$\text{Mult}(n,p)$		$\frac{n!}{x_1! \dots x_k!} p_1^{x_1} \cdots p_k^{x_k} \quad \sum_{i=1}^k x_i = n$	$\begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$	$\begin{pmatrix} np_1(1-p_1) & -np_1p_2 \\ -np_2p_1 & \ddots \end{pmatrix}$	Multivariate Analogue of binomial	Marginals are binomial
Poisson $0 \cup \mathbb{Z}^+$	$\text{Po}(\lambda)$	$e^{-\lambda} \sum_{i=0}^{x} \frac{\lambda^i}{i!}$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ	counts to poisson process	**Normal approx** If λ large, $\text{Poi}(\lambda) = \mathcal{N}(\lambda, \lambda)$
Negative Binomial $0 \cup \mathbb{Z}^+$	$\text{NBin}(r,p)$	$I_p(r, x+1)$	$\binom{x+r-1}{r-1} p^r (1-p)^x$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$	sum of IID geom rvs	**Normal Approx** If $r(1-p)$ large $\text{NBin}(r,p) \approx \mathcal{N}\left(\frac{r(1-p)}{p}, \frac{r(1-p)}{p^2}\right)$
Uniform $S \subset \mathbb{R}$	$\text{Unif}(a,b)$	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$	$\frac{I(a<x<b)}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	Equally likely Outcomes	$\text{Beta}(1,1) = \text{Unif}(0,1)$
Normal $\mathbb{R}$	$\mathcal{N}(\mu, \sigma^2)$	$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$	$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2	Limiting Dist of Central Limit Theorem	$Z \sim \Phi \implies Z^2 \sim \text{Gamma}(1/2, 1/2) \sim \chi_1^2$
Multivariate Normal $\mathbb{R}^k$	$\text{MVN}(\mu, \Sigma)$		$(2\pi)^{-k/2} \lvert\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$	μ	Σ	mv analogue of $\mathcal{N}(\mu, \sigma^2)$	Marginals are $\mathcal{N}(\mu, \sigma^2)$
Student's t $\mathbb{R}$	$\text{Student}(\nu)$	$I_x\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	$0 \quad \nu > 1$	$\begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ \infty & 1 < \nu \le 2 \end{cases}$	Sampling dist of pivotal q $\sqrt{n}(\bar{X}_n - \mu)/\sigma_x)$	$X \sim \mathcal{N}(0,1), Y \sim \chi_\nu^2 \implies X/\sqrt{Y/\nu} \sim t(\nu)$
Chi-square $(0, \infty)$	χ_k^2	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$	k	$2k$	Sampling dist of sample variance; SSQ of IID Normal	$X \sim \mathcal{N}(0,1) \implies X^2 \sim \chi_1^2$
F $(0, \infty)$	$\text{F}(d_1, d_2)$	$I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$	$\frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\text{B}\left(\frac{d_1}{2}, \frac{d_1}{2}\right)}$	$\frac{d_2}{d_2 - 2}$	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$	Ratio of Sum of Squares	$X \sim \chi_\mu^2, Y \sim \chi_\nu^2 \implies (X/\mu)/(Y/\nu) \sim \text{F}(\mu, \nu)$
Exponential $(0, \infty)$	$\text{Exp}(\lambda)$	$1 - e^{-x\lambda}$	$\lambda e^{-x\lambda}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	wait time of poisson process	$\text{Exp}(\lambda) = \text{Gamma}(1, \lambda)$
Gamma $(0, \infty)$	$\text{Gamma}(\alpha, \lambda)$	$\frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	Sum of IID exponential rvs	**Normal Approx**: α large $\mathcal{N}(\alpha/\lambda, \alpha/\lambda^2)$
Beta $(0,1)$	$\text{Beta}(\alpha, \beta)$	$I_x(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	Ratio of gamma rvs	$\text{U}[0,1] = \text{Beta}(1,1)$
Dirichlet Unit Simplex	$\text{Dir}(\alpha)$		$\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$	$\frac{\alpha_i}{\sum_{i=1}^k \alpha_i}$	$\frac{\mathbb{E}X_i(1-\mathbb{E}X_i)}{\sum_{i=1}^k \alpha_i + 1}$	Multivariate analogue of Beta	Marginals are Beta
Weibull $[0, \infty)$	$\text{Weibull}(\lambda, k)$	$1 - e^{-(x/\lambda)^k}$	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$\lambda\Gamma\left(1 + \frac{1}{k}\right)$	$\lambda^2\Gamma\left(1 + \frac{2}{k}\right) - \mu^2$	failure rates	
Pareto	$\text{Pareto}(x_m, \alpha)$	$1 - \left(\frac{x_m}{x}\right)^\alpha \quad x \ge x_m$	$\alpha\frac{x_m^\alpha}{x^{\alpha+1}} \quad x \ge x_m$	$\frac{\alpha x_m}{\alpha - 1} \quad \alpha > 1$	$\frac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)} \quad \alpha > 2$	Long tail -income	

2 Probability and Mathematical Statistics

2.1 Basic Concepts and Distribution Theory

Defn 2.1 (Probability).
Given a measurable space $(\Omega, \mathcal{F})$, if $\mathbb{P}[\Omega] = 1$, $\mathbb{P}[]$ is called a **probability measure** and so $(\Omega, \mathcal{F}, P)$ is a probability space. Sets $f \in \mathcal{F}$ are called events, points $\omega \in \Omega$ are called outcomes, and $P(f)$ is called the probability of f.

Defn 2.2 (Kolmogorov Axioms).
The triple $(\Omega, \mathcal{S}, P)$ is a probability space if it satisfies the following

- **Unitarity**: $\mathbf{Pr}(\Omega) = 1$
- **Non Negativity**: $\forall s \in S, \mathbf{Pr}(a) \geq 0\ \mathbf{Pr}(a) \in \mathbb{R} \wedge \mathbf{Pr}(a) < \infty$
- **Countable Additivity**: If $A_1, A_2, \ldots, \in S$ are *pairwise disjoint*[*i.e.* $\forall i \neq j, A_i \cap A_j = \emptyset$], Then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

Other properties for any event A, B

- $A \subset B \implies \mathbf{Pr}(A) \leq \mathbf{Pr}(B)$
- $\mathbf{Pr}(A) \leq 1$
- $\mathbf{Pr}(A) = 1 - \mathbf{Pr}(A^c)$
- $\mathbf{Pr}(\emptyset) = 0$

Fact 2.1 (Properties of Probability).
For events $A, B \in \mathcal{E}$,

1. $0 \leq \mathbf{Pr}(A) \leq 1$: Events range from never happening to always happening
2. $\mathbf{Pr}(\mathcal{E}) = 1$: Something must happen
3. $\mathbf{Pr}(\emptyset) = 0$: Nothing never happens
4. $\mathbf{Pr}(A) + \mathbf{Pr}(A^c) = 1$: A must either happen or not happen

Defn 2.3 (Random Variable).
$X : \Omega \to \mathbb{R}$ s.t. $\forall x \in \mathbb{R}, \{\omega : X(\omega) \leq x\} \in \mathcal{F}$, where Ω is the sample space and $\mathcal{F}$ is the event space.
i.e. a RV is a mapping/function from the sample space (or per some authors, event space) to the real line.

Example 2.2 (Continuous Random Variable).

- Sample space is $\mathbb{R}$
- Event space is $\mathcal{B}(\mathbb{R})$: the Borel σ-algebra on the real line
- P_x defined so that $\forall A \in \mathcal{B}(\mathbb{R})$,

$$P_x(A) = P_\omega(\omega \in \Omega : x(\omega) \in A) =: P_\omega(x^{-1}(A))$$

Defn 2.4 (Demorgan's Laws, Conditional Probability).

- DM: $(A \cap B) = (A^C \cup B^C)^C; (A \cup B) = (A^C \cap B^C)^C$
- **Inclusion-Exclusion Rule**: $(A \cup B) = (A^C \cap B^C)^C = P(A) + P(B) - P(A \cap B)$
- **Conditional Probability**: $P(A|B) = P(A \cap B)/P(B)$

Theorem 2.3 (Bayes Rule).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Equivalently,

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_{\theta' \in \Theta} f(x|\theta')f(\theta')d\theta'} = \frac{\underbrace{f(x|\theta)}_{\text{likelihood}}\ \underbrace{f(\theta)}_{\text{prior}}}{\underbrace{f(x)}_{\text{data}}}$$

Defn 2.5 (Statistical Independence).
$A \perp\!\!\!\perp B \Leftrightarrow P(A \cap B) = P(A)P(B), P(A|B) = P(A)$

2.2 Densities and Distributions

Defn 2.6 ((Cumulative) Distribution Function).
$\mathbb{F} : \mathbb{R} \to [0, 1]$

$$\mathbb{F}(x) = \mathbf{Pr}(X \leq x) = \int_{-\infty}^{x} p(x)dx$$

Similarly, $\mathbb{F}(x-) := \mathbf{Pr}(X < x)$, so $\mathbf{Pr}(X = x) = \mathbb{F}(x) - \mathbb{F}(x-)$.
Properties of CDFs:

1. **Bounded on [0, 1]**: $\lim_{x\to\infty}\mathbb{F}(x)=1;\lim_{x\to-\infty}=0$
2. **Nondecreasing**: if $x_1<x_2$, then $\mathbb{F}(x_1)\leq\mathbb{F}(x_2)$
3. **Right Continuous**: $\lim_{h\to0+}\mathbb{F}(x+h)=\mathbb{F}(x)$
4. $\lim_{h\to0+}\mathbb{F}(x-h)=\mathbb{F}(x-)=\mathbb{F}(x)-\mathbf{Pr}\left(X=x\right)=\mathbf{Pr}\left(X<x\right)$

Suppose $\mathbb{F}'(x)$ exists $\forall x\in\mathbb{R}$ and

$$\int_{-\infty}^{\infty}\mathbb{F}'(x)dx<\infty$$

then $\mathbb{F}$ is **absolutely continuous** with density function $\mathbb{F}'=\mathsf{f}\left(\cdot\right)$
The sample analogue of the CDF is the *Empirical* CDF (ECDF). An ECDF for $X_1,\ldots,X_n\in\mathbb{R}$ is

$$\widehat{\mathbb{F}}(x)\equiv\mathbb{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\leq x}$$

for $-\infty<x<\infty$.

Defn 2.7 (Probability Density / mass Function).
A density is $\mathsf{f}:\mathbb{R}\to\mathbb{R}$ such that

$$\mathbb{F}(x)=\int_{-\infty}^{x}p(t)dt\ \ ,-\infty<x<\infty$$

which defines the density / PMF

$$\mathsf{f}\left(x\right):=\underbrace{\mathbb{F}'(x)}_{\text{Continuous Version}}\equiv\underbrace{\mathbf{Pr}\left(X=x\right)}_{\text{Discrete version}}$$

wherever $\mathbb{F}'(\cdot)$ exists.
Since $\mathbb{F}$ is nondecreasing, f is **nonnegative** and must have

$$\int_{-\infty}^{\infty}\mathsf{f}(x)dx=1$$

Fact 2.4 (Integration w.r.t. a distribution function).
Suppose X is a random-variable with distribution function $\mathbb{F}$. Then we expect that for any set $\mathcal{A}\subset\mathbb{R}$

$$\mathbf{Pr}\left(X\in\mathcal{A}\right)=\int_{\mathcal{A}}\mathsf{d}\mathbb{F}(x)$$

This is a Lebesgue-Stieltjes integral of $X(\omega)$ with respect to measure P.

If X has an absolutely continuous distribution, this integral simplifies to the familiar form of a Riemann-Stieltjes integral

$$\mathbb{F}(x)=\int_{-\infty}^{x}\mathsf{d}\mathbb{F}(t)=\int_{-\infty}^{x}f(t)\mathsf{d}t$$

and more generally

$$\int_{-\infty}^{\infty}g(x)\mathsf{d}\mathbb{F}(x)=\int_{-\infty}^{\infty}g(x)\mathsf{f}\left(x\right)\mathsf{d}x$$

Defn 2.8 (Quantile Function / Inverse-CDF).
Since real valued R.V. can be characterised by its $\mathbb{F}$ s.t. $\mathbb{F}\left(x\right):=\mathbf{Pr}\left(X\leq x\right)$, we can invert it. In other words, we can ask for the point x_p s.t. $\mathbf{Pr}\left(X\leq x_p\right)=\tau$ for any $\tau\in[0,1]$. This defines the quantile function
$\mathbb{Q}:(0,1)\to\mathbb{R}$ where

$$\mathbb{Q}_x(\tau)\equiv\mathbb{F}^{-1}(\tau):=\inf\left\{x:\mathbb{F}\left(x\right)\geq\tau\right\}$$

is called the τth quantile of F. The associated loss-function is the check function

$$\rho_\tau(u)=u(\tau-\mathbb{1}_{u\leq0})=\mathbb{1}_{u>0}\tau\left|u\right|+\mathbb{1}_{u\leq0}(1-\tau)\left|u\right|$$

If the distribution of Y is continuous, one can show that the $\tau-$th quantile of the distribution of $Y_i=:Q_\tau$ minimises the distance between Y_i and $y\in\mathbb{R}$, where the distance is defined as the check function.

Properties of quantile functions

1. $\mathbb{Q}(\mathbb{F}(x))\leq x\ ,-\infty<x<\infty$
2. $\mathbb{F}(\mathbb{Q}(t))\geq t\ ,0<t<1$
3. $\mathbb{Q}(t)\leq x\Leftrightarrow\mathbb{F}(x)\geq t$
4. If $\mathbb{F}^{-1}$ exists, then $\mathbb{Q}(t)=\mathbb{F}^{-1}(t)$
5. if $t_1<t_2,\mathbb{Q}(t_1)\leq\mathbb{Q}(t_2)$

Fact 2.5 (Equivariance of quantiles under monotone transformations).
Let $g(.)$ be a nondecreasing function. Then, for a r.v. Y,

$$Q_\tau[g(Y)]=g[Q_\tau(Y)]$$

i.e. the quantiles of $g(Y)$ coincide with transformed quantiles of Y.

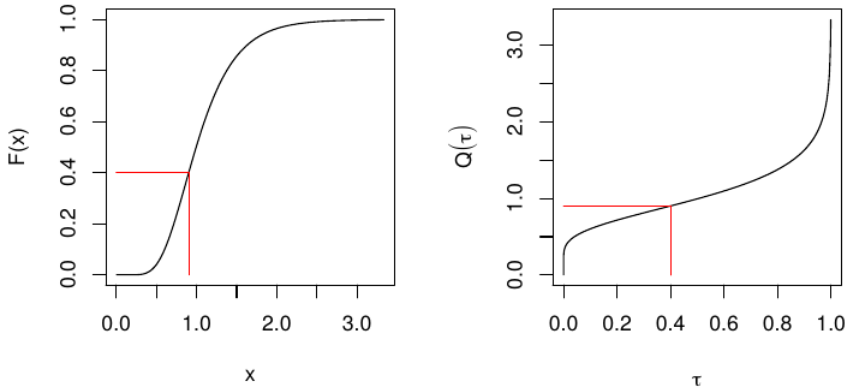


Figure 1: CDF and Quantile function. Rotate and flip CDF to get QF

Fact 2.6 (Lorenz Curve).
let Y be a positive random variable (e.g. income) with distribution function $\mathbb{F}_Y$ and mean $\mu < \infty$; then the Lorenz curve mayb be written in terms of the quantile function $Q_Y(\tau)$

$$\lambda(t) = \mu^{-1} \int_0^t Q_Y(\tau) d\tau$$

which describes the proportion of total wealth owned by the poorest t proportion of the population. Gini's mean difference ('gini coefficient') can be expressed as

$$\gamma = 1 - 2 \int_0^1 \lambda(t) dt$$

which is twice the area between the $45°$ line and the Lorenz curve.

2.2.1 Multivariate Distributions

Defn 2.9 (Random Vectors).
A $p-$random vector is a map $\mathbf{X} : \Omega \to \mathbb{R}^p$, $\mathbf{X}(\omega) := (X_1(\omega), \ldots, X_p(\omega))'$ such that each X_i is a random variable.
Joint CDF of $\mathbf{X}$ is

$$\mathbb{F}(\mathbf{x}) := \mathbb{P}[\mathbf{X} \leq \mathbf{x}] := \mathbb{P}[X_1 \leq x_1, \ldots, X_p \leq x_p]$$

If X is continuous, the joint pdf is

$$\mathsf{f}(\mathbf{x}) = \frac{\partial^p}{\partial x_1 \ldots \partial x_p} \mathbb{F}(\mathbf{X})$$

The marginals of $\mathbb{F}$ and f are

$$\mathbb{F}_{X_i}(x_i) := \mathbb{P}[X_i \leq x_i] = \mathbb{F}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)$$
$$\mathsf{f}_{X_i}(x_i) := \frac{\partial}{\partial x_i} \mathbb{F}_{X_i}(x_i) = \int_{\mathbb{R}^{p-1}} f(\mathbf{x}) d\mathbf{x}_{-i}$$

The conditional CDF and PDF of $X_1 | (X_2, \ldots, X_p)$ are defined as

$$\mathbb{F}_{X_1|\mathbf{X}_{-1}}(x_1) := \mathbb{P}[X_1 \leq x_1 | \mathbf{X}_{-1} = \mathbf{x}_{-1}]$$
$$\mathsf{f}_{X_1|\mathbf{X}_{-1}=\mathbf{x}_{-1}}(x_1) := \frac{\mathsf{f}(\mathbf{x})}{\mathsf{f}_{\mathbf{X}_{-1}}(\mathbf{x}_{-1})}$$

Defn 2.10 (Marginalization of $f(x, y)$).

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

Defn 2.11 (Conditional Distribution).

$$f(y|x) = \frac{\mathsf{f}(x, y)}{\mathsf{f}_x(x)}$$

To get marginal for x, we can integrate out y.

$$f_x(x) = \int f_{x|y}(x|y) f_y(y) dy$$

Defn 2.12 (Independent Random Variables).
two r.v.s X and Y are said to be independent if

- Joint density can be factored into marginals: $f_{X,Y}(X, Y) = f_x(X) f_y(Y)$
- $\text{Cov}[X, Y] = 0$
- $\rho(X, Y) = 0$
- $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$

2.3 Moments

For a random variable x with support $[\underline{x}, \bar{x}]$

Defn 2.13 (N-th raw moment).
$\mu_j =:= \mathbb{E}x^n$

Defn 2.14 (N-th central moment).
$\mu_j := \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^j\right]$

Defn 2.15 (Expectation and Variance).
The expectation is the Lebesgue-Stieltjes integral of r.v. $X(\omega)$ with respect to measure $\mathbb{P}$.
Common notation for expectation includes

- $\mathbb{E}\left[X\right]$
- $\mathbb{E}X$
- $\int_\Omega X(\omega) d\mathbb{P}(\omega)$
- $\int_\Omega X(\omega) d\mathbb{P}(d\omega)$
- $\int X d\mathbb{P}$

$$\mathbb{E}\left[X\right] := \int_{\underline{x}}^{\bar{x}} x \mathsf{d}\mathbb{F}(x) \underbrace{=}_{\text{If Absolute Continuity holds}} \int_{\underline{x}}^{\bar{x}} x \mathsf{f}\left(x\right) \mathsf{d}x$$

$$\begin{aligned}\mathbb{V}\left[X\right] &:= \int_{\underline{x}}^{\bar{x}} (X - \mathbb{E}(X))^2 \mathsf{d}x \\ &= \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}(X^2) - (\mathbb{E}X)^2\end{aligned}$$

Defn 2.16 (Variance Covariance Matrix).
For vector-valued RVs, this translates to

$$\begin{aligned}\mathbb{V}\left[\mathbf{X}\right] &= \mathbb{E}(\mathbf{X}\mathbf{X}') - \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{X})' \\ \text{Cov}\left[\mathbf{X}, \mathbf{Y}\right] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})'] = \mathbb{E}\left[\mathbf{X}\mathbf{Y}\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[\mathbf{Y}\right]\end{aligned}$$

Defn 2.17 (Skewness and Kurtosis).

$$\begin{aligned}\text{Skewness} \equiv \gamma &:= \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\sigma^3} \\ \text{Kurtosis} \equiv \kappa &:= \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\sigma^4} - 3\end{aligned}$$

Fact 2.7 (Linear functions of a vector valued RV).
For any (well-behaved) vector rv $\boldsymbol{x}$, $\text{Cov}\left[\mathbf{A}\boldsymbol{x} + b\right] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ where $\boldsymbol{\Sigma}$ is the covariance matrix of the random vector $\boldsymbol{x}$.
For Normal quantities where $\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$,

$$\begin{aligned}\mathbf{A}\boldsymbol{x} + \boldsymbol{y} &\sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{y}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\right) \\ \boldsymbol{\Sigma}^{-1/2} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right) \\ (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) &\sim \chi^2_n\end{aligned}$$

Theorem 2.8 (Law of the Uncoscious Statistician (LOTUS)).
let $Y = r(X)$ is a transformation of a random variable X. Then,

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[r(X)\right] = \int r(x) dF(x) = \int r(x) f(x) dx$$

Example 2.9 (Properties of combinations of RV).

- $\mathbb{E}\left[aX + bY\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[y\right]$; Expectation is a **linear operator**
  - $\mathbb{E}\left[(\sum_i a_i X_i)\right] = \sum_i a_i \mathbb{E}\left[X_i\right]$
  - $\mathbb{E}\left[(\prod_i X_i)\right] = \prod_i \mathbb{E}\left[X_i\right]$
  - **For $p-$random vector $\mathbf{X}$,** $\mathbb{E}\left[\mathbf{A}\mathbf{X} + \mathbf{b}\right] = \mathbf{A}\mathbb{E}\left[\mathbf{X}\right] + \mathbf{b}$ for $q \times p$ matrix $\mathbf{A}$ and $\mathbf{b} \in \mathbb{R}^q$
- **Variance**
  - $\mathbb{V}\left[aX\right] = a^2\mathbb{V}\left[X\right]$
  - $\mathbb{V}\left[aX + bY\right] = a^2\mathbb{V}\left[X\right] + b^2\mathbb{V}\left[Y\right] + 2ab\text{Cov}\left[X, Y\right]$
  - $\text{Cov}\left[X, X\right] = \mathbb{V}\left[X\right]$
  - $\forall a, b, c, d \in \mathbb{R},\ \text{Cov}\left[aX + c, bY + d\right] = ab\text{Cov}\left[X, Y\right]$
  - $\text{Cov}\left[X + W, Y + Z\right] = \text{Cov}\left[X, Y\right] + \text{Cov}\left[X, Z\right] + \text{Cov}\left[W, Y\right] + \text{Cov}\left[W, Z\right]$
  - **For random vector $\mathbf{X}$,** $\mathbb{V}\left[\mathbf{A}\mathbf{X} + \mathbf{b}\right] = \mathbf{A}\mathbb{V}\left[\mathbf{X}\right]\mathbf{A}^\top$ for $q \times p$ matrix $\mathbf{A}$ and $\mathbf{b} \in \mathbb{R}^q$

Defn 2.18 (Moment Generating Function).
If X is nonnegative, we know $\mathbb{E}\left[\exp(tX)\right] < \infty \forall t \leq 0$. Then, we define the **Laplace Transform**

$$\mathcal{L}(t) = \mathbb{E}\left[\exp(-tX)\right];\ t \geq t$$

Since this is limited to nonnegative RVs, one generalises them to **Moment Generating Functions**

$$M_X(t) = \mathbb{E}\left[\exp(tX)\right]$$

e.g. Standard Normal MGF $e^{t^2/2}$
For a given r.v. with MGF $M_x(t), |t| < \delta$ for some $\delta > 0$, $\mathbb{E}\left[X^n\right]$ exists and is finite $\forall n = 1, 2, \ldots$ and

$$M_x(t) = \sum_{j=0}^{\infty} t^j \frac{\mathbb{E}\left[X^j\right]}{j!}$$

and $\mathbb{E}\left[X^n\right] = M_x^{(n)}(0)$.

Moments from MGF The kth derivative of the m.g.f. evaluated at $t = 0$ is the kth (uncentered) moment of X.

$$\frac{\partial^k M_x}{\partial t^k}|_{t=0} = \mathbb{E}\left[X^k\right]$$

Defn 2.19 (Cumulant + Cumulant Generating Function).
Let X be a real-valued scalar rv and $M_x(t)$ be its moment generating function. The *cumulant-generating function* of X is defined as

$$K_X(t) = \log M_x(t), \quad |t| < \delta$$

the CGF may be expanded to the form

$$K_x(t) = \sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j \quad , |t| < \delta$$

where $\kappa_1, \kappa_2, \ldots$ are constants that depend on the distribution of X and are called **Cumulants**. Cumulants can be obtained by differentiating the CGF

$$\kappa_j = \frac{\partial^j}{\partial t^j} K_x(t)\Big|_{t=0} \ ;\ j = 1, 2, \ldots$$

Defn 2.20 (Characteristic Function).
The characteristic function of a random variable X is the function

$$\phi(t) := \mathbb{E}\left[\exp(itX)\right] \ , -\infty < t < \infty \ , i = \sqrt{-1}$$
$$\phi(t) = \mathbb{E}\left[\cos(tX)\right] + i\mathbb{E}\left[\sin(tX)\right]$$

If X has a moment generating function M, then it can be shown that $M(it) = \phi(t)$. ϕ, unlike the MGF, is always well defined, and shares properties of the MGF.

Defn 2.21 (Order Statistics).
$X_1, \ldots, X_n \sim_{iid} \mathsf{f}_x(x)$ with $\mathbb{F}_X(x)$. $X_{(k)}$ is the $k-$th order statistic (in ascending order)

$$\mathsf{f}_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} \mathbb{F}_X(x)^{k-1} \ (1 - \mathbb{F}_X(x))^{n-k} \mathsf{f}(x)$$

Defn 2.22 (Correlation Coefficient).

$$\rho[X, Y] = \frac{\mathrm{Cov}\left[X, Y\right]}{\sigma_x \sigma_y} \ \in [-1, 1] \text{ by Cauchy Schwarz}$$

- $\rho[X, Y] = 1 \Leftrightarrow \exists a, b \in \mathbb{R}$ with $b > 0$ s.t. $Y = a + bX$
- $\rho[X, Y] = -1 \Leftrightarrow \exists a, b \in \mathbb{R}$ with $b > 0$ s.t. $Y = a - bX$

Defn 2.23 (Entropy).
Cover (1999, Chap 2-4)
For a (discrete) random variable X with pmf $\mathsf{f}(x_i)$, the entropy $H(X)$ is

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log_b\left(p(x)\right)$$

where p_i is $\mathbf{Pr}\left(X = x\right) \ \forall i \in \text{Supp}\left[X\right]$. By convention, the log is taken with base 2.
Properties:

- $H(X) \geq 0$
- $H_b(X) = \log_b a H_a(X)$
- Conditioning reduces entropy: $H(X|Y) \leq H(X)$; with equality IFF $X \perp\!\!\!\perp Y$
  - generalisation : $H(X_1, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$ with equality IFF X_is are independent
- $H(X) \leq \log |\mathcal{X}|$ with equality IFF X is uniformly distributed in $\mathcal{X}$

- $H(p)$ is concave in p

Defn 2.24 (Relative Entropy / Kullback-Leibler Distance).
Relative entropy of pmf p w.r.t. pmf q

$$D(p||q) := \sum_x p(x) \log \frac{p(x)}{q(x)}$$

This is *not* a conventional metric because it is not symmetric.

Defn 2.25 (Mutual Information).

$$I(X;Y) := \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \underbrace{p(x,y)}_{\text{joint}} \log \frac{\overbrace{p(x,y)}^{\text{joint}}}{\underbrace{p(x)p(y)}_{\text{product of marginals}}}$$

$$\begin{aligned}
H(X) &= \mathbb{E}_p\left[\log \frac{1}{p(X)}\right] \\
H(X,Y) &= \mathbb{E}_p\left[\log \frac{1}{p(X,Y)}\right] \\
H(X|Y) &= \mathbb{E}_p\left[\log \frac{1}{p(X|Y)}\right] \\
I(X;Y) &= \mathbb{E}_p\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \\
D(p||q) &= \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right]
\end{aligned}$$

This is the KL divergence between the joint and product of marginals.
Properties

- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$
- **Chain rules**
  - $H(X_1, X_2, \ldots X_n) = \sum_{i=1}^n H(X_i, X_{i-1}, \ldots X_1)$
  - $I(X_1, \ldots, X_n; Y) = \sum_{i=1}^n I(X_i; Y|X_1, \ldots, X_{i-1})$

Defn 2.26 (Copula).
For a pair of r.v.s X, Y with joint distribution $\mathbb{G}(x,y)$ and marginal distribution functions $\mathbb{F}(x)$ and $\mathbb{H}(y)$, a copula function

$$\mathbb{C} : [0,1]^2 \to [0,1];\ \mathbb{C}(u,v) = \mathbb{G}(\mathbb{F}^{-1}(u), \mathbb{H}^{-1}(v))$$

where $\mathbb{C}$ has the following properties

1. $\mathbb{C}(u,0) = \mathbb{C}(0,v) = 0\ \forall u, v \in [0,1]$
2. $\mathbb{C}(u,1) = \mathbb{C}(1,v) = 1\ \forall u, v \in [0,1]$
3. $\forall\ u_1 < u_2 \wedge v_1 < v_2 \in [0,1], \mathbb{C}(u_2,v_2) - \mathbb{C}(u_2,v_1) - \mathbb{C}(u_1,v_2) + \mathbb{C}(u_1,v_1) \geq 0$

Defn 2.27 (Frechet Bounds).

$$\max\{\mathbb{F}(x), \mathbb{H}(y) - 1, 0\} \leq \mathbb{G}(x,y) \leq \min\{\mathbb{F}(x), \mathbb{H}(y)\}$$

Important in partial identification lit. The upper bound his occurs when X and Y are comonotonic, that is, when Y can be expressed as a deterministic, non-decreasing function of X. The lower bound is achieved when X and Y are countermonotonic, so Y is a deterministic, non-increasing function of X. These two very special cases correspond to the situations in which all of the mass of the copula function is concentrated on a curve connecting opposite corners of the unit square. These special cases correspond to rank correlation of $+1$ and -1 respectively. The other important special case is independent X and Y , which obviously corresponds to $C(u,v) = uv$.

2.4 Transformations of Random Variables

2.4.1 Useful Inequalities

Basic question : given a random variable X with expectation $\mathbb{E}[X]$, how likely is X to be close to its expectation, and how close is it likely to be? This implies putting bounds on quantities of the form $\mathbf{Pr}(X \geq \mathbb{E}[X] \pm t)\ t \geq 0$.

Theorem 2.10 (Cauchy Schwartz Inequality).
For random n-vectors $\mathbf{a}, \mathbf{b}$,

$$\left\|\mathbf{a}^\top \mathbf{b}\right\| \leq \|\mathbf{a}\| \|\mathbf{b}\|$$

for RVs X, Y with $\mathbb{E}\left[X^2\right] < \infty \wedge \mathbb{E}\left[Y^2\right] < \infty$

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$$

$$\text{Cov}[X,Y]^2 \leq \sigma_x \sigma_y$$

Theorem 2.11 (Jensen's Inequality).
Let Y be a random function and $g(\cdot)$ be a concave function. If $\mathbb{E}[Y]$ and $\mathbb{E}[g(Y)]$ exist, then $\mathbb{E}[g(Y)] \leq g(\mathbb{E}[Y])$.
Similarly, if $f : \mathbb{R} \to \mathbb{R}$ is convex, $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Theorem 2.12 (Markov's Inequality).

$$\mathbf{Pr}(|X| \geq t) \leq \frac{\mathbb{E}[X]}{t} \ \forall t > 0 \qquad \text{Equivalent}$$
$$\mathbf{Pr}(\psi(X) \geq t) \leq \frac{\mathbb{E}[\psi(X)]}{t}$$

Where $\psi(.)$ is a nonnegative, nondecreasing function; in the basic form $\psi = I$.
Equivalently, for $\epsilon, r \geq 1$

$$\mathbb{P}[|X| > \epsilon] \leq \frac{\mathbb{E}[|X|^r]}{\epsilon^r}$$

Theorem 2.13 (Chebychev's Inequality).
Special case of Markov's inequality.
Let X be any r.v. w $\mathbb{E}[X] = \mu < \infty$ and $\mathbb{V}[X] = \sigma^2 < \infty$. Then, $\forall \epsilon > 0$

$$\mathbf{Pr}(|X - \mu| > \epsilon) \leq \frac{\mathbb{V}[X]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

This implies that **averages of random variables with finite variance converge to their mean**.

Theorem 2.14 (Kolmogorov Inequality).
Let $X_i, i = 1, \ldots, n$ be independent random variables with $\mathbb{E}[X_i] = 0$ having finite second-order moments. Then, for $\varepsilon > 0$,

$$\mathbf{Pr}\left(\max_{1 \leq j \leq n} X_i \geq \varepsilon\right) \leq \frac{\sum_j \mathbb{E}[X_j]^2}{\varepsilon^2}$$

Theorem 2.15 (Chernoff Inequality).
$X \sim \mathcal{N}(0, 1)$ Let $g(X)$ be an absolutely continuous function of X having finite variance. Then

$$\mathbb{E}\left[[g'(X)^2]\right] \geq \mathbb{V}[g(X)]$$

equality IFF $g(X)$ is linear in X.

Theorem 2.16 (Holder's Inequality).
Let X be a r.v. with range $\mathcal{X}$ and let g_1, g_2 denote real valued functions on $\mathcal{X}$. Let $p, q > 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathbb{E}[|g_1(X)g_2(X)|] \leq \mathbb{E}[|g_1(X)|^p]^{1/p} \mathbb{E}[|g_2(X)|^q]^{1/q}$$

Theorem 2.17 (Chernoff Bounds).
Let Z be any random variable. Then, $\forall t \geq 0$,

$$\mathbf{Pr}(Z \geq \mathbb{E}[Z] + t) = \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] e^{-\lambda t} = \min_{\lambda \geq 0} M_{Z - \mathbb{E}[Z]}(\lambda) e^{-\lambda t}$$

Where M is the MGF of Z.

Theorem 2.18 (Hoeffding's Inequality).
Let X be a random variable with $a \geq X \geq b$. Then, $\forall s \in \mathbb{R}$,

$$\log \mathbb{E}\left[e^{sX}\right] \leq s\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}$$

2.5 Transformations and Conditional Distributions

$Y = g(X)$ in terms of f_x and $\mathbb{F}_x$.

Defn 2.28 (CDF of transformation).

$$\mathbb{F}_Y(y) = \mathbf{Pr}(Y \leq y) = \mathbf{Pr}(g(X) \leq y) = \mathbf{Pr}\left(X \leq g^{-1}(y)\right) = \mathbb{F}_X\left(g^{-1}(y)\right)$$

Defn 2.29 (Change of Variables technique for PDF of transformation).
Density of a transformation $y = g(X)$ of a random variable x:

$$f_Y(y) = f_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

If $X \in \mathbb{R}^n$,

$$\mathsf{f}_y(y) = \mathsf{f}_x\left(g^{-1}(y)\right) \left|\det J_{g^{-1}}(y)\right| = \mathsf{f}_x\left(g^{-1}(y)\right) \left|\det J_g(g^{-1}(y))\right|^{-1}$$

Example 2.19 (finding pdf of transformation).
Let X have $\mathsf{f}(x) = 3x^2$, and we want to find $Y = X^2$.
Then, $g^{-1}(y) = y^{1/2}$, and $\frac{\partial g^{-1}(y)}{\partial y} = (1/2)y^{-1/2}$
Plugging into the expression above, we get

$$f_Y(y) = f_x(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$
$$= 3(y^{1/2})^2 \left| \frac{1}{2} y^{-1/2} \right| = \frac{3}{2} y \times y^{-1/2} = \frac{3}{2} y^{1/2}$$

Defn 2.30 (Conditional Expectations).
For jointly continuous X, Y with joint pdf f, the *conditional expectation* of Y given $X = x$ is a *function*

$$\mathbb{E}[Y|X = x] = \int y \mathsf{d}\mathbb{F}_{Y|x}(y|X) = \int y \, \mathsf{f}_{Y|X}(y|x) \, \mathsf{d}y$$

Conditional Expectation of *function* h of random variables is

$$\mathbb{E}[h(X,Y)|X = x] = \int_{-\infty}^{\infty} h(x,y) \, f_{Y|X}(y|x) \, dy \; \forall x \in \text{Supp}[X]$$

Defn 2.31 (Conditional Variance).
For r.v.s X, Y, conditional variance $\mathbb{V}[Y|X = x]$ is

$$\mathbb{V}[Y|X] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2|X\right] = \mathbb{E}\left[Y^2|X\right] - [\mathbb{E}[Y|X]]^2$$

Theorem 2.20 (Law of Iterated Expectations/ Adam's Law).

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

Theorem 2.21 (Law of Total Variance / ANOVA Theorem / Eve's Law).

$$\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$$

2.5.1 Distributions facts and links between them

Fact 2.22 (Normal Distribution Facts).
Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$

- $\forall a, b \in \mathbb{R} \; a \neq 0, W = aX + b \implies W \sim \mathcal{N}(a\mu_x + b, a^2\sigma_X^2))$
- If $X \perp\!\!\!\perp Y$ and $Z = X + Y$, then $Z = \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

Fact 2.23 (Multivariate Normal).
$\mathcal{N}_p(\mu, \mathbf{\Sigma})$ with μ a $p-$vector of means and $\mathbf{\Sigma}$ a $p \times p$ symmetric and positive definite matrix.

PDF of $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ is

$$\phi_{\mathbf{\Sigma}} = \frac{1}{(2\pi)^{p/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

It also inherits the linearity property of the form

$$\mathbf{A}\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma}) = \mathcal{N}_q\left(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top\right)$$

where $\mathbf{A}$ is $q \times p$

Fact 2.24 (Special Cases of RVs).

- t : let $x \sim \chi_n^2, Y = z/\sqrt{x/n} \to y \sim t_n$
- F : let $x_1 \sim \chi_{n1}^2$, and $x_2 \sim \chi_{n2}^2$; $y = \frac{x_1/n_1}{x_2/n_2} \to Y \sim Fn_1, n_2$
- let $x \sim \mathcal{N}(0, I)$ and A is symmetric, then

$$p(x'Ax) \sim \chi^2(K) \text{ where } K := \text{tr}(A)$$

- Bin(1, p) $\sim$ Bern(p)
- Beta(1, 1) $\sim$ Unif(0,1)
- Gamma(1, λ) $\sim$ Expo(λ)
- $\chi_n^2 \sim$ Gamma$(n/2, 1/2)$
- NBin(1, p) $\sim$ Geom(p)

Fact 2.25 (Misc Distribution Facts).

- Exponential distribution is
  - memoryless: $\mathbf{Pr}(X > s + t|X > s) = \mathbf{Pr}(X > t)$
  - Gaps between Poisson realisations is exponential
  - Scales ($Y \sim \text{Expo}(\lambda) \implies \lambda Y = \text{Expo}(1)$)
  - Order statistics (min, max) of expos are also expo
- if $X \sim$ Gamma(a, λ), it is the distribution of the wait time for the *a realisations* of a Expo(λ) process
- **Discrete distributions:** If $X \sim$
  - *Bernoilli*, coin flip,

- *Binomial*: n coin flips,
- *Geometric*: number of tails before first head when success probability is p
- *Negative* binomial: the number of tails until rth head
- *Poisson*: if rare events occur at rate λ per unit time, the number events that occur in a unit or space of time is X
- *Multinomial*: n items that can fall into k buckets independently w.p. $\boldsymbol{p} = (p_1, \ldots, p_k)$.

Defn 2.32 (Exchangeability).
Real-valued r.v.s $X_1, \ldots, X_n$ are said to have an exchangeable distribution if the distribution of $X_1, \ldots, X_n)$ is the same as the distribution of $X_{i_1}, \ldots, X_{i_n}$ for any permutation $i_1, \ldots, i_n$ of $1, \ldots, n$.

Defn 2.33 (Martingales).
Consider a sequence of random variables $\{X_1, X_2, \ldots\}$ s.t. $\mathbb{E}[|X_n|] < \infty \; \forall \; n = 1$. The sequence $\{X_1, X_2, \ldots\}$ is said to be **martingale** if

$$\forall n, \mathbb{E}[X_{n+1}|X_1, \ldots, X_n] = X_n$$

2.6 Statistical Decision Theory

Define a *statistical decision problem* as a game involving 'nature' and 'decision maker'(DM). In the first stage (data-generation), nature selects a parameter $\theta \in \Theta$ and uses it to generate data according to the distribution $\mathbb{P}_\theta$. In the second stage (decision making), the DM observes the data but not θ, but knows the statistical model used by nature. Based on realised data, the DM would like to take an action $a \in \mathcal{A}$ whose payoff depends on the parameter drawn by nature , which can be modeled by endowing DM with a state-contingent utility $u(a, \theta)$ or loss $\mathcal{L}(a, \theta)$. The DM's **decision problem** is the selection of an action depending on the realisation of the data.

Defn 2.34 (Statistical Problem).
is a tuple

$$(\Theta, \mathcal{A}, u(.), \{P_\theta\})$$

containing a parameter space, action space, utility function, and statistical model. A *decision rule* d is a function $d : \mathcal{X} \to \mathcal{A}$.

Example 2.26 (Estimation Problem).
The action space $\mathcal{A} = \Theta$. The DM needs to decide what is the parameter θ that generated the data. The decision rule for this problem is called an **estimator**. A typical loss function is **quadratic loss** $:= \mathcal{L}(a, \theta) = (a - \theta)^2$.

Example 2.27 (Testing Problem).
Partition the parameter space into Θ_0 [Null hypothesis] and Θ_1 [Alternative hypothesis]. Action space $\mathcal{A} = \{a_0, a_1\}$ defined as choosing null or alternative. Decision rules are called a **test**. Loss function is typically **zero-one loss** $\mathcal{L}(a_1, \theta_0) = \mathcal{L}(a_0, \theta_1) = 1; 0$ otherwise.

Example 2.28 (Inference Problem).
$\mathcal{A} \subseteq \mathbb{R}$ where each action $a[]$ is an interval containing the best candidate values for θ. Decision rules here are **confidence sets**.

Defn 2.35 (Risk Function).
of decision rule d is

$$R(\theta; d) = \mathbb{E}_{P_\theta}[\mathcal{L}(d(X), \theta)]$$

A decision rule d is **dominated** by d' if $R(\theta; d') \leq R(\theta; d)$. Decision rules that are not dominated are called **admissable**.

Example 2.29 (James-Stein Estimator).
Given (possibly correlated) jointly normal r.v's $Y_1, \ldots, Y_n$ with $y_i \sim \mathcal{N}(\mu_i, 1)$, and would like to estimate the $n-$ vector $\boldsymbol{\mu}$ under squared loss

$$\mathcal{L}(\widehat{\mu}, \mu) = \sum_{i=1}^{n} (\widehat{\mu}_i - \mu)^2 = \|\widehat{\mu} - \mu\|_2$$

The MLE for each μ is just the (unbiased) vector $\boldsymbol{Y}$ itself, but the estimator

$$\widehat{\mu}_i = \left(1 - \frac{n-2}{\sum_i Y_i^2}\right) Y_i$$

has better $\mathcal{L}$ than the MLE whenever $n \geq 3$.

Defn 2.36 (Bayes Risk).
given probability distribution π on Θ (defined as a **prior**), we define the Bayes risk of a decision rule d as

$$r(\pi, d) = \int_\Theta R(\theta, d) d\pi(\theta) = \mathbb{E}_\pi[R(\theta; R)]$$

A decision rule d^* is said to be **Bayes Rule** w.r.t. prior π and class of decision rules $\mathcal{D}$ if $r(\pi, d^*) = \inf_d r(\pi, d)$ [i.e. it minimises Bayes Risk].

Defn 2.37 (Minimax).
a decision rule d_0 is said to be minimax (relative to a class $\mathcal{D}$ of decisions) if

$$\sup_{\theta \in \Theta} R(\theta, d_0) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d)$$

minimax rule protects DM against worst-case situations.

2.7 Estimation

Defn 2.38 (Sample Statistic).
For iid random variables $X_1, X_2, \ldots, X_n$, a sample statistic is a function

$$T_{(n)} = h_{(n)}(X_1, X_2, \ldots, X_n)$$

where $h_{(n)} : \mathbb{R}^n \to \mathbb{R}\ \forall n \in \mathbb{N}$

Defn 2.39 (Unbiasedness).
An estimator $\hat{\theta}$ is unbiased if $\mathbb{E}\left[\hat{\theta}\right] = \theta$

Defn 2.40 (Consistency).
An estimator $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{p} \theta$.

Defn 2.41 (Asymptotic Normality).
An estimator $\hat{\theta}$ is asymptotically normal iff

$$\sqrt{n}(\hat{\theta}(\mathbf{X}) - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\left[\hat{\theta}\right]\right)$$

where $\mathbb{V}\left[\hat{\theta}\right]^{-1}$ is called the asymptotic efficiency.

Defn 2.42 (Sampling Variance of an Estimator).
For an estimator $\hat{\theta}$, the sampling variance is $\mathbb{V}\left[\hat{\theta}\right]$.

Defn 2.43 (Mean Squared Error(MSE) of an estimator).
For an estimator $\hat{\theta}$, the MSE in estimating θ is

$$\text{MSE} = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \underbrace{\left[\mathbb{E}\left[\hat{\theta}\right] - \theta\right]^2}_{\text{Bias}} + \underbrace{\mathbb{V}\left[\hat{\theta}\right]}_{\text{Variance}}$$

$$\underset{c \in \mathbb{R}}{\operatorname{argmin}}\, \mathbb{E}\left[(X - c)^2\right] = \mathbb{E}\left[X\right]$$

Fact 2.30 (Properties of Mean and Variance).

- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are unbiased estimators of $\mathbb{E}\left[X\right]$ and $\mathbb{V}\left[X\right]$ respectively.
- Both are asymptotically normal
- If $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$
  - $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - $\frac{n-1}{\sigma^2} S_X^2 \sim \chi^2_{n-1}$
  - $\bar{X}$ and S_X^2 are independent
  - $$\frac{\bar{X} - \mu}{\sqrt{S_X^2/n}} \sim t_{n-1}$$

2.8 Hypothesis Testing

Defn 2.44 (Test statistic).
A **test statistic**, similar to an estimator, is a real valued function $S_n := T(X_1, \ldots, X_n)$ of the data sample. It is a random variable. A **test** $t : \text{Domain}(T_n) \to \{0, 1\}$.
standard normal test statistic:

$$s = \left|\frac{\hat{\theta} - \theta_0}{\omega}\right| \leq z_{(1-\alpha/2)}$$

where $\omega = \sqrt{\sigma^2/n} = \frac{1}{n-k}\sum {\hat{u}_i}^2$

- Null $H_0 : \theta \in \Theta_0$ is held as true unless data provides sufficient evidence against it. typically $\theta = 0$ ('simple' hypothesis)
- Alternative $H_1 : \theta \in \Theta_1$. Held to be true IFF null is found false.

Θ_0, Θ_1 chosen by the econometrician.
Let $S \in \mathcal{S}$ be a test statistic and its support. A decision rule is a partition of $\mathcal{S}$ in to acceptance and rejection regions such that $\mathcal{S} = \mathcal{A} \cup \mathcal{R}$.

	Null is **True**	Null is **False**
Reject	α **Type 1 error**	Power
¬ Reject	$1 - \alpha$	1 - Power **Type 2 Error**

Defn 2.45 (Power).
$\Pr(\text{reject } H_0 \mid H_1 \text{ is true}) : \pi(\theta) = P_\theta(S \in \mathcal{R})$.

Defn 2.46 (Size of Test).
is the largest probability of type-1 error.

$$\sup_{\theta\in\Theta_0}(\theta) = \sup_{\theta\in\Theta_0} P_\theta(S\in\mathcal{R})$$

Defn 2.47 (Two-sided Normal Approximation-Based Confidence interval).

$$\begin{aligned} CI_{1-\alpha}(\beta) &:= \{\hat{\beta}\in\mathbb{R}_k : \mathbf{Pr}\left(\beta\in CI_{1-\alpha}\right) = 1-\alpha\} \\ &= \left[\hat{\theta} - z_{(1-\alpha/2)}\sqrt{\widehat{V}[\hat{\theta}]},\ \hat{\theta} + z_{(1-\alpha/2)}\sqrt{\widehat{V}[\hat{\theta}]}\right] \end{aligned}$$

where z_c denotes the c^{th} quantile of the standard normal Φ s.t. $\Phi(z_c) = c\ \forall c\in(0,1)$.

Defn 2.48 (Asymptotically valid two-tailed P-value).
P-value = $2[1-\Phi(|s|)]$; in words - smallest critical value under which H_0 would be rejected

Defn 2.49 (Asymptotically valid one-tailed P-value).
One sided P-value = $1-\Phi(s)$ or $\Phi(s)$

2.9 Convergence Concepts

We estimate $\widehat{\theta}$ from data, and hope that it is close to true parameter θ_0. How close is $\widehat{\theta}$ to θ_0? Basic idea of asymptotics is to take the taylor expansions and show **asymptotic normality**: which is that the distribution of $\sqrt{n}(\widehat{\theta}-\theta_0)\to\mathcal{N}(0,1)$ as $n\to\infty$.

Defn 2.50 (Modes of Convergence: Probability, Mean Squared, Distribution).
Amemiya (1985, Chapter 3)
A sequence of r.v.s $\{X_n\}, n=1,2,\ldots$ is said to

- $X_n\to_p X$ - **converge in probability** if $\forall\epsilon,\delta>0, \exists N$ s.t. $\forall n>N, \mathbf{Pr}\left(|X_n-X|<\epsilon\right)<1-\delta$. Equivalent notation: plim $(X_n)=X$
- $X_n\to_M X$ - **converge in mean square** if $\lim_{n\to\infty}\mathbb{E}\left[X_n-X\right]^2=0$.
- $X_n\to_d X$ - **converge in distribution** if $\mathbb{F}_n$ of X_n converges to the distribution function $\mathbb{F}$ of X at every continuity point of $\mathbb{F}$. We call $\mathbb{F}$ the limit distribution of $\{X_n\}$.

Relations between convergence concepts: $M\to P\to D$

- $\mathbb{E}\left[X_n\right]^2\to 0 \implies X_n\xrightarrow{p}0$
- $X_n\xrightarrow{p}X \implies X_n\xrightarrow{d}X$
- $X_n\xrightarrow{a.s.}X \implies X_n\xrightarrow{p}X$
- $X_n\xrightarrow{d}\alpha \implies X_n\xrightarrow{p}\alpha$ (α constant)

2.9.1 Laws of Large Numbers

Basic form

$$\frac{1}{n}\sum_{i=1}^{n}(z_i-\mathbb{E}\left[z\right]_i)\xrightarrow{p}0$$

Theorem 2.31 (Chebyshev Law of Large Numbers).
$X_1,\ldots,X_n$ are IID random variables such that $\mathbb{E}\left[X_1\right]=\mu$, $\sigma^2:=\mathbb{V}\left[X_1\right]<\infty$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{p}\mathbb{E}\left[X\right]_1$$

Theorem 2.32 (Strong Law of large Numbers).
For IID $\{X_i\}$ with finite variance σ_i^2

$$\overline{X}\xrightarrow{a.s.}\mu\equiv\overline{X}-\mu\xrightarrow{a.s.}0$$

Theorem 2.33 (Glivenko-Cantelli).
Let $X_i, i=1,\ldots,n$ be an iid sequence with distribution $\mathbb{F}$ on $\mathbb{R}$. The empirical distribution function is the function of x defined by

$$\widehat{\mathbb{F}_n}(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le x}$$

for a given $x\in\mathbb{R}$, apply the SLLN to the sequence $\mathbb{1}_{X_i\le x}$ to assert

$$\widehat{\mathbb{F}}_n(x)\xrightarrow{a.s.}\mathbb{F}\left(x\right)$$

Similarly, G.C. asserts

$$\sup_{x\in\mathbb{R}}\left|\widehat{\mathbb{F}}_n(x)-\mathbb{F}\left(x\right)\right|\xrightarrow{a.s.}0$$

In words, for random samples from a continuous distribution $\mathbb{F}$, the empirical distribution $\widehat{\mathbb{F}}$ is consistent. By extension, so are the sample quantiles $\widehat{\mathbb{F}}^{-1}(\tau)$. This is important for inference in quantile regression.

2.9.2 Central Limit Theorem

Theorem 2.34 ((Lindberg-Levy) Central Limit Theorem).
$X_1, \ldots, X_n$ are IID with $\mathbb{E}[X_i] = \mu, \mathbb{V}[X_i] = \sigma^2$, for a general class of X_n,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

Equivalently,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Another way to state this is to define $Z_n := \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$, $z := \frac{y-\mu}{\sigma/\sqrt{n}}$, and $\mathbb{F}_{Z_n} := \mathbf{Pr}(Z_n \leq z)$. Then,

$$\forall z \in \mathbb{R}, |\mathbb{F}_{Z_n}(z) - \Phi(z)| \to 0 \text{ as } n \to \infty$$

Informally, **for large n, $\overline{Y}_n$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.**

Example 2.35 (CLT + Slutzky for asymptotic distribution of test statistic).
Z test: Under the null that $\mathbb{E}[x] = \theta_0$,

$$Z_n(\theta_0) := \frac{\sqrt{n}(\hat{\mu}_n - \theta_0)}{s_n} \xrightarrow{d} \mathcal{N}(0,1)$$

because $s_n^2 \xrightarrow{p} \mathbb{V}[X]$. Reject if $Z_n(\theta_0) \notin \left(z_{\alpha/2}, z_{1-\alpha/2}\right)$.

2.9.3 Tools for transformations

Theorem 2.36 (Continuous Mapping Theorem I).
$X_n \to_d X; h(.)$ is continuous. Then, $h(X_n) \xrightarrow{d} h(x)$

Theorem 2.37 (Continuous Mapping Theorem II).
$X_n \to_p X; h(.)$ is continuous. Then, $h(X_n) \xrightarrow{p} h(x)$

Theorem 2.38 (Slutsky's Theorem).
Let X_n, Y_n be sequences of scalar/vector random elements. If $X_n \xrightarrow{d} X$ and $Y \xrightarrow{p} c$, then

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n Y_n \xrightarrow{d} Xc$
- $X_n / Y_n \xrightarrow{d} X/c$ given $c \neq 0$

Theorem 2.39 (Delta Method).
Let $g : \mathbb{R}^k \to \mathbb{R}$, let θ be a point in the domain of g, and let $\{\widehat{\theta}_n\}$ be a sequence of random vectors in $\mathbb{R}^k$. If

- $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$
- g is continuously differentiable, i.e. $\nabla g(\theta)$ exists and is continuous

where $\nabla g(\theta) := \frac{\partial}{\partial \theta} g(\cdot) = \begin{bmatrix} g_n'(\theta)_1 \\ \vdots \\ g_n'(\theta)_n \end{bmatrix}$

Then we have

$$\sqrt{n}(g(\widehat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \nabla g(\theta) \Omega \nabla g(\theta)^\top\right)$$

Scalar version

$$\sqrt{n}\{g(t_n) - g(\theta)\} \xrightarrow{d} \mathcal{N}\left(0, g'(\theta)^2 \sigma^2\right)$$

Equivalently,

$$\frac{\sqrt{n}(g(\hat{\mu}_n) - g(\mu))}{|g'(\mu)| \sqrt{\mathbb{V}[X]}} \xrightarrow{d} \mathcal{N}(0,1)$$

Defn 2.51 (Orders of Magnitude).

For functions $u(x), v(x)$,

- $u(x) = \mathsf{O}(v(x)), x \to L$ denotes $|u(x)/v(x)|$ remains bounded as $x \to L$.
- $u(x) = \mathsf{o}(v(x)), x \to L$ denotes $\lim_{x \to L} u(x)/v(x) = 0$
- $u(x) \sim v(x), x \to L$ denotes $\lim_{x \to L} u(x)/v(x) = 1$

A function $f(n)$ is of **constant order** or **of order 1** if $\exists\, c > 0$ s.t. $\frac{f(n)}{c} \to 1$ as $n \to \infty$. Equivalently, if $f(n) \to c$ as $n \to \infty$. We can then write $f(n) = \mathsf{O}(1)$ (read *of the same order as*).

Defn 2.52.
Stochastic Orders of Magnitude
White (2014) definition using sequences.
For deterministic sequences $\{a_n\}, \{b_n\}$,

1. $\exists \Delta < \infty$ s.t. $a_n/b_n \to 0$ for sufficiently large n, we say $a_n = \mathsf{o}(b_n)$ [**tending to zero in probability**; a_n is smaller than b_n asymptotically]
$$a_n = \mathsf{o}(b_n) :\Leftrightarrow \lim_{n\to\infty} \frac{a_n}{b_n} = 0$$
2. $\exists \Delta < \infty$ s.t. $a_n/b_n \leq \Delta$ for sufficiently large n, we say $a_n = \mathsf{O}(b_n)$ [**bounded in probability, not larger than** b_n asymptotically; a_n **does not decrease slower than** b_n]
$$a_n = \mathsf{O}(b_n) :\Leftrightarrow \lim_{n\to\infty} \frac{a_n}{b_n} \leq C \text{ for } C > 0$$

- A sequence $\{a_n\}$ is $\mathsf{O}(n^\lambda)$ (*read - at most of order* n^λ), if $n^{-\lambda} a_n$ is bounded. $Z_n = O_p(n^\lambda) \Leftrightarrow \forall \delta, \exists \Delta(\delta) < \infty \wedge n^*(\delta)$ s.t. $\mathbf{Pr}\left(\left|\frac{Z_n}{n^\lambda}\right| > \Delta\right) < \delta \; \forall \geq n^*(\delta)$
  When $\lambda = 0, \{a_n\}$ is bounded and we write $a_n = \mathsf{O}(1)$.
  - If $w_N = O_p(1)$, $\{w_N\}$ is stochastically bounded, i.e. not explosive as $N \to \infty$. Formally, for any constant $\epsilon > 0, \exists \delta_\epsilon$ s.t.
  $$\sup_N \mathbb{P}\left[|w_N| > \delta_\epsilon\right] < \epsilon$$
  - Any random sequence converging in distribution is $O_p(1)$, which implies $N^{-1/2} \sum_{i=1}^n \{z_i - \mathbb{E}[z]\} = O_p(1)$.
  - For an estimator a_N for a parameter α, in most cases, we have $\sqrt{N}(a_N - \alpha) = O_p(1)$: a_N is $\sqrt{N}$-consistent. The convergence rate is therefore $N^{-1/2}$.
- A sequence $\{a_n\}$ is $\mathsf{o}(n^\lambda)$ (*read - of order smaller than* n^λ) if $n^{-\lambda} a_n \to 0$. $b_n = \mathsf{o}(n^\lambda) \implies b_n = \mathsf{O}(n^\lambda)$
  - In other words, when $w_N \xrightarrow{p} 0$, it is also denoted as $w_N = o_p(1)$. For $\overline{z}_N$, by LLN we thus have $\overline{z}_N - \mathbb{E}[z] = o_p(1)$.

Sums and Products

- $o_p(1) + o_p(1) = o_p(1); O_p(1) + O_p(1) = O_p(1)$
- $o_p(1) \times o_p(1) = o_p(1); O_p(1) \times O_p(1) = O_p(1);$
- $o_p(1) + O_p(1) = O_p(1); o_p(1) \times O_p(1) = o_p(1)$

Example 2.40 (Consistency of MM Estimator).
$$\hat{\beta} = \beta + \underbrace{\frac{1}{n}\sum_{i=1}^n x_i x_i'}_{\xrightarrow{p} (Q)^{-1}} \underbrace{\frac{1}{n}\sum_{i=1}^n x_i' u_i}_{\xrightarrow{p} 0} = \beta + \underbrace{O_p(1) \times o_p(1)}_{o_p(1)} \xrightarrow{p} \beta$$

2.10 Parametric Models

Defn 2.53 (Parametric Model).
For r.v Y and random vector $\mathbf{X}$ of length K, a parametric model is a set of functions: $\mathcal{P} : \mathbb{R}^{K+1} \to \mathbb{R}$ indexed by a **parameter vector** $\boldsymbol{\theta}$ of length τ
$$\mathcal{P} = \{P(y, \mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$
where $\boldsymbol{\Theta} \subset \mathbb{R}^\tau$ is called the *parameter space* [this guarantees existence as long as $\boldsymbol{\Theta}$ is a compact set]. The model is said to be *true* if $\exists \boldsymbol{\theta} \in \boldsymbol{\Theta}$ s.t.
$$f_{Y|X}(y|\boldsymbol{X}) = g(y, \boldsymbol{X}; \boldsymbol{\theta})$$
A parametrisation is **identifiable** if there is a *unique* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that corresponds with each $P \in \mathcal{P}$. Equivalently, $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \implies P(\cdot, \boldsymbol{\theta}_1) \neq P(\cdot, \boldsymbol{\theta}_2)$

Defn 2.54 (Regression Model).
Consider a parametric model where $\mathcal{Y} \subset \mathbb{R}^d$ and $\mathcal{P} := \{P(\cdot, \lambda) : \lambda \in \Lambda\}$. $Y_1, \ldots, Y_n$ are independent such that $\forall j = 1, \ldots, n$, the distribution of $Y_j \in \mathcal{P}$ corresponding with parameter λ_j (i.e. independent but not identically distributed).
$\mathbf{x}_1, \mathbf{x_2}, \ldots, \mathbf{x}_n$ is a known sequence of nonrandom vectors such that $\forall j = 1, \ldots, n$, $\exists h(\cdot)$ s.t. $\lambda_j = h(\mathbf{x}_j, \theta)$ for some $\theta \in \Theta$. Thus, the distribution of Y_j depends on the value of x_j, θ, and the function h.
In a regression, h is known/assumed, while θ is an unknown parameter.

Example 2.41 (Classical Linear Model).
the parametric setup is
$$g(y, \boldsymbol{X}, (\boldsymbol{\beta}, \sigma)) = \phi(y; (\boldsymbol{X}\boldsymbol{\beta}, \sigma^2))$$
and the parameter space $\boldsymbol{\Theta} = \left\{(\boldsymbol{\beta}, \sigma) \in \mathbb{R}^{K+2} : \sigma \geq 0\right\}$

Example 2.42.
For a binary choice model where $y \in \{0, 1\}$, the parametric model is

$$g(y, \boldsymbol{X}; \boldsymbol{\beta}) = \begin{cases} 1 - h(\mathbf{X}\boldsymbol{\beta}) & \text{if } y = 0 \\ h(\mathbf{X}\boldsymbol{\beta}) & \text{if } y = 1 \end{cases}$$

and the mean function $h : \mathbb{R} \to [0,1]$ is known. Common choices of h are logit ($h(\boldsymbol{X}\boldsymbol{\beta}) = \exp(\boldsymbol{X}\boldsymbol{\beta})/(1+\exp(\boldsymbol{X}\boldsymbol{\beta}))$) or normal CDF Φ.

Theorem 2.43 (Sufficient Statistic / Fisher-Neyman Theorem).
Let X have pdf $p(x, \theta)$. Then, the statistics $\phi(x)$ are sufficient for θ IFF the density can be written as

$$p(x|\theta) = h(x) g_\theta(\phi(x))$$

where $h(x)$ is a distribution independent of θ and g_θ captures all the dependence on θ via sufficient statistics $\phi(x)$. Equivalently, the bayesian interpretation is that $\phi(x)$ is sufficient if the posterior $p(\boldsymbol{\theta}|x) = p(\boldsymbol{\theta}|\phi(x))$.

2.11 Robustness

Write estimators as $\widehat{\theta} = \theta_n(\mathbb{F}_n)$ where $\mathbb{F}_n(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i \le y}$ is the empirical distribution fn.
In this setup,

- Mean: $\theta_n = \int y d\mathbb{F}_n(y)$
- Median: $\theta_n = \mathbb{F}_n^{-1}(\frac{1}{2})$
- Trimmed Mean =

$$\theta_n = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} \mathbb{F}_n^{-1}(u) du$$

The mapping $\theta_n(\cdot)$ induces a probability distribution for the estimator $\widehat{\theta}_n$ under $\mathbb{F}$ which we denote $L_{\mathbb{F}}(\theta_n)$. $\mathbb{F}_n \to \mathbb{F} \wedge \widehat{\theta}_n \to \theta_\infty(\mathbb{F})$

Defn 2.55 (Prokhorov Distance).
Metric on the collection of probability measures on a given metric space.
Let $\mathcal{A}$ denote the Borel sets on $\mathbb{R}$ $\forall A \in \mathcal{A}$ and $A^\varepsilon := \left\{x \in \mathbb{R} | \inf_{y \in \mathcal{A} | x-y| \le \varepsilon}\right\}$
The prokhorov distance between $\mathbb{F}$ and $\mathbb{G}$ distributions is given by

$$\pi(\mathbb{F}, \mathbb{G}) = \inf\{\varepsilon | \mathbb{F}[A] \le \mathbb{G}[A^\varepsilon] + \varepsilon \ \forall A \in \mathcal{A}\}$$

Defn 2.56 (Hampel Robustness).
The sequence of estimators $\{\theta_n\}$ is robust at $\mathbb{F}$ iff $\forall\, \varepsilon > 0$ $\exists \delta > 0$ s.t. $\forall n$,

$$\pi(\mathbb{F}, \mathbb{G}) < \delta \implies \pi(L_{\mathbb{F}}(\theta_n), L_G(\theta_n)) < \epsilon$$

This is a continuity requirement on the mapping $\theta(.)$. An estimator is robust at $\mathbb{F}$ if small departures from $\mathbb{F}$ induce small departures in the distribution of $\widehat{\theta}$ measured by the Prokhorov distance.

Defn 2.57 (Influence function).
also called the *Frechet derivative* of the functional θ_n.

$$\text{IF}_{\theta_n, \mathbb{F}}(x) := \lim_{\varepsilon \to 0} \frac{\theta_n(\mathbb{F}_\varepsilon) - \theta_n(\mathbb{F})}{\varepsilon}$$

where $\mathbb{F}_\varepsilon = (1-\varepsilon)\mathbb{F} + \varepsilon\delta_x$.

Example 2.44 (Nonrobustness of mean and variance).
Mean: $T(F) = \int x dF(x)$, $T(F_\varepsilon) = (1-\varepsilon)\mu(F) + \varepsilon\delta_x$, so $IF_{(F)}(x) = x - \mu(F)$
Variance $T(F) = \int (x-\mu)^2 dF(x)$.

$$IF_{T,F}(x) = \lim_{t \to 0} \frac{(1-\varepsilon)\sigma^2(F) t (x-\mu)^2 \sigma^2(F)}{\varepsilon} = (x-\mu)^2 - \sigma^2$$

Since IF$(x) \to \infty$ as $x \to \infty$, ε contamination noises things up enormously.

2.12 Identification

A *data generating process*(DGP) is a complete specification of the stochastic process generating the observed data. Knowledge of the DGP allows one to compute the likelihood of any realisation of the data but is conceptually distinct since it provides a description of the *structure* (/ mechanism) that gives rise to the distribution.
A **Model** $\mathcal{M}$ is a family of theoretically possible DGPs. A model can be

1. *fully parametric*: indexed by a finite number of parameters
2. *non-parametric*: indexed by an infinite dimensional parameter (ie an unknown function)
3. *semi-parametric*: indexed by a finite-dimensional vector of parameters and an infinite-dimensional nuisance function

Example 2.45 (OLS as a semiparametric model).
The simplest semiparametric model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ \mathbb{E}[\varepsilon_i | \boldsymbol{x}_i] &= 0 \\ \boldsymbol{\beta} &\in \mathbb{R}^k \end{aligned}$$

This model is deemed 'semi-parametric' because it contains a finite dimensional parameter $\boldsymbol{\beta}$ and an infinite dimensional joint distribution for $\varepsilon_i, \boldsymbol{x}_i$ left unspecified other than for the conditional mean assumption $\mathbb{E}[\varepsilon_i|\boldsymbol{x}_i] = 0$.

Example 2.46 (Index Models - Canonical semi-parametric models).

$$\begin{aligned} y_i &= g(\boldsymbol{x}_i'\boldsymbol{\beta}) + \varepsilon_i \\ \mathbb{E}[\varepsilon_i|\boldsymbol{x}_i] &= 0 \\ \boldsymbol{\beta} &\in \mathbb{R}^k \\ g(.) &: \mathbb{R} \to \mathbb{R} \text{ is monotone increasing} \end{aligned}$$

where we are interested in β. Functions $\{g(.), \mathbb{F}_{\varepsilon|X}(.), \mathbb{F}_X\}$ are nuisance

Example 2.47 (Generic Non-parametric model).

$$\begin{aligned} Y_i &= g(\boldsymbol{x}_i, \varepsilon_i) \\ \boldsymbol{x}_i &\perp \varepsilon_i \\ g(.,.) &: \mathbb{R}^2 \to [0,1] \text{ is monotone increasing in both arguments} \end{aligned}$$

Unrestricted marginals of $\varepsilon_i, \boldsymbol{x}_i$. Interested in function $g(.,.)$ such as $h(\boldsymbol{x}) = \mathbb{E}_\varepsilon[g(\boldsymbol{x}_i, \varepsilon_i)]$

Most models can be written in the form

$$\boldsymbol{Y}_i = \boldsymbol{g}(\boldsymbol{U}_i),\ \boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathbb{F}_U(\boldsymbol{u})$$

where $\boldsymbol{Y}_i$ is a vector of observables, $\boldsymbol{U}_i$ is a vector of unobserved r.v.s, with distribution function $\mathbb{F}_U(\boldsymbol{u})$ and $\boldsymbol{g}(.)$ is a vector-valued function. We call the pair of functions

$$\boldsymbol{\theta} := (\boldsymbol{g}(.), \mathbb{F}_U(\cdot))$$

the *structure*. There is a 1:1 mapping between a particular DGP and a particular choice of structure. A model space $\mathcal{M}$ can be represented in terms of a family Θ of structures.

Defn 2.58 (Identification).
While structure $\boldsymbol{\theta}$ uniquely identifies the distribution of observed variables, the reverse isn't necessarily true. *Identification* is the study of which structures are consistent with the joint distribution of observed variables. Let $\mathbb{F}_y(\boldsymbol{y})$ denote the distribution function governing the observed variables and $\mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{y})$ denote the distribution function implied by a particular structure $\boldsymbol{\theta}$. The *identified* set of structures is

$$\Omega(\mathbb{F}_Y, \Theta) = \{\boldsymbol{\theta} \in \Theta : \mathbb{F}_\theta(.) = \mathbb{F}_y(.)\}$$

The structure is *point identified* if $\Omega(\mathbb{F}_Y, \Theta)$ is a singleton.

Defn 2.59 (Observational Equivalence).
Two structures $\boldsymbol{\theta}', \boldsymbol{\theta}''$ are said to be *observationally equivalent* if

$$\mathbb{F}_{\boldsymbol{\theta}'}(\boldsymbol{y}) = \mathbb{F}_{\boldsymbol{\theta}''}(\boldsymbol{y})\ \forall \boldsymbol{y} \in \mathbb{R}^k$$

i.e. the distribution function implied by the two structures is identical. . The structure $\boldsymbol{\theta}'$ is globally point identified if there is no other $\boldsymbol{\theta}$ in the model space with which it is observationall equivalent.

Defn 2.60 (Partial Identification).
If for some $\widetilde{\boldsymbol{\theta}} \in \Theta, \Omega(\mathbb{F}_{\widetilde{\boldsymbol{\theta}}}(.), \Theta)$ is a subset of the family of Θ but not a singleton, the structure $\widetilde{\boldsymbol{\theta}}$ is said to be *partially identified* because some (but not all) competing structures have been ruled out. The identified set for a feature $\mu(\boldsymbol{\theta})$ can be written as

$$\left\{\mu(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega(\mathbb{F}_{\widetilde{\boldsymbol{\theta}}}(.), \Theta)\right\}$$

Defn 2.61 (Ceteris Paribus Effect).
Consider the model $Y_i = f(X_i, U_i; \phi)$ where (Y_i, X_i) are observed scalars, U_i is an unobserved scalar, ϕ is a parameter vector, and $f(.,.;\phi)$ is a function. The model implies a set of counterfactual values $f(x,u)$ that the outcome Y_i would take under various realisations of the random variables X_i, U_i. The *causal effect of changing* X_i from x' to x'' for individual i can be written as

$$\Delta_i(x'', x') = f(x'', U_i, \phi) - f(x', U_i, \phi)$$

If we can identify ϕ and establish that $X_i \perp\!\!\!\perp U_i$, we can identify a distribution of causal effects $\delta_i(x'', x')$.

Defn 2.62 (Statistical Functional).
In a model defined by (finite or infinite)-dimensional parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which in turn indexes the set of distributions of all observed and counterfactual quantities

$$\mathcal{P}_\Theta := \left\{P_\theta(Y, \{Y^D\}_{D \in \mathcal{D}}, \mathbf{X}) : \theta \in \Theta\right\}$$

A functional is the map $\psi(\cdot) : \mathcal{P}_\Theta \mapsto \mathbb{R}$. In causal inference, the functionals of interest are called **estimands**.

3 Linear Regression

3.1 Simple Linear Regression

3.1.1 OLS in Summation Form

Stipulate model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\mathbb{E}[\varepsilon_i|X_i] = 0, \mathbb{V}[\varepsilon_i|X_i] = \sigma^2$.

Fact 3.1 (BLP Least Squares Estimands and Estimators).
Consistent inference for $\boldsymbol{\beta}$ with only assumption being $\{y_i, \boldsymbol{x}_i\}_{i=1}^N$ is IID with well defined moments.

$$\beta_1 := \frac{\text{Cov}[X,Y]}{\mathbb{V}[X]} \implies \hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$$
$$\beta_0 := \mathbb{E}[Y] - \frac{\text{Cov}[X,Y]}{\mathbb{V}[X]}\mathbb{E}[X] \implies \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2}$$

As an application of the law of total variance, we can construct $\hat{Y} = \hat{\beta}_0 + \hat{\beta} X$ and $U = Y - \hat{Y}$. Then, the variance decomposition for $\mathbb{V}[Y]$ is

$$\mathbb{V}\left[\widehat{Y}\right] = (\hat{\beta}_1)^2 \mathbb{V}[X] = \left[\frac{\text{Cov}(X,Y)}{\mathbb{V}[X]}\right]^2 \mathbb{V}[X] = \frac{\rho_{XY}^2 \sigma_X^2 \sigma_Y^2}{\sigma_X^2} = \rho_{XY}^2 \sigma_Y^2$$
$$\mathbb{V}[U] = \sigma_Y^2 - \mathbb{V}\left[\widehat{Y}\right] = (1 - \rho_{XY}^2)\sigma_Y^2$$

Theorem 3.2 (Properties of Least Squares Estimators).
Let $\hat{\beta}^\top = (\hat{\beta}_0, \hat{\beta}_1)^T$ denote least squares estimators. The conditional means and variances are

$$\mathbb{E}\left[\hat{\beta}|X\right] = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$$
$$\mathbb{V}\left[\hat{\beta}|X\right] = \frac{\sigma^2}{n s_X^2}\begin{pmatrix}\frac{1}{n}\sum_i X_i^2 & -\bar{X} \\ -\bar{X} & 1\end{pmatrix}$$

where $s_X^2 = n^{-1}\sum_i (X_i - \bar{X})^2 =: n^{-1} S_{xx}$. This simplifies to

- $\mathbb{V}\left[\hat{\beta}_0\right] = \sigma^2(1/n + \bar{x}^2/S_{xx})$.
- $\mathbb{V}\left[\hat{\beta}_1\right] = \sigma^2/S_{xx} = \frac{\sigma^2}{n \cdot \mathbb{V}[X]}$; under Heteroskedasticity, this is $\mathbb{V}\left[\hat{\beta}_1\right] = \frac{\mathbb{V}[[x_i - \bar{x}]u_i]}{n\mathbb{V}\left[(X_i - \bar{X})\right]^2}$
- $\text{Cov}\left[\hat{\beta}_0, \hat{\beta}_1\right] = \sigma^2(-\bar{x}/S_{xx})$

With more than 1 predictor, the variance is

$$\mathbb{V}\left[\hat{\beta}_j\right] = \frac{\sigma^2}{TSS_j(1 - R_j^2)}$$

where R_j^2 is the R^2 from the regression of X_j on the other Xs and an intercept, and $TSS_j = \sum_i (x_{ij} - \bar{x}_j)^2$. This denominator is called the **Variance Inflation Factor**.

3.1.2 Prediction

Say we have a model $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 X$ from data $(X_1, Y_1, \ldots X_n, Y_n)$. We see a new observation $X = x_0$ and want to predict y_0. An estimate of the outcome $\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Variance of the prediction is

$$\mathbb{V}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right] = \mathbb{V}\left[\hat{\beta}_0\right] + x_0^2 \mathbb{V}\left[\hat{\beta}_1\right] + 2x_0 \text{Cov}\left[\hat{\beta}_0, \hat{\beta}_1\right] = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_{xx}}\right]$$

Theorem 3.3 (Prediction Interval for OLS).
Variance of prediction error $e_f := y_0 - \hat{y}_0$ is

$$\mathbb{V}[e_0] = \mathbb{V}[\varepsilon_0] + \mathbb{V}[\mathbb{E}[y_0|x_0] - \hat{y}_0] = \sigma^2 + \mathbb{V}[\hat{y}_0]$$

$$\hat{\xi}^2 = \mathbb{V}[\hat{y}_0 - y_0] = \hat{\sigma}^2\left(1 + \frac{\sum_i (x_i - x_0)^2}{n\sum_i (x_i - \bar{x})^2}\right) = \hat{\sigma^2}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{xx}}\right)$$

Example 3.4 (Simple linear regression in matrix form).
Partition design matrix s.t. $X = \begin{bmatrix}1 & x\end{bmatrix}$

$$X'X = \begin{bmatrix}1'1 & 1' \\ x'1 & x'x\end{bmatrix} \implies (X'X)^{-1} = \begin{bmatrix}S_{xx}/\Delta & -S_x/\Delta \\ -S_x/\Delta & n/\Delta\end{bmatrix}$$

where $S_{xx} = \sum x_i^2, S_x = \sum x_i, \Delta = n\sum (x_i - \bar{x})^2 = N\sum_i (x_i^2) - (\sum_i (x_i))^2$.
Equivalent expression

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n^2 \widehat{\mathbb{V}[x]}}\begin{bmatrix}n\widehat{\mathbb{V}[x]}\overline{x}^2 & -n\overline{x} \\ -n\overline{x} & n\end{bmatrix}; \mathbf{X}'\boldsymbol{y} = \begin{bmatrix}n\overline{y} \\ n\widehat{\text{Cov}}(x,y) + n\overline{x}\overline{y}\end{bmatrix}$$

3.2 Classical Linear Model

Reference: White (2014, Chap 1), Marmer lecture notes.

3.2.1 Assumptions

1. **Linearity** : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. **Strict Exogeneity** : $\mathbb{E}[\varepsilon|X] = 0$ almost surely
   - Replaced with $\mathbb{E}[\varepsilon] = 0$ *when estimating with 'fixed' instead of random Xs*
   - **A2*** Cross moment of residuals and regressors is zero, X is orthogonal to ε : $\mathbb{E}[X_i\varepsilon_i] = 0$.
3. **Spherical error variance** : $\mathbb{V}[\varepsilon_i|X] = \sigma^2; \mathbb{E}[\varepsilon\varepsilon'|X] = \sigma^2 I_n$
   - Replaced with $\mathbb{V}[\varepsilon] = \sigma^2\mathbf{I}_n$ *when estimating with 'fixed' instead of random Xs*
4. **Full Rank**: No multicollinearity - $\text{rank}(X) = k$
5. **Spherical Errors**: $\varepsilon|X \sim N(0, \sigma^2 I_n)$
   - $\varepsilon \sim N(0, \sigma^2)$ for *fixed-regressors* case.
6. $(Y_i, X_i) : i = 1, ..., n$ are i.i.d.
   - $\varepsilon_i, \ldots, \varepsilon_n$ assumed IID for *fixed-regressors* case.

- (A1-A5) define the **Classical Normal regression model**

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$$

- (A1-A4) sufficient for unbiasedness + Gauss-Markov.
- Under A1, A2, A4, $\widehat{\beta}$ is unbiased (i.e. $\mathbb{E}\left[\widehat{\beta}\right] = \beta$)

Replace A3 with $\varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, then

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

3.2.2 Optimisation Derivation

OLS minimises $\langle\varepsilon, \varepsilon\rangle$

$$\begin{aligned}&\min_{\beta\in\mathbb{R}^K}(y - X\beta)'(y - X\beta)\\ &= y'y - \underbrace{y'X\beta - \beta'X'y}_{\text{scalar, combined}} - \beta'X'X\beta\end{aligned}$$

Taking FOCs and solving yields

$$\frac{\partial\langle\varepsilon, \varepsilon\rangle}{\partial\beta} = -2\beta X'y - 2X'X\beta \implies \hat{\beta} = (X'X)^{-1}X'y$$

With fixed regressors,

$$V(\beta) = \sigma^2(X'X)^{-1}$$

where, under homoskedasticity, $\widehat{\sigma}^2 = \frac{e'e}{n-k}$, where $e = y - X\beta$

3.3 Finite and Large Sample Properties of $\widehat{\beta}, \widehat{\sigma}^2$

Property 3.5 (Finite: Unbiased).
- under **fixed regressor assumption** - that $\boldsymbol{X}$'s are nonrandom. Otherwise, the conditioning is ill-defined. Under A(1-4),

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}|X\right] &= \mathbb{E}\left[\left(X^\top X\right)^{-1}X^\top(X\beta + \varepsilon)\right]\\ &= \beta + \mathbb{E}\left[\left(X^\top X\right)^{-1}X^\top\varepsilon\right] = \beta && \text{2nd term 0 by A2}\\ \mathbb{E}\left[\widehat{\beta}\right] &= \mathbb{E}\left[\mathbb{E}\left[\widehat{\beta}|X\right]\right] && \text{by Law of iterated expectations}\end{aligned}$$

Otherwise, the expectation operator cannot pass through a ratio.
Alternate statement of bias without fixed regressors:

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\sum_{i=1}^n \boldsymbol{x}_i y_i\right]\\ &\neq \left(\sum_{i=1}^n \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\sum_{i=1}^n \mathbb{E}\left[\boldsymbol{x}_i y_i\right] = \beta\end{aligned}$$

Property 3.6 (Finite: Variance).

$$\mathbb{V}\left[\hat{\beta}|X\right] = \left(X^\top X\right)^{-1} X^\top \underbrace{\mathbb{E}\left[\varepsilon\varepsilon'|X\right]}_{\sigma^2 \text{ by A2}} X \left(X^\top X\right)^{-1} = \sigma^2 \left(X^\top X\right)^{-1}$$

Under (A1-A5),

$$\widehat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$$

Property 3.7 (Finite: Best Linear Unbiased Estimator (BLUE) - Gauss Markov Thm).
OLS coefficient efficient in the class of LUEs. For any other Linear unbiased estimator b,

$$a'\left(\mathbb{V}\left[\hat{\beta}|X\right] - \mathbb{V}\left[b|X\right]\right) a \geq 0 \; \forall \; a \in \mathbb{R}^k$$

Fact 3.8 (Large-Sample Assumptions).

- **A7** $\mathbb{E}\left[X_i X_i^\top\right]$ is finite, positive definite
- **A8** $\mathbb{E}\left[X\right]_{i,j}^4 < \infty \; \forall j = 1, \ldots, k$
- **A9** $\mathbb{E}\left[\varepsilon_i\right]^4 < \infty$
- **A10** $\mathbb{E}\left[\varepsilon_i^2 X_i X_i^\top\right]$ is positive definite

Property 3.9 (Large: Consistent - $\hat{\beta} \xrightarrow{p} \beta$).
Under A1, A2*, A6, and A7.

$$\widehat{\beta} - \beta = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i \varepsilon_i$$

$\because \mathbb{E}\left[x_i u_i\right] = 0$, we can apply Slutsky's Theorem and LLN to write

$$\text{plim}_{\, n\to\infty}(\widehat{\beta} - \beta) = \left(\text{plim}\, \frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \text{plim}\, \frac{1}{n}\sum_{i=1}^n x_i \varepsilon_i = \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1} \mathbb{E}\left[x_i \varepsilon_i\right] = 0$$

Property 3.10 (Large: Asymptotically Normal).
Under A1, A2*, A6, A7 - A10.

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}\right) \equiv \mathcal{N}\left(0, \underbrace{\left(M_{XX}\right)^{-1}}_{N^{-1}X'X} \underbrace{M_{X\Omega X}}_{N^{-1}\sum_i \hat{\varepsilon}_i x_i x_i'} \underbrace{\left(M_{XX}\right)^{-1}}_{N^{-1}X'X}\right)$$

Ω is generally unknown. We can replace it with a consistent estimator $\widetilde{\Omega}$, which is the diagonal matrix $[\widehat{\varepsilon}_{ii}]$

Defn 3.1 (Huber-White Sandwich 'Robust' SEs).
This a plug-in estimate of an asymptotic approximation of the standard error.

$$\begin{aligned}\text{Asym}\, \mathbb{V}\left[\hat{\beta}\right] &:= \frac{1}{n} \underbrace{\left(\sum_i x_i x_i'\right)^{-1}}_{\mathbf{A}^{-1}} \underbrace{\left(\sum_i \hat{\varepsilon}_i^2 x_i x_i'\right)}_{\mathbf{B}} \underbrace{\left(\sum_i x_i x_i'\right)^{-1}}_{\mathbf{A}^{-1}} \\ &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \hat{\Omega} \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\end{aligned}$$

where
$\hat{\Omega} := \text{diag}(\hat{e}_1^2, \ldots, \hat{e}_n^2)$
For the vanilla univariate linear regression, this similifies to

$$\mathbb{V}\left[\widehat{\beta}\right] = \frac{\mathbb{E}\left[\varepsilon^2 (X - \mathbb{E}\left[X\right])^2\right]}{n\mathbb{V}\left[X\right]^2}$$

which can be replaced with sample analogues and residuals to compute the robust SE.

Property 3.11 ($\hat{\sigma}^2$ is unbiased).
Since $\hat{\sigma}^2 = \frac{e'e}{n-k}$, this follows from trace of M_x. Recall that $e = M\varepsilon$, so

$$\mathbb{E}\left[e'e\right] = \mathbb{E}\left[\varepsilon' \mathbf{M}_x \varepsilon\right] = \text{tr}((\mathbf{M}_x) I \sigma^2) = (n-k)\sigma^2$$

Theorem 3.12 (Wierstrass Approximation Theorem).
Let $f : [a,b] \to \mathbb{R}$ be continuous. Then $\forall \varepsilon > 0$, $\exists p$ s.t. $\forall x \in [a,b], |f(x) - p(x)| < \varepsilon$

Defn 3.2 (Polynomial Approximation of CEF).
Let X, Y be r.v.s and suppose $\mathbb{E}\left[Y|X = x\right]$ is continuous and $\text{supp}[X] = [a,b]$. Then, $\forall \varepsilon > 0, \exists K \in \mathbb{N}$ s.t. $\forall\; K' \geq K$,

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y|X\right] - g(X, X^2, \ldots X^{K'})\right)^2\right] < \varepsilon$$

where $g(X, X^2, \ldots X^{K'})$ is the BLP of Y given $\mathbf{X}$ and higher orders.

Defn 3.3 (Polynomial least squares Sieve Estimator).
For iid r.v.s $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$, the polynomial least squares sieve estimator of the CEF is

$$\widehat{\mathbb{E}}[Y|X=x] = \sum_{k=1}^{J_n} \hat{\beta}_k x^k$$

where

$$\widehat{\beta} = \underset{\mathbf{b}\in\mathbb{R}^{J_n+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{J_n} b_k X_i^k \right)$$

where $J_n \to \infty$ and $\frac{1}{n} J_n \to 0$. Asymptotics of the estimator allow for increasing flexibility; as n grows, so does flexibility. As long as 'flexibility' grows slowly relative to n, the estimator will be consistent.

Defn 3.4 (Inference for conditional mean $\widehat{m}(\mathbf{X})$).

$$\widehat{\mathbf{m}}(\mathbf{X}) = \mathbf{X}^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$
$$\widehat{\mathbf{m}}(\mathbf{X}) = \text{MVN}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{X}^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top\right) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{P}_X)$$

3.4 Geometry of OLS

Define 2 matrices that are

- **positive semidifinite** $x'Ax \geq 0 \; \forall x \in \mathbb{R}^k$ (conformable x)
- **symmetric** $A' = A$
- **idempotent** $AA = A$

Defn 3.5 (Projection Matrices).
Matrices that project a vector y into a subspace S. For OLS,

- $L := \text{span}(\mathbf{X}) := \{\mathbf{X}b : b \in \mathbb{R}^k\}$ is the linear space spanned by the columns of $\mathbf{X}$.
- $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ - **Hat Matrix** - projector into columns space of $\mathbf{X}$
  - $\text{rank}(P) = \text{trace}(P) = p$.
  - p eigenvalues of 1 and $n-p$ zero eigenvalues
  - $0 \leq h_{ii} \leq 1$
  - Prediction for observation i is simply $M_{i\cdot}y$ where $M_{i\cdot}$ is the $i$the row of the hat matrix
- $\mathbf{M}_x = \mathbf{I}_n - \mathbf{P}_x = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ - **Annihilator Matrix** - projector into $\text{span}(X)^\perp$ (orthogonal subspace of X)
  - $\text{rank}(M) = \text{trace}(M) = \text{trace}(I-P) = \text{trace}(I) - \text{trace}(P) = n-k$

This generates

- **Fitted Value**: $\hat{y} = P_x Y$
- **Residual** : $e = M_x Y$

Theorem 3.13 (Frisch-Waugh-Lovell Theorem).
Let X_1, X_2 be partitions of X containing first K_1 and $K - K_1$ columns respectively, and β_1, β_2 be the corresponding coefficients in β. Further, let M_1, P_1 be the projection and residualiser matrices for X_1. Then,

$$\hat{\beta}_2 = \left(X_2^\top M_1 X_2\right)^{-1} X_2^\top M_1 y$$

IoW, one can estimate coefficients for X_2 by first residualising X_2 and y on X_1

3.4.1 Partitioned Regression

Choose $k_1, k_2; k_1 + k_2 = k$ s.t. $X = [X_1 \quad X_2]$. Then, the normal equation is

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}$$

yields the FWL solutions

$$\hat{\beta}_1 = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2'\mathbf{y} = \left(\underbrace{\mathbf{X}_1'\mathbf{M}_2'}\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y} = \left(\mathbf{X}^{*\prime}{}_1\mathbf{X}^*{}_1\right)^{-1}\mathbf{X}^{*\prime}{}_1\mathbf{y}$$

3.5 Relationships between Exogeneity Assumptions

main model: $y_i = \beta_0 + \beta_1 x_1 + u_i$.

1. $\mathbb{E}[u] = 0$ is technical assumption; not meeting it only affects the constant term β_0.
2. **Zero Covariance**: Key assumption for consistency of β_1: $\text{Cov}[u,x] = 0$. Assumption (1) implies that $\text{Cov}[u,x] = \mathbb{E}[ux] - \mathbb{E}[u]\mathbb{E}[x] \implies \mathbb{E}[ux] = 0$
3. **Mean independence**: $\mathbb{E}[u|x] = \mathbb{E}[u] = 0$
   - Mean independence implies zero covariance $\mathbb{E}[ux] = \mathbb{E}[x\mathbb{E}[u|x]] = \mathbb{E}[x\mathbb{E}[u]] = \mathbb{E}[x]\mathbb{E}[u]$. Since $\text{Cov}[u,x] = \mathbb{E}[ux] - \mathbb{E}[u]\mathbb{E}[x]$, $\mathbb{E}[u|x] = \mathbb{E}[u] \implies \mathbb{E}[ux] = \mathbb{E}[u]\mathbb{E}[x] \implies \text{Cov}[x,u] = 0$

- Zero covariance **does not** imply mean independence

4. $u \perp\!\!\!\perp x$ if $f(u,x) = f(u)f(x)$.

Violations of the zero conditional mean assumption $\mathbb{E}[\varepsilon|\boldsymbol{x}]$ usually arise one of three ways:

- **Omitted Variables Bias**: if unobserved variable $\boldsymbol{q}$ is correlated with both y and $\boldsymbol{x}$, failing to include it in the regression results in $\text{Cov}(\boldsymbol{x}, \varepsilon) \neq 0$
- **Measurement Error**
- **Simultaneity**

3.6 Residuals and Diagnostics

Defn 3.6 (Leverage).
Since $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$, h_{ij} can be interpreted as the weight associated with the datum (x_j, y_j). Diagonal elements h_{ii} measures how much impact y_i has on $\hat{y}_i$, and is therefore called **leverage**.

Fact 3.14 (Variance of $\hat{\varepsilon}_i$).
Since $\hat{\varepsilon} = My = (I - H)y$, $\mathbb{V}[\hat{\varepsilon}_i] = \sigma^2(1 - h_i)$.

Defn 3.7 (Standardised and Studentised Residuals).

$$e_i^{std} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}$$

Studentised residuals often omit the observation in question and estimate

$$e_i^{stu} = \frac{y_i - \hat{y}_i}{s^{(-i)}\sqrt{1 - h_{ii}}}$$

Defn 3.8 (Cook's Distance).
Let $\hat{\beta}^{(-i)}$ be the estimate of β with (x_i, y_i) omitted.
Cook's distance of x_i, y_i is defined as

$$D_i = \frac{d_i'(\mathbf{X}^\top \mathbf{X})d_i}{ks^2} \text{ where } d_i = \hat{\beta}^{(-i)} - \hat{\beta}$$

and p is the rank of $M_x = k$ and $s^2 = \hat{\sigma}^2$. $D_i > 1$ is often interpreted as an influential point.

3.7 Other Least-Squares Estimators

Defn 3.9 (Robust Regression).
When the data has some very high-leverage points, *Huber's Robust Regression* is an alternative to OLS.

$$\widetilde{\boldsymbol{\beta}} = \underset{\mathbf{b}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i'\boldsymbol{b})$$

where the $\rho(.)$ term is **Huber's Loss Function**

$$\rho(u) = \begin{cases} u^2 & |u| < c \\ 2c|u| - c^2 & |u| \geq c \end{cases}$$

which looks like square error for 'small' errors and absolute error for 'large' errors. It is continuous in c and has a continuous first derivative, which helps with optimisation. c is a tuning parameter.
Implemented in R using `MASS:rlm`.

Defn 3.10 (Weighted Least Squares).
Minimise WMSE

$$\text{WMSE}(\boldsymbol{\beta}, \omega_1, \ldots, \omega_n) = \frac{1}{n} \sum_{i=1}^{n} \omega_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$$

which yields the estimator

$$\widehat{\boldsymbol{\beta}}_{WLS} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}\mathbf{W}\mathbf{y}$$

Defn 3.11 (Generalised least squares).
If covariance matrix of errors is known: $E(\varepsilon\varepsilon'|X) = \boldsymbol{\Omega}$

$$\hat{\beta}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$$

$$\mathbb{V}(\hat{\beta}_{GLS}) = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

Defn 3.12 (Restricted OLS).
maximise

$$L(b, \lambda) = (Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) + 2\lambda(R\beta - r)$$

where R and r are the restriction matrix and vector respectively.

3.8 Measures of Goodness of Fit

- TSS = Total Sum of Squares $:= \|y\|^2$
- ESS = Explained Sum of Squares $:= \|Py\|^2$
- RSS = Residual Sum of Squares $:= \|My\|^2$

$$\begin{aligned}\langle y, y\rangle = \langle \hat{y} + e\rangle &= (X\beta + e)'(X\beta + e) \\ &= \beta' X' X\beta + e'e \qquad \text{Other terms disappear bc } x\beta \perp e \\ (y'y - n\bar{y}^2) &= \beta' X' X\beta - n\bar{y}^2 + e'e \\ TSS &= ESS + RSS\end{aligned}$$

Defn 3.13 (R^2).

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^n (Y - \bar{Y})^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n (\hat{Y} - Y)^2}{\sum_{i=1}^n (Y - \bar{Y})^2}$$

Defn 3.14 (Adjusted R^2).

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$$

Defn 3.15 (Mallow's Cp).

$$Cp = \frac{RSS + 2(k+1)\hat{\sigma}^2}{k}$$

Defn 3.16 (Akaike Information Criterion (AIC)).

$$AIC = \ln\left(\frac{e'e}{n}\right) + \frac{2k}{n}$$

Defn 3.17 (Bayesian Information Criterion (BIC)).

$$BIC = \ln\left(\frac{e'e}{n}\right) + \frac{k\ln(n)}{n}$$

Defn 3.18 (F statistic).

$$\text{F Stat} = (R\hat{\beta} - r)'(s^2 R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)/q$$

Equivalently,

$$\text{F Stat} = \frac{(\text{TSS} - \text{RSS})/(k-1)}{\text{RSS}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}$$

Defn 3.19 (Wald Statistic).

$$W_n = nh(\hat{\beta}_n)' \left(\frac{\partial h(\hat{\beta}_n)}{\partial \beta'} \hat{V}_n \frac{\partial h(\hat{\beta}_n)'}{\partial \beta}\right) nh(\hat{\beta}_n)$$

reject H_0 if $W_q > \chi^2_{q,1-\alpha} = F/q$

3.8.1 Model Selection

Defn 3.20 (Generalisation Error).
$G = \mathbb{E}\left[(Y - \widehat{m}(\mathbf{x}))^2\right]$
for a **new** data point $(\mathbf{x}, y)$.
This is different from **in-sample training error**

$$T = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{m}(\mathbf{x}_i))^2$$

Usually, $T < G$.

Defn 3.21 (Generalized / Leave-one-out Cross-Validation).
Let $\widehat{y}_i^{-(i)}$ be the predicted value when we leave out $(\mathbf{x}_i, y_i)$ from the dataset. The LOOCV is

$$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^n \left(y_i - \widehat{y}_i^{-(i)}\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \widehat{y}_i}{1 - H_{ii}}\right)^2$$

since $\text{tr}(\mathbf{H}) = p + 1$, the average of $H_{ii} = (p+1)/n =: \gamma$. Then,

$$\text{LOOCV} \approx \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \widehat{y}_i}{1 - \gamma}\right)^2 \approx \text{training error} + \frac{2\widehat{\sigma}^2}{n}(p+1)$$

For ridge, the trace is

$$\text{tr}(\mathbf{H}) = \sum_{j=1}^J \frac{\lambda_j}{\lambda_j + \lambda}$$

where λ_j is the j-th Eigenvalue of $\mathbf{\Sigma} = \mathbf{X}^\top \mathbf{X}$. For linear regression, λ is zero, and the trace is simply the sum of eigenvalues of the vcov.

3.9 Multiple Testing Corrections

T-tests perform inference on **one** hypothesis. Suppose one is interested in testing K hypotheses (e.g. $\beta_1 = 0, \ldots, \beta_k = 0$).

Fact 3.15 (Probability of rejecting any nulls when k independent true nulls are tested).

- Test size = 0.05
- Probability (No rejections) = 0.95^k
- Probability (Any (incorrect) rejections) = $1 - 0.95^k \to 1$ as k gets big

Defn 3.22 (Family Wise Error Rate (FWER) adjustments).

Define $\mathcal{M}^0 := \{i : H_i \text{ is true}\}$ and $\mathcal{R} := \{i : H_i \text{ is rejected}\}$. The FWER is defined as follows

$$\text{FWER} = \mathbf{Pr}\left(\mathcal{M}^0 \cap \mathcal{R} \neq \emptyset\right)$$

Equivalently, let N_1 be the number of type I errors (rejections of nulls when null is true). Then, **FWER** = $\mathbf{Pr}(N_1 > 0) = 1 - \mathbf{Pr}(N_1 - 0)$.
FWER $= 1-(1-\alpha)^k$ where α is the size of the test. FWER for $k = 10 = 1 - .95^{10} \approx 0.4$.
We want procedures for which FWER $\leq \alpha$.

- **Bonferroni correction** - If testing J hypotheses : Critical value $\tau = \alpha/J$
- **Holms-Bonferroni stepdown method**
  - Order k p-values from smallest to largest $p_{(1)}, \dots p_{(k)}$
  - If $p_{(1)} > \alpha/k$, stop. Fail to reject all. Else, Reject $H_{(1)}$ if $p_{(1)} < \alpha/k$. Proceed.
  - If $p_{(2)} > \alpha/(k-1)$, stop.
  - rinse and repeat until you stop rejecting because $p_{(j)} > \alpha/(k-(j-1))$ or all k have been rejected.
- Computational methods such as Romano and Wolf (2005), Westfall and Young

These methods all **modify test sizes**; instead we might want an adjusted p-value for each hypothesis.

Defn 3.23 (Joint Confidence Bands).
relies on convergence of distribution over rectangles. Suppose $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim^a \mathcal{N}(\boldsymbol{\beta}, \mathbf{V}/n)$. We want to construct a band $[\mathbf{a}, \mathbf{b}] = ([a_k, b_k])_{k=1}^K$ such that

$$\mathbf{Pr}(\boldsymbol{\beta} \in [\mathbf{a}, \mathbf{b}]) = \mathbf{Pr}(\beta_k \in [a_k, b_k] \ \forall k) \to 1 - \alpha$$

These bands take the form of $[a_k, b_k] = \left[\widehat{\beta}_k - c\sqrt{\mathbf{V}_{kk}/n}, \widehat{\beta}_k + c\sqrt{\mathbf{V}_{kk}/n}\right]$ where c is chosen such that

$$\mathbf{Pr}\left(\left\|\mathcal{N}\left(0, \mathbf{S}^{-1/2}\mathbf{V}\mathbf{S}^{-1/2}\right)\right\|_\infty \leq 0\right) = 1 - \alpha$$

where $\mathbf{S} = \text{diag}(\mathbf{V})$. This is chosen by simulation plugging in $\widehat{\mathbf{V}}$.

Defn 3.24 (False Discovery Proportion (FDP)/ Rate (FDR)).

- setup
  - data $X \sim P \in \mathcal{P}$
  - nulls $H_1, \dots, H_m \subseteq \mathcal{P}$
  - $p-$ values $p_1(X), \dots, p_m(X)$. Not independent.
- $\mathcal{H}_0 = \{i : H_i \text{ is true}\}$
- $\mathcal{R} = \{i : H_i \text{ is rejected}\}$; $R = |\mathcal{R}|$
- $V = |\mathcal{R} \cap \mathcal{H}_0|$

$$\textbf{FDP} = \frac{V}{R \vee 1} \quad \textbf{FDR} = \mathbb{E}[\text{FDP}] \leq \alpha$$

Benjamini-Hochberg Procedure

- Order $p-$ values $p_{(1)} \leq \dots \leq p_{(m)}$
- $BH(\alpha)$ rejects R hypotheses

$$R(X) = \max\left\{r : p_{(r)} \leq \frac{\alpha r}{m}\right\}$$

- **Data-dependent** rejection threshold

$$\text{Reject } H_i \Leftrightarrow p_i \leq \frac{\alpha R(X)}{m} =: \tau(\alpha; X)$$

Adjusted P-value / BH q-value $q_i(X) = \min\{\alpha : X_i \text{ is rejected by BH}(\alpha)\}$.
`p.adjust(pvals, method = "BH")`
Anderson (2008) adjustment - Rescale p values by number of hypotheses / p-value rank, and adjust for non-monotonicity.

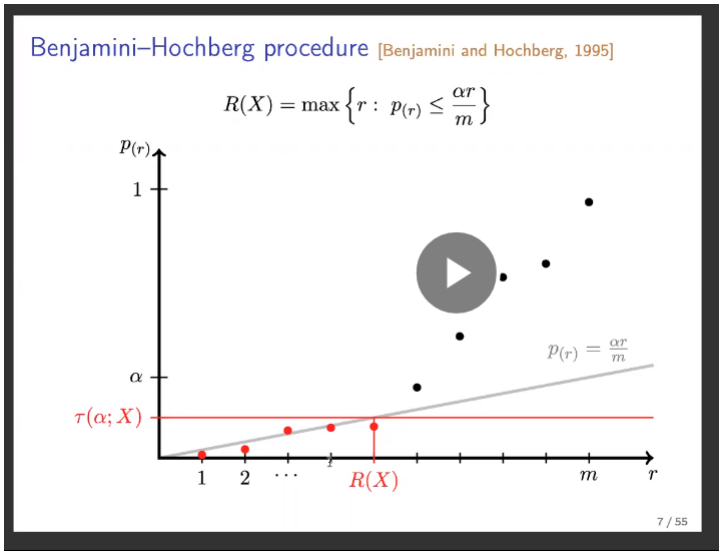


Figure 2: BH Procedure visual - data-dependent slope gray line

3.10 Quantile Regression

Notes based on Koenker (2005).

Defn 3.25 (Conditional Quantile Function).
Instead of CEF, we may be interested in the conditional quantile function at quantile τ. Define the conditional CDF of y_i

$$\mathbb{F}\left(y|\boldsymbol{x_i}\right) = \mathbf{Pr}\left(Y_i < y|\boldsymbol{x_i}\right)$$

The quantile regression model assumes that the $\tau-$th conditional quantile of y_i given $\boldsymbol{x_i}$ is a parametric function of x_i and is given by $Q_\tau(\tau|\boldsymbol{x}) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau$, where β_τ tells us the impact of x on a conditional quantile.
The conditional quantile function at quantile τ is $Q_\tau(y_i|x_i) = \mathbb{F}_y^{-1}(\tau|x_i)$.

Fact 3.16 (Relation to Heteroskedasticity).
Let $y_i = \boldsymbol{x}_i'\beta + \varepsilon_i \ ; \ \varepsilon_i|\boldsymbol{x}_i \sim \mathcal{N}\left(0, \sigma^2(\boldsymbol{x}_i)\right)$. The $\tau - th$ conditional quantile function $Q_\tau(\boldsymbol{x}_i)$ satisfies

$$\tau = \mathbf{Pr}\left(y_i \leq Q_\tau(\boldsymbol{x_i})|\boldsymbol{x}_i\right) = \mathbb{F}\left(Q_\tau(\boldsymbol{x_i})|\boldsymbol{x}_i\right)$$

Let z_τ denote the $\tau-$the quantile of the standard normal distribution. Since

$$\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma(\boldsymbol{x}_i)}|\boldsymbol{x}_i \sim \mathcal{N}\left(0,1\right) \qquad \text{we have}$$

$$\begin{aligned}\tau &= \mathbf{Pr}\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma(\boldsymbol{x}_i)} \leq z_\tau|\boldsymbol{x}_i\right) \\ &= \mathbf{Pr}\left(Y_i \leq \boldsymbol{x_i}'\boldsymbol{\beta} + \sigma(\boldsymbol{x_i})z_\tau|\boldsymbol{x_i}\right)\end{aligned}$$

This implies that the $\tau-$th conditional quantile of the distribution of y_i is given by

$$Q_\tau(\boldsymbol{x_i}) = \boldsymbol{x_i}'\boldsymbol{\beta} + \sigma(\boldsymbol{x_i})z_\tau$$

The marginal effect of $\boldsymbol{x}_i$ on the $\tau-$th quantile of y_i is therefore given by

$$\frac{\partial Q_\tau(\boldsymbol{x_i})}{\partial \boldsymbol{x}_i} = \boldsymbol{\beta} + \frac{\sigma(\boldsymbol{x}_i)}{\boldsymbol{x_i}}z_\tau$$

If the errors are homoskedastic (i.e. $\sigma(\boldsymbol{x}_i = \sigma)$, the effect of $\boldsymbol{x}_i$ is the same for all $\tau \in (0,1)$ and coincides with the effect on the conditional mean of y_i. Moreover, since $z_\tau < 0$ if $\tau < 0.5 \wedge z_\tau > 0$ if $\tau > 0.5$, the contribution of the second term $\frac{\partial\sigma(\boldsymbol{x_i})}{\partial \boldsymbol{x_i}}$ has opposite effects on the upper and lower quantiles.

Defn 3.26 (Quantile Regression Estimator $\hat{\beta}_\tau$).
$\widehat{\beta}(\tau)$ solves

$$\min_{\beta \in \mathbb{R}^p} R(\boldsymbol{\beta}) := \mathbb{E}\left[\rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\right]$$

This objective function is piecewise linear and continuous, and differentiable except at the points at which one or more residuals $y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ are zero. At these points, only Gateaux derivatives exist, see details in Koenker (2005, Chap 2-3).
The objective function (reframed as a linear program and solved numerically) is

$$Q_n(\beta_\tau) = \sum_{i:y_i \geq x_i'\beta}^{N} \tau\left|y_i - x_i'\beta_\tau\right| + \sum_{i:y_i < x_i'\beta}^{N}(1-\tau)\left|y_i - x_i'\beta_\tau\right|$$

- Since the median is robust to outliers in y, QR is a useful check relative to OLS when there are high-leverage outliers in y.
- The slope parameter $\beta(\tau)$ is interpreted as the slope of the relationship between the τth quartile of y and X.

Fact 3.17 (Asymptotic Distribution of $\hat{\beta}_q$).
$\sqrt{N}(\hat{\beta}_q - \beta_q) \xrightarrow{d} \mathcal{N}\left(0, (A)^{-1}B(A)^{-1}\right)$

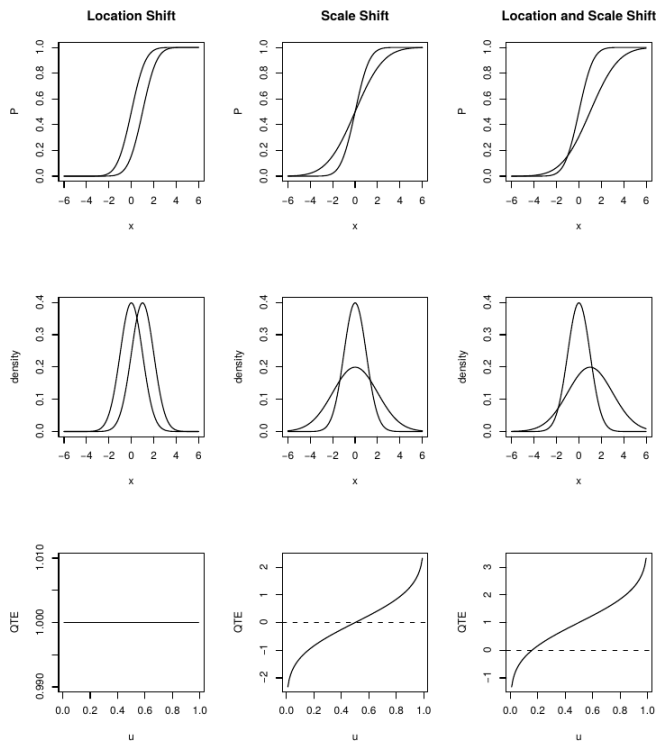


Figure 3: Examples of Treatment Effects on CDF, QF, and QTE from Koenker (2005)

where $A := \text{plim } \frac{1}{N}\sum_i f_{u_q}(0|x_i)x_i x_i'$ and $B := \text{plim } \frac{1}{N}\sum_i q(1-q)x_i x_i'$

Defn 3.27 ((Lehmann-Doksum) Quantile Treatment Effect).
Let $\mathbb{F}$ be the CDF of Y_0 [potential outcome under control] and $\mathbb{G}$ be the CDF of Y_1 [potential outcome under treatment].
Define $\Delta(x)$ as the 'horizontal distance' between $\mathbb{F}$ and $\mathbb{G}$, such that $\mathbb{F}(x) = \mathbb{G}(x + \Delta(x))$, then $\Delta(x) = \mathbb{G}^{-1}(\mathbb{F}(x)) - x$
On changing variables so that $\tau =: \mathbb{F}(x)$, we have the quantile treatment effect

$$\delta(\tau) := \Delta(\mathbb{F}^{-1}(\tau)) = \mathbb{G}^{-1}(\tau) - \mathbb{F}^{-1}(\tau)$$

In this setting, the ATE is obtained by integrating the QTE over τ

$$\overline{\delta} = \int_0^1 \delta(\tau)d\tau = \int \mathbb{G}^{-1}d\tau - \int \mathbb{F}^{-1}d\tau = \mu(\mathbb{G}) - \mu(\mathbb{F})$$

The QTE analogue estimator to difference in means is

$$\widehat{\delta}(\tau) = \widehat{\mathbb{G}}^{-1}(\tau) - \widehat{\mathbb{F}}^{-1}(\tau)$$

where $\widehat{\mathbb{F}}$ denotes an empirical distribution function. The quantile regression analogue is

$$Q_Y(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i$$

The L-D quantile treatment effect is the response necessary to keep a respondent at the same quantile under both control and treatment regimes.

3.10.1 Interpreting Quantile Regression Models

In a transformed model $Q_{h(y)}(\tau|X = x) = h(Q_Y(\tau|X = x)) = \boldsymbol{x}'\boldsymbol{\beta}(\tau)$, for monotone transforms $h(.)$, we get

$$\frac{\partial Q_Y(\tau|X = x)}{\partial x_j} = \frac{\partial h^{-1}(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_j}$$

If we specify $Q_{\log(Y)}$, then $\frac{\partial Q_Y(.)}{\partial x_j} = \exp(x'\beta)\beta_j$.
For practical purposes, suppose we observe two CDFs F and G for treated and untreated groups. Under randomisation, the two CDFs are identical by assumption (since the treatment was randomly administered), so the difference between their medians is the median treatment effect.

3.11 Measurement Error

Fact 3.18 (Classical Measurement error in the outcome does not Bias OLS).
let observed $y = y^* + u_y$ be the true value plus noise. We estimate the regression $y = x\beta + \varepsilon + u_y$. The coefficient is

$$\text{plim } (\hat{\beta}) = \frac{\text{Cov}(y,x)}{\text{Var}(x)} = \frac{\text{Cov}(x+\varepsilon+u_y,x)}{\text{Var}(x)} = \beta + \frac{\text{Cov}(\varepsilon+u_y,x)}{\text{Var}(x)} = \beta$$

The last equality holds **iff measurement error in y is orthogonal to x**. This means this rarely holds in practice.
Also means OLS estimates are more imprecise. The (homoscedastic) variance is now $(\mathbf{X}'\mathbf{X})^{-1}[\sigma_\varepsilon^2 + \sigma_{u_y}^2]$.

Defn 3.28 (Classical measurement error / Error in variables).
True data $X*$ is measured with error, X. Let $y = X^*\beta + U$, and $X = X^* + V$. Then, $y = X\beta + (u - V\beta)$. The error term is correlated with X through measurement error V, so OLS is inconsistent.

Example 3.19 (Measurement Error with a Scalar Regressor).
True regressor: x^*, variance $\sigma^2_{x^*}$ measured with $v \sim \mathcal{N}\left(0, \sigma^2_v\right)$. Then, we underestimate the true coefficient [**Attentuation Bias**]

$$\text{plim } \hat{\beta} = \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_v}\beta = \beta\left(\frac{1-s}{1+s}\right)$$

where $s = \sigma^2_v/\sigma^2_{x^*}$ is the **noise-to-signal** ratio.

Fact 3.20 (Measurement error with correlated regressors).
Adding correlated errors makes attenuation bias from measurement error **worse**.

Fact 3.21 (IVs solve measurement error problem).
$\exists Z_i$ s.t. $\text{Cov}\left[Z_i, X_i^*\right] \neq 0, \text{Cov}\left[Z_i, m_i\right] = 0$, where Z_i is the instrument, X^* is the signal, and m is the measurement error. in the bivariate regression,

$$\hat{\beta}_{IV} = \frac{\text{Cov}\left[Y, Z\right]}{\text{Cov}\left[X, Z\right]} = \frac{\text{Cov}\left[\alpha + \beta X_i^* + e_i, Z_i\right]}{\text{Cov}\left[X_i^* + m, Z_i\right]} = \frac{\beta\text{Cov}\left[X_i^*, Z_i\right]}{\text{Cov}\left[X_i^*, Z_i\right]} = \beta$$

3.12 Missing Data

Defn 3.29 (Missing Data Categories).

- **Missing at Random (MAR)** : missingness in x_i does not depend on its value but may depend on values of x_j $(j \neq i)$
- **Missing Completely at Random (MCAR)**: X_{obs} is a simple random sample of all potentially observable data values. *ignorable for likelihood inference* if this is the case.
- **Not missing at Random (NMAR)** if neither of the above applies.

3.13 Inference on functions of parameters

3.13.1 Bootstrap

Based on Cosma Shalizi's ADAEPoV and notes https://www.stat.cmu.edu/ cshalizi/402/lectures/08-bootstrap/lecture-08.pdf
Statistical quantities of interest, be they means, variances, or more complicated quantities, are functions of the underlying stochastic model (represented by the distribution function), and hence are *statistical functionals*.

The bootstrap principle Say the original data is $\mathbf{X}$. Our parameter estimate from the data is $\widehat{\boldsymbol{\theta}}$. We can simulate *surrogate datasets* called (**bootstrap replications**) by sampling from the data $\mathbf{X}$ and computing a sequence of statistics $\widetilde{t}_1 = T(\widetilde{\mathbf{X}}_1), \ldots, \widetilde{t}_b = T(\widetilde{\mathbf{X}}_M)$
For a reasonable number of replications M, we can $\widehat{\text{Var}}[\widehat{t}]$ as $\mathbb{V}\left[\widehat{t}\right]$
Debiasing: This logic can also be used for debiasing: since $\widehat{t}$ is an estimator for t_0, the sampling distribution of $\widetilde{t}$ is close to that of $\widehat{t}$, and $\widehat{t}$ itself is close to t_0,

$$\mathbb{E}\left[\widehat{t}\right] - t_0 \approx \mathbb{E}\left[\widetilde{t}\right] - \widehat{t}$$

we want to approximate the RHS using what we can calculate (LHS). This requires $\widehat{t} - t_0$ to be *approximately pivotal*.
Key idea: resampling samples from the Empirical CDF, which is consistent for the true CDF $\mathbb{F}$. Since all statistical functionals $t(\mathbb{F})$ are calculated on $\mathbb{F}$, we can get the distribution of t by computing the ECDF of $\widetilde{t}(\cdot)$.
General Bootstrap Algorithm

1. Given data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, draw a bootstrap sample of size N, denoted as $\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*$
2. compute test-statistic $\hat{t}(\theta)_n^*(\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*)$

Repeat steps (1) and (2) B independent times, obtaining B bootstrap replications of the statistic $\hat{\theta}_n$. Compute quantiles / variance of **empirical distribution of** $t(\beta)_N^{(1)}, \ldots, t(\beta)_N^{(M)}$.

Example 3.22 (Bootstrap Standard Error).

$$s^2_{\hat{\theta},Boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2 \text{ where } \bar{\hat{\theta}}^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^*$$

Defn 3.30 (Edgeworth Expansion).
the EE is the expansion of the distribution function around the normal distribution. If we have n IID random variables $X_1, \ldots, X_n$ with density f, mean μ, and variance σ^2. An edgeworth expansion for the CDF of $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ can be written

$$\mathbf{Pr}\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq \omega\right) = \Phi(\omega) + \phi(\omega)\left[\frac{-1}{6\sqrt{n}}\kappa(\omega^2-1) + R_n\right]$$

where nR_n has a boundary.

Defn 3.31 (Jackknife).
Exposition from Efron (1982).
Given an estimate $\widetilde{u} =: T(\widehat{\mathbb{F}})$ based on a sample of n draws from empirical distribution $\mathbb{F}$, define the estimate with the i-th observation left out as

$$\widetilde{u}_{-i} := T(\widehat{\mathbb{F}}_{-i})$$

And let $\widetilde{u}_{(\cdot)} := \sum_{i=1}^{n} \widetilde{u}_{-i}/n$ average these leave-out estimates.
The jackknife estimate of the standard error for $\widetilde{u}$ is the square root of

$$\sigma^2_{\text{jack}} := \frac{n}{n-1}\sum_{i=1}^{n} \frac{\left(\widetilde{u}_{-i} - \overbrace{\sum_{i=1}^{n}\widetilde{u}_{-i}/n}^{u^*}\right)^2}{n-1} = \sum_{i=1}^{n}\frac{(\widetilde{u}-u_{(\cdot)})^2}{n(n-1)}$$

Defn 3.32 (Asymptotically Pivotal Statistic).
A statistic whose limit distribution does not depend on unknown parameters is said to be **asymptotically pivotal**. Estimators are generally not asymptotically pivotal, while standard normal or chi squared test statistics typically are.

Defn 3.33 (Cluster Wild Bootstrap).
With a 'small' number of clusters, conventional clustered bootstrap errors yield over-optimistic variances. CGM Algorigthm for each resampling

- Estimate the main model imposing the null, e.g. to test the stat. significance of a single variable regress y_{ig} on all components of x_{ig} except the variable that has coefficient zero under the null. Construct $\tilde{u}_{ig} = y_{ig} - x'_{ig}\tilde{\beta}_{H0}$
- for each resampling
  - Randomly assing cluster g the weight $d_g = \begin{cases} -1 & \text{w.p. } 0.5 \\ 1 & \text{w.p. } 0.5 \end{cases}$. All observations in cluster g get the same *Rademacher* weights
  - Generage new pseudo-residuals $u^*_{ig} = d_g \times \tilde{u}_i g \implies y^*_{ig} = x'_{ig}\tilde{\beta}_{H0} + u^*_{ig}$.
  - Regress y^*_{ig} on x_{ig} [not imposing the null] and compute w^* [the t-stat with clustered SEs]
- p value for this test is the proportion of times that $|w| > |w^*_b|$ [where $|w|$ is the original sample statistic]

3.13.2 Propogation of Error / Delta Method

Propagation of error for generated quantities
Say we estimate $\widehat{\theta}$, which is a function of some intermediate quantities $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$, which are themselves estimated. Difference in group means, marginal effects in nonlinear models are examples of such generated quantities.

Since $\widehat{\theta} = f(\widehat{\phi_1}, \ldots, \widehat{\phi}_p)$, we derive standard errors for the generated quantity by writing a taylor expansion.

$$\begin{aligned}
\widehat{\theta} \approx f(\phi^*_1, \ldots \phi^*_p) &\approx f(\widehat{\phi}_1, \ldots, \widehat{\phi}_p) + \sum_{i=1}^{p}(\phi^*_i - \widehat{\phi}_i)\left.\frac{\partial f}{\partial \phi_i}\right|_{\phi=\widehat{\phi}} \\
&\approx f(\phi^*_1, \ldots, \phi^*_p) + \sum_{i=1}^{p}(\widehat{\phi}_i - \phi^*_i) f'(\widehat{\boldsymbol{\phi}}) \\
&\approx \theta^* + \sum_{i=1}^{p}(\widehat{\phi}_i - \phi^*_i) f'_i(\widehat{\phi})
\end{aligned}$$

The variance for $\widehat{\theta}$ can be written using a general analogue to $\mathbb{V}[a + bX + cy] = b^2\mathbb{V}[X] + c^2\mathbb{V}[Y] + 2bc\text{Cov}[XY]$. Allowing for covariance between any two parameters in the vector $\widehat{\phi}_i, \widehat{\phi}_j$, we can write the variance as

$$\mathbb{V}\left[\widehat{\theta}\right] \approx \sum_{i=1}^{p}\left(f'_i(\widehat{\phi})\right)^2 \mathbb{V}\left[\widehat{\phi}_i\right] + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} f'_i(\widehat{\phi}) f'_j(\widehat{\phi})\text{Cov}\left[\widehat{\phi}_i, \widehat{\phi}_j\right]$$

The second term is zero if the quantities are uncorrelated.
General Statement
Works by considering a Taylor expansion of QoI $h(x_i, \theta)$.

$$h(z) \approx h(z_0) + h'(z_0)(z - z_0) + o(\|z - z_0\|)$$

If $\tau = h(\beta)$ and $h(\theta) \neq 0$, by Slutsky's thm,

$$h(\boldsymbol{\beta}^*) \xrightarrow{d} \text{MVN}\left(h(\boldsymbol{\beta}), \nabla h(\boldsymbol{\beta}^*)' \boldsymbol{\Sigma} \nabla h(\boldsymbol{\beta}^*)\right)$$

where the gradient is evaluated at MLE estimates and $\boldsymbol{\Sigma}$ is the covariance matrix of the MLE.

$$\nabla h(\boldsymbol{\beta}^*) = \left(\left.\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_1}\right|_{\boldsymbol{\beta}^*}, \left.\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_2}\right|_{\boldsymbol{\beta}^*}, \ldots, \left.\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_K}\right|_{\boldsymbol{\beta}^*}\right)$$

For the scalar case,

$$\frac{\hat{\tau}_n - \tau}{\widehat{se}(\hat{\tau})} \xrightarrow{d} \mathcal{N}(0, 1)$$

where

$$\widehat{se}(\hat{\tau}_n) = \left|g'(\hat{\theta})\widehat{se}(\hat{\theta}_n)\right|$$

3.13.3 Parametric Bootstrap

Assuming the point estimates and variances of the parameters are correct, draw from distribution of parameters and construct quantity of interest (e.g. a marginal effect) for each draw. For m of M simulations,

$$\begin{aligned}\boldsymbol{\beta}^m &\sim \text{MVN}\left(\boldsymbol{\beta}^*, I_N(\boldsymbol{\beta})^{-1}\right)\\ h(\boldsymbol{\beta})^m &= h\left(\boldsymbol{\beta}^m\right)\end{aligned}$$

Then, average them or take their quantiles

$$f\left[h\left(\boldsymbol{\beta}^*\right)\right] = \sum_{m=1}^{M} \frac{f\left(h(\boldsymbol{\beta})^m\right)}{M}$$

3.14 Generalised Method of Moments

Data is $\mathcal{D} := (Y_i, \mathbf{X}_i, \mathbf{Z}_i)_{i=1}^n$ where $Y_i \in \mathbb{R}$, $\mathbf{X}_i$ is a k-vector of regressors, and $\mathbf{Z}_i$ is a $l-$vector of 'instruments'. Need $l \geq k$

Defn 3.34 (Linear GMM).
Model given by $Y_i = \mathbf{X}_i'\boldsymbol{\beta} + U_i$ and $\mathbb{E}\left[\mathbf{Z}_i U_i\right] = 0$
When $k = l$, we can solve for $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}}_n = \left(\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{Z}_i Y_i$$

Let $\mathbf{W}_n$ be a (possibly random) $l \times l$ weight matrix such that $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$. For a given choice of a weight matrix $\mathbf{W}_n$, the GMM estimator solves

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n(\mathbf{W}_n) &= \underset{\mathbf{b}}{\text{argmin}} \left\| \mathbf{W}_n \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i (Y_i - \mathbf{X}_i'\boldsymbol{\beta}) \right\|_2 \\ &= \underset{\mathbf{b}}{\text{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i (Y_i - \mathbf{X}_i'\boldsymbol{\beta}) \right)' \mathbf{W}_n' \mathbf{W}_n \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i (Y_i - \mathbf{X}_i'\boldsymbol{\beta}) \right)\end{aligned}$$

Different choices of $\mathbf{W}$ produce different estimators. Choice of weight matrix only matters in *over-identified case* ($l > k$).

Defn 3.35 (GMM Estimator (general formulation)).
Assume existence (from theoretical model) of r moment conditions for q parameters

$$\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta}_0)\right] = \mathbf{0}$$

where θ_0 is a $q \times 1$ vector, $g(\cdot)$ is a $r \times 1$ vector function with $r \geq q$, and θ_0 denotes the value of θ in the DGP. w includes all observables.
Sample Analogue:

$$\mathbb{E}\left[g(w_i, \theta)\right] \approx \frac{1}{n} \sum_{i=1}^{n} g(w_i, \theta) =: g_N(\theta) \ \in \mathbb{R}^q$$

Define Jacobian

$$\mathbf{D}(\theta) := \mathbb{E}\left[\frac{\partial g(w_i, \theta)}{\partial \theta'}\right], \ q \times k$$

problem is **under-identified** if $\text{rank}(\mathbf{D}) < k$, **just identified** if $\text{rank}(\mathbf{D}) = k$, and **over-identified** when $\text{rank}(\mathbf{D}) > k$.
Evaluated at the maximum,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{n} g(w_i, \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{S})$$

where

$$\mathbf{S} = \mathbb{E}\left[g(w_i, \theta_0) g(w_i, \theta)^\top\right] \ ; q \times q$$

Can insert a positive definite $q \times q$ weighting matrix $\mathbf{W}$ that tells us how much to penalise violations of one moment condition relative to another.
So, GMM estimator $\hat{\theta}_{\text{GMM}}$ minimises a quadratic form

$$\widehat{\theta} = \underset{\theta}{\text{argmin}}\, Q_N(\theta) := g_N(\theta)^\top \underbrace{\mathbf{W}_N}_{q \times q,\ \text{PSD}} g_N(\theta)$$

in sample, this means

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} g_i(\theta) \right)^\top \mathbf{W} \left(\frac{1}{n} \sum_{i=1}^{n} g_i(\theta) \right)$$

Technical conditions for GMM (Newey and McFadden, 1994)

1. $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (parameter space is compact)
2. $\mathbb{E}\left[g(\mathbf{z}_i, \boldsymbol{\theta}_0)\right] = 0$ and $\mathbb{E}\left[g(\mathbf{z}_i, \boldsymbol{\theta})\right] \neq 0 \ \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ (global identification condition)
3. $g_n(\boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}\left[g(\mathbf{z}_i, \boldsymbol{\theta})\right]$
4. $\mathbb{E}\left[g(\mathbf{z}_i, \boldsymbol{\theta})\right]$ is continous
5. $\boldsymbol{\Delta} := \mathbf{D}(\boldsymbol{\theta}_0)\ \mathbf{W}\ \mathbf{D}(\boldsymbol{\theta}_0)^\top$ is invertible

6. $g(\mathbf{z}_i, \boldsymbol{\theta})$ has at least two moments finite and finite derivatives at all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
7. $g_n(\boldsymbol{\theta})$ is twice-continuously differentiable about $\boldsymbol{\theta}_0$
8. $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ (weight matrix goes to constant)
9. $\boldsymbol{\theta_0}$ is not on the boundary of $\boldsymbol{\Theta}$

Asymptotic Distribution

$$\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\theta)$$

Where

$$\mathbf{V}_\theta = (\mathbf{DWD})^{-1}(\mathbf{DWSW'D'})(\mathbf{DWD})^{-1}$$

with $\mathbf{D} := \mathbb{E}\left[\frac{\partial}{\partial \theta'} g_i(\theta)\right]$ and $\mathbf{S} := \mathbb{E}\left[g_i(\theta) g_i(\theta)^\top\right]$ for general weight matrices $\mathbf{W}$.

Defn 3.36 ((2-step) Efficient GMM).
The weight matrix $\mathbf{W}_N$ that minimises the variance of $\widehat{\theta}_{\text{GMM}}$ is $\mathbf{W}_N = \mathbf{S}^{-1}$, which produces

$$\mathbf{V}_\theta = \left(\mathbf{DS}^{-1}\mathbf{D}'\right)^{-1}$$

Want S to be small (sampling variation / noise of the moments) to be small and D to be large (objective function steep around θ_0).
Problem is $\mathbf{S} = \mathbb{E}\left[g(w_i, \theta_0) g(w_i, \theta_0)'\right]$ is unobserved. So use sample analogue

$$\widehat{\mathbf{W}} = \widehat{\mathbf{S}}^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(g(w_i, \widehat{\theta}) - g_N(\widehat{\theta})\right)\left(g(w_i, \widehat{\theta}) - g_N(\widehat{\theta})'\right)\right)^{-1}$$

Steps

1. Pick some initial guess $\mathbf{W}_0 = \mathbf{I}_q$
2. Solve $\widehat{\theta} = \operatorname{argmin}_\theta g_N(\theta)' \mathbf{W}_0 g_N(\theta)$
3. Update $\widehat{\mathbf{W}} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(g(w_i, \widehat{\theta}) - g_N(\widehat{\theta})\right)\left(g(w_i, \widehat{\theta}) - g_N(\widehat{\theta})'\right)\right)^{-1}$
4. Solve $\widehat{\theta}_{\text{GMM}} = \operatorname{argmin}_\theta g_N(\theta)' \widehat{\mathbf{W}} g_N(\theta)$
5. Compute $\mathbf{D}(\widehat{\theta}_{\text{GMM}})$ and $\mathbf{S}(\widehat{\theta}_{\text{GMM}})$

Example 3.23 (Standard Methods nested in GMM).
Most standard methods can be re-expressed as GMM.

- **OLS**: $y_i = x_i'\beta + \epsilon_i$. Exogeneity implies $\mathbb{E}\left[x_i'\epsilon_i\right] = 0$. Can write in terms of observables and parameters as $\mathbb{E}\left[x_i'(y_i - x_i'\beta)\right] = 0$. Yields moment condition $g(y_i, x_i, \beta) = x_i'(y_i - x_i'\beta)$
- **2SLS**: x_i is endogenous, so $\mathbb{E}\left[x_i'\epsilon_i\right] \neq 0$. However, $\exists z_i$ such that $\mathbb{E}\left[z_i'\epsilon_i\right] = 0$, so moment condition for exclusion restriction is $g(y_i, x_i, z_i, \beta) = z_i'(y_i - x_i'\beta)$
- **Maximum Likelihood**: g is simply the score function, so $g(w_i, \theta) = \frac{\partial \log f(w_i, \theta)}{\partial \theta}$

3.14.1 Empirical Likelihood

Notes based on Owen (2001) and Anatolyev and Gospodinov (2011).

Defn 3.37 (Nonparametric Likelihood).
Given $X_1, \ldots, X_n \in \mathbb{R}$ assumed IID with common CDF $\mathbb{F}_0$, the *nonparametric likelihood* of the CDF $\mathbb{F}$ is

$$\mathcal{L}(\mathbb{F}) = \prod_{i=1}^{N}(\mathbb{F}(X_i) - \mathbb{F}(X_i-)) = \prod_{i=1}^{N}\mathbf{Pr}\left(X_i = x\right)$$

The value $\mathcal{L}(\mathbb{F})$ is the probability of getting *exactly* the observed sample values $X_1, \ldots, X_n$ from the CDF $\mathbb{F}$.
$\mathcal{L}(\mathbb{F}) = 0$ for continuous $\mathbb{F}$; for positive nplikelihood, a distribution $\mathbb{F}$ must place positive probability on every one of the observed data values.

Theorem 3.24 (The Empirical CDF (ECDF) maximises the Nonparametric Likelihood).
where

$$\widehat{\mathbb{F}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

where δ_x is a point mass as x.

NPMLE as constrained optimisation Consider a random sample $\{x_i\}_{i=1}^n$ from a population distribution $\mathbb{F}(x)$ with density $\mathsf{f}(x)$. The joint likelihood is given by

$$\prod_{i=1}\mathsf{f}(x_i)$$

instead of assuming a particular parametric form for $f(x|\boldsymbol{\eta})$, we define $\boldsymbol{\eta} := \mathbf{p} = (p_1, \ldots, p_n)$ where $p_i = f(f(x_i)$ denotes a sequence of discrete probability weights assigned to each sample observation observation. The **Nonparametric Maximum Likelihood** estimate of $\mathbf{p}$ solves

$$\max_{\mathbf{p}} \frac{1}{n}\sum_{i=1}^{n}\log(p_i)$$

subject to the constraint
$\sum_{i=1}^{n} p_i = 1$.
The lagrangian for this problem is

$$\mathcal{L}(p_1,\ldots,p_n,\mu) = \frac{1}{n}\sum_{i=1}^{n}\log(p_i) - \mu\left(\sum_{i=1}^{n} p_i - 1\right)$$

which yields the solution $\widehat{p}_i = \frac{1}{n}\ \ , i = 1,\ldots,n$

Defn 3.38 (Empirical Likelihood).
Suppose now we have a model in the form of a system of unconditional moment restrictions

$$\mathbb{E}\left[m(\mathbf{w},\boldsymbol{\theta}_0)\right] = 0$$

where $\boldsymbol{\theta}_0$ is $k\times 1$, $\mathbf{w}$ is a vector of observables, and $m(\mathbf{w},\boldsymbol{\theta})$ is an l vector of moment conditions. This amends the above constrained optimisation problem to

$$\max_{\mathbf{p}} \frac{1}{n}\sum_{i=1}^{n}\log(p_i)$$

subject to the constraints

$$\sum_{i=1}^{n} p_i m(\mathbf{w}_i,\boldsymbol{\theta}) = 0$$

$\sum_{i=1}^{n} p_i = 1$.
The lagrangian for this problem is

$$\mathcal{L}(p_1,\ldots,p_n,\mu) = \frac{1}{n}\sum_{i=1}^{n}\log(p_i) - \boldsymbol{\lambda}^\top\sum_{i=1}^{n} p_i m(\mathbf{w}_i,\boldsymbol{\theta}) - \mu\left(\sum_{i=1}^{n} p_i - 1\right)$$

which, upon tedious rearrangement, yields the system of equations that implicitly define the solutions

$$\frac{1}{n}\sum_{i=1}^{n}\frac{m(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})}{1+\widehat{\boldsymbol{\lambda}}^\top m(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})} = 0$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})'/\partial\boldsymbol{\theta}}{1+\widehat{\boldsymbol{\lambda}}^\top m(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})}\widehat{\boldsymbol{\lambda}} = 0$$

where the solution $\widehat{\boldsymbol{\theta}}$ is called the *empirical likelihood* (*EL*) *estimator* and $\widehat{\boldsymbol{\lambda}}$ is a vector of EL multipliers. The dual of the above problem solve the *EL saddlepoint problem*

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\min_{\boldsymbol{\lambda}}\frac{1}{n}\sum_{i=1}^{n}\left(-\log(1+\boldsymbol{\lambda}^\top m(\mathbf{w}_i,\boldsymbol{\theta}))\right)$$

Generalised Empirical Likelihood Replacing the log above with an arbitrary shape-constrained criterion function $\rho(v)$ ($\rho(0)=0, \rho'(0)=\rho''(0)=-1$) yields the *generalised* empirical likelihood estimator which solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}_n}\frac{1}{n}\sum_{i=1}^{n}\rho(\boldsymbol{\lambda}^\top m(\mathbf{w}_i,\boldsymbol{\theta}))$$

where $\boldsymbol{\Lambda}_n = \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^\top m(\mathbf{w}_i,\boldsymbol{\theta}) \in \Upsilon\right\}$ where Υ is an open set containing zero.

- $\rho(v) = \log(1-v)$, $\Upsilon = (-\infty, 1)$ reduces GEL to the basic EL setup.
- $\rho(v) = -\frac{1}{2}v^2 - v$: Continuously Updated GMM (CUE)
- $\rho(v) = 1 - \exp(v)$: Exponential tilting

The primal of the problem is

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta},\mathbf{p}}{\operatorname{argmin}}\sum_{i=1}^{n} h_n(p_i)$$

where $h_n(\cdot)$ belongs to the Cressie-Read family of divergences

$$h_n(p_i) = \frac{[\gamma(\gamma+1)]^{-1}[(np_i)^{\gamma+1}-1]}{n}$$

Implemented in `momentfit::gel4`.

3.14.2 M-estimation

GMM, but as taught in stats departments.

A $M-$ estimator is a solution for $\boldsymbol{\theta}$ that solves a moment condition

$$\sum_{i=1}^{n} \psi(\mathcal{O}_i, \widehat{\boldsymbol{\theta}}) = 0 \tag{1}$$

where $\mathcal{O}_1, \ldots, \mathcal{O}_n$ are IID obs (of arbitrary length), $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^k$, and $\psi(\cdot)$ is a known $k \times 1$ estimating function that does not depend on i or n.
The moment condition 1 is solved numerically using standard root-finding techniques.
The M-estimator $\widehat{\boldsymbol{\theta}}$ is *consistent* and *asymptotically normal* with asymptotic variance of the following sandwich form

$$\mathbb{V}\left[\mathcal{O}_i, \widehat{\boldsymbol{\theta}}\right] = \left(\mathbf{B}_n(\mathcal{O}_i, \widehat{\boldsymbol{\theta}})\right)^{-1} \mathbf{M}_n(\mathcal{O}_i, \widehat{\boldsymbol{\theta}}) \left(\left(\mathbf{B}_n(\mathcal{O}_i, \widehat{\boldsymbol{\theta}})\right)^{-1}\right)^{\top} \quad \text{where}$$

$$\text{Bread } \mathbf{B}_n(\mathcal{O}_i, \widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} -\psi'(\mathcal{O}_i, \widehat{\boldsymbol{\theta}})$$

$$\text{Meat } \mathbf{M}_n(\mathcal{O}_i, \widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \psi(\mathcal{O}_i, \widehat{\boldsymbol{\theta}})\psi(\mathcal{O}_i, \widehat{\boldsymbol{\theta}})^{\top}$$

Since many (most?) statistical estimation problems are solutions to an optimization problem, M-estimation nests many familiar statistical problems with $\psi(\cdot)$ as the corresponding FOC. For example, for maximum likelihood, $\psi(\cdot)$ is the score equation $\frac{\partial \log f(y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\top}$.

4 Causal Inference

4.1 Foundations, Experiments

4.1.1 Potential Outcomes

Exposition from Athey and Imbens (2016b) and Imbens and D. B. Rubin (2015)
Y_i is the observed outcome, D_i is the treatment with levels $d \in \mathcal{D}$,
potential outcomes denoted $Y_{di}, Y_i^d, Y_i(d)$ (interchangeably).

$$Y_i^{\text{obs}} = Y_i(D_i) = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Equivalently, we have the switching equation

$$Y_i = D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i} = Y_{0i} + \underbrace{(Y_{1i} - Y_{0i})}_{\tau_i} D_i$$

This encodes what is known as the **causal-consistency** assumption (/ SUTVA).
Generally, define a potential outcome (Frölich and Sperlich, 2019)

$$Y_i^d = \varphi(d, \mathbf{X}_i, \mathbf{U}_i)$$

where $\mathbf{X}_i$ is a vector of observed covariates and $\mathbf{U}_i$ is a vector of unobservables, and φ is an unknown measurable function. Typically, we are interested in nonparametric identification of φ or some features of it.

Defn 4.1 (Assignment Mechanism).
Given a population of n units, the assignment mechanism is a row-exchangeable function $\mathbf{Pr}(\mathbf{D}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ taking values on $[0, 1]$ that satisfies

$$\sum_{\mathbf{D} \in \{0,1\}^N} \mathbf{Pr}(\mathbf{D}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 1$$

A unit level assignment probability for unit i is

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \sum_{\mathbf{D}: D_i = 1} \mathbf{Pr}(\mathbf{D}, \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$$

A **finite population propensity score** is

$$e(x) = \frac{1}{N(x)} \sum_{i: \mathbf{X}_i = \mathbf{x}} p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$$

where $N(x) = \#\{i = 1, \ldots, N | \mathbf{X}_i = \mathbf{x}\}$ is the number of units in each stratum defined by $\mathbf{X}_i = \mathbf{x}_i$.

Defn 4.2 (Causal Estimand).
is a row-exchangeable function of potential outcomes, treatment assignment, and

covariates.

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{d})$$

$\mathbf{Y}(0), \mathbf{Y}(1)$ are $n-$ vectors of potential outcomes, $\mathbf{X}$ is a $n \times p$ covariate matrix, and $\mathbf{d}$ is an assignment vector.
The most intuitive estimand is a $n-$ vector $\boldsymbol{\tau} = \mathbf{Y}(1) - \mathbf{Y}(0)$. This is impossible to estimate because of the FPCI, so we instead use summaries, such as its sample average, or subgroup averages.

Defn 4.3 (Fundamental Problem of Causal Inference).
We never see both potential outcomes for any given unit.
Decompositions of Observed Differences:

$$\underbrace{\mathbb{E}[Y_i|D_i=1] - \mathbb{E}[Y_i|D_i=0]}_{\text{observed difference}} = \underbrace{\mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=0]}_{\text{Selection Bias}}$$

$$= \underbrace{\mathbb{E}[Y_1] - \mathbb{E}[Y_0]}_{\text{ATE}} + \underbrace{\mathbb{E}[Y_{0i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=0]}_{\text{Selection Bias}} + \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogeneous Treatment Bias}}$$

where $\pi = \mathbb{E}[D]$ is the share of the sample treated.

Assumption 1 (Identification Assumption: Complete Randomisation).

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i$$

This is a Missing Completely at Random (MCAR) assumption on potential outcomes.

Assumption 2 (Stable Unit Treatment Value Assumption (SUTVA)).
Writing outcomes generated by the switching regression assumes that potential outcomes for any unit do not vary with the treatment assigned to other units. In practice, this is equivalent to a no spillovers assumption.

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp \mathbf{D}_{-i}$$

Equivalently, let $\boldsymbol{D}$ denote a treatment vector for N units, and $\mathbf{Y}(\mathbf{D})$ be the potential outcome vector that would be observed if was based on allocation $\mathbf{D}$. Then, SUTVA requires that for allocations $\mathbf{D}, \mathbf{D}'$,

$$Y_i(\mathbf{D}) = Y_i(\mathbf{D}') \text{ if } D_i = D_i'$$

Intuitively, SUTVA ensures that the 'science table' (Imbens & Rubin 2015) has 2 columns for the two potential outcomes as opposed to 2^n (number of potential outcomes with arbitrary interference).

4.1.2 Treatment Effects

Estimands

- $\tau_{\text{ATE}} := \mathbb{E}(Y_{1i} - Y_{0i})$
- $\tau_{\text{ATT}} := \mathbb{E}(Y_{1i} - Y_{0i}|D_i=1) = \mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=1]$

Under randomisation, $\tau_{ATE} = \tau_{ATT}$, since the treated are a random sample of the population. Under weak(er) assumption of $Y_{0i} \perp\!\!\!\perp D_i$, only τ_{ATT} is identified.

4.1.3 Difference in Means

Defn 4.4 (Difference in Means point estimate).

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D \cdot Y_1}{N_1/N} - \frac{(1-D)\cdot Y_0}{N_0/N}\right) = \frac{1}{N_1}\sum_N D_i Y_i - \frac{1}{N_0}\sum_N (1-D_i)Y_i$$

Defn 4.5 (Variance estimation for difference in means).
Variance of Difference in means estimator is given by

$$\mathbb{V}[\tau]_{\text{DiM}} = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{N}$$

where S_0, S_1 are sample variances of Y^0, Y^1 respectively, and S_{01} is the variance of the *unit level* treatment effect

$$\frac{1}{N-1}\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - \tau)$$

This is **not identifiable** because of the last term. If the treatment effect is constant in the population, the last term is zero.
A (**conservative**) variance estimator is given by

$$\widehat{\mathbb{V}}(\widehat{\tau}_{\text{DiM}}) = \left(\frac{\widehat{\sigma}_1^2}{N_1} + \frac{\widehat{\sigma}_0^2}{N_0}\right)$$

where

$$\widehat{\sigma}_{(d)}^2 = \frac{1}{N_d - 1}\sum_{i:d_i=d}(Y_i - \overline{Y}_d)^2 \quad ; d = 0, 1$$

These variance estimates can be used to construct 95% confidence intervals

$$\mathrm{C}_{0.95}(\tau) = (\widehat{\tau} - 1.96\sqrt{\widehat{\mathbb{V}}}, \widehat{\tau} + 1.96\sqrt{\widehat{\mathbb{V}}},)$$

4.1.4 Regression Adjustment

$$\begin{aligned} Y_i &= \alpha + \tau_{\text{REG}} D_i + \eta_i \\ &= \underbrace{\overline{Y}_0}_{\alpha} + \underbrace{(\overline{Y}_1 - \overline{Y}_0)}_{\tau_{\text{REG}}} D_i + \underbrace{\left\{ (Y_{i0} - \overline{Y}_0) + D_i \cdot \left[(Y_{i1} - \overline{Y}_1) - (Y_{i0} - \overline{Y}_0) \right] \right\}}_{\eta_i} \end{aligned}$$

- $\alpha = \mathbb{E}[Y_{0i}]$
- $\tau = \mathbb{E}\left[Y_{1i} - Y_{0i}\right]$
- $\eta_i = Y_{0i} - \mathbb{E}(Y_{0i})$ [extra terms above come from allowing for heterogeneous TEs]

Selection bias: $\text{Cov}\left[D_i, \eta_i\right] \neq 0$

Example 4.1 (Matrix formulae for randomized regression).
Suppose 50 percent of the population gets the treatment. Let $\mathbf{X}_i' = \begin{bmatrix} D_i & 1 \end{bmatrix}$. Then,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \frac{N}{2} & \frac{N}{2} \\ \frac{N}{2} & N \end{bmatrix} = \frac{N}{2} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \implies (\mathbf{X}'\mathbf{X})^{-1} = \frac{2}{N} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

Similarly,

$$\mathbf{X}'y = \begin{bmatrix} \sum_T Y_i \\ \sum_T Y_i + \sum_c Y_i \end{bmatrix}$$

Therefore

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} \bar{Y}_T - \bar{Y}_c \\ \bar{Y}_c \end{bmatrix}$$

Generalise to p fraction treated
VCV under homoscedasticity

$$\left(\frac{1}{p(1-p)N(N-2)} \right) \begin{bmatrix} 1 & -p \\ -p & p \end{bmatrix} \left(\sum_T \hat{u}^2 + \sum_C \hat{u}^2 \right)$$

VCV under heteroskedasticity

$$\left(\frac{1}{p^2(1-p)^2 N(N-2)} \right) \begin{bmatrix} (1-p)^2 \sum_T \hat{u}^2 + p^2 \sum_C \hat{u}^2 & -p^2 \sum_C \hat{u}^2 \\ -p^2 \sum_C \hat{u}^2 & -p^2 \sum_C \hat{u}^2 \end{bmatrix}$$

Including controls:

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i' \boldsymbol{\beta} + \eta_i$$

Corrects for chance covariate imbalances, improves precision by *removing variation in outcome accounted for by pre-treatment characteristics*.

Fact 4.2.
Freedman (2008) Critique
Regression of the form $Y_i = \alpha + \tau_{reg} D_i + \beta_1 X_i + \epsilon_i$

- $\hat{\tau}_{reg}$ is consistent for ATE but has small sample bias (unless model is true); bias is on the order of $1/n$
- $\hat{\tau}_{reg}$ precision does not improve through the inclusion of controls; including controls is **harmful** to precision if more than $3/4$ units are assigned to one treatment condition

Theorem 4.3 (Lin (2013) fix / response to Freedman critique).
Recommends fitting

$$Y_i = \alpha + \tau_{\text{lin}} D_i + \boldsymbol{\beta}_0 \cdot (\mathbf{X}_i - \bar{\mathbf{X}}) + \boldsymbol{\beta}_1 \cdot D_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}) + \epsilon_i$$

Where the two potential outcomes are stipulated to follow

$$\begin{aligned} Y^1 &= \overline{Y^1} + (\overline{\mathbf{X}} - \overline{X}_1)^\top (\widehat{\boldsymbol{\beta}_0} + \widehat{\boldsymbol{\beta}_1}) \\ Y^0 &= \overline{Y^0} + (\overline{\mathbf{X}} - \overline{X}_0)^\top (\widehat{\boldsymbol{\beta}_0}) \end{aligned}$$

which has same small sample bias, but cannot hurt asymptotic precision even if the model is incorrect and will likely increase precision if covariates are predictive of the outcomes.

4.1.5 Randomisation Inference

Defn 4.6 (Fisher's Exact Test).
sharp null: $Y_{1i} = Y_{0i}\ \forall i$. Implies $H_0 : \mathbb{E}\left[Y_i\right] = \mathbb{E}\left[Y_0\right];\ H_1 : \mathbb{E}\left[Y_i\right] \neq \mathbb{E}\left[Y_0\right]$.
To test sharp null, **set** $Y_1 = Y_0$ **for all units and re-randomize treatment**. Complete randomisation of $2N$ units with N treated. $\binom{2N}{N}$ assignment vectors. P value can be as small as $1/\binom{2N}{N}$.
Ω is the full set of randomisation realisations, and ω is an element in the set (drawn either under complete randomization or binomial randomization), with associated probability $1/\binom{2N}{N}$
One sided P-value : $\mathbf{Pr}\left((\hat{\alpha}(\omega) \geq \hat{\tau}_{\text{ATE}})\right)$

4.1.6 Blocking

Stratify randomisation to ensure that groups start out with identical observable characteristics on blocked factors.

$V[\hat{\tau}_{BR}] < V[\hat{\tau}_{CR}]$ if $\frac{SSR_{\hat{\varepsilon}^*}}{n-k-1} < \frac{SSR_{\hat{\epsilon}}}{n-2}$ where $\hat{\epsilon}$ and $\hat{\epsilon}^*$ are errors from specification omitting and including block dummies respectively.
For J blocks,
Point estimate

$$\hat{\tau}_B = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\tau}_j$$

Variance Randomisations within each block are independent, so the variances are simple means (with squared weights).

$$\text{Var}(\hat{\tau}_B) = \sum_{j=1}^{J} \left(\frac{N_j}{N}\right)^2 \text{Var}(\hat{\tau}_j)$$

Regression Formulation

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

If treatment probabilities vary by block, then weight by

$$w_{ij} = \left(\frac{1}{p_{ij}}\right) D_i + \left(\frac{1}{1-p_{ij}}\right)(1 - D_i)$$

Efficiency Gains from Blocking

- **Complete Randomisation** : $Y_i = \alpha + \tau_{CR} D_i + \epsilon_i$
- **Block Randomisation**: $Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \epsilon_i^*$

$$\text{Var}[\hat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} \left(D_i - \overline{D}\right)^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}$$

$$\text{Var}[\hat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^{n} \left(D_i - \overline{D}\right)^2 \left(1 - R_j^2\right)} \text{ with } \widehat{\sigma}_{\varepsilon^*}^2 = \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^{*^2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon}^*}}{n-k-1}$$

Where R_j^2 is the fit from regressing D on all B_j dummies. Since $R_j^2 \approx 0$ by randomisation,

$$\mathbb{V}[\hat{\tau}_{BR}] < \mathbb{V}[\hat{\tau}_{CR}] \Leftrightarrow \frac{SSR_{\hat{\epsilon}}}{n-k-1} < \frac{SSR_{\hat{\epsilon}}}{n-2}$$

4.1.7 Power Calculations

Basic idea: With large enough samples, $\mathbb{V}\left[\bar{Y}_1 - \bar{Y}_0\right] \approx \frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}$[where $p = N_1/N$ is the share of sample treated]. Set p to minimise overall variance. Yields $p^* = \frac{\sigma_1}{\sigma_1+\sigma_0}$. With homoskedasticity, this is $\frac{1}{2}$ Treatment, $\frac{1}{2}$ control.

Defn 4.7 (Power Function).
$\tau = \mu_1 - \mu_0$ (effect size)
Test for $\tau > (t_{1-\kappa} + t_{\alpha/2} SE(\hat{\beta})$.
For common variance σ,

$$\pi = \mathbf{Pr}(|t| > 1.96) = \Phi\left(-1.96 - \frac{\tau\sqrt{N}}{2\sigma}\right) + \left(1 - \Phi\left(1.96 - \frac{\tau\sqrt{N}}{2\sigma}\right)\right)$$

General formula for Power with **unequal variances**

$$\pi = \Phi\left(1.96 - \frac{\tau}{\sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}}\right) + \left(1 - \Phi\left(1.96 - \frac{\tau}{\sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}}\right)\right)$$

This yields

Defn 4.8 (Minimum Detectable Effect MDE).
Common variance (assumed)

$$\text{MDE}(\tau) = M_{n-2}\sqrt{\frac{\sigma^2}{Np(1-p)}}$$

where $M_{n-2} = t_{(1-\alpha/2)} + t_{1-\kappa}$ = Critical t-value to reject null + t-value for alternative (where $1-\kappa$) is power.
MDES (Minimum Detectable Effect Size in Standard Deviation Units):

$$\text{MDES}(\tau) = M_{n-2}\sqrt{\frac{1}{Np(1-p)}}$$

Fact 4.4 (Typical MDE for $\alpha = 0.05, \kappa = 0.8$, N large).
Multiplier M_{n-2} simplifies to $1.96 + 0.84 \approx 2.8$

$$\text{MDE} \approx (0.84 + 1.96)\text{SE}(\hat{\tau}) \approx 2.8\ \text{SE}(\hat{\tau})$$

Rearrange to get necessary sample size for any given hypothesised MDE and expected variance.

$$N = (z_{1-\kappa} + z_{\alpha/2})^2 \cdot \left(\frac{1}{p(1-p)}\right) \cdot \frac{\sigma^2}{\text{MDE}^2}$$

MDES for Blocking

$$\text{MDES}(\tau_{BR}) = M_{n-k-1}\sqrt{\frac{1-R_B^2}{Np(1-p)}}$$

where R_B^2 is the R-squared from regressing Y on block dummies.

Fact 4.5 (Required Sample Size for rejection probability β, size α, treatment share γ, effect size τ).
To test $H_0 : \mathbb{E}[Y_i(1) - Y_i(0)] = 0$ against the alternative, we look at the T Statistic

$$T = \frac{\overline{Y}_t^{obs} - \overline{Y}_c^{obs}}{\sqrt{S_y^2/N_t + S_y^2/N_c}} \approx \mathcal{N}\left(\frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}, 1\right)$$

Inverting this for size $\alpha/2$ gives us a **required sample size**

$$\text{Required Sample Size} = N = \frac{\left(\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2)\right)}{(\tau/\sigma)^2 \cdot \gamma \cdot (1-\gamma)}$$

typically, $\beta = 0.8$, $\alpha = 0.05, \gamma = 0.5$, so by subtitution

$$N = \frac{\left(\Phi^{-1}(0.8) + \Phi^{-1}(0.975)\right)}{(\tau/\sigma)^2 \cdot 0.5^2}$$

4.2 Selection On Observables

Imbens (2004) typology

- Regression estimators: rely on consistent estimation of $\mu_0(\mathbf{x}), \mu_1(\mathbf{x})$
- Matching estimators
- Propensity score estimators: rely on estimation of $\pi(\mathbf{x})$
- Combination methods (*augmented* IPW, *bias-corrected* Matching, etc)

4.2.1 Regression Anatomy / FWL

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Fact 4.6 (Omitted-Variables Bias Formula).
If structural (**long**) equation is $Y_i = \alpha + \tau D_i + W_i'\gamma + \epsilon_i$, with W_i vector of unobserved, and we estimate **short** $Y_i = \alpha + \rho D_i + \epsilon_i$, then we can write the specification as $y = \tau D_i + \underbrace{W_i'\gamma + \epsilon}_{\nu_i}$

$$\rho = \frac{\text{Cov}[Y_i, D_i]}{\mathbb{V}[D]} = \tau + \gamma'\delta_{WD}$$

equivalently,

$$\text{plim}\ \hat{\tau}_{\text{OLS}} = \tau + \delta\gamma = \tau + \underbrace{\text{plim}\left[\left(N^{-1}D'D\right)^{-1}N^{-1}D'W\right]\gamma}_{\text{Omitted Variables Bias}}$$

Coefficient in Short Regression = Coefficient in long regression + effect of omitted $\times$ regression of omitted on included. This bias can be arbitrarily large.

4.2.2 Identification of Treatment Effects under Unconfoundedness

Assumption 3 (Conditional Independence Assumption and Overlap).

- *Unconfoundedness / Selection on Observables / Ignorability / Conditional Independence Assumption*: $(Y_0, Y_1) \perp\!\!\!\perp D|\mathbf{X}$

– In terms of densities, this is equivalent to the validity of the following density factorisation

$$\begin{aligned}\mathsf{f}_{Y(d),D|\mathbf{X}}(y,d|\mathbf{x}) &= \mathsf{f}_{Y(d)|\mathbf{X}}(y|\mathbf{x})\mathsf{f}_{D|\mathbf{X}}(d|x)\\ &= \mathsf{f}_{Y|D,\mathbf{X}}(y|d,\mathbf{x})\mathsf{f}_{D|\mathbf{X}}(d|x)\end{aligned}$$

- *common support* $0 < \mathbf{Pr}\left(D=1|X\right) < 1$

$$\begin{aligned}\mathbb{E}\left[Y_d\right] &= \int \mathbb{E}\left[Y^d|\mathbf{X}=\mathbf{x}\right] \mathsf{d}P_{\mathbf{x}} && \text{by LIE}\\ &= \int \mathbb{E}\left[Y^d|D=d,\mathbf{X}=\mathbf{x}\right] \mathsf{d}P_{\mathbf{x}} && \text{by unconfoundedness, overlap}\\ &= \int \mathbb{E}\left[Y|D=d,\mathbf{X}=\mathbf{x}\right] \mathsf{d}P_{\mathbf{x}} && \text{by consistency}\end{aligned}$$

The third quantity is estimable using observed data.
Estimators:
Discrete Case: $\mathbf{x}$ has finite values indexed by $k = 1, \ldots, K$ with generic entry $\mathbf{x}_k$

$$\tau_{\text{ATE}} = \sum_{k=1}^{K}(\mathbb{E}\left[Y|D=1,\mathbf{X}=\mathbf{x}_k\right] - \mathbb{E}\left[Y|D=0,\mathbf{X}=\mathbf{x}_k\right])\mathbf{Pr}\left(\mathbf{X}=\mathbf{x}_k\right)$$

$$\tau_{\text{ATT}} = \sum_{k=1}^{K}\mathbb{E}\left[Y|D=1,\mathbf{X}=\mathbf{x}_k\right] - \mathbb{E}\left[Y|D=0,\mathbf{X}=\mathbf{x}_k\right])\ \mathbf{Pr}\left(\mathbf{X}=\mathbf{x}_k|D=1\right)$$

Multi-valued and Continuous Treatments Imbens (2000) and Hirano and Imbens (2004)
Treatment values: $\mathcal{D}$ finite if multi-valued / $\subset \mathbb{R}$ for continuous, with corresponding **dose-responses** $Y_i(d)$. We are interested in dose-response function $\mu(d) = \mathbb{E}\left[Y_i(d)\right]$, and contrasts.
First define **Generalised propensity score** :

$$R := r(d,\mathbf{x}) = \mathsf{f}_{D|\mathbf{X}}\left(d|\mathbf{x}\right)$$

Assumptions:

- Weak unconfoundedness: $Y(d) \perp\!\!\!\perp D|\mathbf{X}=\mathbf{x}\ \ \forall\ D\in\mathcal{D}$
- Conditional density overlap: $\mathsf{f}\left(D=d|\mathbf{X}=\mathbf{x}\right) > 0$

Bias removal using the generalised propensity score:

- Estimate the conditional expectation of the outcome as a function of treatment level d and GPS R as

$$\beta(d,r) = \mathbb{E}\left[Y(d)|r(d,\mathbf{X})=r\right] = \mathbb{E}\left[Y|D=d|R=r\right]$$

- Estimate the dose-response function of the treatment by averaging the conditional expectation *at that particular level of treatment* $\mu(d) = \mathbb{E}\left[\beta(d,r(d,\mathbf{X}))\right]$

Then compute contrasts to get first derivative (MTE)

$$\frac{\partial}{\partial d}\mathbb{E}\left[\mu(d)\right]$$

4.2.3 Estimators of $\mathbb{E}\left[Y^d\right]$

which can be used to construct estimators of ATE$(\widehat{\gamma}_1 - \widehat{\gamma}_0)$, ATT$((\gamma_1 - \widehat{\gamma}_0|D=1)$, and other estimands. reference: Imbens (2004), David Childers' lecture notes.

- **Regression Adjustment**
  - Estimate $\mu_d(\mathbf{x}) = \mathbb{E}\left[Y|D=d,\mathbf{X}=\mathbf{x}\right]$ by nonparametric regression estimator $\widehat{\mu}_d(\mathbf{x})$
  - Average $\widehat{\gamma}_d^{\text{reg}} := \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_d(\mathbf{x})$
  - Since average of predicted treated outcome for the treated is equal to the average predicted outcome for controls, can also write ATE as

$$\widehat{\tau}_{\text{reg}}^{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} D_i\cdot[Y_i - \widehat{\mu}_0(\mathbf{X}_i)] + (1-D_i)[\widehat{\mu}_1(\mathbf{X}_i) - Y_i]$$

  - SATT only requires imputation of one potential outcome

$$\widehat{\tau}_{\text{reg}}^{\text{ATT}} = \frac{1}{N_t}\sum_{i=1}^{n} D_i[Y_i - \widehat{\mu}_0(\mathbf{X}_i)]$$

- **Inverse Propensity Weighting**
  - Estimate propensity score $\pi(\mathbf{x}) = \mathbb{E}\left[D=d|\mathbf{X}=\mathbf{x}\right]$ by conditional probability estimator $\widehat{\pi}(d|\mathbf{x})$
  - Average

$$\widehat{\gamma}_d^{\text{IPW}} := \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\mathbb{1}_{D_i=d}}{\widehat{\pi}(d|\mathbf{x})}$$

- **Augmented Inverse Propensity Weighting / Combination methods**
  - Estimate $\mu_d(\mathbf{x}) = \mathbb{E}[Y|D=d, \mathbf{X}=\mathbf{x}]$ by $\widehat{\mu}_d(\mathbf{x})$
  - Estimate $\pi(\mathbf{x}) = \mathbb{E}[D=d|\mathbf{X}=\mathbf{x}]$ $\widehat{\pi}(d|\mathbf{x})$
  - Average

$$\widehat{\gamma}_d^{\text{AIPW}} := \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_d(\mathbf{x}) + \frac{(Y_i - \widehat{\mu}_d(\mathbf{x}))\mathbb{1}_{D_i=d}}{\widehat{\pi}(d|\mathbf{x})}\right)$$

- Hahn (1998) normalized outcome regression : estimate

$$\widetilde{\mu}_1 = \frac{(\widehat{\mu}_1(\mathbf{x}))}{\widehat{\pi}(\mathbf{x})};\ \ \widetilde{\mu}_0 = \frac{(\widehat{\mu}_0(\mathbf{x}))}{1-\widehat{\pi}(\mathbf{x})}$$

4.2.4 Subclassification / Blocking

Weighted combination of K subclasses of covariate values, which partition the population

$$\hat{\tau}^{\text{ATE}} = \sum_{k=1}^{K}\left(\overline{Y}_1^k - \overline{Y}_0^k\right)\cdot\left(\frac{N^k}{N}\right)$$

$$\hat{\tau}_{ATT} = \sum_{k=1}^{K}\left(\overline{Y}_1^k - \overline{Y}_0^k\right)\cdot\left(\frac{N_1^k}{N_1}\right)$$

4.2.5 Regression Adjustment

A single regression with controls X is potentially problematic because of Simpson's paradox. To account for this in a parametric setup, assume a set of iid subjects $i = 1, \dots n$ we observe a tuple (X_i, Y_i, D_i), comprised of

- **feature vector** $\boldsymbol{X}_i \in \mathbb{R}^p$
- **response** $Y_i \in \mathbb{R}$
- **treatment assignment** $D_i \in \{0, 1\}$

Define conditional response surfaces as

$$\mu_{(d)}(x) := \mathbb{E}[Y_i|\mathbf{X_i} = \mathbf{x}, D_i = d]$$

First pass regression adjustment estimator (using OLS)

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_{(1)}(\boldsymbol{X}_i) - \widehat{\mu}_{(0)}(\boldsymbol{X}_i)\right]$$

where $\widehat{\mu}_{(d)}(\boldsymbol{x})$ is obtained via OLS. This generically doesn't work for regularised regression.
With known propensity score $\pi(\mathbf{X})$ (as in case of regression), an efficient estimator (Hahn, 1998) weights all estimated treatment effects $\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$ by the propensity score:

$$\widetilde{\tau}_{\text{reg}}^{\text{ATT}} = \frac{\sum_{i=1}^{n}\pi(\mathbf{X}_i)[\widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)]}{\sum_{i=1}^{n}\pi(\mathbf{X}_i)}$$

Fact 4.7 (Consistency of Regression estimation of ATE).
Additional Assumptions for consistent estimate of ATE from OLS:

1) Constant treatment effects
2) Outcomes linear in X

$\implies$ τ will provide unbiased and consistent estimates of ATE.

- (2) fails - τ_{OLS} is Best Linear Approximation of *average causal response function* $\mathbb{E}[Y|D=1, X] - \mathbb{E}[Y|D=0, X]$.
- (1) fails - τ_{OLS} is conditional variance weighted average of underlying τs.

Pretend there are m strata of $\mathbf{X}$. Then, OLS estimates

$$\tau_{\text{OLS}} = \sum_{k=1}^{m}\left(\mathbb{E}[Y|X=x_k, D=1] - \mathbb{E}[Y|X=x_k, D=0]\right)\omega_k$$

where the weight

$$\omega_k = \frac{\mathbb{V}[(D|X=x_k]\,\mathbf{Pr}(X=x_k)}{\sum_{r=1}^{m}\mathbb{V}[D|X=x_r]\,\mathbf{Pr}(X=x_r)}$$

τ_{OLS} weighs up groups where the size of the treated and untreated population are roughly equal, and weighs down groups with large imbalances in the size of these two groups.
τ_{OLS} is true effect IFF constant treatment effects holds.

4.2.6 Matching

Regression estimators impute missing potential outcomes by imputing it using $\widehat{\mu}_d(\mathbf{X}_i)$. Matching estimators proceed by by 'imputing' potential outcome using the observed outcome from 'closest' control unit.

Defn 4.9 (Matching Estimators).
Define $\ell_m(i)$ as the index that satisfies

$$\sum_{j:d_j \neq d_i} \mathbb{1}_{\|\mathbf{X}_j - \mathbf{X}_i\| \leq \|\mathbf{X}_l - \mathbf{X}_i\|} = m$$

So, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is $m-$th closest to unit i in terms of covariate values in terms of the norm $\|\cdot\|$. Let $\mathcal{J}_M(i) := \{\ell_1(i), \ldots, \ell_M(i)\}$ denote the indices of the first M matches for unit i. Then, impute potential outcomes as

$$\widehat{Y}_i(0) = \begin{cases} Y_i & \text{if } D_i = 0 \\ \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 1 \end{cases}$$
$$\widehat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

then, the simple matching (with replacement) estimator for ATE is

$$\widehat{\tau}^{ATE}_{\text{Match}} = \frac{1}{n}\sum_{i=1}^{n}[\widehat{Y}(1) - \widehat{Y}(0)]$$

and corresponding ATT

$$\hat{\tau}^{ATT}_{\text{Match}} = \frac{1}{N_1}\sum_{D=1}\left(Y_i - Y_{j(i)}\right)$$

where $M = 1$ corresponds with one-to-one matching and $M > 1$ is many-to-one. Many-to-one matching is not $\sqrt{n}$ consistent (Abadie and Imbens (2006)) and has a bias of $O(N^{-1/k})$ where k is the number of continuous covariates.
Bias-corrected (Abadie-Imbens)

$$\hat{\tau}_{ATT} = \frac{1}{N_1}\sum_{D=1}\left(Y_i - Y_{j(i)}\right) - \left(\hat{\mu}_0\left(X_i\right) - \hat{\mu}_0\left(X_{j(i)}\right)\right)$$

Where $\mu_0(x) = E[Y|X = x, D = 0]$ is the regression function under the control.

Metrics

- Euclidian Distance

$$\|X_i - X_j\| = ED\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)'\left(X_i - X_j\right)}$$

- Stata diagonal distance

$$\text{StataD}\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)' \text{diag}\left(\widehat{\Sigma}_X\right)^{-1}\left(X_i - X_j\right)}$$

  where the normalisation factor is the diagonal element of $\hat{\Sigma}$, the estimated variance covariance matrix.

- Mahalanobis distance (scale-invariant)

$$\text{MD}\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)'\Sigma^{-1}\left(X_i - X_j\right)}$$

Where $\hat{\Sigma}$ is the variance-covariance matrix.

Defn 4.10 (Variance Estimators for Matching).
Matching estimators have a normal distribution in large samples provided that bias is small.
For matching without replacement,

$$\widehat{\sigma}^2_{ATT} = \frac{1}{N_T}\sum_{D_i=1}\left(Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{jm(i)} - \hat{\delta}_{ATT}\right)^2$$

For matching with replacement,

$$\widehat{\sigma}^2_{ATT} = \frac{1}{N_T}\sum_{D_i=1}\left(Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{jm(i)} - \hat{\delta}_{ATT}\right)^2 + \frac{1}{N_T}\sum_{D_i=0}\left(\frac{K_i(K_i-1)}{M^2}\right)\mathbb{V}\left[\epsilon|X_i, D_i = 0\right]$$

where K_i is the number of times observation i is used in a match, and the last error variance term is estimated by matching also. **the bootstrap doesn't work for matching**.

Theorem 4.8 (Balancing Property of the Propensity Score Rosenbaum and D. B. Rubin (1983)).
PScore is a balancing score - Conditioning on Propensity score is equivalent to conditioning on covariates:

$$\mathbf{Pr}\left(D = 1|Y_0, Y_1, \pi(X)\right) = \mathbf{Pr}\left(D = 1|\pi(X)\right) = \pi(X)$$

$$Y_i(0), Y_i(1) \perp\!\!\!\perp D_i | \boldsymbol{X}_i \equiv Y_i(0), Y_i(1) \perp\!\!\!\perp D_i | \pi(\boldsymbol{X}_i)$$

Theorem 4.9 (Efficiency bound and efficient score).
Hahn (1998) defines the semiparametric Efficiency Bound for ATE: the asymptotic variance of any regular estimator of τ of the population ATE obeys

$$\sqrt{n}(\widehat{\tau} + \tau^P) \xrightarrow{d} \mathcal{N}(0, \mathbb{V})$$

where

$$\mathbb{V} \geq \mathrm{V}_{\text{eff}}^{\text{PATE}} := \mathbb{E}\left[\frac{\sigma_1^2(\mathbf{X})}{\pi(\mathbf{X})} + \frac{\sigma_0^2(\mathbf{X})}{1-\pi(\mathbf{X})} + (\tau(\mathbf{X}) - \tau)^2\right]$$

and for PATE (γ)

$$\mathrm{V}_{\text{eff}}^{\text{PATT}} = \mathbb{E}\left[\frac{\pi(\mathbf{X})\,\sigma_1^2(\mathbf{X}))}{p^2} + \frac{\pi(\mathbf{X}^2)\,\sigma_0^2(\mathbf{X})}{p^2(1-\pi(\mathbf{X})} + \frac{(\tau(\mathbf{X}) - \gamma)^2\pi(\mathbf{X})}{p^2}\right]$$

where $\sigma_d^2(\boldsymbol{X}) = \mathbb{V}\left[Y^d | \boldsymbol{X}\right]$, $\tau(X) := \mathbb{E}\left[Y^1 - Y^0 | \boldsymbol{X}\right]$, and $p := \mathbb{E}\left[\pi(\mathbf{X})\right]$.
Any regular estimator whose asymptotic variance achieves this efficiency bound is equal to $\frac{1}{n}\sum_{i=1}^n \psi_i(\mu) + O_P(\sqrt{n})$, where

$$\psi_i(\mu) = \mu(1, \boldsymbol{X}_i) - \mu(0, \boldsymbol{X}_i) + \frac{D_i(Y_i - \mu(1, \boldsymbol{X}_i))}{\pi(\boldsymbol{X}_i)} - \frac{(1-D_i)(Y_i - \mu(0, \boldsymbol{X}_i))}{1-\pi(\boldsymbol{X}_i)}$$

is the **Efficient Influence Function** for estimating τ.
Imbens (2004) shows that

$$\mathrm{V}_{\text{eff}}^{\text{SATE}} = \mathrm{V}_{\text{eff}}^{\text{PATE}} - \underbrace{\mathbb{E}\left[\overbrace{Y_1 - Y_0}^{\widehat{\tau}} - \tau^P\right]^2}_{\text{Variance of treatment effect}}$$

Estimators in this section try to attain the SPEB.

Defn 4.11 (Weighting on the Propensity Score: Horvitz-Thompson Estimands).

$$\tau_{\text{ATE}}^{\text{IPW}} = \mathbb{E}\left[Y \cdot \frac{D - \pi(\mathbf{X})}{\pi(\mathbf{X})(1-\pi(\mathbf{X}))}\right] = \mathbb{E}\left[\frac{YD}{\pi(\mathbf{X})} - \frac{Y(1-D)}{(1-\pi(\mathbf{X}))}\right]$$

and

$$\tau_{\text{ATT}}^{\text{IPW}} = \frac{1}{\mathbf{Pr}\left(D=1\right)}\mathbb{E}\left[Y \cdot \frac{D - \pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right]$$

The counterfactual mean $\mathbb{E}\left[Y^0 | D=1\right] = \mu_0^1$ can be identified as

$$\mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{\rho}\frac{1-D_i}{1-\pi(\mathbf{X}_i)}Y_i\right]$$

where $\rho = \mathbf{Pr}\left(D=1\right)$.

Defn 4.12 (Inverse Probability-Weighted Estimators).

$$\widehat{\tau}_{\text{ipw}}^{\text{ate}} = \frac{1}{n}\sum_{i=1}^n \left(\underbrace{\frac{Y_i D_i}{\widehat{\pi}(\mathbf{X}_i)}}_{\mathbb{E}[Y_1]} - \underbrace{\frac{Y(1-D_i)}{(1-\widehat{\pi}(\mathbf{X}_i))}}_{\mathbb{E}[Y_0]}\right) = \frac{1}{n}\sum_{i=1}^n Y_i\left(\frac{D_i}{\widehat{\pi}(\boldsymbol{X}_i)} - \frac{1-D_i}{1-\widehat{\pi}(\boldsymbol{X}_i)}\right)$$

Hirano, Imbens, and Ridder (2003) normalise both pieces using a Hajek-style adjustment, since extreme values of $\widehat{\pi}$ makes variance explode. Often advisable to trim or use **Hajek weights**, which introduces limited bias at the cost of large decreases in variance.

$$\widehat{\tau}_{\text{ipw2}}^{\text{ate}} = \frac{1}{n}\sum_{i=1}^n \left[\left(\frac{Y_i D_i}{\widehat{\pi}(\mathbf{X}_i)} \Big/ \frac{D_i}{\widehat{\pi}(\mathbf{X}_i)}\right) - \left(\frac{(1-D_i)Y_i}{1-\widehat{\pi}(\mathbf{X}_i)} \Big/ \frac{1-D_i}{1-\widehat{\pi}(\mathbf{X}_i)}\right)\right]$$

Similarly, for the effect on the treated

$$\begin{aligned}\widehat{\tau}_{\text{ipw}}^{\text{att}} &= \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i}{\mathbf{Pr}\left(D_i=1\right)} \cdot \frac{D_i - \widehat{\pi}(X_i)}{(1-\widehat{\pi}(X_i))}\right) \\ &\equiv \frac{1}{n}\sum_{i=1}^n Y_i\left(\frac{D_i}{\widehat{\pi}(\boldsymbol{X}_i)} - \frac{(1-D_i)\widehat{\pi}(\boldsymbol{X}_i)}{(1-\widehat{\pi}(\boldsymbol{X}_i))\mathbf{Pr}\left(D_i=1\right)}\right)\end{aligned}$$

$$\widehat{\tau}_{\text{ipw2}}^{\text{att}} = \left[\frac{1}{N_1}\sum_{i:D_i=1} Y_i\right] - \left[\sum_{i:D_i=0} Y_i \cdot \frac{\widehat{\pi}(\mathbf{X}_i)}{1-\widehat{\pi}(\mathbf{X}_i)} \Big/ \sum_{i:D_i=0}\frac{\widehat{\pi}(\mathbf{X}_i)}{1-\widehat{\pi}(\mathbf{X}_i)}\right]$$

Horvitz-Thompson Estimator as Regression $Y_i = \alpha + \tau D_i + \epsilon_i$ with IPW weights

$$\lambda_i = \sqrt{\frac{D_i}{\pi(X_i)} + \frac{1-D_i}{1-\pi(X_i)}}$$

Defn 4.13 (Weighted Average Treatment Effect Hirano, Imbens, and Ridder (2003)).
define the Weighted ATE (WATE) as

$$\tau_{\text{ATE}} = \frac{\int \mathbb{E}\left[Y^1 - Y^0 | \mathbf{X} = \mathbf{x}\right] g(\mathbf{x}) \mathrm{d}\mathbb{F}(\mathbf{x})}{\int g(\mathbf{x}) \mathrm{d}\mathbb{F}(\mathbf{x})}$$

where $g(\mathbf{x})$ is a weighting function. ATT is constructed when $g(\mathbf{x}) = \pi(\mathbf{x})$ the corresponding estimator is

$$\widehat{\tau}_{\text{WATE}} = \sum_{i=1}^{n} g(\mathbf{x}_i) \left(\frac{Y_i D_i}{\widehat{\pi}(\mathbf{x}_i)} - \frac{Y_i(1-D_i)}{1-\widehat{\pi}(\mathbf{x}_i)} \right) \Big/ \sum_{i=1}^{n} g(\mathbf{x}_i)$$

Defn 4.14 (Overlap Weights (Li, Morgan, and Zaslavsky, 2018)).
Sample drawn from $f(\boldsymbol{X})$, and can represent a target population as $g(\boldsymbol{X}) \propto f(\boldsymbol{X})h(\boldsymbol{X})$ where $h(\cdot)$ is the *tilting function*.
Define $f_d(\boldsymbol{x}) = \mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x} | D = d)$, which gives $f_1(\boldsymbol{x}) \propto f(\boldsymbol{x})\pi(\boldsymbol{x}) \quad ; f_0(\boldsymbol{x}) \propto f(\boldsymbol{x})(1-\pi(\boldsymbol{x}))$
For a given tilting function, to estimate τ_h, weight $f_d(x)$

$$w_1(\boldsymbol{x}), w_0(\boldsymbol{x}) = \frac{h(\boldsymbol{x})}{\pi(\boldsymbol{x})}, \frac{h(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}$$

Target	$h(x)$	Estimand	w_1, w_0
Combined	1	ATE	$\left(\frac{1}{\pi(x)}, \frac{1}{1-\pi(x)}\right)$ [IPW]
Treated	$\pi(x)$	ATT	$\left(1, \frac{\pi(x)}{1-\pi(x)}\right)$
Control	$1-\pi(x)$	ATC	$\left(\frac{1-\pi(x)}{\pi(x)}, 1\right)$
Overlap	$\pi(x)(1-\pi(x))$	ATO	$(1-\pi(x), \pi(x))$

Overlap weights are defined by choosing $h(x)$ that **minimises asymptotic variance of $\widehat{\tau}^h$**. The achieve exact balance on covariates included in the propensity score estimation.

$$\widehat{\tau} = \widehat{\mu}_1^h - \widehat{\mu}_0^h = \sum_i^N \frac{w_1(\mathbf{x}_i) D_i Y_i}{w_1(\mathbf{x}_i) D_i} - \frac{w_0(\mathbf{x}_i)(1-D_i) Y_i}{w_0(\mathbf{x}_i)(1-D_i)}$$

τ^{OW} can be interpreted as treatment effect among population that have good balance on observables.
Implemented in `PSweight`.

Defn 4.15 (Entropy Balancing (Hainmueller, 2012)).
Entropy weights w_i for each control unit are chosen by a reweighting scheme

$$\max_{w_i} H(w) = -\sum_{i:D=0} w_i \log(w_i)$$

subject to balance/moment-condition and normalising constraints

$$\sum_{i:D=0} w_i c_{ri}(\mathbf{X}_i) = m_r \quad r \in 1, \ldots, R$$
$$\sum_{i:D=0} w_i = 1 \text{ and } w_i \geq 0 \; \forall \{i : d_i = 0\}$$

The above problem is convex but has dimensionality of n_0 (nonnegativity) + p (moment conditions) + 1 (normalisation). The dual, on the other hand, only has dimensionality $p+1$ and unconstrained, which is considerably easier to solve using Newton-Raphson.

Defn 4.16 (Covariate Balancing Propensity Score).
Imai and Ratkovic (2014) propose CBPS, which is a method that involves modifying an initial propensity score estimate (e.g. by changing coefficients from a logistic model) iteratively until a balance criterion is reached.
Their basic insight is that when we use a logistic regression to estimate a propensity score, we assert that the pscore takes the form $\pi_\beta(\mathbf{x}_i) = \Lambda(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}$, and maximise the bernoulli log likelihood

$$\sum_{i=1}^{n} d_i \log(\pi_\beta(\mathbf{x}_i)) + (1-d_i) \log(1-\pi_\beta(\mathbf{x}_i))$$

which is then solved by the corresponding score

$$\frac{1}{n} \sum_{i=1}^{n} \frac{d_i \pi'_\beta(\mathbf{x}_i)}{\pi_\beta(\mathbf{x}_i)} - \frac{(1-d_i)\pi'_\beta(\mathbf{x}_i)}{1-\pi_\beta(\mathbf{x}_i)} = 0$$

this score balances a particular function of covariates: $\pi'_\beta(\mathbf{x}_i)$. Alternatively, we could choose that function by specifying a moment condition

$$\mathbb{E}\left[\frac{d_i f(\mathbf{x}_i)}{\pi_\beta(\mathbf{x}_i)} - \frac{(1-d_i)\mathbf{x}_i}{1-\pi_\beta(\mathbf{x}_i)} \right] = 0$$

Analogously for ATT, this moment condition is

$$\mathbb{E}\left[d_i f(\mathbf{x}_i) - \frac{\pi_\beta(\mathbf{x}_i)}{1-\pi_\beta(\mathbf{x}_i} (1-d_i) f(\mathbf{x}_i) \right]$$

When this balance condition is solved independently, the problem is *just-identified*. When it is used in conjunction with the conventional bernoulli likelihood, the problem is *over-identified*. Implemented in `CBPS::CBPS` as well as *balance*.

Defn 4.17 (Covariate-Balancing Scoring Rules Q. Zhao (2016)).

Defn 4.18 (General form of Weighting Estimators : Ben-Michael et al. (2021)).
The estimand is $\mu_1 = \mathbb{E}[Y(1)]$ (with μ_0 defined analogously). The estimator for this quantity is written

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} D_i \widehat{\gamma}(\mathbf{X}_i) Y_i$$

where the weights $\gamma(\cdot)$ are chosen to satisfy the *sample* balance property

$$\frac{1}{n}\sum_{i=1}^{n} D_i \gamma(\mathbf{X}_i) f(\mathbf{X}_i) \approx \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i) \text{ for any bounded } f(x)$$

in words: for every function $f(x)$, the weighting function equates weighted averages of f over the *treated units* to unweighted averages over the *study population*.
The weights are solved by solving an optimisation problem to trade off imbalance and some measure of complexity

$$\widehat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \left\{ \underbrace{\text{imbalance}^2_{\mathcal{M}}(\gamma)}_{\zeta(\cdot)} + \underbrace{\frac{\sigma^2}{n^2}\sum_{D_i}\gamma(\mathbf{X}_i)^2}_{\chi(\gamma\{X_i)\}} \right\}$$

with convex ζ, χ functions.
A common imbalance measure is

$$\text{imbalance}^2_{\mathcal{M}}(\gamma) = \max_{j=1\ldots p} \left| \frac{1}{n}\sum_{i=1}^{n} X_{ij} - \frac{1}{n}\sum_{i=1}^{n} D_i \gamma(\mathbf{X})_i X_{ij} \right|$$

for $\mathcal{M} = \{\beta \cdot x : \|\beta\|_1 \leq 1\}$

4.2.7 Hybrid Estimators

A doubly-robust estimator is consistent if one gets *either* the propensity score $\hat{\pi}$ or the regression $\hat{\mu}$ right.

Defn 4.19 (Augmented IPW Estimators).
Oracle AIPW

$$\widehat{\tau}^*_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_{(1)}(\boldsymbol{X}_i) - \mu_{(0)}(\boldsymbol{X}_i) + D_i\frac{y_i - \mu_{(1)}(\boldsymbol{X}_i)}{e(\boldsymbol{X_i})} + (1-D_i)\frac{y_i - \mu_{(0)}(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\right]$$

Feasible AIPW

$$\begin{aligned}\widehat{\tau}_{\text{AIPW}} &= \frac{1}{N}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)} + \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left[\overbrace{\hat{\mu}_1(\mathbf{X}_i)}^{\text{Regression}} + \overbrace{\frac{D_i(Y_i - \hat{\mu}_1(\mathbf{X}_i))}{\hat{\pi}(\mathbf{X}_i)}}^{\text{IPW}}\right]}_{\text{estimator for } \mathbb{E}[Y_i(1)]} - \underbrace{\left[\hat{\mu}_0(\mathbf{X}_i) + \frac{(1-D_i)(Y_i - \hat{\mu}_0(\mathbf{X}_i))}{1-\hat{\pi}(\mathbf{X}_i)}\right]}_{\text{estimator for } \mathbb{E}[Y_i(0)]}\end{aligned}$$

This is the **Augmented-Inverse-Propensity Weighting Estimator** (AIPW) introduced by Robins, Rotnitzky, and L. P. Zhao (1994) and Hahn (1998). Additional overviews:(Bang and Robins, 2005; Chernozhukov, Chetverikov, et al., 2018). General double-robustness property also shared by *targeted maximum-likelihood estimators*(TMLE) - due to Van Der Laan and D. Rubin (2006).
Similarly, analogous estimator for ATT

$$\widehat{\tau}^{\text{att}}_{\text{aipw}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Y_i - \widehat{\mu}_0(\mathbf{X}_i))D_i}{\widehat{\rho}} - \frac{\widehat{\pi}(\mathbf{X}_i)(1-D_i)\,(Y_i - \widehat{\mu}_0(\mathbf{X}_i))}{\widehat{\rho}(1-\widehat{\pi}(\mathbf{X}_i))}\right)$$

where $\rho = \mathbf{Pr}(D_i = 1)$ and $\widehat{\rho}$ is its empirical analogue.

Defn 4.20 (Cross-Fit AIPW).
The **Cross-fit** version can be stated as

$$\begin{aligned}\widehat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\;& \widehat{\mu}^{-k(i)}_{(1)}(\mathbf{X}_i) - \widehat{\mu}^{-k(i)}_{(0)}(\mathbf{X}_i) \\ &+ D_i\frac{y_i - \widehat{\mu}^{-k(i)}_{(1)}(\mathbf{X}_i)}{\widehat{\pi}^{-k(i)}(\mathbf{X}_i)} - (1-D_i)\frac{y_i - \widehat{\mu}^{-k(i)}_{(0)}(\mathbf{X}_i)}{1-\widehat{\pi}^{-k(i)}(\mathbf{X}_i)}\end{aligned}$$

where $k(i)$ is a mapping that takes an observation and puts it into one of the k folds. $\widehat{\mu}^{-k(i)}_{(1)}$ is an estimator excluding the k^{th} fold.
Define individual treatment effect score as

$$\widehat{\Gamma}_i = \widehat{\mu}_{(1)}(\mathbf{X}_i) - \widehat{\mu}_{(0)}(\mathbf{X}_i) + \frac{D_i(Y_i - \widehat{\mu}_{(1)}(\mathbf{X}_i))}{\widehat{\pi}(\mathbf{X_i})} - \frac{(1-D_i)(Y_i - \widehat{\mu}_{(0)}(\mathbf{X}_i))}{1-\widehat{\pi}(\mathbf{X}_i)}$$

Then, $\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\Gamma}_i$
We can form level-α CIs $\mathcal{I}_\alpha$:

$$\mathcal{I}_\alpha = \widehat{\tau} \pm z_{1-\alpha/2}\widehat{V}^{\frac{1}{2}} \;\; ; \widehat{V} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\widehat{\Gamma}_i - \widehat{\tau})^2$$

`grf` has a forest-based implementation of AIPW

```
cf = causal_forest(X, Y, D)
ate_hat = average_treatment_effect(cf)
```

Defn 4.21 (Double Selection Estimator for High-Dimensional Controls).
Belloni, Chernozhukov, and Hansen (2014) and Chernozhukov, Chetverikov, et al. (2018) partially-linear setup

$$\begin{aligned} y_i &= d_i\tau + g(\boldsymbol{x}_i) + \varepsilon_i & \mathbb{E}\left[\varepsilon_i|\boldsymbol{x}_i, d_i\right] &= 0 \\ d_i &= m(\boldsymbol{x}_i) + \eta_i & \mathbb{E}\left[\eta_i|\boldsymbol{x}_i\right] &= 0 \end{aligned}$$

where d_i is a scalar treatment indicator. Observations are independent but not necessarily identically distributed. We are interested in inference about τ that is robust to mistakes in model-selection.
Approximate g and m with linear combinations of control terms $\boldsymbol{c}_i = P(\boldsymbol{x}_i)$, which may contain interactions and non-linear transformations.
Assume *approximate sparsity* (:= there are only a small number of relevant controls, and irrelevant controls have a high probability of being small).
Naive (incorrect) approach: use LASSO on an eqn of the form

$$y_i = \tau D_i + \boldsymbol{x}_i'\beta + \varepsilon_i \text{ with penalty } h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_j |\beta|_j$$

where the treatment τ is not penalised. This will mean we drop any control that is highly correlated with the treatment if the control is moderately correlated with the outcome. Then, if we use a post-LASSO selection to estimate the treatment effect, the effect will be contaminated with an omitted variable bias.
recommended two-step approach

1. Estimate $y_i = \boldsymbol{c}_i'\boldsymbol{\beta} + \nu_i$ with LASSO, select predictive variables (i.e. those with nonzero coefficients) in $\mathcal{A}$
2. Estimate $d_i = \boldsymbol{c}_i'\boldsymbol{\beta} + \nu_i$ with LASSO, select predictive variables (i.e. those with nonzero coefficients) in $\mathcal{B}$
3. Estimate $y_i = \tau D_i + \boldsymbol{e}_i'\boldsymbol{\kappa} + \upsilon_i$ where $\boldsymbol{c}_i := \mathcal{A} \cup \mathcal{B}$ [i.e. control for variables that are selected in either the first or second regression]

Defn 4.22 (Post-double-selection estimator).
Let $\widehat{l}_1, \widehat{l}_2 \subset \{1, \ldots, p\}$ be the indices of the selected controls for the outcome and treatment respectively.
The post-double-selection estimator is

$$(\check{\tau}, \check{\boldsymbol{\beta}}) = \underset{\tau\in\mathbb{R}, \boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}} \left\{\mathbb{E}_n\left[(y_i - d_i\tau - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right] : \beta_j = 0 \; \forall j \notin \widehat{l}_1 \cup \widehat{l}_2\right\}$$

Can use plugin estimator for variance based on residuals

$$\sigma_n^{-1}\sqrt{\check{\tau} - \tau} \xrightarrow{d} \mathcal{N}(0,1) \implies \sigma_n^2 = \frac{\mathbb{E}\left[v_i^2\psi_i^2\right]}{\mathbb{E}\left[v_i^2\right]^2}$$

where

$$\begin{aligned} \widehat{\psi}_i &= (y_i - d_i\check{\tau} - \boldsymbol{x}_i'\check{\boldsymbol{\beta}})\sqrt{\frac{n}{n-\widehat{s}-1}} \\ \widehat{v}_i &= d_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\ \widehat{\beta} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\mathbb{E}_n\left[(d_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right] : \beta_j = 0 \; \forall j \notin \widehat{l}\right\} \end{aligned}$$

Implemented in `hdm::rlassoEffect(., 'double selection')`

Defn 4.23 (Double-ML - General case with moment conditions).
Let the target parameter τ_0 solve the equation $\mathbb{E}\left[m(Z_i, \tau_0, \beta_0)\right] = 0$ for known score function m, vector of observables $Z_i := \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n$, and nuisance parameter β_0. In fully parametric models, m is simply the score function [derivative of the log-likelihood]. For ATE, $m(Z_i, \tau, \beta) := (Y_i - D_i\tau - \boldsymbol{X}_i'\beta)D_i$.
In naive double-ML settings, $\mathbb{E}\left[\partial_\beta m(Z_i, \tau_0, \beta_0) = \pi_0\mu_1 \neq 0\right]$. So, we replace m with the **Neyman-orthogonal** score ψ s.t.

$$\mathbb{E}\left[\partial_\eta \psi(Z_i, \tau_0, \eta_0) = 0\right].$$

which yields the **Orthogonalised Moment Condition** $\mathbb{E}\left[\psi(Z_i, \tau_0, \eta_0)\right] = 0$ for some real-valued condition $\psi(.)$.
Using a Neyman-orthogonal score eliminates first-order biases arising from the replacement of η_0 with $\widehat{\eta}_0$.

Defn 4.24 (Orthogonal Scores).
Reference: Bach et al. (2021)
Consider data $\mathcal{W} := (Y, D, X)$ with $D \in \{0, 1\}$
Partial linear setup $Y = D\theta_0 + g_0(X) + U;\ D = m_0(X) + V$.
Score function is

$$\psi(\mathcal{W}; \theta, \eta) = (Y - \underbrace{l(X)}_{\mathbb{E}[Y|X]} - \theta(D - \underbrace{m(X)}_{\mathbb{E}[D|X]}))(D - m(X))$$

Partially Linear IV

$$Y - D\theta_0 = g_0(X) + \zeta,\ \mathbb{E}(\zeta \mid Z, X) = 0$$
$$Z = m_0(X) + V,\ \mathbb{E}(V \mid X) = 0$$

Score is

$$\psi(\mathcal{W}, \theta, \eta) := (Y - \underbrace{l(X)}_{\mathbb{E}[Y|X]} - \theta(D - \underbrace{r(X)}_{\mathbb{E}[D|X]}))(Z - \underbrace{m(X)}_{\mathbb{E}[Z|X]})$$

Interactive Regression

$$Y = \underbrace{g_0(D, X)}_{\mathbb{E}[Y|D,X]} + \varepsilon,\ \ \mathbb{E}[\varepsilon|D, X] = 0$$
$$D = \underbrace{m_0(X)}_{\mathbb{P}[D=1|X]} + \xi,\ \ \mathbb{E}[\xi|X] = 0$$

Here, the estimands are

$$\theta_0^{ATE} = \mathbb{E}[g_0(1, X) - g_0(0, X)]$$
$$\theta_0^{ATT} = \mathbb{E}[g_0(1, X) - g_0(0, X)|D = 1]$$

The score function for ATE (Hahn (1998))

$$\psi^{ATE}(Z_i, \theta, \eta) = (g(1, X) - g(0, X)) + \frac{D(Y - g(1, X))}{m(X)} - \frac{(1 - D)(Y - g(0, X))}{1 - m(X)} - \theta$$

The nuisance parameter true value is $\eta_0 = (g_0, m_0)$. For ATET,

$$\psi^{ATT}(Z_i, \theta, \eta) = \frac{1}{\pi_0}\left(D - \frac{m(X)}{1 - m(X)}(1 - D)\right)(Y - g(0, X)) - \frac{D}{\pi_0}\theta$$

Defn 4.25 (Cross-fitting Double-ML).
1. Take a K-fold random partition $(I_k)_{k=1,\dots,K}$ of observation indices $\{1, \dots, n\}$ s.t. each fold I_k has size n/k. For each k, define $I_k^C := \{1, \dots, n\}\ I_k$ as the complement / auxilliary sample.
2. For each $k \in \{1, \dots, K\}$, construct a ML estimator of η_0 using only the auxilliary sample I_k^C; $\widehat{\eta}_k = \hat{\eta}((Z_i)_{i \in I_k^C})$
3. For each $k \in \{1, \dots, K\}$, using the main sample I_k, construct the estimator $\check{\tau}_k$ as the solution of
$$\frac{1}{n/K}\sum_{i \in I_k} \psi(Z_i, \check{\tau}_k, \hat{\eta}_k) = 0$$
4. Aggregate the estimators $\check{\tau}_k$ on each main sample $\check{\tau} = \frac{1}{K}\sum_{k=1}^{K} \check{\tau}_k$

Example 4.10 (Sample-Splitting for Treatment Effects).
Simple implementation of Cross-fitting for Treatment effects

1. Partition the data in two, such that each fold I_1, I_2 has size $n/2$.
2. Using only sample I_1, construct a ML estimator of $g(0, X)$ and $m(X)$,e.g. a feedforward nnet of Y_i on X_i, denoted as $\widehat{g_{I_1}}(x)$, and logit-lasso of D_i on X_i, denoted by $\widehat{m_{I_1}}(x)$.
3. Use the estimators on the hold-out sample I_2 to compute the T.E
$$\check{\tau}_{I_2} = \frac{1}{\sum_{i \in I_2} D_i}\left[D_i - \frac{\widehat{m}_{I_1}(X_i)}{1 - \widehat{m}_{I_1}(X_i)}(1 - D_i)\right](Y_i - \widehat{g}_{I_1}(X_i))$$
4. Repeat (2,3) swapping the roles of I_1 and I_2 to get $\check{\tau}_{I_1}$
5. Aggregate the estimators:
$$\check{\tau} = \frac{\check{\tau}_{I_1} + \check{\tau}_{I_2}}{2}$$

Implemented in `DoubleML`

4.2.8 Augmented Balancing

Loosely: AIPW without the (potentially fraught) inversion of the propensity score step. Exposition based on Bruns-Smith et al (2023)
setup:

- Covariates $X \in \mathcal{X} \subseteq \mathbb{R}^k$, $Y \in \mathbb{R}$ outcome, two populations p and q that are distributions over (X, Y)
  - p is 'source', q is 'target' (e.g. treatment group and overall sample)
- Estimand is $\mathbb{E}_q[Y]$
- **Identification Assumptions**
  1. Conditional Mean Ignorability: $\mathbb{E}_p[Y|X] = \mathbb{E}_q[Y \mid X]$
  2. Population Overlap: $q(x)$ is absolutely continuous w.r.t. $p(x)$

Effect Functionals
Regression Functional

$$\mathbb{E}_q\left[\mathbb{E}_p[Y \mid X]\right] = \mathbb{E}_q\left[\mathbb{E}_q[Y \mid X]\right] = \mathbb{E}_q[Y]$$

Weighting Functional

$$\mathbb{E}_p\left[\frac{dq}{dp}(X)Y\right] = \mathbb{E}_p\left[\frac{dq}{dp}(X)\mathbb{E}_p[Y \mid X]\right] = \mathbb{E}_q\left[\mathbb{E}_p[Y \mid X]\right] = \mathbb{E}_q[Y]$$

Doubly-Robust Functional

$$\mathbb{E}_q\left[\mathbb{E}_p[Y \mid X]\right] + \mathbb{E}_p\left[\frac{dq}{dp}(X)\left\{Y - \mathbb{E}_p[Y \mid X]\right\}\right]$$

Balancing Weights: Rationale

- $\frac{dq}{dp}(X)$ is difficult to estimate using plug-in estimation
- Alternative: weighting for balance ≡ automatic estimation of the Riesz representer

Weighting to minimise covariate imbalance

$$\min_w \left\{ \overbrace{\sup_{f \in \mathcal{F}} \mathbb{E}_p[w(X)f(X)] - \mathbb{E}_q[f(X)]}^{\text{Imbalance}} + \delta \|w\|_2^2 \right\}$$

Direct estimation of the density ratio

$$\min_{f \in \mathcal{F}} \left\{ \mathbb{E}_p\left[\left(f(X) - \frac{dq}{dp}(X)\right)^2\right] \right\}$$

Minimum variance weights that balance $\mathcal{F}$ are also guaranteed to balance *all other* measurable functions in $\mathcal{F}$.

Defn 4.26 (Linear Balancing Weights).
- In linear setting, relevant imbalance is captured entirely by *feature mean* imbalance
- X_p, Y_p are n iid draws from p, X_q are m draws from q
- Define feature map $\phi : \mathcal{X} \to \mathbb{R}^d$; construct gram matrices
  - $\Phi_p := \phi(X_p)$
  - $\overline{\Phi}_q := \widehat{\mathbb{E}}_q[\Phi_q]$
- Let $\mathcal{F} = \left\{f(x) = \theta^\top \phi(x) : \|\theta\| \leq r\right\}$
  - Let $\|\cdot\|_*$ denote dual norm $[\ell_2 : \ell_2, \ell_1 : \ell_\infty]$

$$\widehat{\text{Imbalance}}_{\mathcal{F}(w)} = \left\|w\Phi_p - \overline{\Phi}_q\right\|_*$$

Three Equivalent Representations

$$\begin{aligned}
\text{Penalised Form}: &\min_{w \in \mathbb{R}^n} \left[\left\|w\Phi_p - \overline{\Phi}_q\right\|_*^2 + \delta_1 \|w\|_2^2\right] \\
\text{Constrained form}: &\min_{w \in \mathbb{R}^n} \|w\|_2^2 \quad \text{s.t.} \quad \left\|w\Phi_p - \overline{\Phi}_q\right\| \leq \delta_2 \\
\text{Automatic Form}: &\min_{\theta \in \mathbb{R}^d} \left\{\theta^\top(\Phi_p^\top\Phi_p)\theta - 2\theta^\top\overline{\Phi}_q + \delta_3\|\theta\|\right\}
\end{aligned}$$

Example 4.11 (Equivalence Example : OLS is DR).
OLS is equivalent to a weighting estimator that exactly balances the feature means. Let $\widehat{\beta}_{\text{OLS}} = (\Phi_p^\top\Phi_p)^{-1}\Phi_p^\top Y_p$ be the linear regression fit on p (source sample). Then,

$$\begin{aligned}
\widehat{\mathbb{E}}_q[\Phi_q\widehat{\beta}_{\text{OLS}}] &= \widehat{\mathbb{E}}_p[\widehat{w}_{\text{exact}} \circ Y_p] \\
\widehat{\mathbb{E}}_q[\Phi_q(\Phi_p^\top\Phi_p)^{-1}\Phi_p^\top Y_p] &= \widehat{\mathbb{E}}_p[\overline{\Phi}_q(\Phi_p^\top\Phi_p)^{-1}\Phi_p^\top \circ Y_p]
\end{aligned}$$

Analogue for Ridge

$$\widehat{\mathbb{E}}_q[\Phi_q(\Phi_p^\top\Phi_p + \delta I)^{-1}\Phi_p^\top Y_p] = \widehat{\mathbb{E}}_p[\overline{\Phi}_q(\Phi_p^\top\Phi_p + \delta I)^{-1}\Phi_p^\top \circ Y_p]$$

Proposition 4.12 (Augmented Balancing as Undersmoothed Regression).
$\forall \widehat{\beta}^\lambda_{reg} \in \mathbb{R}^d$, and any linear balancing weight estimator with estimated coefficients $\theta^d \in \mathbb{R}^d$, $\widehat{w}^\delta = \widehat{\theta}\Phi_p^\top$, and $\widehat{\Phi}_q^\delta = \widehat{w}\Phi_p$

$$\widehat{\mathbb{E}}_q[\Phi_q\widehat{\beta}^{\lambda}_{reg}] + \widehat{\mathbb{E}}_p[\widehat{w}^{\delta} \circ (Y_p - \Phi_p\widehat{\beta}^{\lambda}_{reg})] = \widehat{\mathbb{E}}_p[\widehat{w}^{\delta} \circ Y_p] + \widehat{\mathbb{E}}_q\left[\left(\overline{\Phi}_q - \widehat{\Phi}^{\delta}_q\right)\widehat{\beta}^{\lambda}_{reg}\right]$$
$$= \widehat{\mathbb{E}}_q\left[\widehat{\Phi}^{\delta}_q\widehat{\beta}_{OLS} + \left(\overline{\Phi}_q - \widehat{\Phi}^{\delta}_q\right)\widehat{\beta}^{\lambda}_{reg})\right] = \widehat{\mathbb{E}}_q[\Phi_q\widehat{\beta}_{aug}]$$

$$\widehat{\beta}_{aug,j} := (1 - a^{\delta}_j)\widehat{\beta}^{\lambda}_{reg,j} + a^{\delta}_j\widehat{\beta}_{ols,j} \quad \text{where } a^{\delta}_j := \frac{\widehat{\Phi}^{\delta}_{q,j} - \overline{\Phi}_{p,j}}{\overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}}$$

In words: when both outcome and weighting models are linear, the augmented estimator is equivalent to a linear model with coefficients that are element-wise affine combinations of **base learner** $\widehat{\beta}^{\lambda}_{reg}$ and **coefs** $\widehat{\beta}_{OLS}$ **from regressing** Y_p **on** Φ_p

4.2.9 Heterogeneous Treatment Effects with selection on observables

Conditional Average Treatment Effects (CATEs) $(\tau(x) = \mathbb{E}\left[Y^1 - Y^0 | \boldsymbol{X} = \boldsymbol{x}\right])$ are often of great policy interest for targeting those who have largest potential gains. However, conventional methods are prone to a severe risk of fishing from researchers (cf 'conditional effects' in most published work in the social sciences). Instead, recent work proposes to use nonparametric estimators to find subgroups, use sample-splitting for *honesty*.

1. *transformed outcome regression* use outcome transformed w pscore

$$H = \frac{DY}{p(\boldsymbol{X})} - \frac{(1-D)Y}{1-p(\boldsymbol{X})}$$

2. *conditional mean regression* use the fact that under SOO

$$\tau(\boldsymbol{x}) = \mathbb{E}\left[Y_1 | \boldsymbol{X} = \boldsymbol{x}\right] - \mathbb{E}\left[Y_0 | \boldsymbol{X} = \boldsymbol{x}\right] = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x})$$

(1) typically inefficient because of pscore in denominator, so most focus is on (2). Random forests are a flexible method that is widely liked.

Defn 4.27 (Robinson Semiparametric Setup (Robinson, 1988)).
Consider a model for $\tau(x)$ where

$$Y_i(d) = f(\boldsymbol{X}_i) + d \cdot \tau(\boldsymbol{X}_i) + \varepsilon(d), \ \mathbb{P}\left[d_i | \boldsymbol{X}_i\right] = e(x)$$

where $\tau(\boldsymbol{x}) = \psi(\boldsymbol{x})\boldsymbol{\beta}$ for some pre-determined set of basis functions $\psi : \mathcal{X} \to \mathbb{R}^k$. We allow for non-parametric relationships between $\boldsymbol{X}_i, y_i, d_i$, but the treatment effect function itself is parametrised by $\beta \in \mathbb{R}^k$. Robinson (1988) showed that under unconfoundedness, we can rewrite the semiparametric setup above as

$$Y_i - m(\boldsymbol{X}_i) = (d_i - e(\boldsymbol{X}_i))\psi(\boldsymbol{X}_i) \cdot \beta + \varepsilon_i \qquad \text{where}$$
$$m(\mathbf{x}) = \mathbb{E}\left[Y_i | \boldsymbol{X}_i = \mathbf{x}\right] = f(\boldsymbol{X}_i) + e(\boldsymbol{X}_i)\tau(\boldsymbol{X}_i)$$

The *oracle* algorithm for estimating $\boldsymbol{\beta}$ is (1) define $\widetilde{Y}^*_i = Y_i - m(\boldsymbol{X}_i)$ and $\widetilde{Z}^*_i = (d_i - e(\boldsymbol{X}_i)\psi(\boldsymbol{X}_i)$, then estimate residuals-on-residual regression. This procedure is $\sqrt{n}$-consistent and asymptotically normal.
Use cross-fitting to emulate the Oracle.

1. Run non-parametric regressions $Y \sim \boldsymbol{X}$ and $D \sim \boldsymbol{X}$ to get $\widehat{m}(\boldsymbol{x}), \widehat{e}(\boldsymbol{x})$
2. define transformed features $\widetilde{Y}_i = Y_i - \widehat{m}^{-k(i)}(\boldsymbol{X}_i)$, $\widetilde{Z} = (D_i - \widehat{e}^{-k(i)}(\boldsymbol{X_i})\psi(\boldsymbol{X_i})$
3. Estimate $\widehat{\zeta}_b$ by regressing $\widetilde{Y}_i \sim \widetilde{Z}_i$

Defn 4.28 (R-Loss).
To define R-Loss (Athey, J. Tibshirani, and Wager, 2019), under more general setup restate unconfoundedness as follows

$$\mathbb{E}\left[\varepsilon_i(d_i) | \boldsymbol{X}_i, d_i\right] = 0 \text{ where } \varepsilon_i(d) := Y_i(d) - (\mu_{(0)}(\boldsymbol{X}_i) + w\tau(\boldsymbol{X}_i))$$

and follow Robinson's approach to write

$$Y_i - m(\boldsymbol{X}_i) = (D_i - e(\boldsymbol{X}_i))\tau(\boldsymbol{X}_i) + \varepsilon_i$$

R-loss is then written

$$\tau(\cdot) = \underset{\tau'}{\operatorname{argmin}} \left\{\mathbb{E}\left[\left((Y_i - m(\boldsymbol{X}_i) - (d_i - e(\boldsymbol{X}_i))\tau'(\boldsymbol{X}_i)\right)^2\right]\right\}$$

Defn 4.29 (R-Learner, Athey, J. Tibshirani, and Wager (2019)).
Define $e(x) = \mathbf{Pr}\left(D = 1 | X = x\right)$ and $m(x) = \mathbb{E}\left[Y | X = x\right]$ The R-learner consists of the following steps

1. Use any method to estimate the response functions $\widehat{e}(x), \widehat{m}(x)$
2. Minimise R-loss using cross-fitting for nuisance components

$$\widehat{\tau}(.) = \underset{\tau}{\operatorname{argmin}} \left(\left(Y_i - \widehat{m}(\boldsymbol{X_i})\right) - \tau(\boldsymbol{X_i})(D_i - \widehat{e}(\boldsymbol{X_i}))\right)^2 + \Lambda_n(\tau(\cdot))$$

where Λ_n is some regulariser.

Causal forest as implemented by `grf` starts by fitting two separate trees to estimate $\widehat{m}, \widehat{e}$, makes out-of-bag predictions [using cross-fitting] using the two first-stage forests, then grow causal forest via

$$\widehat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(Y_i - \widehat{m}(\boldsymbol{X_i})(D_i - \widehat{e}(\boldsymbol{X_i})))}{\sum_{i=1}^{n} \alpha_i(x)(D_i - \widehat{e}(\boldsymbol{X_i}))^2}$$

where

$$\alpha_i(x) = \frac{1}{B}\sum_b \frac{\mathbb{1}_{\boldsymbol{X_i} \in \mathcal{L}_b(x), i \in B}}{|i : X_i \in \mathcal{L}_b(x), i \in B|}$$

are the learned adaptive weights.

Defn 4.30 (Double-sample / Honest trees (Athey and Imbens, 2016a)).

1. Draw a subsample of size s from the sample with replacement and divide it into disjoint sets $\mathcal{I}, \mathcal{J}; |\mathcal{I}| = |\mathcal{J}| = n/2$.
2. Grow a tree via recursive partitioning, with splits chosen from $\mathcal{J}$ (i.e. without using Y observations from $\mathcal{I}$ sample)
3. Estimate leaf responses using only $\mathcal{I}$ sample

Finally, aggregate all trees over subsamples of size s

$$\begin{aligned}\widehat{\mu}(\boldsymbol{x}, \boldsymbol{Z}_i, \dots, \boldsymbol{Z}_n) &= \binom{n}{s}^{-1} \sum_{1 \le i_1 < \dots < i_s \le n} \mathbb{E}_{\xi \in \Xi}\left[T(\boldsymbol{x}, \xi, \boldsymbol{Z}_{i_1}, \dots, \boldsymbol{Z}_{i_s})\right] \\ &\approx \frac{1}{B}\sum_{b=1}^{B} T(\boldsymbol{x}, \xi_b^*, \boldsymbol{Z}_{b,1}^*, \dots, \boldsymbol{Z}_{b,s}^*) \qquad \text{Bagging}\end{aligned}$$

where ξ summarises randomness in the selection of the variable when growing the tree, $\boldsymbol{Z}_i := (D_i, \boldsymbol{X}_i, Y_i)$ is shorthand for a training sample.
where the base learner

$$T(\boldsymbol{x}; \xi_b^*, \boldsymbol{Z}_{b,1}^*, \dots, \boldsymbol{Z}_{b,s}^*) = \sum_{i \in \{i_{b,1}, \dots i_{b,s}\}} \alpha_{i,b}^*(\boldsymbol{x}) Y_{i,b}^* \;\; ; \alpha_{i,b}^*(\boldsymbol{x}) = \frac{\mathbb{1}_{\boldsymbol{X}_{i,b}^* \in \mathcal{L}_b^*(\boldsymbol{x})}}{\left|i : \boldsymbol{X}_{i,b}^* \in \mathcal{L}_b^*(\boldsymbol{x})\right|}$$

the 'honesty' property is making $\alpha_{i,b}^*(\boldsymbol{x})$ independent of $Y_{i,b}^*$, i.e. do not use the same data to select partition (splits) and make predictions.
Implemented in `causalForest` and `grf`.

4.2.10 Multi-action policy learning

$i = 1, \dots, N$ units, to be assigned to $J+1$ actions $A_i \in \{0, 1, \dots, J\} =: \mathcal{A}$, which has have corresponding rewards $\{Y_i^{(0)}, Y_i^{(1)}, \dots, Y_i^{(J)}\}$. Each observation has covariate $\mathbf{X}_i \in \mathcal{X} \subseteq \mathbb{R}^d$. Define a policy function

$$\pi : \; \mathbf{x} \to \mathcal{A}$$

A given policy assigns each unit to a treatment level. Each policy has a corresponding value function

$$V(\pi) := \mathbb{E}\left[Y(\pi(\mathbf{x}))\right]$$

An optimal policy $\pi^* \in \Pi$ is defined as

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}\left[Y(\pi)\right]$$

Deviations from this optimum is called regret

$$R(\pi) = \mathbb{E}\left[Y(\pi^*)\right] - \mathbb{E}\left[Y(\pi)\right] = V(\pi^*) - V(\pi)$$

Define a CEF as

$$\mu_i(a, \mathbf{x}_i) = \mathbb{E}\left[Y_i^{(a)} \mid \mathbf{x}_i\right]$$

The first-best optimal rule is

$$\pi_i(\mathbf{x}_i) = \arg\max_{a \in \mathcal{J}} \{\mu_i(a, \mathbf{x}_i)\}$$

In the binary action case, this simplifies to $\pi(\mathbf{x}_i) = 1\{\mu(1, \mathbf{x}_i) \ge \mu(0, \mathbf{x}_i)\} = 1\{\tau(x) > 0\}$ which is the *conditional empirical success* (*CES*) rule of Manski (2004).
Under unconfoundedness and Overlap, we can estimate $\hat{\mu}$s and construct an empirical analogue of the value function for a policy π using the following familiar estimators

$$\begin{aligned}\widehat{V}_{RA}(\pi) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i(\pi(\mathbf{x}_i), \mathbf{x}_i) \\ \widehat{V}_{IPW}(\pi) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1[A_i = \pi(\mathbf{x}_i)]}{\widehat{p}_{A_i}(\mathbf{X}_i)} Y_i \\ \widehat{V}_{AIPW}(\pi) &= \frac{1}{n}\sum_{i=1}^{n} \left[\widehat{\mu}_i(\pi(\mathbf{x}_i), \mathbf{x}_i) + \frac{1[A_i = \pi(\mathbf{x}_i)]}{\widehat{p}_{A_i}(\mathbf{x}_i)}\right]\end{aligned}$$

A $\sqrt{n}$-convergent estimator of the value function is the Cross-fit Augmented Inverse Probability Weighted Learning (CAIPWL) estimator of Zhou, Athey, and Wager (2018), which is constructed as a cross-fit analogue of the AIPW estimator.

4.2.11 Sensitivity Analysis

Defn 4.31 (Standardised Difference).
Check balance by computing SDiff for observable confounders

$$\text{Standardised Difference} = \frac{\overline{X}_t - \overline{X}_c}{\sqrt{(s_t^2 + s_c^2)/2}}$$

Example 4.13 (Multiple control groups).
Three valued treatment indicator: $T_i \in \{-1, 0, 1\}$ corresponding with ineligibles, eligible nonparticipants, and participants. We can test unconfoundedness by comparing ineligibles with eligible nonparticipants, i.e. test

$$Y_i \perp\!\!\!\perp \mathbf{1}\{T_i = 0\} | X_i, T_i \in \{-1, 0\}$$

Placebo Outcomes
Covariates included lagged outcomes $Y_{i,-1}, \dots Y_{i,-T}$. Test

$$Y_{i,-1} \perp\!\!\!\perp D_i | Y_{i,-2}, Y_{i,-T}, X_i$$

e.g. Earnings in 1975 in Lalonde

Defn 4.32 (Parametric Sensitivity Analysis (Imbens (2003)).
U is a nuisance parameter.

$$Y_1, Y_0 \perp\!\!\!\perp D | X, U$$

Where $U \sim \mathcal{B}(\pi = 0.5)$, and $U \perp\!\!\!\perp X$. $P(U = 1) = P(U = 0) = 0.5$.
Propensity score is Logistic:

$$P(D = 1|X, U) = \frac{\exp(X\theta + \gamma U)}{1 + \exp(X\theta + \gamma U)}$$

γ indicates strength of relationship between U and $D|X$.
Y is conditionally normal

$$Y|X, U \sim \mathcal{N}(\alpha D + X\beta + \delta U, \sigma^2)$$

δ indicates strength of relationship between U and $Y|X$.
MLE setup
Construct grid of (γ, δ) and calculate the MLE for $\hat{\alpha}(\gamma, \delta)$ by maximising $l(\alpha, \beta, \theta, \gamma, \delta)$ over (γ, δ).
Use 2 partial R^2s:

- $R^2_{Y,par}(\delta)$: Residual variation in outcome explained by U (after partialling out X).
- $R^2_{D,par}(\gamma)$: Residual variation in treatment assignment explained by U (after partialling out X).

Draw threshold contours, should expect most covariates to be clustered around origin.

Rosenbaum (2002)
Tuning parameter $\Gamma \geq 1$ that measures departure from zero hidden bias.
For any two observations i and j with identical covariate values $X_i = X_j$, under unconfoundedness, probability of assignment into treatment should be identical $\pi(X_i) = \pi(X_j)$
Treatment assignment probability may differ due to unobserved binary confounder U. We can bound this by the ratio:

$$\frac{1}{\Gamma} \leq \frac{\widehat{\pi}_i(1 - \widehat{\pi}_j)}{(1 - \widehat{\pi}_i)\widehat{\pi}_j} \leq \Gamma$$

$\gamma = 1 \implies$ No bias. $\Gamma = 2 \implies$ i is twice as likely to be treated than j despite identical x.
Γ is assumed to satisfy

$$\frac{1}{\Gamma} \leq \frac{\mathbf{Pr}(D = d|X = x)/(1 - \mathbf{Pr}(D = d|X_{=}x))}{\mathbf{Pr}(D = d|X = x, Y(d) = y)/(1 - \mathbf{Pr}(D = d|X = x, Y(d) = y))} \leq \Gamma$$

For any given candidate $\Gamma > 1$, estimates of the treatment effect can be computed.
Implemented in `rbounds::hlsens`.

Defn 4.33 (Coefficient Stability Approaches).
Altonji, Elder, Taber (2005)
Only informative if selection on observables is informative about selection on unobservables.
How much does treatment effect *move* when controls are added? Estimate model with and without controls:

- $Y_i = \alpha^F D_i + X\beta + \epsilon$
- $Y_i = \alpha^R D_i + \epsilon$

AET ratio: $\rho = \frac{\hat{\alpha}^F}{\hat{\alpha}^R - \hat{\alpha}^F}$
Want ρ to be as big as possible (i.e. $\hat{\alpha}^R - \hat{\alpha}^F \to 0$ under unconfoundedness).

Defn 4.34 (Oster (2019) - Proportional selection coefficient).
Define proportional selection coefficient

$$\delta = \frac{\text{Cov}[\epsilon, D]}{\mathbb{V}[\epsilon]} / \frac{\text{Cov}[X'\gamma]}{\mathbb{V}[X'\gamma]}$$

Then,

$$\beta^* \approx \tilde{\beta} - \delta\left[\dot{\beta} - \tilde{\beta}\right]\frac{R_{max} - \tilde{R}}{\tilde{R} - \dot{R}} \xrightarrow{p} \beta$$

where

- $\dot{\beta}, \dot{R}$ are from a univariate regression of Y on T
- $\tilde{\beta}, \tilde{R}$ are from a regression including controls
- R_{max} is maximum achievable R^2

Defn 4.35 (q-Robustness Value (Cinelli and Hazlett, 2020)).
True model is $Y = \tau D + \mathbf{X}\boldsymbol{\beta} + \gamma Z + \varepsilon$, but we don't observe Z. We would like to quantify how biased the coefficient from the short regression $\widehat{\tau}_s$ is for the long regression coefficient τ. From OVB FOrmula, we know $\widehat{\tau}_s = \widehat{\tau} + \widehat{\gamma} + \widehat{\delta}$ where $\widehat{\gamma}$ is the conditional association between the omitted Z and Y ('impact') and $\widehat{\delta}$ is the coefficient from regressing Z on D ('imbalance').
The bias from this omission is

$$\left|\widehat{\text{Bias}}\right| = \sqrt{\left(\frac{R^2_{Y\sim Z|D,\mathbf{X}} R^2_{D\sim Z|\mathbf{X}}}{1 - R^2_{D\sim Z|\mathbf{X}}}\right)\frac{\text{sd}(Y^{\perp\mathbf{X},D})}{\text{sd}(D^{\perp\mathbf{X}})}}$$

They then define

$$\text{RV}_q = \frac{1}{2}\left[\sqrt{f_q^4 + 4f_q^2} - f_q^2\right]$$

where $f_q := q\left|f_{Y\sim D|\mathbf{X}}\right|$ where $f_{Y\sim D|\mathbf{X}}$ is the partial Cohen's f of the treatment with the outcome, and q is the proportion of reduction on the treatment coefficient τ that would be deemed problematic.

4.2.12 Partial Identification

the ATE can be decomposed as

$$\begin{aligned}\text{ATE} &= \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] \\ &= \mathbb{E}\left[Y_i(1)|D_i = 1\right]\mathbf{Pr}\left(D_i = 1\right) \; + \; \mathbb{E}\left[Y_i(1)|D_i = 0\right]\mathbf{Pr}\left(D_i = 0\right) \\ &- \mathbb{E}\left[Y_i(0)|D_i = 1\right]\mathbf{Pr}\left(D_i = 1\right) + \mathbb{E}\left[Y_i(0)|D_i = 0\right]\mathbf{Pr}\left(D_i = 0\right)\end{aligned}$$

The terms in red are counterfactual outcomes for which the data contains no information. Bounding approaches involve estimators for these missing quantities.

Defn 4.36 (Agnostic Bounds).
Suppose all we know is $Y^d \in [0, 1]$
w.l.o.g. given bounded support $[\underline{Y}, \overline{Y}]$, we can always min-max rescale to $\frac{Y - \underline{Y}}{\overline{Y} - \underline{Y}}$

$$\begin{aligned}\mathbb{E}\left[Y^1 - Y^0\right] \in &[\{\mathbb{E}\left[Y|D = 1\right]\mathbf{Pr}\left(D = 1\right) - \mathbb{E}\left[Y|D = 0\right]\left(1 - \mathbf{Pr}\left(D = 1\right)\right)\} - \mathbf{Pr}\left(D = 1\right), \\ &\{\mathbb{E}\left[Y|D = 1\right]\mathbf{Pr}\left(D = 1\right) - \mathbb{E}\left[Y|D = 0\right]\left(1 - \mathbf{Pr}\left(D = 1\right)\right)\} + \left(1 - \mathbf{Pr}\left(D = 1\right)\right)]\end{aligned}$$

Width of possible interval learnable from data is $[0, 1]$ at largest, $[-1, 0]$ at smallest, so worst case interval always contains 0. Need theory/assumptions to even get the sign right.

Defn 4.37 (Manski Bounds).
Assume bounded support for the outcome. Replace missing values with **maximum** (y^{UB}) or **minimum** (y^{LB}) of support. These are **worst-case** bounds and yield intervals that are basically uninformative.

$$\begin{aligned}\mathbb{E}\left[Y(1)\right]^{UB} &= \mathbb{E}\left[Y|D = 1\right]\mathbf{Pr}\left(D = 1\right) + y^{UB}\mathbf{Pr}\left(D = 0\right) \\ \mathbb{E}\left[Y(1)\right]^{LB} &= \mathbb{E}\left[Y|D = 1\right]\mathbf{Pr}\left(D = 1\right) + y^{LB}\mathbf{Pr}\left(D = 0\right) \\ \mathbb{E}\left[Y(0)\right]^{UB} &= y^{UB}\mathbf{Pr}\left(D = 1\right) + \mathbb{E}\left[Y|D = 0\right]\mathbf{Pr}\left(D = 0\right) \\ \mathbb{E}\left[Y(0)\right]^{LB} &= y^{LB}\mathbf{Pr}\left(D = 1\right) + \mathbb{E}\left[Y|D = 0\right]\mathbf{Pr}\left(D = 0\right)\end{aligned}$$

And denote $\Delta^{UB} := \mathbb{E}\left[Y(1)\right]^{UB} - \mathbb{E}\left[Y(0)\right]^{LB}$ $\Delta^{LB} := \mathbb{E}\left[Y(1)\right]^{LB} - \mathbb{E}\left[Y(0)\right]^{UB}$
Monotone Treatment Response: assume mean potential outcome under treatment cannot be lower than under control $\mathbb{E}\left[Y(1)\right] \geq \mathbb{E}\left[Y(0)\right] = \Delta \geq 0$. Then

$$\Delta^{LB} = \max(\mathbb{E}\left[Y(1)\right]^{LB} - \mathbb{E}\left[Y(0)\right]^{UB}, 0)$$

Monotone Treatment Selection: subjects select themselves into treatment in a way the mean potential outcomes of the treatment and control groups can be ordered. Positive MTS implies $\mathbb{E}\left[Y(1)|D = 1\right] \geq \mathbb{E}\left[Y(1)|D = 0\right]$ and $\mathbb{E}\left[Y(0)|D = 1\right] \geq \mathbb{E}\left[Y(0)|D = 0\right]$. This implies $\mathbb{E}\left[Y(0)\right]^{LB} = \mathbb{E}\left[Y|D = 0\right]$ and $\mathbb{E}\left[Y(1)\right]^{UB} = \mathbb{E}\left[Y|D = 1\right]$

Theorem 4.14 (Kolmogorov's Conjecture - Sharp bounds on treatment effects).
Let $\tau_i := Y_{1i} - y_{0i}$ denote the treatment effect and $\mathbb{F}$ denote its distribution, and let $\mathbb{F}_1, \mathbb{F}_0$ denote the distributions of outcomes for the two potential outcomes. Then, $\mathbb{F}^L(b) \leq \mathbb{F}(b) \leq \mathbb{F}^U(b)$ where

$$\mathbb{F}^L(b) = \max\left\{\max_y \mathbb{F}_1(y) - \mathbb{F}_0(y-b), 0\right\}$$
$$\mathbb{F}^U(b) = 1 + \min\left\{\min_y \mathbb{F}_1(y) - \mathbb{F}_0(y-b), 0\right\}$$

4.3 Instrumental Variables

SOO Fails/$\mathbb{E}[X_i\epsilon_i] \neq 0$ because of OVB, then $\hat{\beta}_{OLS}$ is no longer consistent. Use Z as instrument for D which isolates variation *unrelated to the omitted variable*.

4.3.1 Traditional IV Framework (Constant Treatment Effects)

Setup

- Second Stage: $Y = \alpha_0 + \alpha_1 D + u_2$
- First Stage: $D = \pi_0 + \pi_1 Z + u_1$
- Reduced Form:

$$\begin{aligned} Y &= \gamma_0 + \gamma_1 Z + u_3 \\ &= \alpha_0 + \alpha_1(\pi_0 + \pi_1 Z + u_1) + u_2 \\ &= (\alpha_0 + \alpha_1\pi_0) + \underbrace{(\alpha_1\pi_1)}_{\gamma_1} Z + (\alpha_1 u_1 + u_2) \end{aligned}$$

Assumption 4 (IV Assumptions).

- **Exogeneity** (as good as random conditional on covariates): $\text{Cov}[u_1, Z] = 0$
- **Exclusion Restriction**: $\text{Cov}[u_2, D] = 0$, Z has no effect on Y except through D.
- **Relevance**: Z affects D

With the above assumptions, we can write

Defn 4.38 (Instrumental Variables Estimator).

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{y}$$

This is equivalent to

Defn 4.39 (Wald Estimator).
With binary treatment and binary instrument, one can write the IV effect as

$$\alpha_1 = \frac{\gamma_1}{\alpha_1} = \frac{\text{Cov}[Y,Z]}{\text{Cov}[Y,D]} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}$$

Defn 4.40 (2SLS Estimator).
With multiple instruments or endogenous variables,

$$\widehat{\alpha}_{2SLS} = (\mathbf{X}'\mathbf{P}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_z\boldsymbol{y}$$

where $\mathbf{P}_z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is $\mathbf{X}$ projected in the column space of $\mathbf{Z}$.

Defn 4.41 (k-Class estimation).

$$\widehat{\alpha}_k = (\mathbf{X}'(\mathbf{I} - k\mathbf{P}_z)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - k\mathbf{P}_z)\boldsymbol{y}$$

which nests 2SLS, LIML, and Fuller's estimator as special cases. Specifically,

- $k = 0 \implies \widehat{\alpha}_k$ is OLS
- $k = 1 \implies \widehat{\alpha}_k$ is 2SLS
- $k = k_{\text{LIML}} \implies \widehat{\alpha}_k$ is LIML
- $k = k_{\text{LIML}} - \frac{b}{n-L-p}; b > 0 \implies \widehat{\alpha}_k$ is Fuller's estimator

here, k_{LIML} is the *minimum* value of k that satisfies

$$\det\begin{pmatrix} \mathbf{y}^\top(\mathbf{I} - k\mathbf{P}_z)\mathbf{y} & \mathbf{y}^\top(\mathbf{I} - k\mathbf{P}_z)\mathbf{X} \\ \mathbf{X}^\top(\mathbf{I} - k\mathbf{P}_z)\mathbf{y} & \mathbf{X}^\top(\mathbf{I} - k\mathbf{P}_z)\mathbf{X} \end{pmatrix} = 0$$

Implemented in `ivmodel`, which takes model fits from `AER::ivreg` and computes LIML / k-class estimates.
Asymptotically, all $k-$ class estimators are consistent for α when $k \to 1, n \to \infty$.

Inference

- Under homoscedasticity, $\mathbb{V}[\hat{\alpha}_{2SLS}] = \sigma^2(\mathbf{X}'\mathbf{P}_z\mathbf{X})^{-1}$
- Under heteroskedasticity,

$$\mathbb{V}(\hat{\beta}_{IV}) = (\mathbf{Z}'\mathbf{P}_z\mathbf{X})^{-1}\mathbf{P}_z\mathbf{X}'\hat{\Omega}\mathbf{P}_z\mathbf{X}(\mathbf{X}'\mathbf{P}_z\mathbf{X})^{-1}; \quad \hat{\Omega} = \text{Diag}[\hat{u}_i^2]$$

Defn 4.42 (Hausman test for exogeneity).
Test statistic and null distribution

$$H := \frac{(\hat{\beta}_{2sls} - \hat{\beta}_{ols})^2}{\widehat{V}(\hat{\beta}_{2sls}) - \widehat{V}(\hat{\beta}_{ols})} \sim \chi^2_1$$

Equivalently, **Assuming the instrument Z is valid**, we can test for whether $\boldsymbol{x}$ is endogenous by estimating the following regression

$$y_i = \boldsymbol{Z}_i'\boldsymbol{\pi} + x_i\alpha_1 + \hat{v}_i\rho_1 + \epsilon_i$$

where $\hat{v}$ are the (fitted) residuals from estimating the first stage regression $x_i = \boldsymbol{Z}_i'\boldsymbol{\psi} + v_i$. A standard t-test for ρ tests *whether x is exogenous assuming $\boldsymbol{Z}_i$ is a valid set of instruments*. [means this test is not that useful in practice]

4.3.2 Weak Instruments

$$\text{plim } \alpha_{IV} = \frac{\text{Cov}[Y,Z]}{\text{Cov}[Z,D]} + \frac{\text{Cov}[Z,u_2]}{\text{Cov}[Z,D]} = \alpha_D + \frac{\text{Cov}[Z,u_2]}{\text{Cov}[Z,D]}$$

Second term non-zero if instrument is not exogenous. Let $\sigma_{u1,u_2} = \text{Cov}[u_1,u_2]$ and $\sigma^2_{u_1} = \mathbb{V}[u_2]$ [variance of first stage error] and F be F statistic of the first-stage. Then, bias in IV is

$$\mathbb{E}[\hat{\alpha}_{IV} - \alpha] = \frac{\sigma_{u_1u_2}}{\sigma^2_{u_2}} \frac{1}{F+1}$$

If first stage is weak, bias approaches $\frac{\sigma_{u_1u_2}}{\sigma^2_{u_2}}$. As $F \to \infty$, $B_{IV} \to 0$.

Defn 4.43 (Anderson-Rubin Robust Confidence Intervals).
When instruments are weak, AR Confidence intervals are preferable to eyeballing F-statistics. Let $\mathbf{M}$ be a $n \times 2$ matrix of $\begin{pmatrix}\mathbf{y} & \mathbf{X}\end{pmatrix}$, and let $a_0 = (\beta_0, 1), b_0 = (1, -\beta_0)$ (where β_0 is typically 0), and

$$\widehat{\boldsymbol{\Sigma}} = \frac{\mathbf{M}^\top \mathbf{P}_z \mathbf{M}}{n - L - p}$$

be an estimator for the covariance matrix for the errors.
and let $\widehat{\mathbf{s}}, \widehat{\mathbf{t}}$ be two-dimensional vectors defined as

$$\widehat{\mathbf{s}} := (\mathbf{Z}^\top\mathbf{Z})^{\frac{1}{2}} \mathbf{Z}^\top \mathbf{M} b_0 (b_0^\top \widehat{\boldsymbol{\Sigma}} b_0)^{-\frac{1}{2}}$$

and

$$\widehat{\mathbf{t}} := (\mathbf{Z}^\top\mathbf{Z})^{\frac{1}{2}} \mathbf{Z}^\top \mathbf{M} \widehat{\boldsymbol{\Sigma}}^{-1} a_0 (a_0^\top \widehat{\boldsymbol{\Sigma}} a_0)^{-\frac{1}{2}}$$

Define the scalars $\widehat{Q}_1 = \widehat{\mathbf{s}}^\top\widehat{\mathbf{s}}, \widehat{Q}_2 = \widehat{\mathbf{s}}^\top\widehat{\mathbf{t}}, Q_3 = \widehat{\mathbf{t}}^\top\widehat{\mathbf{t}}$
based on these scalars, two tests that are fully robust to weak instruments for testing $H_0 : \beta = \beta_0$ - Anderson Rubin test (AR1949) and Conditional Likelihood Test (Moriera 2003)

$$AR(\beta_0) = \frac{\widehat{Q}_1}{L}$$

$$CLR(\beta_0) = \frac{1}{2}\left(\widehat{Q}_1 - \widehat{Q}_3\right) + \frac{1}{2}\sqrt{\left(\widehat{Q}_1 + \widehat{Q}_3\right)^2 - 4\left(\widehat{Q}_1\widehat{Q}_3 - \widehat{Q}_2^2\right)}$$

4.3.3 IV with Heterogeneous Treatment Effects / LATE Theorem

- binary instrument $Z_i \in \{0,1\}$
- binary treatment $D_z \in \{0,1\}$ is potential treatment status given $Z = z$
- potential outcomes: $Y_i(D,Z) = \{Y(1,1), Y(1,0), Y(0,1), Y(0,0)\}$
- heterogeneous treatment effects $\beta_i = Y_i(1) - Y_i(0)$

Defn 4.44 (IV Subpopulations).

- Compliers: $D_1 > D_0, D_0 = 0, D_1 = 0$
- Always takers: $D_0 = D_1 = 1$
- Never Takers : $D_0 = D_1 = 0$
- Defiers: $D_1 < D_0$

Assumption 5 (LATE Thm Assumptions).

- A1: **Independence of Instrument** : $\{Y_0, Y_1, D_0, D_1\} \perp\!\!\!\perp Z$
- A2: **Exclusion restriction** : $Y_i(d,0) = Y_i(d,1) \equiv Y_{di}$ for $d = 0, 1$
- A3: **First Stage**: $\mathbb{E}[D_{1i} - D_{0i}] \neq 0$
- A4: **Monotonicity** / No defiers: $D_{1i} - D_{0i} \geq 0 \; \forall \; i$ or vice versa

Theorem 4.15 (LATE Theorem (Angrist and Imbens (1994))).
Under A1-A4,

$$\alpha_{IV} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]} = \mathbb{E}[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]$$

If A1:A4 are satisfied, the IV estimate is the **Local Average Treatment Effect** for the compliers.

$$\text{LATE} = \text{ATE} + \frac{\text{Cov}\left[\beta_{1i}, \pi_{1i}\right]}{\mathbb{E}\left[\pi_{1i}\right]}$$

So, late is weighted average for people with large π_{1i}; i.e. treatment effect for those whosle probability of treatment is most influenced by Z_i.

Theorem 4.16 (Bloom Result).
IV in Randomized Trials with one-sided noncompliance. Conditional on A1:A4 holding, and $\mathbb{E}\left[D|Z_i = 0\right] = \mathbf{Pr}\left(D = 1|Z = 0\right) = 0$. Then,

$$\frac{\mathbb{E}\left[Y|Z=1\right] - \mathbb{E}\left[Y|Z=0\right]}{\mathbf{Pr}\left(D=1|Z=1\right)} = \frac{\text{ITT}}{\text{Compliance}} = \mathbb{E}\left[Y_1 - Y_0|D=1\right] = ATT$$

Precision for LATE Estimation

$$SE_{\widehat{LATE}} \approx \frac{SE_{\widehat{ITT}}}{\text{Compliance}}$$

4.3.4 Characterising Compliers

PO Model of IV allows for *heterogeneous treatment effects* but **does not** formally identify LATE conditional on X.
Abadie (2003) extends methods by allowing the treatment inducer to be randomized conditionally on the covariates and by allowing the outcome to depend on the covariates besides the treatment intake. The paper also provided semiparametric estimations of the probability of receiving the treatment inducement, which helps to identify the treatment effects in a more robust way.
Need the following assumptions (all conditional on $\mathbf{X}$):

- Independence of instrument: $Z \perp\!\!\!\perp (D(z), Y(z', d))|\mathbf{X}\ \forall z, z',\ d \in \{0,1\}$: SOO w.r.t. instrument.
- Exclusion restriction: $\mathbf{Pr}\left(Y(1,d) = Y(0,d) = Y(d)|\mathbf{X}\right) = 1$
- Monotonicity: $\mathbf{Pr}\left(D(1) \geq D(0)|\mathbf{X}\right) = 1$
- First Stage: $\mathbb{E}\left[D|Z=1,\mathbf{X}\right] - \mathbb{E}\left[D|Z=0,\mathbf{X}\right] \neq 0$
- **Common Support** : $0 < \mathbf{Pr}\left(Z=1,\mathbf{X}\right) < 1$

Specifically, when the treatment inducer Z is as good as randomized after conditioning on covariates X, Abadie proposed a two-stage procedure to estimate treatment effects.

- Estimate the probability of receiving the treatment inducement $P(Z=1|\mathbf{X})$ (preferably using a semiparametric estimator) in order to provide a set of pseudo-weights.
- Second, the pseudo-weights are used to estimate the local average response function (LARF) of the outcome conditional on the treatment and covariates.

The estimated coefficient for the treatment intake D reflects the conditional treatment effect.

Fact 4.17 (Size of Strata).
Given monotonicity, we can identify the proportion of compliers, never-takers, and always-takers respectively.

$$\begin{aligned}
\pi_{\text{compliers}} &= \mathbf{Pr}\left(D_1 > D_0|\mathbf{X}\right) = \mathbb{E}\left[D|\mathbf{X}, Z=1\right] - \mathbb{E}\left[D|\mathbf{X}, Z=0\right] \\
\pi_{\text{always-takers}} &= \mathbf{Pr}\left(D_1 = D_0 = 1|\mathbf{X}\right) = \mathbb{E}\left[D|\mathbf{X}, Z=0\right] \\
\pi_{\text{never-takers}} &= \mathbf{Pr}\left(D_1 = D_0 = 0|\mathbf{X}\right) = 1 - \mathbb{E}\left[D|\mathbf{X}, Z=1\right]
\end{aligned}$$

If nobody in the treatment group has access to the treatment (i.e. $\mathbb{E}\left[D|Z=0\right] = 0$**),** the LATE $=$ ATT.

Fact 4.18 (Proportion of treatment group that are compliers).
By Bayes rule,

$$\begin{aligned}
\mathbf{Pr}\left(D_1 > D_0|D=1\right) &= \frac{\mathbf{Pr}\left(D=1|D_1 > D_0\right)\mathbf{Pr}\left(D_1 > D_0\right)}{\mathbf{Pr}\left(D=1\right)} \\
&= \frac{\mathbf{Pr}\left(Z=1\right)\left[\mathbb{E}\left[D|Z=1\right] - \mathbb{E}\left[D|Z=0\right]\right]}{\mathbf{Pr}\left(D=1\right)}
\end{aligned}$$

Theorem 4.19 (Abadie's Kappa).
Suppose assumptions of LATE thm hold conditional on covariates X. Let $g(\cdot)$ be any measurable real function of Y, D, X with finite expectation. We can show that the expectation of g is a weighted sum of the expectation in the three groups

$$\mathbb{E}[g|\mathbf{X}] = \underbrace{\mathbb{E}[g|\mathbf{X}, D_1 > D_0]\,\mathbf{Pr}(D_1 > D_0|\mathbf{X})}_{\text{Compliers}} +$$
$$\underbrace{\mathbb{E}[g|\mathbf{X}, D_1 = D_0 = 1]\,\mathbf{Pr}(D_1 = D_0 = 1|\mathbf{X})}_{\text{Always takers}}$$
$$+ \underbrace{\mathbb{E}[g|\mathbf{X}, D_1 = D_0 = 0]\,\mathbf{Pr}(D_1 = D_0 = 0\mathbf{X})}_{\text{Never Takers}}$$

Rearranging terms gives us
Then,

$$\mathbb{E}[g(Y, D, \mathbf{X})|D_1 > D_0] = \frac{\mathbb{E}[\kappa \cdot g(Y, D, \mathbf{X})]}{\mathbf{Pr}(D_1 > D_0)} = \frac{\mathbb{E}[\kappa \cdot g(Y, D, \mathbf{X})]}{\mathbb{E}[\kappa]}$$

where

$$\kappa_i = 1 - \frac{D(1-Z)}{1 - \mathbf{Pr}(Z = 1|\mathbf{X})} - \frac{(1-D)Z}{\mathbf{Pr}(Z = 1|\mathbf{X})}$$

This result can be applied to *any characteristic or outcome and get its mean for compliers* by removing the means for never and always takers. Angrist and Pischke (2008, p 181-183) provides overview of estimation. Trick is to construct a weighting scheme with positive weights so that κ_i, which is negative for always-takers and never-takers.
To compute κ, we need $\mathbf{Pr}(Z = 1|\mathbf{X})$, which can be computed using a standard logit/probit or a power-series.
Standard example: average covariate value among compliers:

$$\mathbb{E}[X|D_1 > D_0] = \frac{\mathbb{E}[\kappa X]}{\mathbb{E}[\kappa]}$$

is the weighted average of covariate X using Kappa weights.

Likelihood that Complier has a given value of (Bernoulli distributed) characteristic X relative to the rest of the population is given by

$$\frac{\mathbb{E}[D|Z = 1, X = 1] - \mathbb{E}[D|Z = 0, X = 1]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]} = \frac{\text{FS in Subgroup}}{\text{Overall FS}}$$

Theorem 4.20 (Average Causal Response).
Assume A1-A4 from LATE. Generalise D to take values in the set $\{0, 1, \ldots, \check{D}\}$; Let $Y_{di} := f_i(d)$ denote the potential (or latent) outcome for person i for treatment level d. Then,

$$\frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 1]} = \sum_{d=1}^{\check{D}} \omega_d \mathbb{E}[Y_{di} - Y_{d-1,i}|d_{1i} \geq d > d_{0i}]$$

where the weights

$$\omega_d = \frac{\mathbf{Pr}(d_{1i} > d > d_{0i})}{\sum_{j=1}^{\check{D}} \mathbf{Pr}(d_{1i} \geq j \geq d_{0i})}$$

are non-negative and sum to 1.

Defn 4.45 (Local Average Response Function (LARF)).
CEF of $Y|X, D$ for the subpopulation of compliers: $\mathbb{E}[Y|X, D, D_1 > D_0]$

$$\mathbb{E}[Y|X, D, D_1 > D_0] = \frac{\mathbb{E}[\kappa Y|X, D]}{\mathbb{E}[\kappa]}$$

- Estimate κ
- Estimate $\mathbb{E}[Y|X, D]$ in the whole population, weighting by κ

implemented in `LARF::larf` in R.

Defn 4.46 (Inverse Compliance Score Weighting (Aronow and Carnegie, 2013)).
Treatment is W. First define two additional quantities

- $P_{A,C,i} := \mathbf{Pr}(W_1 > W_0 \cup W_0 = 1|\mathbf{X_i} = \mathbf{x}) = \mathbb{F}(\mathbf{x}_i'\boldsymbol{\theta}_{A,C})$ is the conditional probability that unit i is either a complier *or* an always taker
  - assume that this probability is a function of covariates $\mathbf{X}_i$, with corresponding parameter vector $\boldsymbol{\theta}_{A,C}$ and CDF $\mathbb{F}$ that transforms it to the probability scale [taken to be the normal CDF Φ henceforth, but can be relaxed]
- $P_{A|A,C,i} := \mathbf{Pr}(W_0 = 1|W_1 > W_0 \cup W_0 = 1, \mathbf{X_i} = \mathbf{x}) = \mathbb{F}(\mathbf{x}_i'\boldsymbol{\phi}_{A|A,C})$ is the conditional probability that unit i is an always taker *conditional* on being either a complier or a never taker
  - assume that this probability is a function of covariates with corresponding covariate vector $\boldsymbol{\phi}_{A|A,C}$

Next, they note that the probability of treatment for stratum $\mathbf{X_i} = \mathbf{x_i}$ can be written as

$$\mathbf{Pr}(W = 1|\mathbf{X_i} = \mathbf{x_i}) = \overbrace{\mathbf{Pr}(W_1 > W_i|\mathbf{X_i} = \mathbf{x_i})\,Z_i}^{\text{Compliers assigned to treatment}} + \overbrace{\mathbf{Pr}(W_0 = 1|\mathbf{X_i} = \mathbf{x_i})}^{\text{Always takers}}$$

Using the two conditional probabilties defined above, this can be written as

$$\mathbf{Pr}\left(W=1|\mathbf{X_i}=\mathbf{x_i}\right)=P_{A,C,i}(1-P_{A|A,C,i})Z_i+P_{A,C,i}P_{A|A,C,i}$$

which, for binary treatment W_i lets us write a Bernoulli likelihood for an observation

$$\begin{aligned}\ell_i(P_{A|A,C,i},P_{A,C,i}|W,Z)=&\left(P_{A,C,i}(1-P_{A|A,C,i})Z_i+P_{A,C,i}P_{A|A,C,i}\right)^{W_i}\\&\left(1-P_{A,C,i}(1-P_{A|A,C,i})Z_i-P_{A,C,i}P_{A|A,C,i}\right)^{1-W_i}\end{aligned}$$

Plugging in the definitions of $P_{A,C,i}$ and $P_{A|A,C,i}$ gives us the likelihood and its argmax defines the solution for $\widehat{\boldsymbol{\theta}}_{A,C}$ and $\widehat{\boldsymbol{\phi}}_{A|A,C}$. This is generically a difficult optimisation problem and improving its computation is a promising avenue for future research.

$$\begin{aligned}\mathcal{L}(P_{A|A,C,i},P_{A,C,i}|W,Z)=\prod_{i=1}^{N}&\left(\mathbb{F}(\mathbf{x}_i'\boldsymbol{\theta}_{A,C})(1-\mathbb{F}(\mathbf{x}_i'\boldsymbol{\phi}_{A|A,C})Z_i+\mathbb{F}(\mathbf{x}_i'\boldsymbol{\theta}_{A,C})\mathbb{F}(\mathbf{x}_i'\boldsymbol{\phi}_{A|A,C})\right)^{W_i}\\&\left((1-\mathbb{F}(\mathbf{x}_i'\boldsymbol{\theta}_{A,C}))(1-\mathbb{F}(\mathbf{x}_i'\boldsymbol{\phi}_{A|A,C})Z_i-\mathbb{F}(\mathbf{x}_i'\boldsymbol{\theta}_{A,C})\mathbb{F}(\mathbf{x}_i'\boldsymbol{\phi}_{A|A,C})\right)^{1-W_i}\end{aligned}$$

The maximum likelihood estimates of the two parameter vectors can be plugged into $\mathbb{F}$ to compute individual compliance scores

$$\widehat{P}_{C,i}=\mathbf{Pr}\left(W_1>W_0|\mathbf{X_i}=\mathbf{x_i}\right)=\overbrace{\mathbb{F}(\mathbf{x}_i'\widehat{\boldsymbol{\theta}}_{A,c})}^{\Pr\ i\text{ is C or AT}}\ \overbrace{1-\mathbb{F}(\mathbf{x}_i'\widehat{\boldsymbol{\phi}}_{A|A,c})}^{\Pr\ i\text{ is not AT}\implies i\text{ is C}}$$

The inverse compliance score weighted estimator for the ATE with weights $\omega_{C,i}:=1/\widehat{P}_{C,i}$ is then

$$\widehat{\tau}^{\text{ATE}}_{\text{ICSW}}=\frac{\left(\sum_{i=1}^n\hat{\omega}_{Ci}Z_iY_i\right)/\left(\sum_{i=1}^n\hat{\omega}_{Ci}Z_i\right)-\left(\sum_{i=1}^n\hat{\omega}_{Ci}\left(1-Z_i\right)Y_i\right)/\left(\sum_{i=1}^n\hat{\omega}_{Ci}\left(1-Z_i\right)\right)}{\left(\sum_{i=1}^n\hat{\omega}_{Ci}Z_iW_i\right)/\left(\sum_{i=1}^n\hat{\omega}_{Ci}Z_i\right)-\left(\sum_{i=1}^n\hat{\omega}_{Ci}\left(1-Z_i\right)W_i\right)/\left(\sum_{i=1}^n\hat{\omega}_{Ci}\left(1-Z_i\right)\right)}$$

which is a weighted version of the familiar Wald estimator with a Hajek correction that normalises each expectation by the sum of weights in that treatment group.

4.3.5 Shift Share / Bartik Instruments

SSIV setting from Borusyak, Hull, and Jaravel (2022) and Goldsmith-Pinkham, Sorkin, and Swift (2020) [notation and exposition from PGP's slides]. We want to estimate the causal effect or structural parameter τ in

$$y_l=\tau w_l+\gamma^\top\mathbf{x}_l+\varepsilon_l$$

where Cov $[\varepsilon_l,w_l]\neq 0$ because the 'treatment' w_l is typically a change in an economic quantity (e.g. employment) that is correlated with unobserved shocks to the outcome y_l (e.g. wages). l indexes *locations*.
An accounting identity that decomposes the treatment is

$$w_l=\sum_{k=1}^{K}\overbrace{z_{lk}}^{\text{Location-Industry Shares}}\underbrace{g_{lk}}_{\text{Location-Industry Shifts}}$$

where k indexes *industries*. 2nd accounting identity for location-industry shifts is

$$\underbrace{g_{lk}}_{\text{Location-industry}}=\underbrace{g_k}_{\text{industry}}+\underbrace{\widetilde{g}_{lk}}_{\text{location-industry shocks: unobserved}}$$

As a GMM system

$$\begin{aligned}y_{lt}&=\mathbf{D}_{lt}^\top\boldsymbol{\beta}_0+\tau w_{lt}+\varepsilon_{lt}\\w_{lt}&=\mathbf{D}_{lt}^\top\boldsymbol{\gamma}_0+\psi B_{lt}+\eta_{lt}\qquad\qquad\mathbf{D}_{lt}=\text{Exog controls, FE}\\g_{lkt}&=g_{kt}+\widetilde{g}_{lkt}\\B_{lt}&=\sum_{k=1}^{K}z_{lk0}g_{kt}\end{aligned}$$

$$\left\{\{w_{lt},\mathbf{D}_{lt},\varepsilon_{lt}\}_{t=1}^{T}\right\}_{l=1}^{L}\text{ are IID , }L\to\infty$$

Under constant τ, need

- Exogeneity $\mathbb{E}\left[B_{lt}\varepsilon_{lt}\mid\mathbf{D}_{lt}\right]=0$
- Relevance Cov $\left[B_{lt},w_{lt}\mid\mathbf{D}_{lt}\right]\neq 0$

Defn 4.47 (Bartik Estimator).

$$\widehat{\tau}_{\text{Bartik}}=\frac{\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{K}\overbrace{z_{lkt}}^{\text{Shares}}\overbrace{g_{kt}}^{\text{Shocks}}y_{lt}^{\perp}}{\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{K}z_{lkt}\,g_{kt}\,w_{lt}^{\perp}}$$

- 'shares': focus on z_{lk0} : Goldsmith-Pinkham, Sorkin, and Swift (2020)

- Analogy to DiD: Δ_{gt} = Changes in industry composition g_{kt}

- 'shifts': focus on g_{kt} : Borusyak, Hull, and Jaravel (2022)
  - requires argument for why shocks are randomly assigned

$$\widehat{\tau}_{\text{bartik}} = \sum_k \widehat{\alpha}_k \widehat{\tau}_k$$

with Rotemberg weight

$$\widehat{\alpha}_k = \frac{g_k Z_k^\top W}{\sum_{k=1}^K g_k Z_k^\top W}$$

4.3.6 Marginal Treatment Effects: Treatment effects under self selection

Heckman and Vytlacil (2007) propose the *marginal treatment effect* (MTE) setup that generalises the IV approach for continuous instruments and nests many estimands (and is a generalisation of the Roy (1951) model). It also has a clearer treatment of self-selection.
Exposition based on Cornelissen et al. (2016). Define potential outcomes

$$\begin{aligned} Y_{0i} &= \mu_0(\mathbf{x}_i) + \upsilon_{0i} \\ Y_{1i} &= \mu_1(\mathbf{x}_i) + \upsilon_{1i} \end{aligned}$$

where $\mu_j(\cdot)$ is the conditional mean function and υ_{ji} captures deviations, with $\mathbb{E}[\upsilon_{ji}|\mathbf{x}_i] = 0$.
Treatment assignment assumes a *weakly separable choice model*

$$\begin{aligned} D_i^* &= \mu_d(\mathbf{x}_i, z_i) + \upsilon_i \\ D_i &= \mathbb{1}_{D_i^* \geq 0} \end{aligned}$$

where d_i^* is the *latent* propensity to take the treatment, and is interpreted as the **net gain from treatment** since treatment is only taken up if $D_i^* \geq 0$. z_i is an instrument. υ_i enters the selection equation negatively, and thus represents **latent resistance to treatment**.
The condition $D_i^* \geq 0$ can be rewritten as $\mu_d(\mathbf{x}_i, z_i) \geq \upsilon_i$. Applying the CDF of υ $\mathbb{F}_\upsilon$ to both sides yields

$$\underbrace{\mathbb{F}_\upsilon(\mu_d(\mathbf{x}_i, z_i))}_{\text{Propensity score} =: P(\mathbf{x}_i, z_i)} \geq \underbrace{\mathbb{F}_\upsilon(\upsilon_i)}_{\text{Quantiles of distaste distribution} =: \upsilon_{di}}$$

Both RHS and LHS are distributed on $[0, 1]$. The treatment decision can now be written as

$D_i = \mathbb{1}_{P(\mathbf{x}_i, z_i) \geq \upsilon_{di}}$.
Now, we define treatment effects

$$\begin{aligned} Y_i &= (1 - D_i)Y_{0i} + D_i Y_{1i} \\ &= Y_{0i} + D_i \underbrace{(Y_{1i} - Y_{0i})}_{=: \Delta_i} \\ &= \mu_0(\mathbf{x}_i) + D_i \underbrace{[\overbrace{\mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)}^{\Delta(\mathbf{x}_i)} + \overbrace{\upsilon_{1i} - \upsilon_{0i}}^{\text{Idiosyncratic gain}}]}_{\equiv \Delta_i} + \upsilon_{0i} \end{aligned}$$

Aggregating over different parts of the covariate distribution yields different estimates.

$$\begin{aligned} \text{ATE}(\mathbf{x}) &:= \mathbb{E}[\Delta_i|\mathbf{x}_i = \mathbf{x}] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) \\ \text{ATT}(\mathbf{x}) &:= \mathbb{E}[\Delta_i|\mathbf{x}_i = \mathbf{x}, D_i = 1] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + \mathbb{E}[\upsilon_{1i} - \upsilon_{0i}|D_i = 1] \\ \text{ATU}(\mathbf{x}) &:= \mathbb{E}[\Delta_i|\mathbf{x}_i = \mathbf{x}, D_i = 0] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) + \mathbb{E}[\upsilon_{1i} - \upsilon_{0i}|D_i = 0] \end{aligned}$$

Integrating these over $\mathbf{x}$ yields the conventional estimators. With self-selection based on $D_i = \mathbb{1}_{d_i \geq 0^*}$ typically means ATT > ATE > ATU.

Fact 4.21 (Estimation with Binary Instrument).
The covariate-specific Wald estimator is

$$\text{Wald}(\mathbf{x}) = \frac{\mathbb{E}[Y_i|z_i = z, \mathbf{x}_i = \mathbf{x}] - \mathbb{E}[Y_i|z_i = z, \mathbf{x}_i = \mathbf{x}]}{\mathbb{E}[D_i|z_i = z, \mathbf{x}_i = \mathbf{x}] - \mathbb{E}[D_i|z_i = z, \mathbf{x}_i = \mathbf{x}]}$$

Under the standard A1-A4 from AIR96,

$$\begin{aligned} \text{LATE}(\mathbf{x}) &:= \mathbb{E}[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}, \mathbf{x}_i = \mathbf{x}] \\ &= \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i) + \mathbb{E}[\upsilon_{1i} - \upsilon_{0i}|D_{1i} > D_{0i}, \mathbf{x}_i = \mathbf{x}] \end{aligned}$$

These can be aggregated using the 'saturate and weight' theorem (Angrist and Imbens)

$$\text{IV} = \sum_{\mathbf{x} \in \mathcal{X}} \omega(\mathbf{x}) \text{LATE}(\mathbf{x})$$

with weights

Target parameter	Expression	Weights	
		$\omega_0^\star(u,x,z)$	$\omega_1^\star(u,x,z)$
Average untreated outcome	$E[Y_0]$	1	0
Average treated outcome	$E[Y_1]$	0	1
ATE	$E[Y_1 - Y_0]$	-1	1
ATE given $X=\bar{x}$, where $P[X=\bar{x}]>0$	$E[Y_1 - Y_0 \mid X=\bar{x}]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[x=\bar{x}]}{P[X=\bar{x}]}$
ATT	$E[Y_1 - Y_0 \mid D=1]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u\le p(x,z)]}{P[D=1]}$
ATU	$E[Y_1 - Y_0 \mid D=0]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u> p(x,z)]}{P[D=0]}$
LATE for $z_0 \to z_1$ given $X=x$, where $p(x,z_1)>p(x,z_0)$	$E[Y_1 - Y_0 \mid p(x,z_0)<U\le p(x,z_1), X=x]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[p(x,z_0)<u\le p(x,z_1)]}{p(x,z_1)-p(x,z_0)}$

Figure 4: MTE weights from Mogstad and Torgovitsky (2018)

$$\omega(\mathbf{x}_i) = \frac{\overbrace{p_{\mathbf{x}}}^{\text{Share with } \mathbf{x}_i=\mathbf{x}} \mathbb{V}\left[\overbrace{\widehat{D}_i}^{\mathbb{E}[D_i|\mathbf{x}_i=\mathbf{x},z_i=z]} \middle| \mathbf{x}_i, z_i\right]}{\mathbb{V}\left[\widehat{D}_i\right]}$$

For a continuous instrument, for a pair of instrument values z, z', $\text{LATE}(z,z',\mathbf{x}) = \mathbb{E}[Y_{1i} - Y_{0i} | D_{zi} > D_{z'i}, \mathbf{x}_i = \mathbf{x}]$.

Defn 4.48 (Marginal Treatment Effect (MTE)).

$$\begin{aligned}\text{MTE}(\mathbf{x_i}=\mathbf{x}, V_i = v) &:= \mathbb{E}[Y_{1i} - Y_{0i} | \mathbf{x_i}=\mathbf{x}, V=v] \\ &= \frac{\partial \mathbb{E}[Y_i | \mathbf{x}_i=\mathbf{x}, p(Z,X)=p(z,x)]}{\partial p(z,x)}\end{aligned}$$

MTE is defined as a continuum of treatment effects along the distribution of υ_D.
Define two **marginal treatment response (MTR)** functions

$$m_0(u,\mathbf{x}) = \mathbb{E}[Y_0 | U=u, \mathbf{X}=\mathbf{x}]; \quad m_1(u,\mathbf{x}) = \mathbb{E}[Y_1 | U=u, \mathbf{X}=\mathbf{x}]$$

Many useful parameters are identified using the following expression

$$\beta^\star \equiv E\left[\int_0^1 m_0(u,X)\omega_0^\star(u,X,Z)\mathrm{d}u\right] + E\left[\int_0^1 m_1(u,X)\omega_1^\star(u,X,Z)\mathrm{d}u\right]$$

with weights specified in 4.
Parametric Model: Assuming joint normality for U_0, U_1, V,

$$E[U_{0i} \mid D_i=0, X_i, Z_i] = E[U_{0i} \mid V_i \ge (X_i,Z_i)\beta_d, X_i, Z_i] = \rho_0\left(\frac{\phi((X_i,Z_i)\beta_d)}{1-\Phi((X_i,Z_i)\beta_d)}\right)$$
$$E[U_{1i} \mid D_i=1, X_i, Z_i] = E[U_{1i} \mid V_i < (X_i,Z_i)\beta_d, X_i, Z_i] = \rho_1\left(\frac{-\phi((X_i,Z_i)\beta_d)}{\Phi((X_i,Z_i)\beta_d)}\right)$$

where ρ_0 is the correlation $\rho[U_{0i}, V_i]$, and $\rho_1 = \rho[U_{1i}, V_i]$.
yields MTE estimator

$$\text{MTE}(x,u_D) = E(Y_{1i} - Y_{0i} \mid X_i=x, U_{Di}=u_D) = x(\beta_1-\beta_0) + (\rho_1-\rho_0)\Phi^{-1}(u_D)$$

Defn 4.49 (Control Function IV).
Let $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \overline{\mathbf{x}}$. Write

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\alpha} + D_i \widetilde{\mathbf{x}}_i' \boldsymbol{\theta} + D_i\delta_i + \varepsilon_i$$

where δ_i is a random effect that captures treatment effect heterogeneity . We can rewrite this and by demeaning $\delta_i = \overline{\delta} - \widetilde{\delta}_i$.

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\alpha} + D_i \widetilde{\mathbf{x}}_i^\top \theta + D_i\overline{\delta} + \overbrace{\varepsilon_i}^{\upsilon_{0i}} \tag{2}$$

where $\overline{\delta}$ captures the ATE at means of X, which is the unconditional ATE under the linear specification.
Write the selection equation

$$D_i = \mathbf{x}_i\pi_1^\top + z_i\pi_2 + \nu_i \quad \text{with } \mathbb{E}[\nu_i|\mathbf{x}_i, z_i] = 0$$

Assumptions

- $\mathbb{E}[\varepsilon_i|\nu_i] = \eta\nu_i$: Conventional selection bias.
- $\mathbb{E}\left[\widetilde{\delta}_i|\nu_i\right] = \psi\nu_i$: unobservable part of treatment effect $\widetilde{\delta}_i$ depends linearly on the unobservables that affect treatment selection.

Including $\widehat{\nu}_i$ and $\widehat{\nu}_i D_i$ in eqn 2 yields a consistent estimate of the ATE : $\overline{\delta}$.

4.3.7 High Dimensional IV selection

Chernozhukov, Hansen, and Spindler (2015) setup:

$$y_i = \tau d_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$
$$d_i = \boldsymbol{x}_i'\boldsymbol{\gamma}_0 + \boldsymbol{z}_i'\boldsymbol{\delta}_0 + \upsilon_i$$

where

1. $\boldsymbol{x}_i$ is a vector of p_n^x exogenous controls, including a constant.
2. $\boldsymbol{z}_i$ is a vector of p_n^z instruments
3. d_i is an endogenous variable
4. $p_n^x >> n$ and $p_n^z >> n$

1. Run (post)LASSO of d_i on $\boldsymbol{x}_i, \boldsymbol{z}_i$ to obtain $\widehat{\gamma}, \widehat{\delta}$
2. Run (post)LASSO of y_i on $\boldsymbol{x}_i$ to get $\widehat{\theta}$.
3. Run (post)LASSO of $\widehat{d}_i = \boldsymbol{x}_i'\widehat{\gamma} + \boldsymbol{z}_i'\widehat{\delta}$ on $\boldsymbol{x}_i$ to get $\widehat{\vartheta}$.
4. Construct $\widehat{\rho}_i^y := y_i - \boldsymbol{x}_i\widehat{\theta}$, $\widehat{\rho}_i^d := d_i - \boldsymbol{x}_i'\widehat{\vartheta}$ and $\widehat{v}_i := \boldsymbol{x}_i\widehat{\gamma} + \boldsymbol{z}_i'\widehat{\delta} + \boldsymbol{x}_i'\widehat{\vartheta}$.
5. Estimate $\widehat{\tau}$ by using standard IV regression of $\widehat{\rho}_i^y$ on $\widehat{\rho}_i^d$ with $\widehat{v}_i$ as instrument. Perform inference using score stastics or conventional heteroskedasticity-robust SEs.

implemented in `hdm::rlassoIV(., select.X = T, select.Z = T)`.
Discussion in https://cran.r-project.org/web/packages/hdm/vignettes/hdm.pdf.

4.3.8 Principal Stratification

Treatment comparisons often need to be adjusted for post-treatment variables. Binary treatment $Z_i \in \{0,1\}$. post-treatment **Intermediate** variable $S_i(z_i) \in \{0,1\}$, Outcome $Y_i \in \{0,1\}$. For each individual, the treatment assumes a single value, so only one of the two potential intermediate values are observed. Based on joint potential outcomes of the intermediate variable $(S_i,(0),S_i(1))$, we have 4 strata

$00 = \{i : S_i(0) = 0, S_i(1) = 0\}$	Never Takers
$10 = \{i : S_i(0) = 1, S_i(1) = 0\}$	Defiers
$01 = \{i : S_i(0) = 0, S_i(1) = 1\}$	Compliers
$11 = \{i : S_i(1) = 0, S_i(1) = 1\}$	Always takers

Defn 4.50 (Principal Stratification Frangakis and D. B. Rubin (2002)).
The **basic principal stratification** P_0 w.r.t post treatment variable S is the *partition of units* $i = 1,\ldots,n$ such that, forall units in *any set of* P_0, all units have the same vector of $(S_i(0), S_i(1))$. The principal stratum $G_i \in \{00, 10, 01, 11\}$ to which unit i belongs is not affected by treatment assignment for any principal stratification, so can be considered pre-treatment.

- Treatment Ignorability implies

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp Z_i | S_i(0), S_i(1), \mathbf{X}$$

  (i.e. treatment and control units can be compared conditional on stratum)

- **Principal Causal Effect (PCE)**

$$\tau_{s_0,s_1} := \mathbb{E}\left[Y_i(1) - Y_i(0) | S_i(0) = s_0, S_i(1) = s_1\right]$$

A common example is the
Complier Average Causal Effect (CACE) = Causal Effect on Principal Stratum of Compliers (AIR96)

$$\text{CACE} = \mathbb{E}\left[Y_i(1) - Y_i(0) | S_i(0) = 0, S_i(1) = 1\right]$$

Recall that $G_i = (S_0, S_1)$ concatenated. So, AIR96 in PS terms:

- Monotonicity: $S_1 \geq S_0 \implies \{G_i = 10\}$ must be empty: no defiers.
- Exclusion: $\tau_{11} = \tau_{00}$

Estimation under principal ignorability (Jiang, Yang, and Ding, 2020)

- Treatment ignorability $Z \perp\!\!\!\perp (S_0, S_1, Y_0, Y_1) \mid X$
- monotonicity: $S_1 \geq S_0 \implies G_i = 10$ is not allowed
- principal ignorability

$$\mathbb{E}\left[Y_1 \mid G = 11, X\right] = \mathbb{E}\left[Y_1 \mid G = 01, X\right]$$
$$\mathbb{E}\left[Y_0 \mid G = 00, X\right] = \mathbb{E}\left[Y_1 \mid G = 01, X\right]$$

	S = 0	S = 1
Z = 0	G = 00 or 01	G = 11
Z = 0	G = 00	G = 11 or 01

Disentangle mixture distribution within strata by assuming same conditional expectation across mixture components (complier, never taker, always taker).
Define nuisance functions:

- Treatment probability: $\pi(X) = \mathbf{Pr}(Z = 1 \mid X)$
- Principal Score: $e_g(X) = \mathbf{Pr}(G = g \mid X)$ identified by

$$\begin{aligned} e_{01}(X) &= p_1(X) - p_0(X) \\ e_{00}(X) &= 1 - p_1(X) \\ e_{11}(X) &= p_0(X) \end{aligned}$$

  where $p_z(X) = \mathbf{Pr}(S = 1 \mid Z = z, X)$
- Outcome mean: $\mu_{zs}(X) = \mathbb{E}[Y \mid Z = z, S = s, X]$

Treatment Probability and Principal Score

$$\begin{aligned} \tau_{01} &= \mathbb{E}\left\{\frac{e_{01}(X)}{p_1 - p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{01}(X)}{p_1 - p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\} \\ \tau_{00} &= \mathbb{E}\left\{\frac{1-S}{1-p_1}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{00}(X)}{1-p_1}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\} \\ \tau_{11} &= \mathbb{E}\left\{\frac{e_{11}(X)}{p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{S}{p_0}\frac{1-Z}{1-\pi(X)}Y\right\} \end{aligned}$$

Treatment Probability and Outcome Mean

$$\begin{aligned} \tau_{01} &= \mathbb{E}\left[\frac{SZ/\pi(X) - S(1-Z)/\{1-\pi(X)\}}{p_1 - p_0}\{\mu_{11}(X) - \mu_{00}(X)\}\right] \\ \tau_{00} &= \mathbb{E}\left[\frac{1 - SZ/\pi(X)}{1-p_1}\{\mu_{10}(X) - \mu_{00}(X)\}\right] \\ \tau_{11} &= \mathbb{E}\left[\frac{S(1-Z)/\{1-\pi(X)\}}{p_0}\{\mu_{11}(X) - \mu_{01}(X)\}\right] \end{aligned}$$

Principal Score and Outcome Mean

$$\begin{aligned} \tau_{01} &= \mathbb{E}\left[\frac{p_1(X) - p_0(x)}{p_1 - p_0}\{\mu_{11}(X) - \mu_{00}(X)\}\right] \\ \tau_{00} &= \mathbb{E}\left[\frac{1 - p_1(X)}{1-p_1}\{\mu_{10}(X) - \mu_{00}(X)\}\right] \\ \tau_{11} &= \mathbb{E}\left[\frac{p_0(X)}{p_0}\{\mu_{11}(X) - \mu_{01}(X)\}\right] \end{aligned}$$

Direct and Indirect Effects via Principal Stratification **Direct effect** of Z conditional on S exists if there is a causal effect of Z on Y for observations for whom the treatment does not affect selection S, i.e. principal strata $00, 11$. This is a *zero-first-stage* sample in IV-terms.
The **Indirect Effect** is mediated through S.

Attrition as Selection Bias Let S denote a binary selection indicator for when Y is observed. Let $S(1), S(0)$ denote potential selection states under treatment and nontreatment.

- $S(1) = 0, S(0) = 0$: never-selected
- $S(1) = 1, S(0) = 1$: always selected
- $S(0) = 0, S(1) = 1$: selection compliers
- $S(0) = 1, S(1) = 0$: selection defiers (ruled out by Lee bounds)

Dominance assumption: $\mathbb{E}[Y(1)|S(1) = 1, S(0) = 1] \geq \mathbb{E}[Y(1)|S(1) = 1, S(0) = 0]$ and $\mathbb{E}[Y(0)|S(1) = 1, S(0) = 1] \geq \mathbb{E}[Y(0)|S(1) = 1, S(0) = 0]$. The average potential outcome of the always selected dominates that of compliers under either treatment state.
Then, Zhang and Rubin (2003) bounds are

$$\begin{aligned} \Delta^{UB} &= \mathbb{E}[Y|D = 1, S = 1, Y \geq y^*] - \mathbb{E}[Y|D = 0, S = 1] \\ \Delta^{LB} &= \mathbb{E}[Y|D = 1, S = 1] - \mathbb{E}[Y|D = 0, S = 1] \end{aligned}$$

where y^* is chosen such that the lowest outcomes among those with $D = 1, S = 1$ correspond to the share of compliers among those with $D = 1, S = 1$ are smaller than this value.

Defn 4.51 (Lee Bounds).
Assuming

- **randomisation**: $\{Y(1), Y(0), S(0), S(1), \mathbf{X}\} \perp\!\!\!\perp D$
- **monotonicity**: $S(1) \geq S(0) a.s.$

Lee (2009) focuses on the ATE *among the always observed*

$$\mathbb{E}[Y(1) - Y(0)|S(0) = S(1) = 1]$$

The second quantity: $\mathbb{E}[Y(0)|S(1) = 1, S(0) = 1]$ is point identified. In contrast, the outcome in the treatment group can be either an always-selected's outcome or a selection complier's outcome.

Always selected share among the treated is

$$p_0 = \mathbf{Pr}\left(S(1)=1, S(0)=1 | S(1)=1\right) = \mathbf{Pr}\left(S(0)=1|S(1)=1\right) = \frac{\mathbf{Pr}\left(S=1|D=0\right)}{\mathbf{Pr}\left(S=1|D=1\right)}$$

In the best case, the always-selected comprise the top p_0 quantile of the treatment outcomes. Then the largest possible value of β is

$$\beta_U = \mathbb{E}\left[Y | Y \geq Q_{y|S=1,D=1}(1-p_0), D=1, S=1\right] - \mathbb{E}\left[Y|S=1, D=0\right]$$

The smallest possible one is

$$\beta_L = \mathbb{E}\left[Y | Y \leq Q_{y|S=1,D=1} p_0, D=0, S=1\right] - \mathbb{E}\left[Y|S=1, D=0\right]$$

this can be implemented conditional on covariates by constructing $p_0(x)$ within each x stratum.

4.4 Regression Discontinuity Design

Setup
Treatment (D) changes discontinuously at some particular value x_0 in x [**and nothing else does**], so

$$D_i = \begin{cases} 0 \text{ if } x_i < x_0 \\ 1 \text{ if } x_i \geq x_0 \end{cases}$$

Standard identification assumptions violated by definition because although *unconfoundedness holds trivially* since we have $D_i = \mathbb{1}_{x_i \geq c}$, this also means **overlap is always violated**. Need to invoke continuity to do causal inference.

Defn 4.52 (Sharp Regression Discontinuity Estimand (Hahn et al 2001)).
Identified at $x = c$, i.e. $\tau_c = \mu_{(1)}(c) - \mu_{(1)}(c)$ via

$$\tau_c := \mathbb{E}\left[Y_1 - Y_0 | X = c\right] = \lim_{x \downarrow c} \mathbb{E}\left[Y|X=c\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y|X=c\right]$$

$$\lim_{x\downarrow c} \mathbb{E}\left[y|X\right] - \lim_{x\uparrow c} \mathbb{E}\left[y|X\right] = \tau_{SRDD} + \underbrace{\lim_{x\downarrow c} \mathbb{E}\left[u|X\right] - \lim_{x\uparrow c} \mathbb{E}\left[u|X\right]}_{\approx 0}$$

Identification Assumption 1 (Smoothness of Unobservables).

- Conditional mean function $\mathbb{E}\left[u|X\right]$ is continuous at c
- Mean Treatment effect function $\mathbb{E}\left[\tau_i|X\right]$ is right continuous at c

4.4.1 Estimators

Normalise running variable $c := x_0$. Then, the **linear regression implementation** is the following:

$$Y = \alpha_l + \tau D + \beta_l f(X-c) + (\beta_r - \beta_l) \times D \times g(X-c) + \epsilon$$

where f and g are local or global polynomials. Since the design relies on *identification at infinity* (*i.e. at the cutoff*), choice of polynomial / functional form matters a lot.
Calonico, Cattaneo, Titiunik (2014) recommend local-linear regressions. Older literature relies on global higher-order polynomials, which often yields strange estimates.

Defn 4.53 (Local Linear RD Estimator).

$$\widehat{\tau}_c = \operatorname{argmin} \left\{ \sum_{i=1}^{n} K\left(\frac{|X_i - c|}{h_n}\right) \times \left(Y_i - a - \tau D_i - \beta_{(0)}(Z_i - c)_- - \beta_{(1)}(Z_i - c)_+\right) \right\}$$

Where $K(\cdot)$ is a kernel function. Common choices are the window function $K(x) = \mathbb{1}_{|x|\leq 1}$ or the triangular kernel $K(x) = (1-|x|)_+$

Assumptions for Local Linear Estimator Loosely, we need CEFs $\mu_{(w)}$ to be smooth. More precisely, we need $\mu_{(w)}(x)$ to be twice-differentiable with uniformly bounded second derivative.

$$\left| \frac{d^2}{dx^2} \mu_{(w)}(x) \right| \leq B \; \forall x \in \mathbb{R} \wedge w \in \{0,1\}$$

Taking a taylor expansion around c, we can write the CEFs as

$$\mu_{(w)}(x) = a_{(w)} + \beta_{(w)}(x-c) + \frac{1}{2}\rho_{(w)}(x-c) \;\; \left|\rho_{(w)}(x)\right| \leq Bz^2$$

with $\tau_c = a_{(1)} - a_{(0)}$. The local linear regression with a window kernel can be solved in closed form

$$\widehat{a}_{(1)} = \sum_{c \leq X_i \leq c+h_n} \gamma_i Y_i \;\; , \gamma_i = \frac{\widehat{\mathbb{E}}_{(1)}[(X_i-c)^2] - \widehat{\mathbb{E}}_{(1)}[X_i - c]\cdot(X_i - c)}{\widehat{\mathbb{E}}_{(1)}[(X_i-c)^2] - \widehat{\mathbb{E}}_{(1)}[X_i-c]^2}$$

where $\widehat{\mathbb{E}}(\cdot)$ denote sample averages over the regression window. Then, the error term can be written as

$$\widehat{a}_{(1)} = a_{(1)} + \underbrace{\sum_{c \leq X_i \leq c+h_n} \gamma \rho_{(1)}(X_i - c)}_{\text{Curvature Bias}} + \underbrace{\sum_{c \leq X_i \leq c+h_n} \gamma_i (Y_i - \mu_{(1)}(X_i))}_{\text{Sampling Noise}}$$

Curvature bias bounded by Bh_n^2.

$$\widehat{\tau}_c = \tau_c + \mathsf{O}\left(n^{-2/5}\right) \text{ with} h_n \sim n^{-1/5}$$

This rate is a consequence of working with the 2nd derivative. In general, if we assume $\mu_{(w)}(\cdot)$ has a bounded $k-$th derivative, we can achive $n^{-k/(2k+1)}$ rate using local polynomial regression of order $k-1$ with a bandwidth scaling as $h_n \sim n^{-1/(2k+1)}$.

Defn 4.54 (Minimax Linear Estimation (Imbens and Wager, 2017)).
The local linear regression estimator for τ_c

$$\widehat{\tau}_c = \operatorname{argmin} \sum_{i=1}^{n} K\left(\frac{|Z_i - c|}{h_n}\right)(Y_i - a - \tau W_i - \beta_{(0)}(Z_i - c)_- - \beta_{(1)}(Z_i - c)_+)^2$$

which can be written as a *local linear estimator* $\widehat{\tau}_c = \sum_{i=1}^n \gamma_i Y_i$ where weights γ_i only depend on the running variable Z. Imbens and Wager (2017) show that local linear regression is **not the best estimator in this class**.
Under an assumption that $\left|\mu''_{(w)}(z)\right| \leq B | \{Z_1, \ldots, Z_n\}$, the minimax linear estimator is the one that minimises the MSE $\text{MSE}(\widehat{\tau}_c(\gamma) | \{Z_1, \ldots, Z_n\}) \leq \sigma \|\gamma\|_2^2 + I_B^2(\gamma)$ and is given by

$$\widehat{\tau}_c(\gamma^B) = \sum_{i=1}^{n} \gamma_i^B Y_i \; ; \; \gamma^B = \operatorname{argmin}\left\{\sigma \|\gamma\|_2^2 + I_B^2(\gamma)\right\}$$

These weights can be solved for using quadratic programming.

4.4.2 Fuzzy RD

Discontinuity doesn't deterministically change treatment, but affects *probability of treatment*. Analogue of IV with one-sided non-compliance.

$$P[D_i = 1 | x_i] = \begin{cases} g_0(x_i) \text{ if } x_i < x_0 \\ g_1(x_i) \text{ if } x_i \geq x_0 \end{cases}$$

$g_0(x_i) \neq g_1(x_i)$. Assuming $g_1(x_0) > g_0(x_0)$, the probability of treatment relates to x_i via:

$$E[D_i | x_i] = P[D_i = 1 | x_i] = g_0(x_i) + [g_1(x_i) - g_0(x_i)]T_i$$

where $T_i = \mathbb{1}_{x_i \geq x_0}$:= point of discontinuity

4.4.3 Regression Kink Design

First-derivative version of the fuzzy RD. Continuous treatment, where the treatments are a function of the running variable X with kink at x_0. This implies that the first derivative $\frac{\partial D}{\partial X}$ of continuous treatment D is discontinuous at the threshold. The marginal treatment effect at the threshold is defined as

$$\Delta_{X=x_0}(d_0) = \frac{\partial \mathbb{E}\left[Y(d_0) | X = x_0\right]}{\partial D} = \frac{\lim_{\varepsilon \to 0} \frac{\partial \mathbb{E}[Y | X \in [x_0, x_0 + \varepsilon)]}{\partial X} - \lim_{\varepsilon \to 0} \frac{\partial \mathbb{E}[Y | X \in [x_0 - \varepsilon, x_0)]}{\partial X}}{\lim_{\varepsilon \to 0} \frac{\partial \mathbb{E}[D | X \in [x_0, x_0 + \varepsilon)]}{\partial X} - \lim_{\varepsilon \to 0} \frac{\partial \mathbb{E}[D | X \in [x_0 - \varepsilon, x_0)]}{\partial X}}$$

4.5 Differences-in-Differences

4.5.1 DiD with 2 periods

Binary treatment $d \in \{0, 1\}$, 2 time periods $t \in \{0, 1\}$.
Potential outcomes denoted Y_t^d.

Defn 4.55 (Estimand).
ATT in the 2nd period.

$$\tau_{ATT} := \mathbb{E}\left[Y_1^1 - Y_1^0 | D = 1\right]$$

$\mathbb{E}\left[Y_1^0 | D = 1\right]$ not observed, so must be imputed.

Naive Estimation Strategies

- **Before-After Comparison**: $\tau = \mathbb{E}\left[Y_1^1 | D = 1\right] - \mathbb{E}\left[Y_0^0 | D = 1\right]$
  - assumes $\mathbb{E}\left[Y_1^0 | D = 1\right] = \mathbb{E}\left[Y_0^0 \; D = 1\right]$ (No trending)
- **Post Treatment-Control Comparison**: $\tau = \mathbb{E}\left[Y_1^1 | D = 1\right] - \mathbb{E}\left[Y_1^0 | D = 0\right]$
  - Assumes $\mathbb{E}\left[Y_1^0 | D = 1\right] = \mathbb{E}\left[Y_1^0 | D = 0\right]$ (Random Assignment in the 2nd period)

Both typically untenable in practice, so we need parallel trends.

Defn 4.56 (DiD Estimator).
Sample analogue of
Impute $\mathbb{E}\left[Y_1^0 | D = 1\right]$ with

$$\underbrace{\mathbb{E}\left[Y_0^0 | D = 1\right]}_{\text{Baseline PO for treated}} + \underbrace{\mathbb{E}\left[Y_1^0 | D = 0\right] - \mathbb{E}\left[Y_0^0 | D = 0\right]}_{\text{Change over time in control series}}$$

$$\Delta_{D=1} := \underbrace{\mathbb{E}\left[Y_1^1|D=1\right] - \mathbb{E}\left[Y_0^0|D=1\right]}_{\text{over-time difference for treated unit}} - \underbrace{\mathbb{E}\left[Y_1^0|D=0\right] - \mathbb{E}\left[Y_0^0|D=0\right]}_{\text{over-time difference for control unit}}$$

Defn 4.57 (Parallel Trends Assumption).

$$\underbrace{\mathbb{E}\left[Y_1^0 - Y_0^0|D=1\right]}_{\text{Trend in control PO for Treated}} = \underbrace{\mathbb{E}\left[Y_1^0 - Y_0^0|D=0\right]}_{\text{Trend in control PO for Control}}$$

Often justified using a figure [with transformed y if necessary], or control for time trends [which relies on a strong functional form assumption], or a clear **falsification test** [on a placebo group].
If $\mathbb{E}\left[Y_0^0|D=1\right] = \mathbb{E}\left[Y_0^0|D=0\right]$, this collapses to a Selection-on-observables in the 2nd period assumption.

$$\mathbb{E}\left[Y_1^0|D=1\right] = \mathbb{E}\left[Y_1^0|D=0\right]$$

For a two-period difference, we can also write the standard OLS exogeneity condition in differences form

$$\begin{aligned}\mathbb{E}\left[\Delta \boldsymbol{x}' \Delta \boldsymbol{\epsilon}\right] &= \mathbf{0} \\ \mathbb{E}\left[\boldsymbol{x}_2' \boldsymbol{\epsilon}_2\right] + \mathbb{E}\left[\boldsymbol{x}_1' \boldsymbol{\epsilon}_1\right] - \underbrace{\mathbb{E}\left[\boldsymbol{x}_1' \boldsymbol{\epsilon}_2\right] - \mathbb{E}\left[\boldsymbol{x}_2' \boldsymbol{\epsilon}_1\right]}_{\text{No feedback loop}} &= \mathbf{0}\end{aligned}$$

Which makes a direct link with the **strong exogeneity** assumption in panel data models that asserts that $\epsilon_t \perp\!\!\!\perp \boldsymbol{x}_1, \ldots \boldsymbol{x}_t$.

Regression Estimator
We typically prefer the following regression estimator (for automatic standard errors etc).

$$Y_{it} = \alpha + \gamma \text{Treat}_i + \lambda \text{Post}_t + \tau(\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

Triple Differences (DDD) Estimator
Regular Diff-in-Diff estimate - Diff-in-diff estimate for placebo group.

4.5.2 Nonparametric Identification Assumptions with Covariates

Lechner (2011)
Estimand:

$$\begin{aligned}\tau_{ATT} &:= \mathbb{E}\left[Y_t^1 - Y_t^0|D=1\right] \\ &= \mathbb{E}\left[\underbrace{\mathbb{E}\left[Y_t^1 - Y_t^0|\mathbf{X}=\mathbf{x}, D=1\right]}_{\theta_t(\mathbf{x})}|D=1\right] \\ &= \mathbb{E}_{\mathbf{X}|D=1}\left[\theta_t(\mathbf{x})\right]\end{aligned}$$

Identification Assumptions:

- SUTVA
$$Y_t = DY_t^1 + (1-D)Y_t^0 \quad , t \in \{0,1\}$$
- Covariate exogeneity
$$\mathbf{X}^1 = \mathbf{X}^0 = \mathbf{X}\ x \in \mathcal{X}$$
- No effect before treatment
$$\theta_0(\mathbf{x}) = 0;\ \forall x \in \mathcal{X}$$
- Common Trend
(parallel trends within $\mathbf{x}$ strata)

$$\begin{aligned}&\mathbb{E}\left[Y_1^0|\mathbf{X}=\mathbf{x}, D=1\right] - \mathbb{E}\left[Y_0^0|\mathbf{X}=\mathbf{x}, D=1\right] \\ &= \mathbb{E}\left[Y_1^0|\mathbf{X}=\mathbf{x}, D=0\right] - \mathbb{E}\left[Y_0^0|\mathbf{X}=\mathbf{x}, D=0\right] \\ &= \mathbb{E}\left[Y_1^0|\mathbf{X}=\mathbf{x}\right] - \mathbb{E}\left[Y_0^0|\mathbf{X}=\mathbf{x}\right]\end{aligned}$$

- Common support

$$\begin{aligned}&\mathbf{Pr}\left(T=1, D=1|\mathbf{X}=\mathbf{x}, (T,D) \in \{(t,d),(1,1)\}\right) < 1 \\ &\forall (t,d) \in \{(0,1),(0,0),(1,0)\}\ x \in \mathcal{X}\end{aligned}$$

This allows us to estimate the conditional ATT as the standard DiD within each X stratum.

$$\mathbb{E}\left[Y_1|D=1,X\right] - \mathbb{E}\left[Y_0|D=1,X\right] - \mathbb{E}\left[Y_0 \mid D=0,X\right] - \mathbb{E}\left[Y_0 \mid D=0,X\right]$$

Averaging these over $\mathsf{d}X$ gives us the ATT

$$\tau_1^{\text{ATT}} = \mathbb{E}\left[\{\mu_1(1,X) - \mu_1(0,X)\} - \{\mu_0(1,X) - \mu_0(0,X)\}|D=1,T=1\right]$$

where regression functions $\mu_d(t,x)$ denote conditional expectations for treatment d at time t given covariates x.

Defn 4.58 (Semiparametric Difference-in-Differences).
Abadie (2005)
Denote potential outcomes under treatment and control for unit i as Y_{it}^1 and Y_{it}^0.
For some observed covariates X_i, we are interested in the CATT

$$\tau_0(X_i) := \mathbb{E}\left[Y_{i1}^1 - Y_{i1}^0 | X_i, D_i = 1\right]$$

For identification, we need

1. Conditional parallel trends: $\mathbb{E}\left[Y_{i1}^0 - Y_{i0}^0 | D_i = 1, X_i\right] = \mathbb{E}\left[Y_{i1}^0 - Y_{i0}^0 | D_i = 0, X_i\right]$
2. Overlap: $\exists c > 0$ such that $\mathbb{E}\left[D_i = 1 | X_i\right] > c$ and $\mathbb{E}\left[D_i | X_i\right] < 1 - c$

The Abadie estimand can be defined as

$$\mathbb{E}\left[Y_{i1}^1 - Y_{i1}^0 | X_i, D_i\right] = \mathbb{E}\left[\underbrace{\frac{D_i - \mathbb{E}\left[D_i = 1 | X_i\right]}{\mathbb{E}\left[D_i = 1 | X_i\right]\left(1 - \mathbb{E}\left[D_i = 1 | X_i\right)\right]}}_{\rho_0}(Y_{i1} - Y_{i0}) | X_i\right]$$

Defining $\Delta Y_i := Y_{i1} - Y_{i0}$, we then have

$$\begin{aligned}\mathbb{E}\left[Y_{i1}^1 - Y_{i1}^0 | X_i, D_i\right] &= \mathbb{E}\left[\frac{D_i - \mathbb{E}\left[D_i = 1 | X_i\right]}{\mathbb{E}\left[D_i = 1 | X_i\right]\mathbb{E}\left[D_i = 0 | X_i\right]}\Delta Y_i | X_i\right] \\ &= \mathbb{E}\left[\frac{D_i \Delta Y_i}{\mathbb{E}\left[D_i = 1 | X_i\right]} | X_i\right] - \mathbb{E}\left[\frac{(1 - D_i)\Delta Y_i}{(1 - \mathbb{E}\left[D_i = 1 | X_i\right])} | X_i\right]\end{aligned}$$

This is an IPW Estimator.
Integrating this over $\mathsf{d}P(X|D = 1)$ gives us the ATT

$$\mathbb{E}\left[Y_1^1 - Y_0^0 | D = 1\right] = \mathbb{E}\left[\frac{Y_1 - Y_0}{\mathbf{Pr}\left(D = 1\right)} \cdot \frac{D - \mathbb{E}\left[D = 1 | X_i\right]}{1 - \mathbb{E}\left[D = 1 | X_i\right]}\right]$$

THe full IPW estimator can be written

$$\begin{aligned}\Delta_{D=1,T=1} = \mathbb{E}[Y \cdot \{&\frac{DT}{\Pi} - \frac{D(1-T)\rho_{1,1}(X)}{\rho_{1,0}(X)\Pi} \\ &- \left(\frac{(1-D)T\rho_{1,1}(X)}{\rho_{0,1}(X)\Pi} - \frac{(1-D)(1-T)\rho_{1,1}(X)}{\rho_{0,0}(X)\Pi}\right)\}]\end{aligned}$$

where $\Pi = \mathbf{Pr}\left(D = 1, T = 1\right)$ is the unconditional probability of being treated in the post-treatment period, and $\rho_{d,t}(X) = \mathbf{Pr}\left(D = dT = t | X\right)$ are conditional probabilities of specific treatment-group combinations.
Double-robust version - Zimmert (2020)

$$\begin{aligned}\Delta_{D=1,T=1} = \mathbb{E}[\{&\frac{DT}{\Pi} - \frac{D(1-T)\rho_{1,1}(X)}{\rho_{1,0}(X)\Pi} \\ &- \left(\frac{(1-D)T\rho_{1,1}(X)}{\rho_{0,1}(X)\Pi} - \frac{(1-D)(1-T)\rho_{1,1}(X)}{\rho_{0,0}(X)\Pi}\right)\} \times \\ &(Y - \mu_d(T, X)) + \frac{DT}{\Pi}(\mu_1(1, X) - \mu_1(0, X) - (\mu_0(1, X) - \mu_0(0, X)))]\end{aligned}$$

4.6 Panel Data

Setup: We observe a sample of $i = 1, \ldots, N$ cross-sectional units for $t = 1, \ldots, T$ time periods $\implies$ Data : $\{(y_{it}, \mathbf{x}_{it}') : t = 1, \ldots, T\}_{t=1}^T$
One-way fixed effects and Random effects both use the form

$$y_{it} = \boldsymbol{x}_{it}'\beta + \underbrace{\theta_i + \epsilon_{it}}_{e_{it}} \tag{3}$$

although they make different assumptions about the error.
Error assumptions for panel regressions
(1) **FE**: $\mathbb{E}\left[\epsilon_{it} | \boldsymbol{x}_i, \theta_i\right] = 0 \Leftrightarrow \theta_i \not\perp\!\!\!\perp \boldsymbol{x}_i$.
(2) **RE**: (1) *and* $\mathbb{E}\left[e_{it} | \boldsymbol{x}_i\right] = 0$ [Absorb unobserved unit effect into error term, impose orthogonality it] $\implies \theta_i \perp\!\!\!\perp \boldsymbol{x}_i$. Equivalent to Pooled OLS with FGLS.

4.6.1 Fixed Effects Regression

Identification Assumption

- **Strict Exogeneity** - errors are uncorrelated with lags and leads of x

  $$\mathbb{E}\left[\epsilon_{it} | x_i\right] = \mathbb{E}\left[\epsilon_{it} | x_{i1}, \cdots x_{iT}\right] = 0 \Leftrightarrow \mathbb{E}\left[\boldsymbol{x}_{is}'\boldsymbol{\epsilon}_{it}\right] = \mathbf{0}\ \forall s, t = 1, \ldots T$$

  Equivalent statement for y_{it} is

  $$\mathbb{E}\left[y_{it} | \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}\right] = \mathbb{E}\left[y_{it} | \boldsymbol{x}_{it}\right] = \boldsymbol{x}_{it}'\boldsymbol{\beta}$$

  - Rules out feedback loops i.e. x_{it} correlated with $\epsilon_{i,t-1}$ because Xs are set in response to prior error, e.g. Policing and crime.
- regressors vary over time for at least some i.

Setup
Define an individual fixed effect for individual i

$$A_i = \begin{cases} 1 \text{ if the observation involves unit i} \\ 0 \text{ otherwise} \end{cases}$$

and define the same for each time period for panel data.
If D_{it} is as good as randomly assigned *conditional on* A_i:

$$E[Y_{0it}|A_i, X_{it}, t, D_{it}] = E[Y_{0it}|A_i, X_{it}, t]$$

Then, assuming A_i enter linearly,

$$E[Y_{0it}|A_i, X_{it}, t, D_{it}] = \alpha + \lambda_t + A_i'\gamma + X_{it}'\beta$$

Assuming the causal effect of the treatment is additive and constant,

$$E[Y_{1it}|A_i, X_{it}, t] = E[Y_{0it}|A_i, X_{it}, t] + \rho$$

where ρ is the causal effect of interest.
Then, we can write:

$$\begin{aligned} Y_{it} &= \alpha_i + \lambda_t +_\rho D_{it} + X_{it}'\beta + \epsilon_{it} \\ \epsilon_{it} &= Y_{0it} - E[Y_{0it}|A_i, X_{it}, t] && \text{Error Term} \\ \alpha_i &= \alpha + A_i\gamma && \text{Fixed effect} \end{aligned}$$

Restrictions

- Linear
- Additive functional form
- Variation in D_{it}, over time, for i, must be as good as random

Defn 4.59 (Within Estimator).
Estimate the specification

$$\ddot{y}_i = \ddot{\boldsymbol{x}}_i'\beta + \epsilon_i$$

where $\ddot{k}_i = M_i k_i$ individual demeaned values from pre-multiplying by the Individual specific demeaning operator $\mathbf{M}_i := \mathbf{I}_i - \mathbf{1}_i \left(\mathbf{1}_i'\mathbf{1}_i\right)^{-1} \mathbf{1}_i'$ with every component in eqn 3, which removes the fixed effect θ_i.

Defn 4.60 (First Differences Estimator).
Lag eqn 3 1 period and subtracting gives

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} + \Delta\epsilon_{it}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and so on. This naturally eliminates the time-invariant fixed effect θ_i. The pooled OLS estimation of β in the above regression is called the **first differences (FD)** estimator $\widehat{\boldsymbol{\beta}}_{FD}$.

Fact 4.22 (Efficiency of FE and FD Estimators).

- FE estimator is more efficient under the assumption that ϵ_{it} are serially uncorrelated $\left[\mathbb{E}\left[\boldsymbol{e}_i\boldsymbol{e}_i'|\boldsymbol{x}_i, \theta_i\right] = \sigma_e^2\mathbf{I}_T\right]$
- FD more efficient when ϵ_{it} follows a random-walk.

Fact 4.23 (Equivalence between Within/FE and first differences for 2 periods).
For Individual Fixed Effects/Within estimation, using the regression anatomy formula, write:

$$\hat{\rho}_{\text{FE}} = \frac{\text{Cov}(\ddot{Y}_{it}, \ddot{D}_{it})}{\text{Var}(\ddot{D}_{it})}$$

Since $t = 2$, $\overline{Y}_i = Y_{it} + \frac{\Delta Y_{it}}{2}$ and $\overline{D}_i = D_{it} + \frac{\Delta D_{it}}{2}$

$$\begin{aligned} \hat{\rho}_{\text{FE}} &= \frac{\text{Cov}(\ddot{Y}_{it}, \ddot{D}_{it})}{\text{Var}(\ddot{D}_{it})} \\ &= \frac{\text{Cov}(Y_{it} - \overline{Y}_i, D_{it} - \overline{D}_i)}{\text{Var}(D_{it} - \overline{D}_i))} \\ &= \frac{\text{Cov}(Y_{it} - Y_{it} - \frac{\Delta Y_{it}}{2}, D_{it} - D_{it} - \frac{\Delta D_{it}}{2})}{\text{Var}(D_{it} - D_{it} - \frac{\Delta D_{it}}{2})} \\ &= \frac{-\text{Cov}(\Delta Y_{it}, \Delta D_{it})}{-\text{Var}(\Delta D_{it})} = \frac{\text{Cov}(\Delta Y_{it}, \Delta D_{it})}{\text{Var}(\Delta D_{it})} \\ &= \hat{\rho}_{\text{FD}} \end{aligned}$$

4.6.2 Random Effects

Identification Assumption
Assume $\theta_i \perp\!\!\!\perp X_i \Leftrightarrow \mathbb{E}\left[\theta_i|\boldsymbol{x}_i\right] = \mathbb{E}\left[\theta_i\right] = 0$ - **strong** assumption
In other words, entire error term $e_{it} = \nu_{it} + \theta_i$ is independent of X. *This assumes OLS is consistent but inefficient*, which is why it is of limited use in observational settings.
When there is autocorrelation in time series (i.e. ϵ_t s are correlated over time), GLS estimates can be obtained by estimating OLS on quasi-differenced data. This allows us to estimate the effects of time-invariant characteristics (assuming the independence condition is met).

$$y_{it} - \lambda\bar{y}_i = (x_{it} - \lambda\bar{x}_i)\beta + (1 - \lambda)\theta_i + \nu_{it} - \lambda\bar{\nu}_i$$

where

$$\lambda = 1 - \left[\frac{\sigma_\nu^2}{\sigma_\nu^2 + T\sigma_\theta^2}\right]^{\frac{1}{2}}$$

Assumption 6 (RE FGLS Assumptions).

- Idiosyncratic errors ν_{it} have constant finite variance: $\mathbb{E}\left[\nu_{it}^2\right] = \sigma_\nu^2$
- Idiosyncratic errors ν_{it} are serially uncorrelated: $\mathbb{E}\left[\nu_{it}\nu_{is}\right] = 0 \forall t \neq s$.
- $\mathbb{E}\left[\theta_i^2 | \boldsymbol{x}_i\right] = \sigma_\theta^2$

Under these assumptions, the FGLS matrix $\boldsymbol{\Omega}$ takes a special form

$$\boldsymbol{\Omega} = \sigma_\nu^2 \mathbf{I}_T + \sigma_\theta^2 \boldsymbol{j}_T \boldsymbol{j}_T'$$

where $\boldsymbol{j}_T \boldsymbol{j}_T'$ is a $T \times T$ matrix of 1s. Estimators for the variance components are in Wooldridge (2010, c 10, pp 260-61). A robust estimator of $\hat{\boldsymbol{\Omega}}$ is constructed using pooled OLS residuals $\hat{\boldsymbol{v}}_i$

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i'$$

With this, we can apply the FGLS estimator

$$\widehat{\boldsymbol{\beta}}_{RE} = \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{y}$$

4.6.3 Hausman Test: Choosing between FE and RE

β_{FE} is **assumed to be consistent**. Oft-abused test as a result.

- H0: $\beta_{FE} - \beta_{RE} = 0$
- H0: $\beta_{FE} - \beta_{RE} \neq 0$

$$\left(\widehat{\boldsymbol{\beta}}_{FE} - \widehat{\boldsymbol{\beta}}_{RE}\right)' \left(\widehat{\text{Var}}\widehat{\boldsymbol{\beta}}_{FE}\right] - \widehat{\text{Var}}\left[\widehat{\boldsymbol{\beta}}_{RE}\right])^{-1} \left(\widehat{\boldsymbol{\beta}}_{FE} - \widehat{\boldsymbol{\beta}}_{RE}\right) \xrightarrow{d} \chi_k^2$$

If the error component θ is correlated with x, RE estimates are not consistent. Perform Hausman test for random vs fixed effects (where under the null, $Cov(\theta_i, x_{it}) = 0$)

- When the idiosyncratic error variance $\hat{\sigma}_\nu^2$ is large relative to $T_i\hat{\sigma}_\theta^2$, $\lambda \to 0$ and $\hat{\beta}_{RE} \approx \hat{\beta}_{pool}$. In words, the individual effect is relatively small, so Pooled OLS is suitable.
- When the idiosyncratic error variance $\hat{\sigma}_\nu^2$ is small relative to $T_i\hat{\sigma}_\theta^2$, $\lambda \to 1$ and $\hat{\beta}_{RE} \approx \hat{\beta}_{FE}$. Individual effects are relatively large, so FE is suitable.

4.6.4 Time Trends

Linear Time Trend

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + t + \varepsilon_{it}, \quad t = 1, 2, \ldots, T$$

Time Fixed Effects (a.k.a. Two-way Fixed Effects)

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + t_t + \varepsilon_{it}, \quad t = 1, 2, \ldots, T$$

Unit Specific Time Trends

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + g_i \cdot t + t_t + \varepsilon_{it}, \quad t = 1, 2, \ldots, T$$

4.6.5 Distributed Lag

Define switching indicator D_{it} as 1 if i switched from control to treatment between $t-1$ and t.

$$Y_{ist} = \gamma_s + \lambda_t + \sum_{\tau=0}^{m} \delta_{-\tau} D_{s,t-\tau} + \sum_{\tau=1}^{q} \delta_{+\tau} D_{s,t+\tau} + X_{ist}'\beta + \epsilon_{ist}$$

where the sums on the RHS allow for m lags / post-treatment effects, and q leads / pre-treatment effects. **Leads should be close to 0**.

4.6.6 Staggered Adoption

Let T denote multiple time periods such that $t \in \{0, 1, \ldots, \mathcal{T}\}$, with nobody treated at $t = 0$ and staggered adoption. Let G_t be a dummy that is equal to one if a subject experiences **treatment introduction** in period t (e.g. $G_2 = 1$ implies the treatment is introduced in period 2 in said group).

Fact 4.24 (Inconsistency for ATT (Chaisemartin and D'Haultfœuille, 2020)). Under parallel trends for the untreated potential outcomes, $Y_{g,t}(0)$, the treatment effect $\widehat{\beta}_{\text{FE}}$ in the vanilla two-way fixed effects regression

$$Y_{g,t} = \widehat{\beta}_{\text{FE}} D_{g,t} + \gamma_g + \psi_t + \varepsilon_{gt}$$

can be decomposed as

$$\mathbb{E}\left[\widehat{\beta}_{\text{FE}}\right] = \mathbb{E}\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t}\Delta_{g,t}\right] \; ; \text{ where } \Delta_{g,t} = Y_{g,t}(1) - Y_{g,t}(0)$$

The weights $W_{g,t}$ sum to one and are proportional to and the same sign as

$$N_{g,t} \underbrace{(D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot})}_{\text{pesky weight}}$$

where $D_{g,\cdot}$ is the average treatment of group g across periods (share of periods treated), $D_{\cdot,t}$ is the average treatment at period t across groups, and $D_{\cdot,\cdot}$ is the grand mean of the treatment indicator. These weights can be negative.
This means that $\widehat{\beta}_{\text{FE}}$ is biased for the ATT because $W_{g,t}$ is in not (only) proportional to $N_{g,t}$. $\widehat{\beta}$ is only unbiased when

- the treatment is binary AND
- the treatment is staggered and absorbing (i.e. groups get treated once and stay treated) AND
- there is no variation in treatment timing

Under these conditions, the pesky weight is constant across treated units, so the weights are proportional to $N_{g,t}$.
OR, $\widehat{\beta}_{\text{FE}}$ is also unbiased if $(D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot})$ is uncorrelated with the treatment effects $\Delta_{g,t}$. This is only plausible *when treatment has been randomly staggered*, otherwise, it is entirely plausible that groups with larger treatment effects selected into treatment early, and so on.

Theorem 4.25 (DiD Decomposition Theorem (Goodman-Bacon, 2018)).
Consider a dataset comprising K timing groups ordered by the time at which they first receive treatment and a maximum of one never-treated group U. The OLS estimate from a two-way fixed effects regression is

$$\hat{\beta}_{DD} = \sum_{k \neq U} s_{kU} \hat{\beta}^{DD}_{kU} + \sum_{k \neq U} \sum_{j > k} \left(s_{kj} \hat{\beta}^{DD}_{kj} + \underbrace{s_{jk} \hat{\beta}^{DD}_{jk}}_{\substack{\text{DD estimated with} \\ \text{already treated group}}} \right)$$

where weights depend on sample size and **variance of treatment** within each DD. This maximises the weights of groups treated in the middle of the panel. The Late vs Early comparison is particularly problematic (and is typically incorrect when treatment effects are heterogeneous in time).
Visually, this involves decomposing the setup in fig 5 into its constituent two-way parts fig 6.

Theorem 4.26 (Group-time average treatment effects (Callaway and Sant'Anna, 2020)).
Estimand: **Group-time average treatment effect**

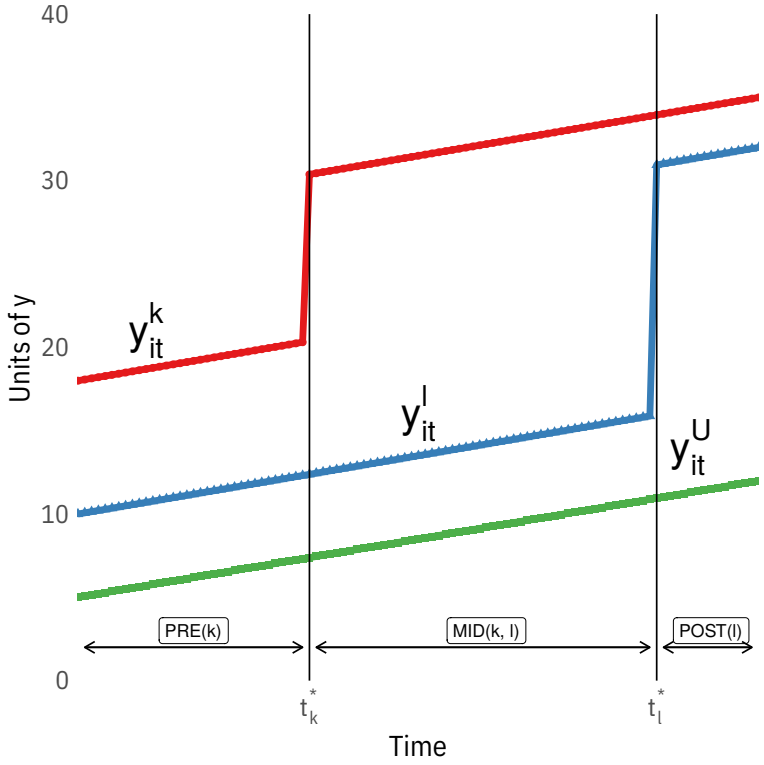


Figure 5: Some Staggered Difference in Differences data

$$ATT(g,t) = \mathbb{E}\left[Y_t(g) - Y_t(\infty) | G_g = 1\right], \ \forall\, t \geq g$$

where $Y_t(g)$ is the potential outcome for group treated at g.
Separate (1) identification, (2) estimation and inference, and (3) aggregation.

- A1: No anticipation $\forall i, t$ and $t < g, g'$, $Y_{it}(g) = Y_{i,t}(g')$
- A2: Parallel trends based on 'never treated' group: $\forall t \in \{2, \ldots, \mathcal{T}\}, g \in \mathcal{G}$ s.t. $t \geq g$, $\underbrace{\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1\right]}_{\text{Trend in group treated at 1}} = \underbrace{\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1\right]}_{\text{Trend in never treated}}$

Estimators for Group-time ATEs

$$ATT^{\text{never}}_{\text{unc}}(g,t) = \mathbb{E}\left[Y_t - Y_{g-1} | G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} | C = 1\right]$$

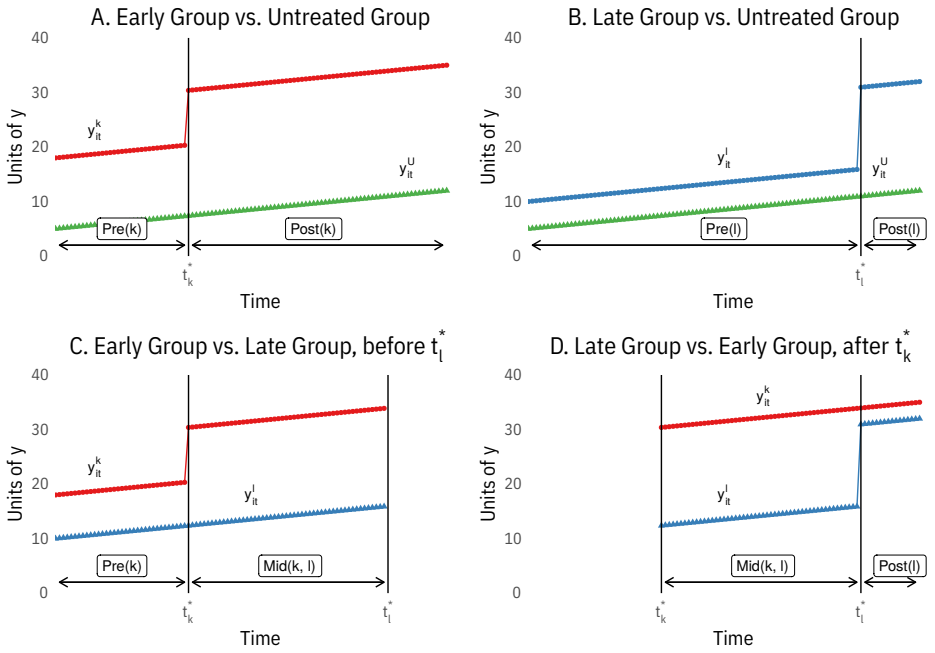


Figure 6: Constituent 2-way Differences in Differences Comparisons

$$ATT^{\text{notyet}}_{\text{unc}}(g,t) = \mathbb{E}\left[Y_t - Y_{g-1} | G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} | D_t = 0, C = 1\right]$$

Aggregation: event-study type estimand.

$$\theta_D(e) = \sum_{g=2}^{\mathcal{T}} \mathbb{1}_{g+e \leq \mathcal{T}} ATT(g, g+e) \mathbf{Pr}\left(G = g | G + e \leq \mathcal{T}, C \neq 1\right)$$

Implemented in `did` and `DRDID`.

Fact 4.27 (Imputation Estimators (IFE / Factor models/ Matrix Completion)).
The negative weighting problem with 2WFE under staggered adoption can be remedied easily by using the following procedure, which is termed Imputation by Liu, Wang, and Xu (2021). This nests the procedures in Xu (2017) and Athey, Bayati, et al. (2017) etc.

- Fit a model for $Y_{it}^{(0)}$ using *only untreated observations* for all units (i.e. untreated periods for units that eventually got treated)
- Impute $\widehat{Y^{(0)}}$ for treated units and treated time periods
- compute $\widehat{\tau}_{it} = Y_{it} - \widehat{Y}_{it}^{(0)} \mid \forall i, t$ where $W_{it} = 1$
- Average for (equal weighting) ATT or average over time for event study

This works well when the outcome model for $Y_{it}^{(0)}$ is good, i.e. when the fixed effects or latent factors are well estimated. This will not work well for short panels.

4.6.7 Changes-in-Changes

Athey and Imbens (2006)
Given a continuous outcome Y and a *monotonicity in unobserved heterogeneity*, CiC allows us to identify both the ATT and Quantile effect on the treated (QTT).
Assume the following about untreated potential outcomes

$$Y_T^0 = \mathcal{H}(U, T) \; U \perp\!\!\!\perp T | D$$

where U is a scalar unobservable or an index of unobservables. $\mathcal{H}(u,t)$ is a general function assumed to be strictly monotonically increasing in values of u for periods $t \in \{0,1\}$. The conditional independence assumption requires that the unobserved heterogeneity is constant over time within treatment groups.
Denote $\mathbb{F}_{Y(d)|dt}(y) = \mathbb{P}\left[Y(d) \leq y | D = d, T = t\right]$ the conditional CDF of potential outcome $Y(d)$, and $\mathbb{F}_{dt}(y) = \mathbb{P}\left[D = d, T = t\right]$ corresponding CDF for observed outcome. Conditional outcome distributions $\mathbb{F}_{01}, \mathbb{F}_{00}, \mathbb{F}_{10}$ are observed. The inverse of the latter is $\mathbb{F}_{dt}^{-1}(y)$, the conditional quantile function. The unobserved CDF is identified as

$$\mathbb{F}_{Y(0)|11}(y) = \mathbb{F}_{10}\left(\mathbb{F}_{00}^{-1}(\mathbb{F}_{01}(y))\right)$$

The QTT at quantile τ is then identified as

$$\Delta_{D=1}(\tau) = \mathbb{F}_{11}^{-1}(\tau) - \underbrace{\mathbb{F}_{(0)|11}(\tau)^{-1}}_{\mathbb{F}_{01}^{-1}\left(\mathbb{F}_{00}(\mathbb{F}_{10}^{-1}(\tau))\right)}$$

and the ATT is identified as

$$\Delta_{D=1} = \mathbb{E}\left[Y | D = 1, T = 1\right] - \mathbb{E}\left[\mathbb{F}_{01}^{-1}(\mathbb{F}_{00}(Y_{10}))\right]$$

Implemented in `qte::CiC`.

4.6.8 Synthetic Control

Original Abadie, Diamond, and Hainmueller (2010) setup.
Observe n_0+1 units in periods $t = 1, \ldots, T$. Unit 1 is treated starting from period $T_0 + 1$, while $2, \ldots, n_0 + 1$ are never treated, and are therefore called the *donor pool*.

$$Y_{it}^{obs} = Y_{it}(D_{it}) = \begin{cases} Y_{it}(0) & \text{if } D_{it} = 0 \\ Y_{it}(1) & \text{if } D_{it} = 1 \end{cases}$$

Since there is only 1 treated unit, the effect of interest

$$\tau_t := Y_{1i}(t) - Y_{0i}(t),\ t = T_0{+}1, \ldots, T$$

Observed data matrix ((Doudchenko and Imbens, 2016))

$$\mathbf{Y}^{obs} := (Y_{it}^{obs})_{t=T,\ldots,1,\ i=1,\ldots,n_0+1} = \begin{pmatrix} Y_{1,T}(1) & Y_{2,T}(0) & \ldots, & Y_{n_0+1,T}(0) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1,T_0+1}(1) & Y_{2,T_0+1}(0) & \ldots, & Y_{n_0+1,T_0+1}(0) \\ Y_{1,T_0}(1) & Y_{2,T_0}(0) & \ldots, & Y_{n_0+1,T_0}(0) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1,1}(1) & Y_{2,1}(0) & \ldots, & Y_{n_0+1,1}(0) \end{pmatrix}$$

FPCI applies; potential outcome matrices are:

$$\mathbf{Y}(0) = \begin{pmatrix} ? & Y_{2,T}(0) & \ldots, & Y_{n_0+1,T}(0) \\ \vdots & \vdots & \vdots & \vdots \\ ? & Y_{2,T_0+1}(0) & \ldots, & Y_{n_0+1,T_0+1}(0) \\ Y_{1,T_0}(1) & Y_{2,T_0}(0) & \ldots, & Y_{n_0+1,T_0}(0) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1,1}(1) & Y_{2,1}(0) & \ldots, & Y_{n_0+1,1}(0) \end{pmatrix}$$

$$\mathbf{Y}(1) = \begin{pmatrix} Y_{1,T}(1) & ? & \ldots, & ? \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1,T_0+1}(1) & & \ldots, & \\ ? & ?) & \ldots, & ? \\ \vdots & \vdots & \vdots & \vdots \\ ? & ? & \ldots, & ? \end{pmatrix}$$

Let $\boldsymbol{X}_{treat}$ be a $p-$vector of a pre-intervention characteristics, and $\mathbf{X}_c$ is a $p \times n_0$ matrix containing the same values for control units. This typically includes pre-treatment outcomes, in which case $p = T_0$, but predictors (even time invariant ones, Z_i) are usually available.

$$\boldsymbol{X}_{treat} := \begin{pmatrix} Y_{1,1}^{obs} \\ Y_{1,2}^{obs} \\ \vdots \\ Y_{1,T_0}^{obs} \\ \boldsymbol{Z}_i \end{pmatrix}$$

Defn 4.61 (Synthetic Control Estimator).
For some $p \times p$ PSD matrix $\mathbf{V}$, define $||\mathbf{X}||_{\mathbf{V}} = \sqrt{\mathbf{X}'\mathbf{V}\mathbf{X}}$, where $\mathbf{V}$ is typically diagonal. Consider weights $\boldsymbol{\omega} = (\omega_2, \ldots, \omega_{n_0+1})$ satisfying

$$\omega_i \geq 0,\ 2, \ldots, n_0 + 1 \qquad \text{(Non-Negativity)}$$
$$\sum_{i \geq 2} \omega_i = 1 \qquad \text{(Sum to 1)}$$

This forces *interpolation*, i.e. the counterfactual cannot take a value greater than the maximal value or smaller than the minimal value of for a control unit. The synthetic control solution $\boldsymbol{\omega}^*$ solves

$$\min_{\boldsymbol{\omega}} ||\mathbf{X}_{treat} - \mathbf{X}_c\boldsymbol{\omega}||_{\mathbf{V}}^2 \quad \text{s.t. Non-negativity, Sum to 1}$$

The Synthetic Control Estimator is then

$$\widehat{\tau}_t := Y_{1,t}^{obs} - \sum_{i=2}^{n_0+1} \omega_i^* Y_{it}^{obs}$$

In contrast, a simple difference-in-differences estimator gives

$$\widehat{\tau}_t^{DID} := Y_{1,t}^{obs} - \left(Y_{1,T_0}^{obs} - \frac{1}{n_0} \sum_{i=2}^{n_0+1} Y_{it}^{obs} - Y_{i,T_0}^{obs} \right)$$

Abadie, Diamond, and Hainmueller (2010) choose $\mathbf{V} = \operatorname{diag} v_1, \ldots, v_p$ using a nested-minimisation of the Mean Square Prediction Error (MSPE) over the **pre**-treatment period

$$\text{MSPE}(\mathbf{V}) := \sum_{t=1}^{T_0} Y_{1,t}^{obs} - \sum_{i=2}^{n_0+1} \omega_i(\mathbf{V}) Y_{it}^{obs}$$

Defn 4.62 (Imbens and Doudchenko representation).
Doudchenko and Imbens (2016) Setup:

$$\mathbf{Y}^{\text{obs}} = \begin{bmatrix} \mathbf{Y}^{\text{obs}}_{\text{t, post}} & \mathbf{Y}^{\text{obs}}_{\text{c, post}} \\ \mathbf{Y}^{\text{obs}}_{\text{t, pre}} & \mathbf{Y}^{\text{obs}}_{\text{c, pre}} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\text{t, post}}(1) & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} \qquad T \times (N+1)$$

$$\mathbf{Y}(0) = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix} = \begin{bmatrix} ? & \mathbf{Y}_{\text{c, post}}(0) \\ \mathbf{Y}_{\text{t, pre}}(0) & \mathbf{Y}_{\text{c, pre}}(0) \end{bmatrix}$$

- relative magnitudes of T and N might dictate whether we impute the missing potential outcome ? using this or this comparison
  - **Many Units and Multiple Periods**: $N >> T_0$, $\mathbf{Y}(0)$ is 'fat', and red comparison becomes challenging relative to blue. So matching methods are attractive.
  - $T_0 >> N$, $\mathbf{Y}(0)$ is 'tall', and matching becomes infeasible. So it might be easier to estimate blue dependence structure.
  - Finally, if $T_0 \approx N$, regularization strategy for limiting the number of control units that enter into the estimation of $Y_{0,T_0+1}(0)$ may be important
- Focus on last period for now: $\tau_{0,T} = Y_{0,T}(1) - Y_{0,T}(0) = Y_{0,T}^{\text{obs}} - Y_{0,T}(0)$
- Many estimators impute $Y_{0,T}(0)$ with the linear structure $\widehat{Y}_{0,T}(0) = \mu + \sum_{i=1}^{n} \omega_i \cdot Y_{i,T}^{\text{obs}}$
  - Methods differ in how μ and ω are chosen as a function of $\mathbf{Y}_{\text{c, post}}^{\text{obs}}$, $\mathbf{Y}_{\text{t, pre}}^{\text{obs}}$, $\mathbf{Y}_{\text{c, pre}}^{\text{obs}}$
- Impose four constraints
  1. **No Intercept**: $\mu = 0$. Stronger than Parallel trends in DiD.
  2. **Adding up** : $\sum_{i=1}^{n} \omega_i = 1$. Common to DiD, SC.
  3. **Non-negativity**: $\omega_i \geq 0 \; \forall \; i$. Ensures uniqueness via 'coarse' regularisation + precision control. Negative weights may improve out-of-sample prediction.
  4. **Constant Weights**: $\omega_i = \overline{\omega} \; \forall \; i$
- DiD imposes 2-4.
- ADH(2010, 2014) impose 1-3
  - 1 + 2 imply 'No Extrapolation'.

Relaxing these assumptions:

- Negative weights
  - If treated units are outliers on important covariates, negative weights might improve fit
  - Bias reduction - negative weights increase bias-reduction rate
- When $N >> T_0$, (1-3) alone might not result in a unique solution. Choose by
  - Matching on pre-treatment outcomes : one good control unit is better than synthetic one comprised of disparate units
  - Constant weights - implicit in DiD
- Given many pairs of (μ, ω)
- prefer values s.t. synthetic control unit is similar to treated units in terms of lagged outcomes
- low dispersion of weights
- few control units with non-zero weights

Optimisation Problem
Ingredients of objective function

- **Balance**: difference between pre-treatment outcomes for treated and linear-combination of pre-treatment outcomes for control
  - $\left\| \mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}} \right\|_2^2 = (\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}})^\top (\mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}})$
- **Sparse and small weights**:
  - sparsity : $\|\omega\|_1$
  - magnitude: $\|\omega\|_2$

$$(\widehat{\mu}^{en}(\lambda, \alpha), \widehat{\omega}^{en}(\lambda, \alpha)) = \underset{\mu, \omega}{\text{argmin}} \; Q(\mu, \omega | \mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha)$$

$$\text{where} \quad Q(\mu, \omega | \mathbf{Y}_{\text{t, pre}}, \mathbf{Y}_{\text{c, pre}}; \lambda, \alpha) = \left\| \mathbf{Y}_{\text{t, pre}} - \mu - \omega^\top \mathbf{Y}_{\text{c, pre}} \right\|_2^2 + \lambda \left(\frac{1 - \alpha}{2} \|\omega\|_2^2 + \alpha \|\omega\|_1 \right)$$

Tailored Regularisation

- don't want to scale covariates $\mathbf{Y}_{\text{c, pre}}$ to preserve interpretability of weights
- Instead, treat each control unit as a 'pseudo-treated' unit and compute $\widehat{Y}_{j,T}(0) = \widehat{\mu}^{\text{en}}(j; \alpha, \lambda) + \sum_{i \neq j} \widehat{\omega}_i(j; \alpha, \lambda) \cdot Y_{i,T}^{\text{obs}}$ where

$$(\widehat{\mu}^{en}(j;\lambda,\alpha),\widehat{\omega}^{en}(j;\lambda,\alpha)) = \underset{\mu,\omega}{\operatorname{argmin}} \sum_{t=1}^{T_0}\left(Y_{j,t} - \mu - \sum_{i\neq 0,j}\omega_i Y_{i,t}\right)^2 + \lambda\left(\frac{1-\alpha}{2}\|\omega\|_2^2 + \alpha\|\omega\|_1\right)$$

pick the value of the tuning parameters $(\alpha^{en}_{opt}, \lambda^{en}_{opt})$ that minimises

$$CV^{en}(\alpha,\lambda) = \frac{1}{N}\sum_{j=1}^{N}(Y_{j,T} - \overbrace{\widehat{\mu}^{en}(j;\alpha,\lambda) - \sum_{i\neq 0,j}\widehat{\omega}_i^{en}(j;\alpha,\lambda)\cdot Y_{i,T}}^{\widehat{Y}_{j,T}(0)})$$

Difference in Differences

- assume (2-4)
- No unique μ, ω solution for $T = 2$, so fix $\omega = \frac{1}{N}$

$$\omega_i^{\text{did}} = \frac{1}{N} \quad \forall i \in \{1,\dots N\}$$
$$\widehat{\mu}^{\text{did}} = \frac{1}{T_0}\sum_{s=1}^{T_0} Y_{0,s} - \frac{1}{NT_0}\sum_{s=1}^{T_0}\sum_{i=1}^{N} Y_{i,s}$$

Best Subset; One-to-one Matching
$(\widehat{\mu}^S, \widehat{\omega}^S) = \operatorname{argmin}_{\mu,\omega} Q(\cdot;\lambda=0,\alpha)$ with $\sum_{i=1}^{N}\mathbb{1}_{\omega_i\neq 0} \leq k$ (=1 for OtO)
Synthetic Control

- assume (1-3) (i.e. $\mu = 0$)
- For $M \times M$ PSD diagonal matrix $\mathbf{V}$

$$(\widehat{\omega}(\mathbf{V}),\widehat{\mu}(\mathbf{V})) = \underset{\omega,\mu}{\operatorname{argmin}}\{(\mathbf{X}_t - \mu - \omega^\top\mathbf{X})^\top\mathbf{V}(\mathbf{X}_t - \mu - \omega^\top\mathbf{X})\}$$
$$\widehat{\mathbf{V}} = \underset{\mathbf{V}=\text{diag}(v_1,\dots,v_M)}{\operatorname{argmin}}\{(\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}(\mathbf{V})^\top\mathbf{Y}_{\text{c, pre}})^\top(\mathbf{Y}_{\text{t, pre}} - \widehat{\omega}(\mathbf{V})^\top\mathbf{Y}_{\text{c, pre}})\}$$

Constrained regression: When $X_i = Y_{i,t};\ 1 \leq t \leq T_0$ (Lagged Outcomes only) $\mathbf{V} = \mathbf{I}_N$ and $\lambda = 0$

Defn 4.63 (Many treated units : Synthetic Difference in Differences).
Arkhangelsky et al. (2021)
Consider a balanced panel with N units and T time periods, where the first N_{co} units are never treated, while $N_{tr} = N - N_{co}$ treated units are exposed after time T_{pre}. We seek to solve for sdid weights $\widehat{\omega}^{\text{sdid}}$ that *align pre-exposure trends in outcomes of unexposed units with those for exposed units*

$$\sum_{i=1}^{N_{co}}\widehat{\omega}^{sdid}Y_{it} \approx N_{tr}^{-1}\sum_{i=N_{co}+1}^{N} Y_{it}$$

we also look for time weights $\widehat{\lambda}_t^{sdid}$ that balance pre-exposure time periods with post-exposure time periods for unexposed units.
Weights are solved using the following optimisation problems

$$\left(\hat{\omega}_0,\hat{\omega}^{\text{sdid}}\right) = \underset{\omega_0\in\mathbb{R},\omega\in\Omega}{\arg\min}\,\ell_{\text{unit}}\,(\omega_0,\omega) \quad \text{where}$$
$$\ell_{\text{unit}}\,(\omega_0,\omega) = \sum_{t=1}^{T_{\text{pre}}}\left(\omega_0 + \sum_{i=1}^{N_{co}}\omega_i Y_{it} - \frac{1}{N_{\text{tr}}}\sum_{i=N_{\text{co}}+1}^{N} Y_{it}\right)^2 + \zeta^2 T_{\text{pre}}\|\omega\|_2^2,$$
$$\Omega = \left\{\omega\in\mathbb{R}_+^N : \sum_{i=1}^{N_{\text{co}}}\omega_i = 1, \omega_i = N_{\text{tr}}^{-1} \text{ for all } i = N_{\text{co}}+1,\dots,N\right\},$$

where $\mathbb{R}_+$denotes the positive real line. We set the regularization parameter ζ as

$$\zeta = (N_{\text{tr}}T_{\text{post}})^{1/4}\,\hat{\sigma} \text{ with } \hat{\sigma}^2 = \frac{1}{N_{\text{co}}\,(T_{\text{pre}}-1)}\sum_{i=1}^{N_{\text{co}}}\sum_{t=1}^{T_{\text{pre}}-1}\left(\Delta_{it} - \bar{\Delta}\right)^2,$$

$$\text{where } \Delta_{it} = Y_{i(t+1)} - Y_{it}, \quad \text{and } \bar{\Delta} = \frac{1}{N_{\text{co}}\,(T_{\text{pre}}-1)}\sum_{i=1}^{N_{\text{co}}}\sum_{t=1}^{T_{\text{pre}}-1}\Delta_{it}.$$

and
We implement this for the time weights $\hat{\lambda}^{\text{sdid}}$ by solving

$$\left(\hat{\lambda}_0, \hat{\lambda}^{\text{sdid}}\right) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\, \ell_{\text{time}}\ (\lambda_0, \lambda) \quad \text{where}$$

$$\ell_{\text{time}}\ (\lambda_0, \lambda) = \sum_{i=1}^{N_{\text{co}}} \left(\lambda_0 + \sum_{t=1}^{T_{\text{pre}}} \lambda_t Y_{it} - \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it} \right)^2,$$

$$\Lambda = \left\{ \lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T_{\text{pre}}} \lambda_t = 1, \lambda_t = T_{\text{post}}^{-1} \text{ for all } t = T_{\text{pre}}+1, \ldots, T \right\}$$

1. Compute regularisation parameter ζ
2. Compute unit weights $\widehat{\omega}^{\text{sdid}}$
3. Compute time weights $\widehat{\lambda}^{\text{sdid}}$
4. Compute the SDID estimator using the following weighted DID regression

$$(\hat{\tau}^{\text{sdid}}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}) = \underset{\tau, \mu, \alpha, \beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau)^2 \widehat{\omega}_i^{sdid} \widehat{\lambda}_t^{sdid} \right\}$$

implemented in `synthdid::synthdid_estimate`

Defn 4.64 (Interactive Fixed Effects (Bai, 2009)).

$$Y_{it} = \delta_{it} D_{it} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + \varepsilon_{it}$$

Where D is the treatment, δ_{it} is the heterogeneous treatment effect for unit i at time t, $\boldsymbol{x}_{it}$ is a $p-$ vector of time-varying controls. $\boldsymbol{f}_t = [f_{1t}, \ldots, f_{rt}]'$ is a $k \times 1$ vector of unknown **common factors**, $\boldsymbol{\lambda}_i = [\lambda_{i1}, \ldots, \lambda_{ir}]'$ is a $r \times 1$ vector of unknown **factor loadings**. This factor component nests standard functional forms

$$\underbrace{U_{it}}_{\text{Confounders}} = \underbrace{\lambda_i}_{\text{Loadings}} \times \underbrace{f_t}_{\text{factors}}$$

- $f_t = 1 \implies \lambda_i \times 1 = \lambda_i$ **unit FEs**
- $\lambda_i = 1 \implies 1 \times f_t = f_t$ **time FEs**
- $f_{1t} = 1, f_{2t} = \xi_t, \lambda_{i1} = \alpha_i, \lambda_{i2} = 1 \implies f_t \times \lambda_i = \alpha_i + \xi_t$ **two-way FEs**.
- $f_t = t \implies \lambda_i \times f_t = \lambda_i \times t$ **Unit-specific linear time trends**
- $\lambda_i = y_{i0}, f_t = \alpha_t \implies \lambda_i \times f_t = \alpha y_{i,t-1} - \nu_{it}$ **Lagged dependent variable**

Steps

1. Get initial value of $\widehat{\beta}$ using within estimator
2. Estimate $\widehat{\lambda}_i, \widehat{f}_t$ using $\widehat{\beta}$
3. Re-estimate $\widehat{\beta}$ using $\widehat{\lambda}_i{}' \widehat{f}_t$
4. Iterate

Drawback - constant effect

Defn 4.65 (Generalized Synthetic Control (Xu, 2017)).
With, N_{CO} control units and N_{TR} treated units, Write DGP for individual unit as

$$\boldsymbol{Y}_i = \mathbf{D}_i \circ \boldsymbol{\delta}_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i \quad ; i \in 1, 2, \ldots N_{CO}, N_{CO}+1, \ldots, N$$

Where $\boldsymbol{Y}_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]'$, $\boldsymbol{D}_i = [D_{i1}, \ldots, D_{iT}]'$, $\mathbf{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}]'$ is $T \times k$, $\mathbf{F} = [f_1, \ldots, f_T]'$ is $T \times r$.
Stack controls together gives

$$\underbrace{\mathbf{Y}_{CO}}_{T \times N_{CO}} = \underbrace{\mathbf{X}_{CO}}_{T \times N_{CO} \times p} \underbrace{\boldsymbol{\beta}}_{p \times 1} + \underbrace{\mathbf{F}}_{T \times N_{CO}} \underbrace{\boldsymbol{\Lambda}_{CO}'}_{N_{CO} \times r} + \boldsymbol{\varepsilon}_{CO}$$

GSC for treatment effects is an out-of-sample prediction method: the treatment effect for unit i at time t is the difference between teh actual outcome and its estimated coutnerfactual $\widehat{\delta}_{it} = Y_{it}(1) - \widehat{Y}_{it}(0)$, where $\widehat{Y_{it}}(0)$ is imputed in three steps.

1. **Estimate an IFE model using only the control group data and estimate $\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{F}}, \widehat{\boldsymbol{\Lambda}}_{CO}$**

$$\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{F}}, \widehat{\boldsymbol{\Lambda}}_{CO} = \underset{\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{F}}, \widetilde{\boldsymbol{\Lambda}}}{\text{argmin}} \sum_{i \in \mathcal{C}} (\mathbf{Y}_i - \mathbf{X}_i\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{F}}\widetilde{\boldsymbol{\Lambda}}_i)'(\mathbf{Y}_i - \mathbf{X}_i\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{F}}\widetilde{\boldsymbol{\Lambda}}_i)$$
$$\text{s.t. } \widetilde{\mathbf{F}}'\widetilde{\mathbf{F}} = \mathbf{I}_r; \quad \widetilde{\boldsymbol{\Lambda}}_{CO}'\widetilde{\boldsymbol{\Lambda}}_{CO} = \text{Diagonal}$$

2. **Estimate Factor loadings for each treated unit by minimising mean-squared error of the predicted treated outcome in pretreatment periods**

$$\widehat{\boldsymbol{\Lambda}}_i = \underset{\widetilde{\boldsymbol{\Lambda}}_i}{\text{argmin}} (\mathbf{Y}_i^0 - \mathbf{X}_i^0\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{F}}^0\widetilde{\boldsymbol{\Lambda}}_i)'(\mathbf{Y}_i^0 - \mathbf{X}_i^0\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{F}}^0\widetilde{\boldsymbol{\Lambda}}_i)'$$
$$= \left(\widehat{\mathbf{F}}^{0\prime}\widehat{\mathbf{F}}^0\right)^{-1} \widehat{\mathbf{F}}^{0\prime}(\mathbf{Y}_i^0 - \mathbf{X}_i^0\widehat{\boldsymbol{\beta}}) \; i \in \mathcal{T}$$

where 0 superscripts denote the pretreatment periods.

3. **Calculate Treated Counterfactuals based on $\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{F}}, \widehat{\boldsymbol{\Lambda}}_i$**

$$\widehat{Y_{it}}(0) = \boldsymbol{x}_{it}'\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\lambda}}_i'\widehat{\boldsymbol{f}}_t \;\; ; i \in \mathcal{T} \; ; t > T_0$$

Choose the number of factors r by cross-validation. Implemented in `gsynth`.

4.6.9 Dynamic Treatment Effects

We may want to estimate the effects of treatment *sequences* ('time-varying exposures'), as in medical settings (Robins 1986, Robins, Hernan, and Brumback (2000)).

2 period example Consider a setting with $t = 1, 2$ and corresponding outcomes Y_t and treatments D_t, where the treatment takes on values $d_1, d_2 \in \{0, 1, \ldots, J\}$, and baseline covariates $\mathbf{X}_0$ and covariates at the end of the first period $\mathbf{X}_1$.
Let $\mathbf{d}_2 := (d_1, d_2) \in \{0, 1, \ldots, J\} \times \{0, 1, \ldots, J\}$. Accordingly, $Y_2(\mathbf{d}_2)$ is the potential outcome realised when treatment is set to sequence $\mathbf{d}_2$. The ATE (contrast) two distinct treatment sequences $\mathbf{d}_2$ vs $\mathbf{d}_2'$ is

$$\Delta(\mathbf{d}_2, \mathbf{d}_2') := \mathbb{E}\left[Y_2(\mathbf{d}_2) - Y_2(\mathbf{d}_2')\right]$$

Estimating this quantity requires a sequential selection on observables assumption

$$Y_2(\mathbf{d}_2) \perp\!\!\!\perp D_1 | \mathbf{X}_0 \text{ and } Y_2(\mathbf{d}_2) \perp\!\!\!\perp D_2 | D_1, \mathbf{X}_0, \mathbf{X}_1, \text{ for } d_1, d_2 \in \{0, 1, \ldots J\}$$
$$\mathbf{Pr}\left(D_1 = d_1 | \mathbf{X}_0\right) > 0 \text{ and } \mathbf{Pr}\left(D_2 = d_2 | D_1, \mathbf{X}_0, \mathbf{X}_1\right) > 0$$

Under these assumptions, dynamic treatment effects can be estimated based on nested conditional means regressions

$$\Delta^{\text{snmm}}(\mathbf{d}_2, \mathbf{d}_2') := \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y_2 | \mathbf{d}_2, \mathbf{X_0}, \mathbf{X_1}\right] | d_1, \mathbf{X}_0\right] - \mathbb{E}\left[\mathbb{E}\left[Y_2 | \mathbf{d}_2', \mathbf{X_0}, \mathbf{X_1}\right] | d_1', \mathbf{X}_0\right]\right]$$

where $\mathbf{d}_2 = (d_1, d_2)$ and $\mathbf{d}_2' = (d_1', d_2')$ denote distinct treatment sequences.
or an IPW estimator

$$\Delta^{\text{ipw}}(\mathbf{d}_2, \mathbf{d}_2') := \mathbb{E}\left[\frac{Y \cdot \mathbb{1}_{D_1 = d_1} \mathbb{1}_{D_2 = d_2}}{p^{d_1}(\mathbf{X}_0) p^{d_2}(D_1, \mathbf{X}_0, \mathbf{X}_1)} - \frac{Y \cdot \mathbb{1}_{D_1 = d_1'} \mathbb{1}_{D_2 = d_2'}}{p^{d_1'}(\mathbf{X}_0) p^{d_2'}(D_1, \mathbf{X}_0, \mathbf{X}_1)} -\right]$$

where $p^{d_1}(\mathbf{X}_0)$ and p^{d_2} are propensity scores in the two periods.
Finally, a double robust estimator is

$$\Delta^{\text{dr}}(\mathbf{d}_2, \mathbf{d}_2') = \mathbb{E}\left[\psi^{\mathbf{d}_2} - \psi^{\mathbf{d}_2'}\right]$$
$$\text{where } \psi^{\mathbf{d}_2} = \frac{\mathbb{1}_{D_1 = d_1} \cdot \mathbb{1}_{D_2 = d_2} \cdot \left(Y_2 - \mu^{Y_2}(\mathbf{d}_2, \mathbf{X}_1)\right)}{p^{d_1}(\mathbf{X}_0) p^{d_2}(D_1, \mathbf{X}_0, \mathbf{X}_1)} + \frac{\mathbb{1}_{D_1 = d_1} \cdot \left(\mu^{Y_2}(\mathbf{d}_2, \mathbf{X}_1) - \nu^{Y_2}(\mathbf{d}_2, \mathbf{X}_0)\right)}{p^{d_1}(\mathbf{X}_0)} + \nu^{Y_2}(\mathbf{d}_2, \mathbf{X}_0)$$

where

$$\mu^{Y_2}(\mathbf{d}_2, \mathbf{X_0}, \mathbf{X_1}) = \mathbb{E}\left[Y_2 | \mathbf{D}_2 = \mathbf{d}_2, \mathbf{X_0}, \mathbf{X_1}\right] \text{ and}$$
$$\nu^{Y_2}(\mathbf{d}_2, \mathbf{X}_0) = \mathbb{E}\left[\mathbb{E}\left[Y_2 | \mathbf{d}_2', \mathbf{X_0}, \mathbf{X_1}\right] | D_1 = d_1', \mathbf{X}_0\right]$$

are (nested) conditional mean outcomes.
If we assume that D_2 is conditionally independent of potential outcomes given pre-treatment covariates $\mathbf{X}_0$ and D_1 (implying that post-treatment $\mathbf{X}_1$ aren't required to control for confounders jointly affecting the second treatment and the outcome).
In this case, the second part of the first SOO assumption can be strengthened to $Y(\mathbf{d}_2) \perp\!\!\!\perp D_2 | D_1, \mathbf{X}_1$. This simplifies

$$\psi^{\mathbf{d}_2} = \frac{\mathbb{1}_{D_1 = d_1} \cdot \mathbb{1}_{D_2 = d_2} \cdot \left(Y_2 - \mu^{Y_2}(\mathbf{d}_2, \mathbf{X}_0)\right)}{p^{d_1}(\mathbf{X}_0) p^{d_2}(d_1, \mathbf{X}_0)} + \mu^{Y_2}(\mathbf{d}_2, \mathbf{X}_0)$$

implemented in `causalweight::dyntreatDML`.

Generalisation to arbitrary panels (Blackwell and Glynn, 2018; Hernán, Brumback, and Robins, 2001)
Let D_{it} denote treatment status at time t, and collect them into a $t-$ vector for each unit to form a **Treatment History** $\mathbf{D}_i := (D_{i1}, D_{i2}, \ldots, D_{iT})$. A **partial treatment history** up to time t is denoted $\mathbf{D}_{i,1:t}$. Time varying covariates are arranged analogously $X_{it}, \mathbf{X}_{it}, \mathbf{X}_{i,1:t}$.
Potential outcomes are defined on *treatment histories* and rely on the standard consistency assumption / SUTVA, which assumes that the potential outcome for the same observed history $Y_{it} := Y_{it}(\mathbf{d}_{1:t})$ when $\mathbf{D}_{i,1:t} = \mathbf{d}_{1:t}$. This generates 2^t potential outcomes for the outcome in period t, which permits many hypothetical comparisons.
The estimand typically of interest the *average causal effect* of a treatment history

$$\tau(\mathbf{d}_{1:t}, \mathbf{d}_{1:t}') := \mathbb{E}\left[Y_{it}(\mathbf{d}_{1:t}) - Y_{it}(\mathbf{d}_{1:t}')\right]$$

Define potential outcomes just intervening on the last j periods as $Y_{it}(\mathbf{D}_{i,1:t-j-1}, \mathbf{d}_{t-j:t})$, which is the 'marginal' potential outcome if the treatment history runs its natural course up to $t - j - 1$ and set the last j lags to $\mathbf{d}_{t-j:t}$.
This allows us to define a *contemporaneous treatment effect* (CET)

$$\tau_c(t) = \mathbb{E}\left[Y_{it}(\mathbf{D}_{i,1:t-1}, 1) - Y_{it}(\mathbf{D}_{i,1:t-1}, 0)\right] = \mathbb{E}\left[Y_{it}(1) - Y_{it}(0)\right]$$

The $j-$*step lagged effect* is defined analogously

$$\tau_l(t,j) := \mathbb{E}\left[Y_{it}(\mathbf{D}_{i,1:t-j-1}, 1, \mathbf{0}_j) - Y_{it}(\mathbf{D}_{i,1:t-j-1}, 0, \mathbf{0}_j)\right]$$

and the *step response function* (SRF) describes how this effect varies by time period and distance between the shift and the outcome

$$\tau_s(t,j) = \mathbb{E}\left[Y_{it}(\mathbf{1}_j) - Y_{it}(\mathbf{0}_j)\right]$$

These effects are (clunkily) parametrised in an autoregressive distributed-lag (ADL) models of the form

$$Y_{it} = \beta_0 + \alpha Y_{i,t-1} + \beta_1 D_{it} + \beta_2 D_{i,t-1} + \varepsilon_{it}$$

with assumption $\varepsilon_{it} \perp\!\!\!\perp \mathbf{D}_{i,s}\ \forall t, s$. This implies the following form for potential outcomes

$$Y_{it}(\mathbf{d}_{1:t}) = \beta_0 + \alpha Y_{i,t-1}(\mathbf{d}_{1:t-1}) + \beta_1 D_{it} + \beta_2 D_{i,t-1} + \varepsilon_{it}$$

hence, changes in d_{t-1} can have both a direct and indirect effect on Y_{it}.

Identification Assumption 2 (Baseline Randomisation).

$$\{Y_{it}(\mathbf{d}_{1:t}) : t = 1, \ldots, T\} \perp\!\!\!\perp \mathbf{D}_{i,1:t} | X_{i,0}$$

This relates to linear panel models of the form

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 D_{i,t-1} + \eta_{it}$$

where *strict exogeneity* $\mathbb{E}\left[\eta_{it} | \mathbf{D}_{i,1:T}\right] = \mathbb{E}\left[\eta_{it}\right] = 0$ is assumed.

Identification Assumption 3 (Sequential Ignorability).
For every treatment history $\mathbf{d}_{1:T}$ and period t,

$$\{Y_{is}(\mathbf{d}_{1:s}) : s = 1, \ldots, T\} \perp\!\!\!\perp \mathbf{D}_{i,1:t} | \mathbf{V_{it}}$$

where $\mathbf{V}_{it}$ is a set of covariates such as $\{Y_{i,t-1}, D_{i,t-1}, \mathbf{X}_{it}\}$.
This relates to *sequential exogeneity* in panel models

$$\mathbb{E}\left[\varepsilon_{it} | \mathbf{D}_{i,1:t}, \mathbf{X}_{i,1:t} Y_{i,1:t-1}\right] = \mathbb{E}\left[\varepsilon_{it} | D_{it}, \mathbf{V}_{it}\right] = 0$$

Under sequential ignorability, an ADL approach would be to write the outcome regression with time-varying covariates

$$Y_{it} = \beta_0 + \alpha Y_{i,t-1} + \beta_1 D_{it} + \beta_2 D_{i,t-1} + \mathbf{X}_{it}'\delta + \varepsilon_{it}$$

This generates post-treatment bias because $\mathbf{X}_{it}$ may be affected by $\mathbf{D}_{i,1:t-1}$.

Defn 4.66 (Structural Nested Mean Models (SNMM)).
Define the impulse response functions ('blip-down' functions) as

$$b_t(\mathbf{d}_{1:t}, j) := \mathbb{E}\left[Y_{it}(\mathbf{d}_{1:t-j}, \mathbf{0}_j) - Y_{it}(\mathbf{d}_{1:t-j-1}, \mathbf{0}_{j+1}) | \mathbf{D}_{1:t-j} = \mathbf{d}_{1:t-j}\right]$$

which is the effect of a change from 0 to d_{t-j} in terms of the treatment on the outcome at time t, conditional on treatment history up to time $t-j$.
These functions are parametrised as a function of lag length

$$b_t(\mathbf{d}_{1:t}, j; \boldsymbol{\gamma}) = \gamma_{1j} d_{t-j} + \gamma_{2j} d_{t-j} d_{t-j-1} \ldots$$

This then allows us to construct **blipped-down / demediated outcomes**

$$\widetilde{Y}_{it}^{j} = Y_{it} - \sum_{s=1}^{j-1} \gamma_s D_{i,t-s}$$

Intuitively, this transformation subtracts off the effects of j lags of treatment, creating an estimate of the counterfactual level of the outcome at time t if the treatment had been set to 0 for j periods before t. Under sequential ignorability, the transformed outcome $\widetilde{Y}_{it}^{j}$ has the same expectation as the counterfactual $Y_{it}(\mathbf{d}_{1:t-j}, \mathbf{0}_j)$, and can be used to construct $\widetilde{Y}_{it}^{j+1}$ by modelling the relationship between $\widetilde{Y}_{it}^{j}$ and $D_{i,t-j}$ to estimate the lagged effect for $j+1$. This is recursive, hence the 'nested'.
Sequential g-estimation can be used to estimate effects. Suppose we're interested in the contemporaneous effect and the first-lagged effect and we adopt an impulse response function $b_t(\mathbf{d}_{1:t,j;\gamma}) = \gamma_j d_{t-j}$ for both these effects. We assume sequential ignorability conditional on $\mathbf{V}_{it} := \{D_{i,t-1}, Y_{i,t-1}, \mathbf{X}_{it}\}$. Sequential g-estimation proceeds as follows

1. For $j = 0$ regress the un-transformed outcome on $\{D_{it}, D_{i,t-1}, Y_{i,t-1}, \mathbf{X}_{it}\}$ as in an ADL model. If this is correctly specified, we estimate the blip-down parameter γ_0 (contemporaneous effect) correctly.
2. We use $\widehat{\gamma}_0$ to construct the one-lag blipped-down outcome $\widehat{Y}_{it}^{1} = Y_{it} - \widehat{\gamma}_0 D_{it}$
3. This blipped-down outcome would be regressed on $\{D_{i,t-1}, D_{i,t-2}, Y_{i,t-2}, \mathbf{X}_{i,t-1}\}$ to estimate the next blip-down parameter γ_1 (the first lagged effect)
4. (**repeat for further lags**, standard error estimated via **block-boostrap**)

Defn 4.67 (Marginal Structural Models).
To specify a marginal structural model, we choose a potential outcome lag length and write a model for the marginal model of those potential outcomes in terms of treatment history

$$\mathbb{E}\left[Y_{it}(\mathbf{d}_{1:t})\right] = g(\mathbf{d}_{1:t};\beta)$$

for example, for a contemporaneous and two lagged effects, we write $\mathbb{E}\left[Y_{it}(\mathbf{d}_{t-2:t})\right] = g(\mathbf{d}_{t-2:t};\beta)$, marginalising over further lags and covariates.
The average causal effect is then

$$\tau^{\text{msm}} := g(\mathbf{d}_{1:t};\beta) - g(\mathbf{d}'_{1:t};\beta)$$

This motivates an IPW approach where weights are constructed as

$$\widehat{\text{SW}}_{it} := \prod_{t=1}^{t} \frac{\widehat{\mathbb{P}}(D_{it}|D_{i,t-1},\widehat{\gamma})}{\widehat{\mathbb{P}}(D_{it}|\mathbf{X}_{it},Y_{i,t-1},D_{i,t-1},\widehat{\alpha})}$$

where the denominator of each term is the product of the predicted probability of observing unit i's observed treatment status conditional on covariates that satisfy conditional ignorability. Multiplying this over time produces the probability of seeing this unit's treatment history conditional on the past.
These weights can be used in a regression of the form

$$g(\mathbf{d}_{t-t:2,\beta}) = \beta_0 + \beta_1 d_t + \beta_2 d_{t-1} + \beta_3 d_{t-2}$$

4.7 Decomposition Methods

Basic idea of decomposition

$$\begin{aligned}\mathbb{F}_M(y) - \mathbb{F}_F(y) &= \int \mathbb{F}_M(y|x) f_M(x)dx - \int \mathbb{F}_F(y|x) f_F(x)dx \\ &= \int [\mathbb{F}_M(y|x) - \mathbb{F}_F(y|x)] f_M(x)dx + \int \mathbb{F}_M(y|x)[f_M(x) - f_F(x)]dx\end{aligned}$$

4.7.1 Oaxaca-Blinder Decomposition

$$\overline{y}_A - \overline{y}_B = \overline{x}'_A\beta_A - \overline{x}'_B\beta_B = \overline{x}'_B(\beta_A - \beta_B) + (\overline{x}_A - \overline{x}_B)'\beta_A$$

We consider two groups, A and B, and an outcome Y, and a vector of predictors $\boldsymbol{x}$. Main question for decomposition is how much of the mean outcome difference [or another summary statistic / quantile of CDF] is accounted for by group differences in the predictors $\boldsymbol{x}$. The Oaxaca-Blinder decomposition refers to the following decompositions:

$$R = \mathbb{E}\left[Y_A\right] - \mathbb{E}\left[Y_B\right] = \mathbb{E}\left[\boldsymbol{x}_A\right]'\boldsymbol{\beta}_A - \mathbb{E}\left[\boldsymbol{x_B}\right]'\boldsymbol{\beta}_B$$

Aggregate Decomposition

$$= \underbrace{\left(\mathbb{E}\left[\boldsymbol{x}_A\right] - \mathbb{E}\left[\boldsymbol{x}_B\right]\right)'\boldsymbol{\beta}_A}_{\text{Explained}} + \underbrace{\mathbb{E}\left[\boldsymbol{x}_B\right]'\left(\boldsymbol{\beta}_A - \boldsymbol{\beta}_B\right)}_{\text{Unexplained}}$$

Decomposition from B's PoV (Threefold Decomposition)

$$= \underbrace{\left\{\mathbb{E}\left[\boldsymbol{x}_A\right] - \mathbb{E}\left[\boldsymbol{x}_B\right]\right\}'\boldsymbol{\beta}_B}_{\text{endowments}} + \overbrace{\mathbb{E}\left[\boldsymbol{x}_B\right]'\left(\boldsymbol{\beta}_A - \boldsymbol{\beta}_B\right)}^{\text{discrimination}} + \underbrace{\left(\mathbb{E}\left[\boldsymbol{x}_A\right] - \mathbb{E}\left[\boldsymbol{x}_B\right]\right)'\left(\boldsymbol{\beta}_A - \boldsymbol{\beta}_B\right)}_{\text{interaction}}$$

Stipulating a non-discriminatory coefficient $\boldsymbol{\beta}^*$ (Twofold Decomposition)

$$= \underbrace{\left\{\mathbb{E}\left[\boldsymbol{x}_A\right] - \mathbb{E}\left[\boldsymbol{x}_B\right]\right\}'\boldsymbol{\beta}^*}_{\text{Explained}} + \overbrace{\underbrace{\left\{\mathbb{E}\left[\boldsymbol{x}_A\right]'\left(\boldsymbol{\beta}_A - \boldsymbol{\beta}^*\right) + \mathbb{E}\left[\boldsymbol{x}_B\right]'\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_B\right)\right\}}_{\mathbb{E}[\boldsymbol{x}_A]'\ \underbrace{\boldsymbol{\delta}_A}_{:=\boldsymbol{\beta}_A - \boldsymbol{\beta}^*}\ -\mathbb{E}[\boldsymbol{x}_B]'\boldsymbol{\delta}_B}}^{\text{Unexplained}}$$

Detailed Decomposition
To examine the 'contribution' of each variable to the observed gap, estimate

$$y_i = \sum_{j=1}^{k} x_{ji}\beta_j + \sum_{j=1}^{k} d_i x_{ji}\delta_j + \varepsilon_i \; ; d_i := \begin{cases} 1 \text{ if } \; i \in B \\ 0 \text{ otherwise} \end{cases}$$

so, β_j is the coefficient for group A, and $\beta_j + \delta_j$ is the coefficient for group B. A t-test for δ_j is used to establish whether a variable is a source of the observed gap. The contribution of each variable to th explained part is

$$c_k^* = \frac{(\overline{x}_k^A - \overline{x}_k^B)\widehat{\beta}_k^A}{(\overline{\boldsymbol{x}}^A - \overline{\boldsymbol{x}}^B)\widehat{\boldsymbol{\beta}}^A}$$

Defn 4.68 (Oaxaca-Blinder-Kitagawa as a Regression imputation estimator).
Let outcome models be linear $Y_i = X_i\boldsymbol{\beta}_1 + \nu_{1i}$ if $W_i = 1$ and $Y_i = X_i\boldsymbol{\beta}_0 + \nu_{0i}$ if $W_i = 0$ where $\mathbb{E}\left[\nu_{1i}\right] = \mathbb{E}\left[\nu\right]_{0i} = 0$.
The difference in means decomposition is

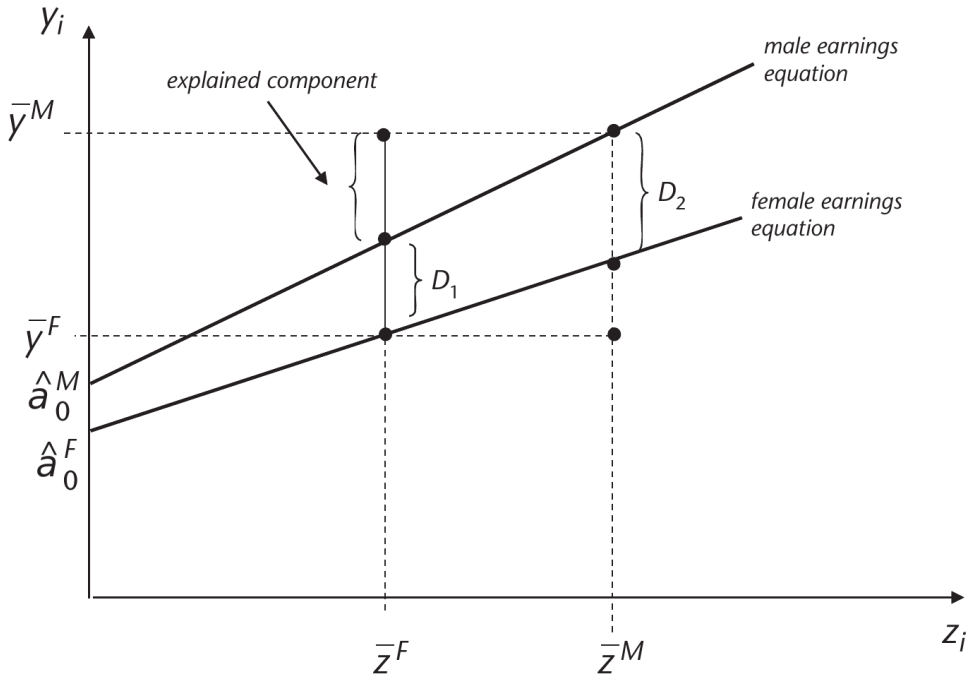


Figure 7: Oaxaca decomposition where D_1 is the 'discrimination' piece (Bazen, 2011). $D_1 \neq D_2$ generically unless two groups have the same slope (which is practically never the case)

$$\begin{aligned}
\mathbb{E}[Y|W=1] - \mathbb{E}[Y|W=0] &= \mathbb{E}[\mathbf{X}|W=0]\boldsymbol{\beta}_1 - \mathbb{E}[\mathbf{X}|W=0]\boldsymbol{\beta}_0 \\
&= \mathbb{E}[\mathbf{X}|W=1](\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\mathbb{E}[\mathbf{X}|W=1] - \mathbb{E}[\mathbf{X}|W=0])\boldsymbol{\beta}_0 \\
&= \mathbb{E}\left[Y^1 - Y^0|W=1\right] + \mathbb{E}\left[Y^0|W=1\right] - \mathbb{E}\left[Y^0|W=0\right] \\
&= \underbrace{\tau_{PATT}}_{\text{Unexplained component}} + \underbrace{\mathbb{E}\left[Y^0|W=1\right] - \mathbb{E}\left[Y^0|W=0\right]}_{\text{Explained Component}}
\end{aligned}$$

Sloczynski: SATT can be estimated by running the following regression:

$$Y_i = \alpha + \tau W_i + \mathbf{X}_i'\boldsymbol{\beta} + \psi W_i(\mathbf{X}_i - \overline{\mathbf{X}}_1) + \varepsilon_i$$

Kline (2011) shows that this is 'doubly robust' and equivalent to a reweighting estimator based on the weights

$$w(\mathbf{x}) := \frac{\mathrm{d}\mathbb{F}_{\mathbf{x}|W=1}(\mathbf{x})}{\mathrm{d}\mathbb{F}_{\mathbf{x}|W=0}(\mathbf{x})} = \frac{1-\rho}{\rho}\frac{e(\mathbf{x})}{1-e(\mathbf{x})}$$

where $\rho := \mathbf{Pr}(W_i = 0)$ is the treated share.

4.7.2 Distributional Regression

Section based on Chernozhukov, Fernández-Val, and Melly (2013). Reference papers:

- https://arxiv.org/abs/0904.0951
- https://ocw.mit.edu/courses/economics/14-382-econometrics-spring-2017/lecture-notes/MIT14_382S17_lec7.pdf
- https://cran.r-project.org/web/packages/Counterfactual/vignettes/vignette.pdf

Let F_{X_k} denote the distribution of job-relevant characteristics (education, experience, etc.) for men when $k = m$ and for women when $k = w$. Let $F_{Y_j|X_j}$ denote the conditional distribution of wages given job-relevant characteristics for group $j \in \{w, m\}$, which describes the stochastic wage schedule that a given group faces. Using these distributions, we can construction $F_{<j|k>}$, the distribution of wages for group k facing group j's wage schedule as

$$F_{\langle j|k\rangle}(y) = \int F_{Y_j|X_j}(y|x)dF_{X_k}(x),\ y \in T$$

For example, $F_{\langle Y_0|X_0\rangle}$ is the distribution of wages for men who face men's wage schedule, and $F_{Y_1|Y_1}$ is the distribution of wages for women who face women's wage schedule, which are both observed distributions. We can also study $F_{\langle 0|1\rangle}$,

the counterfactual distribution of wage for women if they would face the men's wage schedule $F_{Y_0|X_0}$.

$$F_{Y\langle 0|1\rangle}(y) \equiv \int_{\mathcal{X}_1} F_{Y_0|X_0}(y|x)dF_{X_1}(x)$$

is the counterfactual distribution constructed by integrating the conditional distribution of wages for men with respect to the distribution of characteristics for women.
We can Interpret $F_{Y\langle 0|1\rangle}$ as the distribution of wages for women in the absence of gender discrimination, although it is predictive and cannot be interpreted as causal without further (strong) assumptions.

$$F^{\leftarrow}_{Y\langle 1|1\rangle} - F^{\leftarrow}_{Y\langle 0|0\rangle} = \underbrace{\left[F^{\leftarrow}_{Y\langle 1|1\rangle} - F^{\leftarrow}_{Y\langle 0|1\rangle}\right]}_{structure} + \underbrace{\left[F^{\leftarrow}_{Y\langle 0|1\rangle} - F^{\leftarrow}_{Y\langle 0|0\rangle}\right]}_{composition}$$

Assumptions for Causal Interpretation
Under conditional exogeneity / selection on observables, CE can be interpreted as causal effects. Sec 2.3 in ECTA 2013 paper spells this out in detail. Let $(Y_j^* : j \in \mathcal{J})$ be the vector of potential outcomes for various values of a policy $j \in \mathcal{J}$, and X be a vector of covariates. Let J denote the random variable that describes the realised policy and let $Y := Y_J^*$ be denote the realised outcome variable. When J is not randomly assigned, the distribution of $Y|J = j$ may differ from the distribution of Y_j^*. However, under conditional exogeneity, the distribution of $Y|X, J = j$ and $Y_j^*|X$ agree, and the observed conditional distributions have a causal interpretation, and so do counterfactual distributions generated from these conditionals by integrating out X.
Let $F_{Y_j^*|J}(y|k)$ denote the distribution of the potential outcome Y_j^* in the population with $J = k \in \mathcal{J}$. The causal effect of exogenously changing the policy from l to j on the distribution of the potential outcome in the population with the realised policy $J = k$ is $F_{Y_j^*|J}(y|k) - F_{Y_l^*|J}(y|k)$. Under conditional exogeneity, for any $j, k, \in \mathcal{J}$, the counterfactual distribution $F_{Y\langle j|k\rangle}(y)$ exactly corresponds to $F_{Y_j^*|J}(y|k)$, and hence the causal effect of exogenously changing the policy from l to j in the population with $J = k$ corresponds to the CE of changing the conditional distribution from l to j, that is

$$F_{Y_j^*|J}(y|k) - F_{Y_l^*|J}(y|k) = F_{Y\langle j|k\rangle}(y) - F_{Y\langle l|k\rangle}(y)$$

Conditional exogeneity assumption for this section:

$$(Y_j^* : j \in \mathcal{J} \amalg J|X$$

$\mathcal{K}$ groups that partition the sample. For each population $k, \exists X_k \in \mathbb{R}^d$ and outcome Y_k. Covariate vector is observable in all populations, but the otucome is only observatble in populations $j \in \mathcal{J} \subset \mathcal{K}$. Let F_{X_k} denote the covariate distribution in the population $k \in K$, and $F_{Y_j|X_j}$ and $Q_{Y_j|X_j}$ denote the conditional distribution and quantile functions in population $j \in \mathcal{J}$. We denote the support of X_k by $\mathcal{X}_k \subset \mathbb{R}^{d_x}$ and the region of interest Y_j by $\mathcal{Y}_j \subseteq \mathbb{R}$. We refer to j as the reference population and k as the counterfactual population.
The reference and counterfactual populations in the wage example correspond to different groups. We can also generate coutnerfactual populations by artificially transforming a reference population. We can think of X_k as being created through a known transformation of X_j:

$$X_k = g_k(X_j), \text{ where } g_k : \mathcal{X}_j \to \mathcal{X}_k$$

Counterfactual distribution and quantile functions are formed by combining the conditional distribution in population j with the covariate distribution in population k, namely:

$$F_{Y\langle j|k\rangle}(y) \equiv \int_{\mathcal{X}_k} F_{Y_j|X_j}(y|x)dF_{X_k}(x), \; y \in \mathcal{Y}_|$$
$$Q_{Y\langle j|k\rangle}(\tau) \equiv F^{\leftarrow}_{Y\langle j|k\rangle}(\tau), \; \tau \in (0,1)$$

where $(j,k) \in \mathcal{J}\mathcal{K}$ and $F^{\leftarrow}_{Y\langle j|k\rangle}(\tau) = \inf\{y \in \mathcal{Y}_j : F_{Y\langle j|k\rangle}(y) \geq \tau\}$ is the left-inverse function of $F_{Y\langle j|k\rangle}$.
The main interest lies in the quantile effect (QE) function, defined as the difference of the two counterfactual quantile functions over a set of quantile indexes $\mathcal{T} \subset (0,1)$

$$\Delta(\tau) = Q_{Y\langle j|k\rangle}(\tau) - Q_{Y\langle j|j\rangle}(\tau), \; \tau \in \mathcal{T}$$

Estimation of Conditional distribution

$$F_{Y_j|X_j}(y|x) \equiv \int_{(0,1)} 1\{Q_{Y_j|X_J}(u|x) \leq y\}du$$

- `method = "qr"` default implements

$$\hat{F}_{Y_j|X_j}(y|x) = \varepsilon + \int_{(\epsilon,1-\epsilon)} 1\{x'\hat{\beta}_j(u) \leq y\}du$$

where ε is a small constant that avoids estimation of tail quantiles, and $\hat{\beta}(u)$ is the quantile regression estimator

$$\hat{\beta}_j(u) = \text{argmin}_{b\in\mathbb{R}^{d_x}} \sum_{i=1}^{n_j} [u - 1\{Y_{ji} \leq X'_{ji}b\}][Y_{ji} - X'_{ji}b]$$

- `method = "logit"` implements the distribution regression estimator of the conditional distribution with the logistic link function

$$\hat{F}_{Y_j|X_j}(y|x) = \Lambda(x'\hat{\beta}(y))$$

where Λ is the standard logistic CDF and $\hat{\beta}(y)$ is the distribution regression estimator

$$\hat{\beta}(y) \equiv \text{argmax}_{b\in\mathbb{R}^{d_x}} \sum_{i=1}^{n_j} [1\{Y_{ji} \leq y\} \log \Lambda(X_{ij}'b) + 1\{Y_{ij} > y\} \log \Lambda(-X_{ji}'b)]$$

4.8 Causal Directed Acyclic Graphs

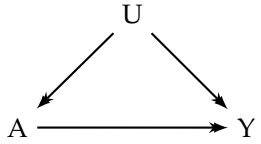
based on http://www.stat.cmu.edu/ cshalizi/350/lectures/31/lecture-31.pdf ,Pearl (2009), Morgan and Winship (2014), Cunningham (2020).
For an undirected graph between $X, Y,$ and Z, there are four possible directed graphs:

- $X \to Y \to Z$ (a chain)
- $X \leftarrow Y \leftarrow Z$ (another chain)
- $X \leftarrow Y \to Z$ (a fork on Y)
- $X \to Y \leftarrow Z$ (collision on Y)

With the fork or either chain, we have $X \perp\!\!\!\perp Z|Y$. However, With a collider, $X \not\perp\!\!\!\perp Z|Y$.
Causal effect of X on Y is written $\mathbf{Pr}(Y \mid \text{do}(X = x))$. Basic idea is condition on *adequate controls* (i.e. not every observed control). Here, controlling for U is unnecessary and would bias the estimate of $\mathbf{Pr}(Y \mid \text{do}(X = x))$.

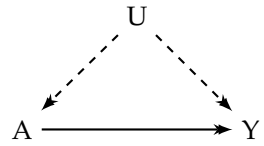
4.8.1 Basics / Terminology

Defn 4.69 (Backdoor Path; Confounder ≈ Omitted Variable).
A backdoor path is a non-causal path from A to Y. They are 'backdoor' because they flow backwards out of A: all of these paths point into A.

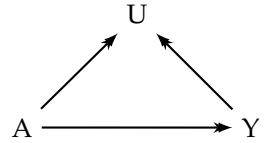


Here, $A \leftarrow U \to Y$, where U is a common cause for treatment and the outcome. So, U is a **confounder**.

A worse problem arises with the following DAG, where dotted lines indecate that U is unobserved. Because U is unobserved, this backdoor path is *open*.



Defn 4.70 (Collider).



Colliders, when left alone, always close a backdoor path. Conditioning on them, however, opens a backdoor path, and yields biased estimates of the causal effect of A on Y.

- Common colliders are post-treatment controls $A \to C \leftarrow Y$
- Another insidious type of collider is of the form $A \leftarrow \cdots \to C \leftarrow \cdots \to Y$, where C is typically a lagged outcome.

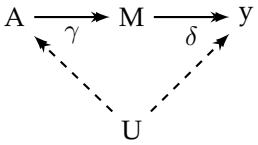
Defn 4.71 (Back Door Criterion).
Vector of *measured* controls $\boldsymbol{S}$ satisfies the **backdoor criterion** if (i) S blocks every path from A to Y that has an arrow **into** A (i.e. **blocks the back door**) and (ii) no node in S is a descendent of A. Then,

$$\mathbf{Pr}(Y|do(A = a)) = \sum_s \mathbf{Pr}(Y|A = a, S = s)\,\mathbf{Pr}(S = s)$$

Which is the same as the subclassification estimator. The conditional Expectation $\mathbb{E}[Y|A = a, S = s]$ can be computed using a nonparametric regression / ML algorithm of choice.

Defn 4.72 (Frontdoor Criterion).
$\boldsymbol{M}$ satisfies the **frontdoor criterion** if (i) M blocks all directed paths from A to Y, (ii) there are no unblocked back-door paths from A to M, and (iii) A blocks all backdoor paths from M to Y.
Then,

$$\mathbf{Pr}(Y|do(A)) = \underbrace{\sum_M \mathbf{Pr}(M = m|A = a)}_{\mathbf{Pr}(M|do(A))} \underbrace{\sum_{a'} \mathbf{Pr}(Y|A = a', M = M)\, Pr(A = a')}_{\mathbf{Pr}(Y|M,\, do(A))}$$



The above DAG in words

1. The only way A influences Y is through M, so there is no arrow bypassing M between X and Y. In other words, M *intercepts all directed paths from* A *to* Y.
2. Relationship between A and M is not confounded by unobservables - i.e. *no back-door paths between* A *and* M.
3. Conditional on A, the relationship between M and Y is not confounded, i.e. *every backdoor path between* M *and* Y *has to be blocked by* A.

With a single mediator M that is not caused by U, the ATE can be estimated by multiplying estimates $\widehat{\gamma} \times \widehat{\delta}$ (Bellemare, Bloem, and Wexler, 2020).
The FDC estimates the ATE because it decomposes a reduced-form relationship that is not causally identified into two causally identified relationships.
Implementation through linear regressions:

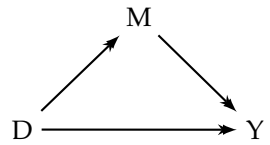
$$M_i = \kappa + \gamma A_i + \omega_i \tag{4}$$
$$Y_i = \alpha + \delta M_i + \psi A_i + \nu_i \tag{5}$$

Since $\mathbb{E}[M|A] = \gamma$ is identified, $\text{Cov}[\omega_i A_i] = 0$ in 4. $\text{Cov}[M, \nu] = 0$ in 5. Assume $\psi = 0$. Then, write

$$\tau_{\text{FDC}} = \mathbb{E}[Y|do(A)] = \widehat{\delta} \times \widehat{\gamma}$$

4.8.2 Mediation Analysis

(Imai, Keele, and Yamamoto, 2010) Pearl (2001), Robins(2003)
Consider SRS where we observe $(D_i, M_i, \mathbf{X}_i, Y_i)$, where D_i is a treatment indicator, M_i is a mediator, $\boldsymbol{X}_i$ is a vector of pre-treatment controls, and Y_i is the outcome. The supports are $\mathcal{M}, \mathcal{X}, \mathcal{Y}$ respectively. $\mathbf{X}$s are partialled out.
Let $M_i(d)$ denote potential value for the mediator under treatment status $D_i = d$.
The outcome $Y_i(d, m)$ is the potential outcome for unit i when $D_i = d, M_i = m$.
The observed variables can be written as $M_i = M_i(D_i), Y_i = Y_i(D_i, M_i(D_i))$.



d, a used interchangeably for treatment.

Assumption 7 (Sequential Unconfoundedness of Treatment, Mediator).

$$\{Y_i(d, m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x \qquad \text{Random assignment of D} \tag{6}$$
$$Y_i(d, m) \perp\!\!\!\perp M_i(d) | D_i, X_i = x \qquad \text{No outcome mediation} \tag{7}$$

$\forall d', d \in \{0, 1\}$ and $(m, x) \in \mathcal{M} \times \mathcal{X}$
This requires the treatment to be conditionally independent of the potential mediator states and outcomes given X, ruling out unobserved confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on the covariates. (5) postulates independence between the counterfactual outcome and mediator values 'across-worlds'.
Effectively, Need M to be randomly assigned (approx).

Defn 4.73 (Natural Indirect Effect).

$$\text{NIE}_i(d) \equiv \delta_i(d) := Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

Difference in Y *holding treatment status constant, and varying the mediator*. Sample Average: **Average Causal Mediation Effect (ACME)**

$$\overline{\delta}(d) := \mathbb{E}[\delta_i(d)] = \mathbb{E}[Y_i(d, M_i(1)) - Y_i(d, M_i(0))]$$

Defn 4.74 (Natural Direct Effect).

$$\text{NDE}_i(d) \equiv \theta_i(d) := Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$

Difference in Y *holding mediator constant, and varying the treatment*.

Defn 4.75 (Total Causal Effect / Treatment Effect Decomposition).

$$
\begin{aligned}
\tau_i &= Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \\
&= \underbrace{Y_i(1, M_i(1)) - Y(0, M_i(1))}_{\theta_i(1)} - \underbrace{Y_i(0, M_i(1)) - Y(0, M_i(0))}_{\delta_i(0)} \\
&= \underbrace{Y_i(1, M_i(0)) - Y(0, M_i(0))}_{\theta_i(0)} - \underbrace{Y_i(1, M_i(1)) - Y(1, M_i(0))}_{\delta_i(1)} \\
&= \underbrace{\delta_i(d)}_{\text{indirect effect}} + \underbrace{\theta_i(1-d)}_{\text{direct effect}} \\
&= \textbf{NDE + NIE defined on opposite treatment states}
\end{aligned}
$$

Defn 4.76 (Controlled Direct Effect (Acharya, Blackwell, and Sen, 2016)).
NDE conditions on potential mediator effects.For CDE, we **set mediator at a prescribed value** m.

$$\text{CDE}_i(d, d', m) = Y_i(d, m) - Y_i(d', m) \; m \in \mathcal{M}$$

Difference between NDE and CDE is *what value mediator is fixed at*. Restated:

$$\psi_i(d, d', m) = Y_i(d, m) - Y_i(d', m)$$

Effect of *changing the treatment while fixing the value of the mediator at some level* m.

$$\overline{\psi}(d, d', m) = \mathbb{E}\left[Y_i(d, m) - Y_i(d', m)\right]$$

Theorem 4.28 (ATE decomposition (VanderWeele and Tchetgen-Tchetgen(2014))).
Decomposing total effect with binary mediator

$$
\begin{aligned}
\tau(d, d') = &\underbrace{\text{ACDE}(d, d', 0)}_{\text{Direct Effect}} + \underbrace{\text{ANIE}(d, d')}_{\text{Indirect Effect}} \\
&+ \underbrace{\mathbb{E}\left[M(a')[\text{CDE}(d, d', 1) - \text{CDE}(d, d', 0)]\right]}_{\text{Interaction}}
\end{aligned}
$$

Fact 4.29 (Parametric Setup for ACME estimation).
Assume linear models for mediator $M = \psi T + U_m$ and $Y = \beta T + \gamma M + U_Y$.
Then fit the following regressions

$$Y_i = \alpha_1 + \underbrace{\tau}_{\text{Total effect}} D_i + \varepsilon_{i1} \tag{8}$$

$$M_i = \alpha_2 + \psi D_i + \varepsilon_{i2} \tag{9}$$

$$Y_i = \alpha_3 + \underbrace{\beta}_{\text{Direct Effect}} D_i + \gamma M_i + \varepsilon_{i3} \tag{10}$$

Baron and Kenny (1986) suggest testing $\tau = \psi = \beta = 0$. If all nulls rejected, Mediation effect $\overline{\delta} = \psi\gamma$. Equivalently, mediation effect is $\tau - \beta = \psi \times \gamma$. Estimate variance using bootstrap / delta method.

Fact 4.30 (Semiparametric Estimation).
Assume selection on observables w.r.t. D, M.
Huber(2014)
Average direct effect identified by

$$\theta(d) = \mathbb{E}\left[\left(\frac{Y \cdot D}{\mathbf{Pr}\left(D = 1|M, X\right)} - \frac{Y \cdot (1 - D)}{1 - \mathbf{Pr}\left(D = 1|M, X\right)}\right) \cdot \frac{\mathbf{Pr}\left((D = d|M, X)\right)}{\mathbf{Pr}\left(D = d|X\right)}\right]$$

Average Indirect Effect identified by

$$\delta(d) = \mathbb{E}\left[\frac{Y \cdot \mathbb{1}_{D=d}}{\mathbf{Pr}\left(D = d|M, X\right)}\left(\frac{\mathbf{Pr}\left(D = 1|M, X\right)}{\mathbf{Pr}\left(D = 1\right)} - \frac{1 - \mathbf{Pr}\left(D = 1|M, X\right)}{1 - \mathbf{Pr}\left(D = 1|X\right)}\right)\right]$$

implemented in `causalweight::medweight`.
https://cran.r-project.org/web/packages/causalweight/vignettes/bodory-huber.pdf

5 Semiparametrics and Nonparametrics

based on Tsiatis (2007), Wasserman (2006), and Kennedy (2015)

5.1 Semiparametric Theory

Observations $Z_1, \dots, Z_n$ that take values in a measurable space $(\mathcal{Z}, \mathcal{B})$ with distribution P_0 . A statistical *model* $\mathcal{P}$ is a collection of probability measures on the sample space, which is assumed to contain the data distribution P_0.
The general goal is estimation and inference for some target parameter $\psi_0 = \psi(P_0) \in \mathbb{R}^p$ where $\psi = \psi(P)$ can be viewed

- A *nonparametric* model $\mathcal{P}$ is a collection of all probability distributions
- A *parametric* model is a model that can be smoothly indexed by a Euclidian vector $\theta \in \mathbb{R}^q$ with $\psi \subseteq \theta$.
- A *semiparametric* model is one that contains both parametric and nonparametric parts.

5.1.1 Empirical Processes Background

A *stochastic process* is a collection of random variables $\{X(t), t \in \mathcal{T}\}$ on the same probability space indexed by an arbitrary index set $\mathcal{T}$. An *empirical process* is a stochastic process based on a random sample.

Defn 5.1 (empirical distribution function).

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le t}$$

$\mathbb{F}_n(t)$ is unbiased and has variance

$$\mathbb{V}\left[\widehat{\mathbb{F}}_n(t)\right] = \frac{1}{n^2}\mathbb{V}\left[n\widehat{\mathbb{F}}_n(t)\right] = \frac{\mathbb{F}(x)(1-\mathbb{F}(x))}{n}$$

This can be generalised to an **empirical measure** over a random sample $X_1, \dots, X_n$ of independent draws from a probability measure P on an arbitrary sample space $\mathcal{X}$. The empirical measure is defined as

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

where δ_x is a dirac delta that assigns mass 1 at x and 0 otherwise.
For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we denote $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.
Setting $\mathcal{X} = \mathbb{R}, \mathbb{F}_n$ can be re-expressed as the empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$ where $\mathcal{F} := \{\mathbb{1}_{x \le t}, t \in \mathbb{R}\}$.

Defn 5.2 (Empirical Process).
Given an empirical distribution $\mathbb{F}_n$, the corresponding empirical process is the rescaled gap

$$\mathbb{Z}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - \mathbb{F}(x))$$

Theorem 5.1 (Glivenko Cantelli Theorem).
the Kolmogorov-Smirnov statistic D_n

$$D_n := \|\mathbb{F}_n - \mathbb{F}\|_\infty \equiv \sup_{t \in R} |\mathbb{F}_n(t) - \mathbb{F}(t)| \overset{a.s.}{\to} 0$$

and

$$\|\mathbb{F}_n - \mathbb{F}\|_\infty = O_p(\log n/n) = O_p(\sqrt{1/n})$$

A class of $\mathcal{F}$ measureable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is said to be a $P-$Glivenko-Cantelli class if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{a.s.}{\to} 0 \; ; \; Pf := \int_{\mathcal{X}} f(x) P(\mathrm{d}x)$$

Theorem 5.2 (Dvoretzky-Kiefer Wolfowitz (DKW) inequality).

$$\forall \varepsilon > 0, \mathbb{P}\left[\sup_x \left|\mathbb{F}(x) - \widehat{\mathbb{F}}_n(x)\right| > \varepsilon\right] \le 2\exp(-2n\varepsilon^2)$$

This allows us to construct a confidence set. For example, let $\varepsilon_n^2 = \frac{\log(2/\alpha)}{2n}$. The nonparametric DKW confidence band is

$$(\widehat{\mathbb{F}}_n - \varepsilon_n, \widehat{\mathbb{F}}_n + \varepsilon_n)$$

Defn 5.3 (Statistical Functional and Plug-in Estimator).
A statistical functional $T(\mathbb{F})$ is any function of $\mathbb{F}$. Examples include the mean, $\mu := \int x \mathrm{d}\mathbb{F}(x)$, variance $\sigma^2 := (x-\mu)\mathrm{d}\mathbb{F}(x)$, and the median $m = \mathbb{F}^{-1}(1/2)$
A **plug-in** estimator of $\theta =: T(\mathbb{F})$ is defined by $\widehat{\theta}_n := T(\widehat{\mathbb{F}}_n)$.

Plugin estimator for a linear functional A functional of the form $\int a(x)\mathrm{d}\mathbb{F}(x)$ is called a **linear functional**. A plug-in estimator for it is

$$T(\widehat{\mathbb{F}}_n) = \int a(x)\mathrm{d}\widehat{\mathbb{F}}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$$

Defn 5.4 (Gateaux derivative).
is defined as

$$L_{\mathbb{F}}(x) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)\mathbb{F} + \varepsilon\delta_x) - \theta(\mathbb{F})}{\varepsilon}$$

The empirical influence function uses the empirical distribution function

$$\varphi(x) = \lim_{\varepsilon \to 0} \frac{\widehat{\theta}(\overbrace{(1-\varepsilon)\widehat{\mathbb{F}}_n + \varepsilon\delta_x}^{=:\widehat{\mathbb{F}}_\epsilon(x)}) - \widehat{\theta}(\widehat{\mathbb{F}})}{\varepsilon}$$

The influence function $L_{\mathbb{F}}(x)$ behaves like the *score function* in parametric estimation because of the MLE analogues

- $L_{\mathbb{F}}(x)\mathsf{d}\mathbb{F}(x) = 0$
- $\mathbb{V}\left[T(\widehat{\mathbb{F}}_n)\right] \approx \int L^2_{\mathbb{F}}(x)\mathsf{d}\mathbb{F}(x)/n.$

Defn 5.5 (Hadamard Differentiability).
Gateaux differentiability is too week to ensure that functionals converge $T(\widehat{\mathbb{F}}) \to T(\mathbb{F})$.
A function T is Hadamard differentiable if, for any sequence $\varepsilon \to 0$, and D_n satisfying $\sup_x |D_n(x) - D(x)| \to 0$, we have

$$\frac{T(\mathbb{F} + \varepsilon_n D_n) - T(\mathbb{F})}{\varepsilon_n} \to L_{\mathbb{F}}(T; D)$$

If T is Hadamard differentiable, $T(\widehat{\mathbb{F}}) \xrightarrow{p} T(\mathbb{F})$

Functional Delta Method If $T(\mathbb{F})$ is a linear functional,

$$\int L_{\mathbb{F}}(x)\mathsf{d}\mathbb{G}(x) = T(\mathbb{G}) - T(\mathbb{F})$$

which is similar to the fundamental theorem of calculus, but for functional calculus.

$$\sqrt{n}\left(T(\widehat{\mathbb{F}}) - T(\mathbb{F})\right) \xrightarrow{d} \mathcal{N}\left(0, \underbrace{\int L^2(x)\mathsf{d}\mathbb{F}(x)}_{=:\gamma^2}\right)$$

This allows us to construct standard errors as $\widehat{\tau}/\sqrt{n}$

5.1.2 Influence Functions

We are concerned with the statistical model where $Z_1, \ldots, Z_n$ are random vectors and the density of Z is assumed to belong to the class $\{p_Z(z; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega\}$. The parameter $\boldsymbol{\theta}$ can be decomposed into $(\boldsymbol{\beta}^\top, \boldsymbol{\eta}^\top)^\top$, where $\boldsymbol{\beta}^{q\times 1}$ is the parameter of interest and $\boldsymbol{\eta}$ is the *nuisance parameter* (which may be finite or infinite dimensional). For simplicity, assume $\boldsymbol{\eta}^{r\times 1}$, so the dimension of $\dim(\boldsymbol{\theta}) = p = q + r$.

Defn 5.6 (influence function).
Reasonable estimators $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ are *asymptotically linear*, such that there exists a random vector $\boldsymbol{\varphi}^{q\times 1}(Z)$ such that $\mathbb{E}\left[\boldsymbol{\varphi}(Z)\right] = \mathbf{0}^{q\times 1}$ and $\mathbb{E}\left[\boldsymbol{\varphi}(Z)\boldsymbol{\varphi}(Z)^\top\right]$ is finite and non-singular.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\varphi}(Z_i) + o_p(1)$$

Equivalently,

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = \frac{1}{n}\sum_{i=1}^n \varphi(Z) + o_p(1/\sqrt{n})$$

where φ has mean zero ($\mathbb{E}\left[\varphi(Z)\right] = 0$) and finite variance ($\mathbb{E}\left[\varphi(Z)^{\otimes 2}\right] < \infty$). This is called the **influence function** because $\varphi(Z_i)$ is the *influence* of the $i-$th observation on $\widehat{\beta}_n$.
By CLT, an estimator $\widehat{\boldsymbol{\beta}}$ with influence function φ is asymptotically normal with

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\varphi}(Z_i) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}^{q\times 1}, \mathbb{E}\left[\boldsymbol{\varphi}\boldsymbol{\varphi}^\top\right]\right)$$

By Slutsky's theorem, the corresponding estimator

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}^{q\times 1}, \mathbb{E}\left[\boldsymbol{\varphi}\boldsymbol{\varphi}^\top\right]\right)$$

Theorem 5.3 (Influence function uniqueness).
Any asymptotically linear estimator has a unique influence function (Tsiatis, 2007, chapter 3)

Example 5.4 (Examples of Influence functions).
Consider a setting where $Z_1, \ldots, Z_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$. The maximum likelihood estimators are $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\widehat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n (Z_i - \widehat{\mu}_n)^2$. They are RAL because

$$\sqrt{n}(\widehat{\mu}_n - \mu_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \underbrace{(Z_i - \mu_0)}_{\varphi(Z_i)}$$

Similarly,

$$\sqrt{n}(\widehat{\sigma}_n^2 - \sigma_0^2) = 1/\sqrt{n}\sum_{i=1}^{n} \underbrace{\left\{(Z_i - \mu_0)^2 - \sigma_0^2\right\}}_{\varphi(Z_i)} + \underbrace{\sqrt{n}(\widehat{\mu}_n - \mu_0)}_{o_p(1)}$$

covariance: $\varphi(x) = (x - \mathbb{E}[X])(y - \mathbb{E}[y]) - \text{Cov}[X, Y]$
linear regression:

$$\varphi(x, y) = \frac{x - \mathbb{E}[X]}{\mathbb{V}[X]}\left\{(y - \mathbb{E}[Y]) - \beta(x - \mathbb{E}[x])\right\}$$

M-estimators solve $\mathbb{E}[g(X, \theta)] = 0$ where g is a score function. The influence function is

$$\varphi_\theta(x) = \mathbb{E}[\nabla_\theta g(X, \theta)]^{-1} g(x, \theta)$$

which nests both MLE and GMM estimators.
https://j-kahn.com/files/influencefunctions.pdf

Theorem 5.5 (Geometry of Regular Asymptotically Linear Estimators).
An estimator $\widehat{\beta}_n$ is said to be regular if, for each θ^*, $\sqrt{n}(\widehat{\beta}_n - \beta)$ has a limiting distribution that does not depend on the local DGP. This rules out degenerate estimators such as the super-efficient Hodges estimator.
Regularity allows us to write $Z \sim p_Z(z, \theta)$, $\theta = (\beta^\top, \eta^\top)$. Now define the **score vector** for a single observation

$$S_\theta(z, \theta_0) = \left.\frac{\partial \log p_Z(z, \theta)}{\partial \theta}\right|_{\theta=\theta_0}$$

which is the p-dimensional vector of derivatives of the log-likelihood with respect to the parameters θ. This can further be partitioned

$$S_\theta(Z, \theta_0) = \left\{S_\beta^\top(Z, \theta_0), S_\eta^\top(Z, \theta_0)\right\}^\top \qquad \text{where}$$
$$S_\beta(Z, \theta_0) = \left.\frac{\partial \log p_Z(z, \theta)}{\partial \beta}\right|_{\theta=\theta_0}^{q\times 1}$$
$$S_\eta(Z, \theta_0) = \left.\frac{\partial \log p_Z(z, \theta)}{\partial \eta}\right|_{\theta=\theta_0}^{r\times 1}$$

These can be collected into a matrix of partial derivatives

$$\mathbf{\Gamma}(\theta)_{q\times p} = \frac{\partial \beta(\theta)}{\partial \theta}^\top$$

Let $\widehat{\beta}_n$ be a **Regular, Asymptotically Linear (RAL)** estimator with influence function $\varphi(Z)$. Then, the following hold

$$\mathbb{E}\left[\varphi(Z) S_\theta^\top(Z, \theta_0)\right] = \mathbf{\Gamma}(\theta)$$

$$\mathbb{E}\left[\varphi(Z) S_\beta^\top(Z, \theta_0)\right] = \mathbf{I}_{q\times q}$$

$$\mathbb{E}\left[\varphi(Z) S_\eta^\top(Z, \theta_0)\right] = \mathbf{0}_{q\times r}$$

Fact 5.6 (Converting an influence function into an estimator).
Let φ be a function satisfying the above RAL conditions, and for each β we have an estimator $\widehat{\eta}_n(\beta)$ such that $\sqrt{n}\left\|\widehat{\eta}_n(\beta) - \eta_0\right\|_{\max}$ is bounded in probability. Define

$$m(Z; \beta, \eta) = \varphi(Z) - \mathbb{E}_{Z\sim p(\cdot;\beta,\eta)}[\varphi(Z)]$$

and let $\widehat{\beta}$ be the solution of

$$\sum_{i=1}^{n} m(Z_i; \beta, \widehat{\eta}_n(\beta)) = 0$$

then $\widehat{\beta}_n$ will be an RAL estimator with influence function $\varphi(Z)$.

Fact 5.7 (Robust estimators have bounded influence functions.).

Fact 5.8 (Influence Functions and Variance).
If $\mathbb{E}[\varphi_\theta(x)] = 0$, we can write

$$\text{Cov}[\varphi_{\theta_1}(x), \varphi_{\theta_2}(x)] = \mathbb{E}[\varphi_{\theta_1}(x)\varphi_{\theta_2}(x)]$$

Say we have an estimate $\widehat{\theta}$ from a random sample. We can look at this sample as a series of $\varepsilon-$ contaminations to the true distribution, each of which puts $\frac{1}{n}$ weight on the derivative. Then, for large enough N, we can represent the difference between $\widehat{\theta}$ and θ as a Taylor expansion

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \varphi_\theta(x_i) + \text{higher order terms}$$

the higher order terms converge in probability to zero, which implies

$$\sqrt{n}(\widehat{\theta}-\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\theta(x_i)+o_p(1)$$

Practically, one can compute variances of complicated structural problems by computing the empirical equivalents of influence functions and stacking. Given estimators $\theta_1,\dots,\theta_M$ and observations $i=1,\dots,N$, create matrix with rows corresponding to observations and columns corresponding to estimators

$$\mathbf{\Phi}=[\varphi_{\theta_1},\dots,\varphi_{\theta_N}]_{N\times M}$$

Since the distribution of each estimator is the same as $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{\theta_j}(x_i)$, the variance can be computed as

$$\mathbf{V}=\frac{1}{N}(\mathbf{\Phi}^\top\mathbf{\Phi})$$

Fact 5.9 ($M-$ and $Z-$ estimation).
Distinction from Kosorok (2008)

- approximate **Maximisers** of data-dependent processes are known as **M-estimators**.
- approximate **Zeroes** of data-dependent processes are known as **Z-estimators**. e.g. $U_n(\beta)=\mathbb{P}_n[X(Y-X'\beta)]$. $Z\subset M$

Defn 5.7 (M-estimator ('Maximum-likelihood-like' / Extremum)).
$\hat{\theta}$ is an estimator that maximises a scalar objective function that is a sum of N subfunctions

$$\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}Q_N(\boldsymbol{\theta})=\frac{1}{N}\sum_{i=1}^{N}q(\boldsymbol{w}_i,\boldsymbol{\theta})$$

MLE is a type of extremum estimator. So is GMM.

5.1.3 Tangent Spaces

Assume target parameter ψ is scalar. Influence functions reside in Hilbert space $L_2(P)$ of measureable functions $g:\mathcal{Z}\to\mathbb{R}$ with $Pg^2=\int g^2\mathrm{d}P=\mathbb{E}\left[g(Z)^2\right]<\infty$ equipped with covariance inner product $\langle g_1,g_2\rangle=P\langle g_1g_2\rangle$.

Defn 5.8 (Tangent Space for parametric models).
the tangent space $\mathcal{T}$ for parametric models indexed by the real-valued parameter $\theta\in\mathbb{R}^{q+1}$ is the linear subspace of $L_2(P)$ spanned by the score vector

$$\mathcal{T}=\left\{bS_\theta^\top(Z;\theta_0):\ b\in\mathbb{R}^{q+1}\right\}$$

where $S_\theta(Z,\theta_0)=\frac{\partial\log p(z;\theta)}{\partial\theta}|_{\theta=\theta_0}$ is the score-function. If we can decompose $\theta=(\psi,\eta)$, we can decompose the tangent space as well and write $\mathcal{T}=\mathcal{T}_\psi\oplus\mathcal{T}_\eta$. In this formulation, $\mathcal{T}_\eta$ is known as the *nuisance tangent space*. Influence functions for ψ reside in the **orthogonal complement** of the nuisance tangent space denoted by

$$\begin{aligned}\mathcal{T}_\eta^\perp&:=\{g\in L_2(P):P(gh)=0\ \ \forall\ h\in\mathcal{T}_\eta\}\\&=\{g\in L_2(P):h-\Pi(h\mid\mathcal{T}_\eta),\ h\in L_2(P)\}\end{aligned}$$

where $\Pi(g|\mathcal{S})$ denotes projections of g on the space $\mathcal{S}$.

- The subspace of influence functions is the set of elements $\phi\in\mathcal{T}_\eta^\perp$ that satisfy $P(\phi S_\psi)=1$.
- The *efficient influence function* is that with the smallest covariance $P(\phi^2)$ and is given by $\phi_{\text{eff}}=P(S_{\text{eff}}^2)^{-1}S_{\text{eff}}$
  - S_{eff} is the efficient score $S_{\text{eff}}=S_\psi-\Pi(S_\psi|\mathcal{T}_\eta)=\Pi(S_\psi|\mathcal{T}_\eta^\perp)$.
- All RAL estimators have influence functions ϕ that reside in $\mathcal{T}_\eta^\perp$ with $P(\phi\mathcal{S}_\psi)=1$

Defn 5.9 (Tangent spaces for semiparametric models).
A parametric submodel $\mathcal{P}_\varepsilon$ is indexed by a real valued parameter ε ($\mathcal{P}_\varepsilon=\{P_\varepsilon:\varepsilon\in\mathbb{R}\}$) is a set of distributions contained in the larger model $\mathcal{P}$, which also contains the truth $P_0\in\mathcal{P}_\varepsilon$. A typical example of a parametric submodel is

$$p_\varepsilon(z)=p_0(z)\left\{1+\varepsilon g(z)\right\}$$

where $\mathbb{E}\left[g(Z)\right]=0$ and $\sup_z|g(z)|<M,|\varepsilon|<1/M$, so $p\varepsilon(z)\geq0$. The parametric submodel is sometimes indexed by g such that $P_\varepsilon=P_{\varepsilon,g}$
The tangent space $\mathcal{T}$ for semiparametric models is defined as the closure of the linear span of scores of the parametric submodels. IoW, we define scores on parametric submodels P_ε as $S_\varepsilon(z)=\frac{\partial\log p_\varepsilon(z)}{\partial\varepsilon}|_{\varepsilon=0}$, and construct parametric submodel tangent spaces $\mathcal{T}_\varepsilon=\left\{b^\top S_\varepsilon(Z):b\in\mathbb{R}\right\}$.
Similarly, the nuisance tangent space $\mathcal{T}_\eta$ is the set of scores in $\mathcal{T}$ that do not vary the target parameter

$$\mathcal{T}_\eta=\left\{g\in\mathcal{T}:\frac{\partial\psi(P_{\varepsilon,g})}{\partial\varepsilon}|_{\varepsilon=0}=0\right\}$$

In nonparametric models, the tangent space is the whole Hilbert space of mean-zero functions.

As before, the efficient influence function is the influence function with the smallest covariance and is defined as the projection $\phi_{\text{eff}} = \Pi(\phi|\mathcal{T})$. It can also be defined as the pathwise derivative of the target parameter in the sense that $P(\phi S_\varepsilon) = \frac{\partial \psi(P_\varepsilon)}{\partial \varepsilon}|_{\varepsilon=0}$.

Nonparametric Delta Method

$$\frac{T(\widehat{\mathbb{F}}_n) - T(\mathbb{F})}{\underbrace{\frac{1}{n}\sum_{i=1}^{n} L^2(X_i)/\sqrt{n}}_{\widehat{\tau}}} \approx \mathcal{N}(0,1)$$

Fact 5.10 (Confidence Bands).
Let $\mathfrak{F}$ be a class of distribution functions $\mathbb{F}$, let θ be the quantity of interest, and C_n be a set of possible values of θ which depends on the data $X_1, \ldots, X_n$.

- C_n is a **finite-sample** $1-\alpha$ **confidence set** if

$$\inf_{F \in \mathfrak{F}} \mathbb{P}_F(\theta \in C_n) \geq 1 - \alpha \ \ \forall \ \ n$$

- C_n is a **uniform asymptotic** $1-\alpha$ confidence set if

$$\liminf_{n\to\infty} \inf_{F \in \mathfrak{F}} \mathbb{P}_F(\theta \in C_n) \geq 1 - \alpha$$

- C_n is a **pointwise asymptotic** $1-\alpha$ confidence set is

$$\forall F \in \mathfrak{F} \liminf_{n\to\infty} \mathbb{P}_F(\theta \in C_n) \geq 1 - \alpha$$

Finite sample confidence set $\succ$ Uniform asymptotic confidence sets *succ* pointwise asymptotic confidence set.
Informally, a true confidence band comprises of two functions that bracket the c.d.f. at all points with probability $1-\alpha$. This should be contrasted to the pointwise bands in common use, which bracket the c.d.f. with probability $1-\alpha$ *at any given point*.

Defn 5.10 (Gaussian Process).
A gaussian process $\mathcal{G}$ indexed by a set $\mathcal{A}$ is a collection $\{\mathcal{G}(x)\}_{x\in\mathcal{A}}$ of Gaussian random variables such that $\forall x_1, \ldots, x_k \in \mathcal{A}, (G(x_1), \ldots, G(x_k))' \sim$ MVN. The function $m(x) := \mathbb{E}[G(x)]$ is a *mean function* and $C(x,y) := \text{Cov}[G(x), G(y)]$ is the covariance function. A *centered gaussian process* is one whose mean function is identically zero.

$$\forall x, y \in \mathcal{A}, \ \mathbb{E}\left[|G(x) - G(y)|^2\right] = C(x,x) + C(y,y) - 2C(x,y)$$

Example 5.11 (Brownian Motion / Wiener Process).
is a centered gaussian process indexed by $[0,\infty)$ whose covariance function is given by

$$C(s,t) := \min(s,t) \ \ , s,t \geq 0$$

Theorem 5.12 (Donsker Theorem).
$\mathbb{Z}_n \to \mathbb{Z} \equiv \mathbb{U}(F)$; $D(\mathbb{R}, \|\cdot\|_\infty)$ where $\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$, which means it is a zero-mean Gaussian process with covariance function $\mathbb{E}[\mathbb{U}(s)\mathbb{U}(t)] = s \wedge t - st \ \ , s,t \in [0,1]$
which means for any bounded continuous function $g : D(\mathbb{R}, \|\cdot\|_\infty) \to \mathbb{R}$, we have

$$\mathbb{E}[g(\mathbb{Z}_n)] \xrightarrow{d} \mathbb{E}[g(Z)]$$

and

$$g(\mathbb{Z}_n) \xrightarrow{d} g(Z)$$

5.2 Semiparametric Theory for Causal Inference

notes from Kennedy (2015) (epi-ish notation). Treatement denoted by A ('action') and controls denoted by L.
Target parameter (ψ) : which may be an ATE $\mathbb{E}\left[Y^1 - Y^0\right]$ or a risk ratio $\mathbb{E}\left[Y^1\right]/\mathbb{E}\left[Y^0\right]$ and so on.
Assumptions

1. **Consistency/SUTVA**: $A = a \implies Y = Y^a$.
2. **Unconfoundedness**: $Y^a \perp\!\!\!\perp A|L$
3. **Overlap**: $\mathbf{Pr}(A = a \mid L = l) \geq \delta > 0 \ \ \forall\, p(L = l) > 0$.

Then the ATE can be written

$$\psi = \int_{\mathcal{L}} (\mathbb{E}[Y \mid L = l, A = 1] - \mathbb{E}[Y \mid L = l, A = 0])\, d\mathbb{F}_L$$

This is the outcome regression (econometrics), subclassification estimator (statistics), g-formula(epidemiology), backdoor criterion estimator (DAGology).
We suppose the data is $Z := \langle L, A, Y \rangle$ and its distribution P_0 admits to the following factorisation:

$$p(z) = p(y \mid l, a)\; p(a \mid l)\; p(l)$$

In semiparametric causal settings, one typically imposes parametric assumptions on the treatment mechanism leaving the outcome mechanism unspecified. For example, for the ATE, one might write

$$p(z; \eta, \alpha) = \underbrace{p(y \mid l, a; \eta_y)}_{\text{Nonparametric}} \overbrace{p(a \mid l; \alpha)}^{\text{Parametric}}\; p(l; \eta_l)$$

where $\alpha \in \mathbb{R}^q$ but $\eta = \langle \eta_y, \eta_l \rangle$ represents an infinite-dimensional parameter that does not restrict the conditional distribution of the outcome given covariates and treatment $p(y \mid l, a)$ and the marginal covariate distribution $p(l)$.

Example 5.13 (Influence functions for causal estimands).
For IPW estimator

$$\widehat{\psi}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{AY}{\pi(L)} - \frac{(1-A)L}{1-\pi(L)}$$

The influence function is clearly

$$\phi_{IPW}(Z) = \frac{AY}{\pi(L)} - \frac{(1-A)Y}{1-\pi(L)} - \psi_0$$

since $\widehat{\psi}_{IPW} - \psi_0 = \frac{1}{n} \sum_{i=1}^{n} \phi_{ipw}(Z)$ exactly.
In an observational study, $\pi(l)$ needs to be estimated, and suppose we do so with a corectly specified parametric model $\pi(l; \alpha)$, $\alpha \in \mathbb{R}^q$ so that $\widehat{\alpha}$ solves some moment condition $\mathbb{P}_n\left(S(Z; \widehat{\alpha})\right) = 0$. Then, we have $\widehat{\theta} = (\widehat{\psi}^*_{IPW}, \widehat{\alpha}^\top)^\top$ which solves $\mathbb{P}_n\left(m(Z; \widehat{\theta})\right)$ where

$$m(z; \theta) = \begin{pmatrix} \psi_{ipw}(Z; \psi, \alpha) \\ S(Z; \alpha) \end{pmatrix}$$

Under standard regularity conditions, we have

$$\widehat{\theta} - \theta_0 = \mathbb{P}_n \left(\mathbb{E}\left[\frac{\partial m(Z; \theta_0)}{\partial \theta} \right]^{-1} m(Z; \theta_0) \right) + o_p(1/\sqrt{n})$$

Example 5.14 (Efficient Influence function for ATE).
For $\psi = \mathbb{E}\left[Y^1 - Y^0\right] = \mathbb{E}\left[\mu(L, 1) - \mu(L, 0)\right]$. Under a nonparametric model where the distribution of P is left unrestricted, the efficient influence function for ψ is given by

$$\phi(Z; \psi, \eta) = m_1(Z; \eta) - m_0(Z; \eta) - \psi$$

where

$$m_a(Z, \eta) = m_a(Z; \pi, \mu) = \frac{\mathbb{1}_{A=a}(Y - \mu(L, a))}{a\pi(L) + (1-a)(1-\pi(L))} + \mu(L, a)$$

Suppose the estimator $\widehat{\eta}$ converges to some $\overline{\eta} = (\overline{\pi}, \overline{\mu})$. Then, $\mathbb{P}(m(Z, \overline{\eta})) = \mathbb{P}(m(Z; \eta_0)) = \psi_0$
Given $\mathbb{P}(f(Z)) = \int f(z) d\mathbb{P}$ to denote expectations of $f(Z)$ for a new observation Z and the decomposition

$$\widehat{\psi} - \psi_0 = (\mathbb{P}_n - \mathbb{P}) m(Z; \widehat{\eta}) - \mathbb{P}(m(Z; \widehat{\eta}) - m(Z; \eta_0)) \tag{11}$$

the first term can be shown to admit to the following result

$$(\mathbb{P}_n - \mathbb{P}) m(Z; \widehat{\eta}) = (\mathbb{P}_n - \mathbb{P}) m(Z; \eta_0) + o_p(1/\sqrt{n})$$

so that $(\mathbb{P}_n - \mathbb{P}) m(Z; \widehat{\eta})$ is asymptotically equivalent to its limiting version $(\mathbb{P}_n - \mathbb{P}) m(Z; \eta_0)$. This requires that $\mathcal{M} = \{m(; \eta) : \eta \in \mathcal{H}\}$ is a Donsker class, where $\mathcal{H}$ is a function class containing the nuisance estimator $\widehat{\eta}$, or that $\mathcal{H}$ itself is Donsker.
We can then expand 11 to

$$\widehat{\psi} - \psi_0 = (\mathbb{P}_n - \mathbb{P})\, m(Z; \overline{\eta}) + \mathbb{P}\{m(Z; \hat{\eta}) - m(Z; \overline{\eta})\} + o_p(1/\sqrt{n})$$

By the fact that $\widehat{\pi}$ is bounded away from 0, 1 and Cauchy Schwartz, the middle term $|\mathbb{P}(m(Z; \widehat{\eta}) - m(Z; \overline{\eta}))|$ is bounded from above by

$$\sum_{a \in \{0,1\}} \|\pi_0(L) - \widehat{\pi}(L)\|\, \|\mu_0(L, a) - \widehat{\mu}(L, a)\|$$

So, if $\widehat{\pi}$ is based on a correctly specified model, we only need $\widehat{\mu}$ to be consistent to make the product term $\mathbb{P}(m(Z; \widehat{\eta}) - m(Z; \overline{\eta})) = o_p(1/\sqrt{n})$ asymptotically negligible.

5.3 Nonparametric Density Estimation

Defn 5.11 (Histogram Estimator).
Given a vector of mutually exclusive bins $B_1, \ldots, B_j$ that partition $\operatorname{supp} X$, and $\nu_1, \ldots, \nu_j$ be the corresponding counts in each bin, the histogram estimator is defined as

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{\nu_j/n}{h} \mathbb{1}_{x \in B_j}$$

Defn 5.12 (Smoothing Kernel).
A **Smoothing kernel** $K : \mathbb{R} \to \mathbb{R}$ that satisfies the following properties

1. $K(x)$ is symmetric around zero and continuous
2. Integral properties
   (a) $\int K(x)dx = 1$: integrates to one
   (b) $\int xK(x)dx = 0$
   (c) $\int |K(x)|\, dx < \infty$
3. Decay. Either:
   (a) $K(x) = 0$ if $|x| \geq x_0$ for some cutoff x_0 OR
   (b) $|x|\, K(x) \to 0$ as $|x| \to \infty$
4. $\int x^2 K(x)dx = \kappa$ where κ is a constant

3(a) is usually preferred over 3(b), which allows us to truncate the domain of the function to [-1, 1] for convenience.
Higher-order Kernels are kernels whose first nonzero moment is the pth moment. These kernels can increase rates of convergence if $f(x)$ is more than twice differentiable, and can take negative values.

Defn 5.13 ((Rosenblatt-Parzen) Kernel Density Estimator).
Given a smoothing kernel and and a bandwidth $h > 0$, the kernel density estimator is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right)$$

$\mathbb{E}\left[\hat{f}_k(x)\right] = f(x) + \mathsf{O}\left(h^2\right)$; bias decreases as h gets smaller.

$$\mathbb{V}\left[\hat{f}_k(x)\right] = \frac{f(x) \int K(x)^2}{nh} + o(\frac{1}{nh}) \approx \frac{f(x)}{nh} \int K(x)^2 dx$$

vanishes as $nh \to \infty$.
Higher density implies higher variance - more data in a neighbourhood makes the *density estimation problem harder*.
Haerdle et al (2004) find 'optimal' bin-width h for n observations is

$$h_{opt} = \left(\frac{24\sqrt{\pi}}{n}\right)^{1/3}$$

5.3.1 Conditional Density and Distribution Function Estimation

Conditional Density

$$\widehat{g}(y|x) = \frac{\widehat{f}(x, y)}{\widehat{f}(x)}$$

Conditional CDF

$$\widehat{\mathbb{F}}(y|x) = \frac{\frac{1}{n} \sum_{i=1}^{n} G(\frac{y - y_i}{h_0}) K_h(\mathbf{x}_i, \mathbf{x})}{\widehat{f}(x)}$$

where $G(.)$ is a kernel CDF (typically Normal), h_0 is a the smoothing parameter associated with y, and $K_h(\mathbf{x}_i, \mathbf{x})$ is a product kernel.
This can be inverted to get the

Conditional Quantile Function

$$\widehat{q}_\alpha(x) = \inf\left\{y : \widehat{\mathbb{F}}(y|x) \geq \alpha\right\} =: \widehat{F}^{-1}(\alpha|x)$$

Conditional Mode

$$\widehat{m}(x) = \max_y \widehat{g}(y|x)$$

where $\widehat{g}(y|x)$ is the kernel estimator of the conditional density.

5.4 Nonparametric Regression

Given a random pair $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, the regression function is

$$m_0(\mathbf{x}) = \mathbb{E}\left[Y|\mathbf{X} = \mathbf{x}\right]$$

We aim to approximate this with $\widehat{\mathbf{m}(\cdot)}$ when estimating nonparametric regressions.

$$\widehat{m}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) y_i$$

where the weights w_i are estimated using different methods.

Defn 5.14 (K-Nearest Neighbours Regression).
Fix an integer $k \geq 1$.

$$\widehat{m}(\mathbf{x}) = \frac{1}{k} \sum_{i \in \mathfrak{N}_k(\mathbf{x})} y_i$$

where $\mathfrak{N}_k(\mathbf{x})$ is the k-neighbourhood of $\mathbf{x}$ which contains the indices of the k closest points.

Defn 5.15 (Nadaraya-Watson / Kernel regression).

$$\widehat{m}_h(x) = \mathbf{S}_x' \boldsymbol{y} = \sum_{i=1}^{N} w_i(x) y_i$$

where K is a kernel and the weights $w_i(x)$ are given by

$$w_i(\mathbf{x}) := \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)}$$

where $K(\cdot)$ is a kernel-function that assigns a value that is lower the closer x_i is to x, and h is the bandwidth. The estimate of $\mathbb{E}[Y|X = x]$ at x is a weighted average of y_i's 'near x'.
Stated differently,

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K(\frac{x - x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x - x_i}{h})}$$

$\widehat{m}(x)$ is consistent and asymptotically normal. Assuming X has density f, Its variance is

$$\mathbb{V}[\widehat{m}(\mathbf{x})] = \frac{\sigma^2(\mathbf{x})}{f(\mathbf{x}) \cdot nh} \int K^2(u) du + o\left(\frac{1}{nh}\right)$$

Where $\sigma^2(\mathbf{x}) = \mathbb{V}[Y|X = \mathbf{x}]$
the bias for this estimator is

$$\text{bias}[\widehat{m_h}(x)] = \mathbb{E}[\widehat{m}_h(x)] - m(x) \sim h^2 \left(\frac{C_1}{2} m''(x) + C_2 m'(x) \frac{f'(x)}{f(x)}\right)$$

A Nadaraya-Watson estimator with a **uniform kernel** is called a **Regressogram**.

Defn 5.16 (Local Linear Regression / `loess`).
Define the loss function

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - m - (\boldsymbol{x}_i - \boldsymbol{x})'\beta)^2 K\left(\frac{x - x_i}{h}\right)$$

$\widehat{m}(x)$ is obtained by regressing Y on $X - x$, with weights equal to $\sqrt{K(\frac{x - x_i}{h})}$. This estimator is consistent for $\mathbb{E}[Y|X = x]$, with the same rate of convergence as NW.

NW fits a Kernel-weighted constant (0-th order polynomial) near $\mathbf{x}$; LLR fits a straight line.

5.5 Semiparametric Regression

A partially linear model is given by

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + g(\mathbf{z}_i) + \varepsilon_i \; i = 1, \ldots, n$$

where $\mathbf{x}_i$ is a $p-$ vector, z_i is a scalar, and $g(.)$ is not specified. Standard exogeneity assumption $\mathbb{E}[\varepsilon|\mathbf{x}_i, z_i] = 0$.
Robinson (1988) estimator is $\sqrt{n}$-consistent.

Defn 5.17 (Robinson Estimator).

$$\underbrace{y_i - \mathbb{E}[y_i|z_i]}_{\widetilde{y}_i} = \underbrace{(\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i|z_i])}_{\widetilde{\mathbf{x}}_i} \boldsymbol{\beta} + \varepsilon_i$$

where $\widetilde{y}_i$ and $\mathbf{x}_i$ are estimated using Kernel regression.

5.5.1 Index Models

Defn 5.18 (Single Index Model).

$$Y = g(\mathbf{x}'\boldsymbol{\beta}) + \varepsilon$$

with $\mathbb{E}[\varepsilon|\mathbf{x}] = 0$. The term $\mathbf{x}'\boldsymbol{\beta}$ is called a **single index** because it is a scalar. $g(.)$ is left unspecified, hence 'semiparametric'.

Klein and Spady's Semiparametric Binary Model

$$\ell(\boldsymbol{\beta}, h) = \sum_{i=1}^{n} ((1 - y_i) \log(1 - \widehat{g}_{-i}(\mathbf{x}_i'\boldsymbol{\beta})) + y_i \log(\widehat{g}_{-i}(\mathbf{x}_i'\boldsymbol{\beta})))$$

5.6 Splines

Defn 5.19 (Regression Splines).
Given inputs $x_1, \ldots, x_n$ and responses $y_1, \ldots, y_n$, fit functions f that are $k-$th order splines with knots at some chosen locations $t_1, \ldots, t_p$. This means expressing

$$f(x) = \sum_{j=1}^{p+k+1} \beta_j g_j(x)$$

where $\beta_1, \ldots, \beta_{p+k+1}$ are coefficients and $g_1, \ldots, g_{p+k+1}$ are basis functions for $k-$splines over the knots $t_1, \ldots, t_p$.
This is equivalent to the standard regression problem when we define

$$\mathbf{G} = g_j(x_i),\ i = 1, \ldots, n\ , j = 1, \ldots, p+k+1$$
$$\widehat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^{p+k+1}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{G}\boldsymbol{\beta}\|_2^2$$

which gives the standard OLS solution $\widehat{\beta} = \left(\mathbf{G}^\top \mathbf{G}\right)^{-1} \mathbf{G}^\top \mathbf{y}$. *Regression splines are linear smoothers*.

Regression splines exhibit erratic boundary behaviour . One solution to this is to force lower degrees at the extremities.

- f is of order k on each $[t_1, t_2], \ldots, [t_{p-1}, t_p]$
- f is of degree $(k-1)/2$ on $(-\infty, t_1]$ and $[t_p, \infty)$
- f is continuous and has continuous derivatives of orders $1, \ldots, k-1$

Defn 5.20 (Smoothing Splines).
Given a simple non-parametric regression problem $y_i = g(x_i) + \varepsilon_i$, with $x_i \in [0,1]$, the problem to be solved is

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \underbrace{\lambda \int_0^1 g''(x)^2 dx}_{\text{roughness penalty}}$$

The solution to the above problem is a **cubic smoothing spline**, which is a piecewise cubic polynomial: a function with continuous first and second derivatives, whose third derivative may take discrete jumps at designated points, called 'knots'. On each segment $[x_i, x_{i+1})$, we may write $s(x)$ as a cubic polynomial.

Defn 5.21 (Generalised Additive Model (GAM)).
Semiparametric model of the form $y = f(x) + \varepsilon$, where $f(x)$ is typically implemented using basis-splines

$$\mathbb{E}[y|x] = \sum_j^J g_j(x_{ij})$$

or 'thin-plate' splines. Estimated using backfitting [`mgcv` in R].

5.6.1 Reproducing Kernel Hilbert Spaces

Defn 5.22 (Hilbert Space).
A **Hilbert Space** is an abstract vector space endowed with an inner product. Let $\mathcal{X}$ be an arbitrary set and $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$, endowed by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The evaluation functional over the hilbert space of functions $\mathcal{H}$ is a linear function that evaluates each function at a point x:

$$L_x : f \to f(x),\ \forall f \in \mathcal{H}.$$

A **Reproducing Kernel Hilbert space** is a Hilbert space (complete inner product space) with extra structure such that the map L_x is continuous at any $f \in \mathcal{H}$

$$\exists C > 0 \text{ s.t. } |L_x(f)| = |f(x)| \leq C \|f\|_{\mathcal{H}} \quad \forall f \in \mathcal{H}$$

Example 5.15 (L2 Space).
Most common example of RKHS is L_2 space

$$L_2[0,1] = \left\{ f : [0,1] \to \mathbb{R} : \int f^2 < \infty \right\}$$

endowed with the inner product

$$\langle f, g \rangle = \int f(x) g(x) dx$$

with corresponding norm

$$||f|| = \sqrt{\langle f, f \rangle} = \sqrt{\int f^2(x) dx}$$

Defn 5.23 (Mercer Kernel).
A RKHS is defined by a Mercer kernel $K(x, y)$ that is symmetric and positive definite, which means that for any function f,

$$\int \int K(x,y) f(x) f(y) dx\, dy \geq 0$$

The main example is the Gaussian kernel

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2} \right)$$

Given a kernel K, let $K_x(\cdot)$ be the function obtained by fixing the first coordinate at x s.t. $K_x(y) = K(x, y)$. For the Gaussian kernel, K_x is a Normal, centered at x. We can create functions by taking linear combinations of the kernel. Let $\mathcal{H}_0$ denote all such functions

$$\mathcal{H}_0 = \left\{ f : \sum_{j=1}^{k} \alpha_j K_{x_j}(x) \right\}$$

which can be used to define an inner product

$$\langle f, g \rangle = \langle f, g \rangle_K = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j)$$

Theorem 5.16 (Representer Theorem / The Reproducing Property of RKHS).
Let $f(x) = \sum_i \alpha_i K_{x_i}(x)$. Then,

$$\langle f, K_x \rangle = \sum_i \alpha_i K(x_i, x) = f(x)$$

This also implies

$$\langle K_x, K_y \rangle = K(x, y)$$

This implies that K_x is the *representer* of the evaluation functional. The completion of $\mathcal{H}_0$ with respect to $\|\cdot\|_K$ is denoted by $\mathcal{H}_K$ and is called the RKHS generated by K.

Defn 5.24 (RKHS Regression).
Define $\widehat{m}$ to minimise

$$R := \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i))^2 + \lambda \|m\|_K^2$$

By the representer thm, $\widehat{m}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(x_i, \mathbf{x})$. Plugging this into the above definition,

$$R = \|Y - \mathbb{K}\alpha\|^2 + \lambda \alpha^\top \mathbb{K} \alpha$$

The minimiser over α is

$$\widehat{\alpha} = (\mathbb{K} + \lambda \mathbf{I})^{-1} Y$$

and $\widehat{m}(x) = \sum_j \widehat{\alpha}_j K(X_i, x)$

Defn 5.25 (Support Vector Machines).
Support Vector Machines. Suppose $Y_i \in \{-1, +1\}$. Recall the the linear SVM minimizes the penalized hinge loss:

$$J = \sum_i \left[1 - Y_i \left(\beta_0 + \beta^T X_i\right)\right]_+ + \frac{\lambda}{2} \|\beta\|_2^2$$

The dual is to maximize

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle$$

subject to $0 \le \alpha_i \le C$ The RKHS version is to minimize

$$J = \sum_i [1 - Y_i f(X_i)]_+ + \frac{\lambda}{2} \|f\|_K^2$$

Defn 5.26 (Linear Smoothers).
Estimators of the form $\widehat{m}(x) = \sum_i w_i(x) Y_i$ with weights $w_i(x)$ that don't depend on Y_i are known as **linear smoothers**. Fittedvalues are $\widehat{\mu} = \mathbf{S}\mathbf{y}$ for some matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ depending on inputs and tuning parameters.
The effective degrees of freedom of a linear smoother is

$$\nu = df(\widehat{\mu}) = \sum_{i=1}^{n} \mathbf{S}_{ii} = \operatorname{tr} \mathbf{S}$$

Defn 5.27 (Wavelet).
A wavelet is a function ψ such that

$$\left\{ 2^{j/2} \psi(2^j \cdot - k); j, k \in \mathbb{Z} \right\}$$

is an Orthonormal basis for L_2 space. ψ is called a 'mother wavelet', which can be constructed from a 'father wavelet' φ.

5.7 Gaussian Processes

Based on Williams and Rasmussen (2006), Murphy (2012), and Scholkopf and Smola (2018).

5.7.1 Bayesian Linear Regression

Center $\boldsymbol{y}, \mathbf{X}$ such that $y = \boldsymbol{y} - \bar{y}\mathbf{1}_N$. Likelihood is $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}\left(\boldsymbol{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N\right)$. Choose normal prior $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \propto \exp(-\frac{1}{2\sigma^2} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2)$ or

Conjugate prior is also Gaussian, which we denote by $p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$. Then, the posterior is given by
This can be decomposed as follows

$$p(\boldsymbol{\beta}|\mathbf{X}, \boldsymbol{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)\mathcal{N}\left(\boldsymbol{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N\right) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_N, \boldsymbol{\Sigma}_N) \quad \text{Where}$$
$$\boldsymbol{\beta}_N = \boldsymbol{\Sigma}_N\left(\boldsymbol{\Sigma}_0\right)^{-1}\boldsymbol{\beta}_{\mathbf{0}} + \frac{1}{\sigma^2}\boldsymbol{\Sigma}_N\mathbf{X}'y$$
$$\boldsymbol{\Sigma}_N = \sigma^2(\sigma^2\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

Fact 5.17 (Bayes - Ridge Equivalence).
Let likelihood be $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$ and let prior on coefficients $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_p)$.

$$p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\frac{1}{\sigma^2}\left(\mathbf{A}\right)^{-1}\mathbf{X}\mathbf{y}, \left(\mathbf{A}\right)^{-1}\right)$$
$$\text{where } \mathbf{A} := \sigma^{-2}\mathbf{X}\mathbf{X}' + \Sigma_p^{-1}$$

The predictive distribution is

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(f_*|\mathbf{x}_*, \widetilde{\boldsymbol{\beta}})p(\widetilde{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{y})d\widetilde{\boldsymbol{\beta}}$$
$$= \mathcal{N}\left(\frac{1}{\sigma^2}\mathbf{x}_*'\left(\mathbf{A}\right)^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*'\left(\mathbf{A}\right)^{-1}\mathbf{x}_*\right)$$

Realistic case where σ^2 is unknown we can show that posterior has the form [Murphy pp 237]

$$p(\boldsymbol{\beta}, \sigma^2, \mathcal{D}) = \text{NIG}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\beta}_N, \boldsymbol{\Sigma}_N, a_N, b_N)$$
$$\boldsymbol{\beta}_N = \boldsymbol{\Sigma}_N((\boldsymbol{\Sigma}_0)^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\boldsymbol{y})$$
$$\boldsymbol{\Sigma}_N = \left(\boldsymbol{\Sigma_0}^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}$$

Defn 5.28 (Kernel Trick).
Let $\boldsymbol{\phi}(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^M$ be a mapping from D dimensional input space to M dimensional feature / basis space. Let $\Phi(\mathbf{X})$ be the aggregation of columns $\boldsymbol{\phi}(\mathbf{x})$ for all observations. Then, the same formulation as above applies, with $\mathbf{X}, \mathbf{x}$ replaced by $\boldsymbol{\phi}, \boldsymbol{\Phi}$, which gives the posterior predictive distribution

$$f_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\phi_*^\top\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\mathbf{y},$$
$$\phi_*^\top\Sigma_p\phi_* - \phi_*^\top\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\Phi^\top\Sigma_p\phi_*)$$

where $K = \Phi^\top\Sigma_p\Phi$.
In the above expression, the feature space always enters in the form or $\Phi^\top\Sigma_p\Phi$, $\phi_*\Sigma\phi$, or $\phi_*^\top\Sigma\phi_*$. This means we can define a kernel $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top\Sigma_p\boldsymbol{\phi}(\mathbf{x}')$, which gives us an equivalent dot-product representation $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x}) \cdot \boldsymbol{\psi}(\mathbf{x}')$ where $\boldsymbol{\psi} = \boldsymbol{\Sigma}_p^{1/2}\boldsymbol{\phi}(\mathbf{x})$.
If an algorithm is defined solely in terms of inner products in input space, then it can be 'lifted' into feature space by replacing the occurrences of those inner products by $k(\mathbf{x}, \mathbf{x}')$. This is called the *kernel trick*.
By replacing $\langle x_i, x_j\rangle$ with $K(x_i, x_j)$, we turn a linear procedure into a nonlinear one without adding much computation.

Example 5.18 (Kernel Trick for Ridge Regression).
We know that the ridge coefficient vector is given by

$$\widehat{\boldsymbol{\beta}}_\lambda = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$$

It can equivalently be written as

$$\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}$$

where $\mathbf{X}\mathbf{X}^\top$ is a $n \times n$ matrix whose i, j elements are $\langle\mathbf{x}_i, \mathbf{x}_j\rangle$. Similarly, $\mathbf{x}^\top\mathbf{X}^\top$ is a $n-$dimensional vector with $i-$th element $\langle\mathbf{x}, \mathbf{x}_i\rangle$. This turns the computation of ridge coefficients into the computation of inner products between $p-$ dimensional covariate vectors.
Now, replace the inner product $\langle\mathbf{x}_i, \mathbf{x}_j\rangle$ with $K(\mathbf{x}_i, \mathbf{x}_j)$, where $K(\cdot)$ is a known function. This yields

$$\mathbf{x}^\top\mathbf{X}^\top = (K(\mathbf{x}, \mathbf{X}_1), \ldots, K(\mathbf{x}, \mathbf{X}_n))$$
$$\mathbf{X}\mathbf{X}^\top = \mathbf{K} = (K(\mathbf{X}_i, \mathbf{X}_j))_{1 \le i,j \le n}$$

which turns the prediction into **kernel ridge regression**

$$\widehat{\mathbf{Y}} = \mathbf{K}\left(\mathbf{K} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}$$

where one can use one of many kernel functions

- **linear kernel** : $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
- **polynomial kernel**: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^d, d = 2, 3, \ldots$
- **gaussian kernel**: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2), \ \gamma > 0$ (also known as **radial basis kernel**)
- **laplacian kernel**: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|), \ \gamma > 0$

Defn 5.29 (Gaussian Process).
A Gaussian Process is a *collection of random variables, any finite number of which have a joint Gaussian Distribution*.
A GP is completely specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. A GP of a real process $f(\mathbf{x})$ is specified as follows

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x})\right] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[f(\mathbf{x} - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \qquad \text{then} \\ f(\mathbf{x}) &\sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \end{aligned}$$

Example 5.19 (Bayesian Linear Regression as a GP).
let $f(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\beta}$ with prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_p)$. Then, the mean and covariance functions are

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{x})\right] &= \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}\left[\boldsymbol{\beta}\right] = 0 \\ \mathbb{E}\left[f(\mathbf{x}) f(\mathbf{x})'\right] &= \boldsymbol{\phi}(\mathbf{x}) \mathbb{E}\left[\boldsymbol{\beta}\boldsymbol{\beta}^\top\right] \boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\Sigma}_\mathbf{p} \boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

Square-Exponential covariance / Gaussian Kernel

$$\operatorname{Cov}\left[f(\mathbf{x}_p), f(\mathbf{x}_q)\right] = k(\mathbf{x}_p, \mathbf{x}_q) = \exp\left(-\frac{1}{2} \left|\mathbf{x}_p - \mathbf{x}_q\right|^2\right)$$

6 Maximum Likelihood

Setup Let $\{Z\}_{i=1}^N$ be a sequence of iid rv's with common CDF $F(z|\theta^*)$. We want to estimate $\theta^* \in \Theta \subset \mathbb{R}^k$.
Since the rv's are IID, we can write

$$f(\mathbf{y}|\theta^*) = \prod_{i=1}^N f(y_i|\theta^*)$$

Defn 6.1 (Likelihood Function).
L: $\Theta \to \mathbb{R}$

$$L(\theta|\mathbf{y}) := \prod_{i=1}^N f(y_i|\theta)$$

We usually work with the log of this $\ell(\theta|y) := \sum_{i=1}^N \log f(y_i|\theta)$, and drop the conditioning on y though strictly speaking $f(\mathbf{y}, \mathbf{X}|\theta) = f(\mathbf{y}|\mathbf{X}|\theta) f(\mathbf{X}|\theta)$

Defn 6.2 (Maximum Likelihood Estimator).
is the estimator that maximises the conditional log-likelihood estimator.

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} \mathcal{L}(\theta) = \arg\max_{\theta \in \Theta} \sum_{i=1} \log f(Z_i|\theta)$$

solves the first-order conditions

$$\frac{1}{N} \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{1}{N} \sum_i \frac{\partial \ell(y_i|x_i, \theta)}{\partial \theta} = 0$$

Example 6.1 (Linear Regression).
conditional density:

$$f(y_i|x_i, |\beta, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right)$$

Log likelihood:

$$\begin{aligned} \ell(\beta, \sigma^2) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \\ &= -\frac{N}{2}\boldsymbol{RSS}(\beta) - \frac{N}{2}\log(2\pi\sigma^2) \qquad = \end{aligned}$$

Maximising this w.r.t. b, s^2 yields $\hat{\beta} = (X'X)^{-1} X'y$, $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{N}$.

Defn 6.3 (Score Function).
gradient vector $\mathcal{S}(\theta) := \nabla \ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$

$$\mathcal{S}(z, \theta) := \frac{\partial \ell(\theta)}{\partial \theta}(z; \theta)$$

Evaluated at θ^*, this is the efficient score.
local maximum that solves the FOCs $\mathcal{S}(z, \theta) = 0$, IOW

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\log f(Z_i|\theta)}{\partial \theta} = \mathbf{0}$$

Defn 6.4 (Fisher Information / Information Matrix Equality).

$$\mathcal{I}(\theta) := \mathbb{E}\left[\mathcal{S}(\theta)\mathcal{S}(\theta)'\right] = \mathbb{E}\left[\frac{\partial \ell(\theta)}{\partial \theta}\frac{\partial \ell(\theta)}{\partial \theta'}\right] = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}\right]$$

$$\mathcal{A} = \mathcal{B} := \mathbb{V}\left[s_i(\theta^*)\right] = \mathbb{E}\left[s_i(\theta^*)s_i(\theta^*)'\right] = \mathbb{E}\left[\frac{\partial}{\partial \theta} s_i(\theta^*)\right]$$

Variance estimate $\mathbb{V}\left[\hat{\theta}\right] = \frac{1}{n}\mathcal{A}^{-1}$;
Estimated as

$$\mathcal{I}(\hat{\theta}) = -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta'}\Big|_{\theta = \theta_{MLE}}$$

Example 6.2 (IM for OLS).

For OLS, parameter vector $\theta = (\beta', \sigma^2)') = (\beta', \gamma)'$.
Scores:

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\gamma} X'(y - X\beta)$$

$$\frac{\partial \ell}{\partial \gamma} = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2} \underbrace{(y - X\beta)'(y - x\beta)}_{:=s}$$

$$\gamma = s/n = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\beta)$$

Information Matrix / Variance

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} X'X & 0 \\ 0' & \frac{n}{2\sigma^4} \end{pmatrix} \implies CRLB := (I(\theta))^{-1} = \begin{pmatrix} \sigma^2 (X'X)^{-1} & 0 \\ 0' & \frac{2\sigma^4}{n} \end{pmatrix}$$

6.1 Properties of Maximum Likelihood Estimators

Property 6.3 (Consistency).

As $N \to \infty$, probability of missing the true parameter goes to zero.

$$P(|\hat{\theta}_n - \theta| > \epsilon) \xrightarrow{p} 0 \; \forall \epsilon > 0$$

Property 6.4 (Asymptotic Normality).

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1})$$

Equivalently,

$$\hat{\theta}_{MLE} \sim^a \mathcal{N}\left(\theta, -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}\right]^{-1}\right)$$

conditions: (1) $\theta \in \Theta$, (2) ℓ is twice-differentiable

Property 6.5 (Efficiency).
Variance of MLE is the **Cramer-Rao Bound**; the asymptotic variance of the MLE is at least as small as that of any other consistent estimator.

Theorem 6.6 (Cramer-Rao Inequality / Bound).
Let the pdf of the r.v. X be $f_X(x|\theta)$ for some $\theta_0 \in \Theta$. Let $\widehat{\theta}$ be an unbiased estimator for θ_0. Suppose the derivative $\partial/\partial\theta$ can be passed under the integral $\int f(x|\theta)dx$ and $\int \theta(x) f(x|\theta)dx$ and suppose the fisher information

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log f}{\partial \theta \partial \theta'}(X|\theta)\right]$$

is finite. Then,

$$\mathbb{V}\left[\widehat{\theta}(X)\right] \geq \mathcal{I}(\theta_0)^{-1}$$

If $X_1, X_2, \ldots, X_n$ are iid with common density $\mathrm{f}_x(x|\theta)$, the implied bound on the variance is $N\mathbb{V}\left[\widehat{\theta}(X)\right] \geq \mathcal{I}(\theta_0)^{-1}$

Defn 6.5 (Asymptotic Efficiency).
Let $X_1, X_2, \ldots$ be iid random variables with common density $\mathsf{f}_x(x|\theta)$. A sequence of estimates $\widehat{\theta}_N$, a function $X_1, X2, \ldots, X_N$ that satisfies

$$\sqrt{N}(\widehat{\theta}_{ml} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta)^{-1}\right)$$

whatever the true value of $\theta \in \Theta$ is, is said to be asymptotically efficient.

Defn 6.6 (Semiparametric Efficiency Bound).
Suppose $X_1, X_2, \ldots$ are iid with density $X \sim f(x|\theta, h(.))$ where $h(.)$ is an unknown function. Next, we pretend to know the infinite dimensional parameter $h()$ up to a finite dimensional parameter γ, in which we have a fully parametric finite-dimensional parameter θ, thus we can calculate the Cramer-Rao Bound.

$$f(x|\theta, \gamma) = f(x|\theta, h(\gamma))$$

Partitioning the information matrix for $(\theta', \gamma')'$ and its inverse in

$$\mathcal{I}(\theta, \gamma) = \begin{bmatrix} \mathcal{I}_{\theta\theta'} & \mathcal{I}_{\theta\gamma'} \\ \mathcal{I}_{\gamma'\theta} & \mathcal{I}_{\gamma\gamma'} \end{bmatrix} \text{ and } \mathcal{I}(\theta, \gamma)^{-1} = \begin{bmatrix} \mathcal{I}^{\theta\theta'} & \mathcal{I}^{\theta\gamma'} \\ \mathcal{I}^{\gamma\theta'} & \mathcal{I}^{\gamma\gamma'} \end{bmatrix}$$

The Cramer-Rao bound implies that

$$ASV(\widehat{\theta}) \geq \mathcal{I}^{\theta\theta'} = (\mathcal{I}_{\theta\theta'} - \mathcal{I}_{\theta\gamma'}(\mathcal{I}_{\gamma\gamma'})^{-1}\mathcal{I}_{\gamma\theta'})^{-1}$$

This is true for **any** parametrisation of the unknown function $h(.)$. The lowest possible variance for any estimator for θ that does not use knowledge of $h(.)$ has to be at least as high as the lowest variance we can get if we know more, that is, the Cramer-Rao bound for any parametric submodel. So, the *semiparametric efficiency bound* is the largest lower-bound we can get for any parametric submodel.
Suppose we have a candidate Estimator $\widehat{\theta}$ and a given parametrisation $h(x; \gamma)$. Then,

$$(\mathcal{I}_{\theta\theta'} - \mathcal{I}_{\theta\gamma'}(\mathcal{I}_{\gamma\gamma'})^{-1}\mathcal{I}_{\gamma\theta'})^{-1} \leq \text{Semiparametric Efficiency Bound} \leq ASV(\widehat{\theta})$$

For any estimator we can calculate the left hand side, for any parametrization we can calculate the right hand side, so if we find an estimator and a parametrization that the two are equal we have found the efficiency bound.

Theorem 6.7 (Equivariance of the MLE).
Let $\tau = g(\theta)$ where g is bijective, continuous, and differentiable. Let $\hat{\theta}_n$ be the MLE of θ. Then, $\hat{\tau}_n = g(\hat{\theta}_n$ is the MLE of τ.

Defn 6.7 (Marginal Effect $\frac{\partial \mathbb{E}[y|x]}{\partial x}$).
For a regression of the form $\mathbb{E}[y|x] = g(x'\beta)$, one can estimate multiple 'marginal effects'. For the special case where $\mathbb{E}[y|x] = x'\beta$, $\frac{\partial \mathbb{E}[y|x]}{\partial x} = \beta$, but this is not generically true.

- **Average Marginal Effect (AME)**: $:= \frac{1}{N}\sum_i \frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_i}$
- **Marginal Effect at Mean (MEM)** $\frac{\partial \mathbb{E}[y|x]}{\partial x}|_{\bar{x}}$

Defn 6.8 (Factorisation Theorem / Sufficient Statistics).
Suppose $t(x)$ is a sufficient statistic for θ. Then,

$$\prod_i f(x_i|\theta) = g(t(x), \theta)\, h(x)$$

$\hat{\theta}(x)$ depends on the data x only through $t(x)$, the sufficient statistic.

6.2 QMLE / Misspecification / Information Theory

If model is misspecified, $f(\cdot|x_i, \theta) \neq p_0(\cdot|x_i) \ \forall\, \theta \in \Theta$
The MLE converges to the best fitting θ for the population (**pseudo-true value**)

$$\theta^\star = \arg\max_{\theta \in \Theta} \ \text{plim} \frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta)$$

For the **linear exponential family**, the quasi-MLE is consistent even when the density is partially misspecified.

6.2.1 Robust Standard Errors

Asymptotic distribution of QMLE

$$\sqrt{n}\left(\hat{\theta} - \theta^\star\right) \xrightarrow{d} \mathcal{N}\left(0, \hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)$$

where

$$\hat{A} = -\mathbb{E}\left[\mathsf{H}(\hat{\beta})\right] = \mathbb{E}\left[\frac{\partial \mathsf{S}_i(\theta)}{\partial \theta'}\right]|_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}}$$

$$\hat{B} = \mathbb{E}\left[\mathsf{S}_i(\theta^\star)\,\mathsf{S}_i(\theta^\star)'\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta} \times \frac{\partial \ell_i}{\partial \theta'}|_{\hat{\theta}}$$

Defn 6.9 (Kullback-Liebler Distance).
let $f(y|\theta)$ be the assumped joint density, and let $h(y)$ be the true density. Then,

$$\text{KL}[h(\cdot)||f(\cdot)] := \mathbb{E}_h\left[\log\left(\frac{h(y)}{f(y|\theta)}\right)\right] = \int_{-\infty}^{\infty} h(t)\log\left(\frac{h(t)}{f(t)}\right)dt$$

Minimised when $\exists\theta_0$ s.t. $h(y) = f(y|\theta_0)$. **QMLE** minimises distance between $f(y|\theta)$ and $h(y)$. Notation $\text{KL}(h, f)$ denotes 'information lost when f is used to approximate h'.
Discrete version illustrates links to Entropy

$$\text{KL}[p||q] := \sum_{j=1}^{J} p_j \log\frac{p_j}{q_j} = \sum_{j=1}^{J} p_j \log p_j - \sum_{j=1}^{J} p_j \log q_j = -\mathbb{H}(p) + \underbrace{\mathbb{H}(p,q)}_{\text{cross entropy}}$$

$KL(p||q) \geq 0$ and with equality IFF $p = q$.
KL 'distance', unlike Euclidian Distance, is not the same between f, g as g, f; i.e. it is directional.

Defn 6.10 (Akaike Information Criterion (AIC)).
Akaike showed that using K-L model selection entails finding a good estimator for

$$\mathbb{E}_{y,h}\left[\mathbb{E}_{x,h}\left[\log(f(x|\widehat{\theta}(y)))\right]\right]$$

where x, y are independent, random sampels from the same distribution and expectations are taken **w.r.t. the true distribution** h. Estimating this quantity for each model $\mathfrak{f}_i$ is biased upwards. An approximately unbiased estimator of the above target quantity is
For a general class of maximum-likelihood models,

$$\text{AIC} = -2\log\mathcal{L}(\widehat{\boldsymbol{\theta}}|\mathbf{y}) + 2K$$

For linear regression models, this simplifies to

$$AIC = n\log\widehat{\sigma}^2 + 2K \;\; ;\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}\widehat{\varepsilon}^2}{n}$$

Defn 6.11 (Bayesian Information Criterion (BIC)).

$$BIC = \ln\left(\frac{e'e}{n}\right) + \frac{k\ln(n)}{n}$$

6.3 Testing

To test the hypothesis $H_0 : \alpha = 0$ against the alternative, there are three classical tests.

We partition the parameter K-vector θ into two parts $(\theta_0', \theta_1')'$ s.t. the dimensions of the two sub-vectors s.t. $K_0 + K_1 = K$. θ_1 is a nuisance parameter: its value is not restricted under the null.
Let $\hat{\theta}_u := (\hat{\theta}_{0u}, \hat{\theta}_{1u})$ be the unrestricted MLEs. If we are testing the restriction $\theta_0 = 0$, then the restricted parameter vector is $\hat{\theta}_R := (0, \hat{\theta}_{1r})$. IOW, test $\boldsymbol{h}(\theta_0) = 0$.

Defn 6.12 (Likelihood Ratio Test).
If null is true, ℓ at restricted model $((0, \theta_{1r})$ should not be much smaller than ℓ at the unrestricted model $((\theta_{0u}, \theta_{1u})$.

$$LR := 2 \times (\ell(\hat{\theta}_u) - \ell(\hat{\theta}_R))$$

Under the null, $LR \sim \chi^2_{K_0}$ (where K_0 is the number of restrictions being tested).

Defn 6.13 (Lagrange Multiplier Test / Score Test).
If the limiting ℓ is maximised at $\theta_0 = 0$, the derivative of the ℓ wrt θ_0 at that point should be close to zero.

$$LM := \mathcal{S}(\hat{\theta}_R)'[\mathcal{I}^{-1}(\hat{\theta}_R)]\mathcal{S}(\hat{\theta}_R)$$

Under the null, $LR \sim \chi^2_{K_0}$ (where K_0 is the number of restrictions being tested).

Defn 6.14 (Wald Test).
Unrestricted estimates of θ_0 should be close to zero.

$$W := N \cdot \hat{\theta}_{0u}'\left(\hat{\mathcal{I}}^{00}\right)^{-1}\hat{\theta}_{0u}$$

Where $\hat{\mathcal{I}}^{00}$ is the top-left of the information matrix (corresponding with the restricted parameters). Under the null, $W \sim \chi^2_{K_0}$.
alternatively, $W = h(\hat{\theta}_u)'\Omega^{-1}h(\hat{\theta}_u)$
where $h_1(\theta)\ldots h_{K_0}(\theta)$ are restrictions,

$$\Omega = \left(\frac{\partial h(\theta)}{\partial\theta}\right)'\mathbb{V}\left[\theta_u\right]\left(\frac{\partial h(\theta)}{\partial\theta}\right)$$

evaluated at $\hat{\theta}_u$.

Defn 6.15 (Pseudo-R^2).
McFadden's Pseudo-R^2

$$R^2_{bin} := 1 - \frac{\ell(\hat{\beta})}{\ell(\bar{y})}$$

6.4 Binary Choice

Defn 6.16 (Linear Probability Model).
Estimate the probability using OLS $\mathbf{Pr}(y=1|\boldsymbol{x}) = \mathbf{X}\boldsymbol{\beta}$. $\mathbb{V}[y|\boldsymbol{x}] = \mathbf{X}\boldsymbol{\beta}(1-\mathbf{X}\boldsymbol{\beta})$, so **heteroskedasticity is mechanically present** unless all coefficients are zero.

Defn 6.17 (Logit and Logistic Functions).

$$\text{Logit(p)} = \log\left(\frac{p}{1-p}\right)$$

$$\text{Logistic(x)} = \frac{1}{1+\exp(-x)} = \frac{e^x}{1+e^x}$$

Logistic regression fits $\text{logit}(p_i) = x_i'\beta$, where *logit* is the link function that scales $x_i'\beta$ onto the probability scale. Alternatively, one can use $\Phi(0,1)$.

Defn 6.18 (Logit Parametrisation).
For $Y_i \in \{0,1\}$, assume latent index model $Y_i^* = X_i'\beta + \epsilon_i$; $Y_i := \mathbb{1}_{Y_i^*>0}$. Y_i is bernoulli, so $\mathcal{L} = \prod_{i=1}^N \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$.
Symmetric CDFs. Let $\pi_i = \mathbb{E}[Y_i|X_i] = \mathbf{Pr}(Y_i=1|X_i) = \mathbb{F}(X_i'\beta) = 1-\mathbb{F}(-X_i'\beta)$

- Probit: $\mathbb{F}(u) = \Phi(u)$
- Logit:
  - $\mathbb{F}(u) = \Lambda(u) = \frac{1}{1+\exp(-u)} = \frac{\exp(u)}{1+\exp(u)}$
  - $f(u) = \Lambda'(u) = (1-e^{-u})^{-2}e^{-u}$

Model	$\pi_i = \mathbf{Pr}(y=1\|x)$	Marginal Effect: $\frac{\partial p}{\partial x_i}$
Logit	$\Lambda(x'\beta) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$	$\Lambda(x'\beta)[1-\Lambda(x'\beta)]\beta_j$
Probit	$\Phi(x'\beta)$	$\phi(x'\beta)\beta_j$
Clog-log	$C(x'\beta) = 1-\exp(-\exp(x'\beta))$	$\exp(-\exp(x'\beta))\exp(x'\beta)\beta_j$
LPM	$x'\beta$	β_j

$$\ell(\beta) = \frac{1}{n}\sum\left(Y_i \log \mathbb{F}(X_i'\beta) + (1-Y_i)\log(1-\mathbb{F}(X_i'\beta))\right)$$

Fact 6.8 (Score and QoIs for binary choice).
Let $f_i := \mathsf{f}(x_i'\beta)$; $F_i = \mathbb{F}(x_i'\beta)$ be the density and CDF evaluated at $\boldsymbol{x}_i'\boldsymbol{\beta}$.

$$\boldsymbol{s}_i(\boldsymbol{\theta}) = \frac{f_i \boldsymbol{x}_i'[y_i - F_i]}{F_i(1-F_i)}$$

Sample Score solves

$$\sum_{i=1}^N \left(\frac{y_i}{F_i}f_i x_i - \frac{1-y_i}{1-F_i}f_i x_i\right) = 0$$

Variance:

$$\widehat{\text{Var}}[\hat{\beta}] = \left(\sum_{i=1}^N \frac{f_i^2 x_i x_i'}{F_i(1-F_i)}\right)^{-1}$$

$\mathbb{V}[y_i|x_i] = F_i(1-F_i)$
Marginal effect:

$$\frac{\partial \mathbf{Pr}(y_i=1|\boldsymbol{x_i})}{\partial \boldsymbol{x_i}} = \mathsf{f}(\boldsymbol{x}_i'\boldsymbol{\beta})\,\boldsymbol{\beta}$$

Example 6.9 (Logistic Regression).

$$Q(\theta) = \ell(\theta) = \frac{1}{n}\sum_i (y_i \log\Lambda(x_i'\theta) + (1-y_i)\log[1-\Lambda(x_i'\theta)])$$

Since for the logistic CDF, $\Lambda'(v) = \lambda(v) = \Lambda(v)(1-\Lambda(v))$, the score and hessian can be written as

$$\mathsf{S}(\theta) = [y_i - \Lambda(x_i'\theta)]x_i$$
$$\mathsf{H}(\theta) = -\Lambda(x_i'\theta)[1-\Lambda(x_i'\theta)]x_i x_i' = -\lambda(x_i'\theta)x_i x_i'$$

6.5 Discrete Choice

In many Additive Random Utility Models (ARUMs), $\mathbb{F}(\epsilon_1 - \epsilon_0)$ is logistic for multivariate extensions to logit. This assumes that the errors themselves are distributed **Gumbel/ type 1 extreme-value distribution**

Defn 6.19 (Gumbel Distribution).

$$\mathsf{f}(\epsilon) = \exp(-\epsilon)\exp(-\exp(-\epsilon)) \quad -\infty < \epsilon < \epsilon$$

and $\mathbb{F}(\epsilon) = \exp(-\exp(-\epsilon))$.

6.5.1 Ordered

Random utility with multiple cutoffs $\phi_1 \dots \phi_J$, where $\phi_1 = 0, \phi_J = \infty$.

Defn 6.20 (Ordered Logit).
Define y_i^* latent variable, and

$$y_i = \begin{cases} 0 \text{ if } & -\infty(= & \psi_0) < y_i^* \leq \psi_1 \\ 1 \text{ if } & & \psi_1 < y_i^* \leq \psi_2 \\ \vdots & & \vdots \\ J \text{ if } & & \psi_{J-1} < y_i^* \leq \infty(= \psi_J) \end{cases}$$

which means

$$\Pr\left(y_i \leq j | \boldsymbol{x}_i\right) = \frac{\exp\left(\psi_j - \boldsymbol{x}_i'\beta\right)}{1 + \exp\left(\psi_j - \boldsymbol{x}_i^\top \beta\right)}$$

Defn 6.21 (Ordered Probit).

$$\Pr\left(y_i \leq j | \boldsymbol{x}_i\right) = \Phi\left(\psi_j - \boldsymbol{x}_i'\beta\right) \Leftrightarrow \mathbf{Pr}\left(y = k-1 | \boldsymbol{x}_i\right) = \Phi(\alpha_k - \mathbf{X}\boldsymbol{\beta}) - \Phi(\alpha_{k-1} - \mathbf{X}\boldsymbol{\beta})$$

Both specifications yield a likelihood that is simply the product of binary logit/probit models that switch between adjacent categories for each observation.

$$\ell(\boldsymbol{\beta}, \psi | \mathbf{Y}, \boldsymbol{X}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \mathbb{1}_{y_i = j} \log\left(\mathbb{F}\left(\psi_j - X_i'\beta\right) - \mathbb{F}\left(\psi_{j-1} X_i'\beta\right)\right)$$

Marginal effects are of the form

$$\left.\frac{\partial \mathbf{Pr}\left(Y = j\right)}{\partial x_j}\right|_{\bar{x}} = \hat{\beta}_j \left(\mathsf{f}\left(\hat{\psi}_j - \bar{x}'\hat{\beta}\right) - \mathsf{f}\left(\hat{\psi}_{j-1} - \bar{x}'\hat{\beta}\right)\right)$$

6.5.2 Unordered

Multinomial distribution $:= p(y_i) = \prod_{j=1}^{J} \pi_j^{\mathbb{1}_{y_j = j}}$

$$\ell(\boldsymbol{\pi} | \boldsymbol{Y}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \mathbb{1}_{ij} \log \pi_j$$

Defn 6.22 (Multinomial Logit).

$$\pi_{ij} = \mathbf{Pr}\left(y_i = j | \boldsymbol{x}_i\right) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\left[1 + \sum_{j=1}^{J} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right]} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{\sum_{k=2}^{J} \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}_k\right)}$$

where we adopt normalisation $\mathbf{Pr}\left(y = 0 | \boldsymbol{x}_i'\boldsymbol{\beta}\right) = \frac{1}{\left[1 + \sum_{j=1}^{J} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right]}$ for identification.
Coefficient interpretation:

$$\frac{p_j(\boldsymbol{x}_i, \boldsymbol{\beta})}{p_0(\boldsymbol{x}, \boldsymbol{\beta})} = \exp(\boldsymbol{x}\boldsymbol{\beta}_j) \Leftrightarrow \log\left[\frac{p_j(\boldsymbol{x}_i, \boldsymbol{\beta})}{p_h(\boldsymbol{x}, \boldsymbol{\beta})}\right] = \boldsymbol{x}_i'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)\ \forall j, h \in 1, \ldots J$$

which implies that the log-odds ratio is linear in $\boldsymbol{x}$.

Defn 6.23 (Conditional Logit).
Permits incorporation of choice-varying predictors X_{ij}, nests MNL.

$$\pi_{ij} = \Pr\left(Y_i = j | X_{ij}\right) = \frac{\exp\left(X_{ij}'\beta\right)}{\sum_{k=2}^{J} \exp\left(X_{ik}'\beta\right)}$$

Log likelihood of the form

$$\ell = \sum_{i=1}^{n} \left(\sum_{h=1}^{M} \mathbb{1}_{ij}[x_i'\beta_h] - \log\left(\sum_{l=1}^{M} \exp x_i'\beta_l\right)\right)$$

Defn 6.24 (IIA).
Relative risk π_{ij}/π_{ik} independent of other choices $\neg\{j, k\}$; choices are series of pairwise comparisons. $p_j(\boldsymbol{x}_j)/p_h(\boldsymbol{x}_h) = \exp[(\boldsymbol{x}_j - \boldsymbol{x}_h)\boldsymbol{\beta}]$. IoW $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$.

Defn 6.25 (Multinomial probit).
$\boldsymbol{\epsilon}_i \sim_{\text{iid}} \text{MVN}\left(0, \Sigma_J\right)$

$$\pi_{ij} = \int_{-\infty}^{-\ddot{X}_{1j}^\top \beta} \cdots \int_{-\infty}^{-\ddot{X}_{Jj}^\top \beta} \phi\left(\ddot{\epsilon}_{1j}, \ldots, \ddot{\epsilon}_{Jj}\right) d\ddot{\epsilon}_{1j} \cdots d\ddot{\epsilon}_{Jj}$$

where $\ddot{X}_{kl} = X_{ik} - X_{il}; \ddot{\epsilon}_{kl} = \epsilon_{ik} - \epsilon_{il}$

6.6 Counts and Rates

6.6.1 Counts

Defn 6.26 (Poisson Regression).
$f(y|\lambda) = \lambda^y \exp(-\lambda)/y!$
Poisson specification: $\lambda = \exp(x_i'\beta)$. Yields log density

$$\log f(y|x, \beta) = y_i \exp(x_i'\beta) - x_i'\beta - \log y!$$

Score:

$$\boldsymbol{s}_i(\boldsymbol{\theta}) = -\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i' y_i \boldsymbol{x}_i' = \boldsymbol{x}_i'(y_i - \exp(\boldsymbol{x}_i'\beta))$$

solves

$$\sum x_i' \left(y_i - \exp(x_i'\beta)\right) = 0$$

Hessian

$$\mathsf{H}(\beta) = \frac{\partial s(\beta)}{\partial \beta} = -\exp\left(x_i'\beta\right) x_i x_i' \implies \text{Avar}(\hat{\beta}) = \left(\sum_{i=1}^{n} \exp\left(x_i'\beta\right) x_i x_i'\right)^{-1}$$

Assumes $\lambda := \exp(x_i'\beta) = \mathbb{E}[Y|X] = \mathbb{V}[Y|X]$.
Marginal Effect: Since $\mathbb{E}[y|x] = \exp x'\beta$ for poisson, $\frac{\partial \mathbb{E}[y|x]}{\partial x_j} = \mathbb{E}[y|x]\,\beta_j$. Parameters can be interpreted as **semi elasticities**, since

$$\beta = \frac{\partial \mathbb{E}[y|x]}{\partial x} \times \frac{1}{\mathbb{E}[y|x]} = \frac{\partial \log \mathbb{E}[y|x]}{\partial x}$$

Defn 6.27 (Overdispersed poisson).

$$\mathrm{E}(Y_i|X_i) = \lambda_i = \exp(X_i'\beta) \quad ; \quad \mathrm{Var}(Y_i|X_i) = V_i = \sigma^2\lambda_i; \sigma^2 > 1$$

Defn 6.28 (Zero Inflated Poisson (ZIP)).
define a bernoulli $\pi_i = 1w.p.\theta_i$ for $y = 0$ observation, and specify separate models for zero and nonzero data, with potentially different covariates on θ and λ. Yields the following (difficult to maximise) likelihood

$$\mathcal{L} = \prod_{i=1}^{N} \left(\theta_i + (1-\theta_i)\exp(-\lambda_i)\left((1-\theta_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right)^{1-\pi_i}\right)$$

Defn 6.29 (Negative Binomial Regression).

$$p(y_i) = \frac{\Gamma\left(\frac{\lambda}{\sigma^2-1} + y_i\right)}{y_i!\Gamma\left(\frac{\lambda}{\sigma^2-1}\right)}\left(\frac{\sigma^2-1}{\sigma^2}\right)^{y_i}\left(\sigma^2\right)^{\frac{-\lambda}{\sigma^2-1}}$$

$\mathbb{E}[Y_i] = \lambda; \mathbb{V}[Y] = \lambda\sigma^2$.

Fact 6.10 (NB2 Likelihood).
Let $\mu_i = \exp(x_i'\beta)$, $r_i = \alpha/(\alpha+\mu_i)$, $q_i = \alpha\mu_i^{2-p}$

$$\ell = \sum_{i=1}^{n} \log\Gamma(y_i + q_i) - \log\Gamma(q_i) - \log\Gamma(y_i+1) + q_i + \log r_i + y_i \log(1-r_i)$$

6.6.2 Rates

- Survival: $S(y) := 1 - \mathbb{F}(y)$
- Hazard: $\lambda(y) = h(y) := \frac{f(y)}{1-F(y)} = \frac{f(y)}{S(y)}$;
- Cumulative Hazard $\Lambda(y) := \int_0^y \lambda(s)ds = -\log S(y)$.

Defn 6.30 (Kaplan-Meier).

$$\widehat{S}(t_j) = \prod_{k=j}^{J}\left(1 - \hat{\lambda}(t_k)\right) = \prod_{k=j}^{J}\frac{r_k - d_k}{r_k}$$

Proportional Hazard Models
Conditional hazard rate $\lambda(t|x)$ can be factored as

$$\lambda(t|\mathbf{x},\beta) = \underbrace{\lambda_0(t)}_{\text{baseline hazard}}\underbrace{\phi(\mathbf{x},\beta)}_{\exp(x'\beta)}$$

baseline hazard ($= 1$ for exponential and $\alpha y^{\alpha-1}$ for weibull).

Parametric Model	Hazard	Survival
Exponential	γ	$\exp(-\gamma t)$
Weibull	$\gamma\alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha}$
Generalised Weibull	$\gamma\alpha t^{\alpha-1}$	$[1-\mu\gamma t^{\alpha}]^{1/\mu}$
Gompertz	$\gamma\exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t-1}))$

For survival models with censoring, Likelihood is often written as

$$\mathcal{L}(\theta) = \prod_i f(t_i|\theta)^{d_i} S(t_i|\theta)^{1-d_i}$$

where d_i is a right-censoring indicator and t_i is the observed time.

Example 6.11 (Weibull Example).
Weibull Density: $f(y) = \gamma\alpha y^{\alpha-1}\exp\left(-\gamma y^{\alpha}\right)$, $y, \alpha, \gamma > 0$. $\mathbb{E}[y] = \gamma^{-1/\alpha}\Gamma(\alpha^{-1}+1)$.
Specify $\gamma = \exp(x'\beta)$, so $\mathbb{E}[y|x] = \exp(-x'\beta/\alpha)\Gamma(\alpha^{-1}+1)$. Then, the log-likelihood is

$$\ell(\theta) = \frac{1}{N}\sum_i \{x_i'\beta + \log\alpha + (\alpha-1)\log y_i - \exp(x_i'\beta)y_i^{\alpha}\}$$

FOCs are

$$N^{-1}\sum_i\{1-\exp(x_i'\beta)y_i^\alpha\}x_i=0$$
$$N^{-1}\{\alpha^{-1}+\log y_i-\exp(x_i'\beta)\,y_i^\alpha\log y_i\}=0$$

Model **needs to be correctly specified to be consistent**. Unlike OLS or poisson.

6.7 Truncation and Censored Regressions

Fact 6.12 (Truncated Distribution Density).
If a c.r.v $y\sim f(y)$ and is truncated at c,

$$f(y|y>c)=\frac{f(y)}{\mathbf{Pr}\,(y>c)}=\frac{f(y)}{1-\mathbb{F}(c)}$$

For the truncated normal distribution where $y\sim\mathcal{N}\left(\mu_0,\sigma_0^2\right)$ is truncated at c,
$\mathbb{E}\left[y|y>c\right]=\mu_0+\sigma_0\lambda(v)$, $\mathbb{V}\left[y|y>c\right]=\sigma_0^2\{1-\lambda(v)[\lambda(v)-v]\}$
where $v=(c-\mu_0)/\sigma_0$ and $\lambda(v)=\frac{\phi(v)}{1-\Phi(v)}$ is the **inverse Mills ratio / Hazard function**.

6.7.1 Tobit Regression

Censored Y_i s.t. $y_i^*=\beta' x_i+\epsilon_i$ $\epsilon_i\sim\mathcal{N}\left(0,\sigma^2\right)$ and $y_i=\begin{cases}y_i^* & \text{if } y_i^*>c\\ c & \text{if } y_i^*\le c\end{cases}$ [i.e. y is censored from below at c].
Truncated MLE maximises

$$\log L_n(\theta)=\sum_{i=1}^n\left(\log f(y_i|\mathbf{x}_i,\theta)-\log\left[1-\mathbb{F}(c|\mathbf{x}_i,\theta)\right]\right)$$

with f and $\mathbb{F}$ denoting the density and distribution of y^* respectively.
Type-I Tobit assumes y^* is normally distributed, which gives us the following likelihood

$$L=\prod_0[1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma)]\prod_1\sigma^{-1}\phi[(y_i-\mathbf{x}_i'\boldsymbol{\beta})/\sigma]$$

6.7.2 Censored Regression

Consider a model $y_i=x_i'\beta+\epsilon_i;\ \epsilon_i|x_i\sim\mathcal{N}\left(0,\sigma^2\right)$ and y_i is **not observed if** $y_i>c$.

Yields log-likelihood

$$\ell(y_i|x_i;\beta,\sigma^2)=\left(-\frac{1}{2}\log(\sigma^2)-\frac{1}{2}\left(\frac{y_i-x_i'\beta}{\sigma}\right)^2\right)-\log\left(1-\Phi\left(\frac{c-x_i'\beta}{\sigma}\right)\right)$$

6.8 Generalised Linear Models Theory

Semi-robust likelihoods belong to the **Linear exponential Family** of the following form:

Defn 6.31 (Random Component).
Response observations y_i are realisations of random variables Y_i with densities of the form

$$f(y|\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right)$$

$\theta\subset\mathbb{R}$ is called the *canonical / natural parameter*, $\phi\subset\mathbb{R}^+$ is the *dispersion parameter*.
$\mathbb{E}\left[Y|\theta,\phi\right]=b'(\theta)$, $\mathbb{V}\left[Y|\theta,\phi\right]=a(\phi)b''(\theta)$

$$f\left(y_i\right)=\exp\left\{\frac{y_i\mu_i-\frac{1}{2}\mu_i^2}{\sigma^2}-\frac{y_i^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right\}$$

Defn 6.32 (Systematic Component).
Linear predictor $\eta_i:=X_i'\beta$ specifying the variation in Y accounted for by known covariates.

Defn 6.33 (Link Function).
g is a transformation of the mean that addresses scaling. It is so called because it links the expected value of the response variable $\mathbb{E}\left[Y|\theta,\phi\right]=\mu_i=b'(\theta_i)$ to the explanatory covariates.

$$g(\mu_i)=\eta_i=X_i'\beta\implies\mu_i=g^{-1}(X_i'\beta)$$

Since $\mu_i=b'(\theta_i)$, under a **canonical link** ($g(\mu_i)=\theta_i(\mu_i)$), $\theta_i=X_i'\beta$.

6.8.1 ML estimation

log likelihood

$$\mathcal{L}(\theta,\phi|y)=\sum_{i=1}^N\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right)$$

Score function

$$\mathcal{S}(\beta, y) = \sum_{i=1}^{N} \frac{\partial \ell_i}{\beta_j} = \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$
$$= \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{\mathbb{V}[\mu_i]} \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} x_{ij}$$

The FoC can be written as

$$\frac{\partial \log \mathcal{L}(\theta, \phi | \boldsymbol{y})}{\partial \beta} = \mathbf{X}' (\mathbf{W})^{-1} [\boldsymbol{y} - \widehat{\boldsymbol{y}}] = \mathbf{0}$$

where $\mathbf{W}$ is a weight matrix (which depends on β). The fitted value

$$\widehat{y} = m(\boldsymbol{x}) = \mathbb{E}[y | \boldsymbol{X} = \boldsymbol{x}] = g^{-1}(\boldsymbol{x}'\boldsymbol{\beta})$$

By a first-order taylor expansion, define

$$z = g(\widehat{y}) + (y - \widehat{y}) \nabla g(\widehat{y})$$

This gives us an update rule

$$\widehat{\beta}_{k+1} = \left(\mathbf{X} (\mathbf{W_k})^{-1} \mathbf{X} \right)^{-1} \mathbf{X} (\mathbf{W}_k)^{-1} \boldsymbol{z}_k$$

repeat until convergence $\widehat{\beta}_\infty$.

Model	Density	Link
OLS	Gaussian	Identity
Logistic	Binomial	Logistic
Logistic	Binomial	Normal
Poisson	Poisson	Log

7 Machine Learning

7.1 Supervised Learning

Every Supervised ML algorithm essentially involves a function class $\mathcal{F}$ and a regulariser $R(f)$ that expresses the complexity of the representation. Then, two steps

1. conditional on a level of complexity, choose best in-sample loss-minimising function

$$\min \underbrace{\sum_{i=1}^{n} L(f(x_i), y_i)}_{\text{in-sample loss}} \text{ over } \underbrace{f \in F}_{\text{function class}} \text{ subject to } \underbrace{R(f) \leq c}_{\text{complexity restriction}}$$

2. Estimate the 'optimal' level of complexity using empirical tuning

Fact 7.1 (Discriminative vs Generative ML).

	Discriminative Model	Generative Model
Goal	Directly estimate $\mathbb{E}[y\|x]$	Estimate $\mathbf{Pr}(x\|y)$ to deduce $\mathbf{Pr}(y\|x)$
What is Learned	Decision Boundary	Probability Distribution of the data
Examples	Regressions, SVM	GDA, Naive Bayes

Defn 7.1 (Loss Functions).
$L : (z, y) \in \mathbb{R} \times Y \mapsto L(z, y) \in \mathbb{R}$ that takes as inputs predicted value z and real data value y and outputs how different they are.

- **Least Squares**: $\frac{1}{2}(y - z)^2$
- **Logistic**: $\log(1 + \exp(-yz))$
- **Hinge**: $\max(0, 1 - yz)$
- **Cross Entropy**: $-[y \log z + (1 - y) \log(1 - z)]$

Table 1: Handy Dandy reference from Mullainathan and Spiess (2017, table 2)

Function class $\mathcal{F}$	**Regulariser / Tuning Parameters** $R(f)$
Global / Parametric Predictors	
Linear $\beta' x$ and generalisations	Subset selection $\|\beta\|_0 = \sum_{j=1}^k \mathbb{1}_{\beta_j \neq 0}$ LASSO $\|\beta\|_1 = \sum_{j=1}^k \|\beta_j\|$ Ridge $\|\beta\|_2^2 = \sum_{j=1}^k \beta_j^2$ Elastic Net $\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2$
Local/Nonparametric predictors	
Decision / Regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest	Number of trees, Number of variables used in aach tree, size of bootstrap sample, complexity (above)
Nearest Neighbours	Number of Neighbours
Kernel Regression	Kernel Bandwidth
Mixed Predictors	
Neural Networks (including Deep, Convolutional)	Number of layers, number of neurons per layer, connectivity between neurons
Splines	Number of knots, order
Combining Predictors	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples, individual tuning parameters
Boosting: linear combination of predictions of residual	learning rate, number of iterations, individual tuning parameters
Ensemble: weighted combination of different predictions	Ensemble weights, individual tuning parameters

Fact 7.2 (Curse of Dimensionality).
take a unit hypercube in dimension p and we put another hypercube within it that captures a fraction r of observations within the cube. Each edge will be $e_p(r) = r^{1/p}$. For moderately high dimensions $p = 10$, $e_{10}(0.01) = 0.63; e_{10}(0.1) = 0.8$. Need 80% data to cover 10% of sample.
Define d(p,N) as distance from the origin to the closest point. $n = 500, p = 10 \implies d = 0.52$ [closest point closer to the boundary than to the origin].

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}$$

7.1.1 Regularised Regression

In general, we want to impose a penalty for model complexity in order to minimise MSE [trade off some bias for lower variance].

Defn 7.2 (Ridge Regression).
Estimate the following regression

$$f(\beta, \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{N} (y_i - X_i'\beta) + \lambda \sum_{j=1}^{J} \beta^2$$

$$\hat{\beta}^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \mathbf{X}'\boldsymbol{y} \equiv \hat{\beta}_j^{Ridge} = \frac{\hat{\beta}}{1+\lambda}$$

where $\mathbf{X}$ is a standardized design matrix [s.t. all Xs have unit variance]. Let $\mathbf{X} = \mathbf{UDV}'$ be the SVD of $\mathbf{X}$.
Then, ridge coefficients can also be written as

$$\hat{\boldsymbol{\beta}}_\lambda^{Ridge} = \mathbf{V}(\mathbf{D}^2\lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}'\boldsymbol{y} = \sum_{j=1}^{p} \frac{d_j}{d_j^2 + \lambda} \langle \mathbf{U}_j, \mathbf{Y} \rangle \mathbf{V}_j$$

This can be used to compute the ridge coeffficient efficiently for a fine grid of λs.

- Compute SVD of $\mathbf{X}$ and save $\mathbf{U}, \mathbf{D}, \mathbf{V}$
- Compute and store $\mathbf{w}_j = \frac{1}{d_j} \langle \mathbf{U}_j, \mathbf{Y} \rangle \mathbf{V}_j$ for $j = 1, \ldots, p$
- For each $\lambda_m, m = [M]$
  - compute $\gamma_j = \frac{d_j^2}{d_j^2 + \lambda_m}$
  - compute $\widehat{\beta}_{\lambda_m} = \sum_{j=1}^{p} \gamma_j \mathbf{w}_j$

The solution vector is 'biased' towards the leading right singular vectors of $\mathbf{X}$, which gives it the property of a 'smoothed' Principal Components regression.

Fact 7.3 (Degrees of Freedom for Ridge (and other semi-parametric methods)).
For Ridge regression,

$$\text{dof}(\lambda) = \sum_j \frac{\lambda_j}{\lambda_j + \lambda}$$

Where λ_js are the eigenvalues of the Covariance Matrix.
More generally, for any smoother matrix $\widehat{\mathbf{W}}$, $df(\hat{\mu}) = tr(\widehat{\mathbf{W}})$, which may not be an integer for semi/non-parametric smoothers. In the special parametric case of OLS, $\widehat{\mathbf{W}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so the DoF is simply k.

Defn 7.3 (Lasso Regression).
Consider the objective function

$$J(\beta, \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i'\beta) + \lambda \sum_{j=1}^{J} \|\beta_j\|_1$$

fit using sequential coordinate descent. Coefficient vector is soft-thresholded:

$$\beta_j^{\text{lasso}} = \text{sgn}(\hat{\beta}_j)\max\left(\left|\hat{\beta}_j\right| - \lambda, 0\right)$$

where $\mathbf{X}$ is a standardized design matrix [s.t. all Xs have unit variance], and || is the l_1 norm.

In both cases, pick tuning parameter λ using cross-validation.

Defn 7.4 (Penalised Maximum Likelihood Regression).
ML analogue to LASSO. Define

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmin}} \left(-\ell(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{X}) + \lambda \left\|\boldsymbol{\theta}^{\boldsymbol{P}}\right\|_1\right)$$

where

$$\left\|\boldsymbol{\theta}^P\right\|_1 = \sum_{k}^{|\boldsymbol{\theta}^P|} \left|\boldsymbol{\theta}_k^P\right|$$

Defn 7.5 (Elastic Net (Zou and Hastie 2005)).
Combines ridge regression and the lasso by adding a ℓ_2 penalty to the LASSO's objective function

$$J(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2$$

Defn 7.6 (Principal Components LASSO (Tay, Friedman, Tibshirani 2018)).
Generalise the ℓ_2 penalty to a class of penalty functions of the form

$$\boldsymbol{\beta}^T \mathbf{V}\mathbf{Z}\mathbf{V}^T\boldsymbol{\beta}$$

where $\mathbf{Z}$ is a diagonal matrix whose diagonal elements are functions of the squared singular values.
Define the following objective function

$$J(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\theta}{2}\mathbf{V}\mathbf{D}_{d_1^2-d_j^2}\mathbf{V}^T\boldsymbol{\beta}$$

Where $\mathbf{D}_{d_1^2-d_j^2}$ is a $m \times m$ diagonal matrix with diagonal entries equalling $d_1^2 - d_1^2, d_1^2 - d_2^2 \ldots$. This penalty term gives no weight to the component of $\boldsymbol{\beta}$ that aligns with the first right singular vector of $\mathbf{X}$ (i.e. the first principal component. This gives it better predictive accuracy in some settings.
Comparing principal-coordinate predictions of ridge and pcLASSO:

$$\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \sum_{j=1}^{m} \frac{d_j^2}{d_j^2+\theta} u_j u_j^T \mathbf{y}$$
$$\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{pcL}} = \sum_{j=1}^{m} \frac{d_j^2}{d_j^2+\theta(d_1^2-d_j^2)} u_j u_j^T \mathbf{y}$$

The latter corresponds to a more aggressive form of shrinkage towards the leading singular vectors.

7.1.2 Classification

Training sample $(\boldsymbol{x}_i, y_i)$ where $y \in \mathcal{Y} := \{-1, +1\}$ (can relabel to Bernoulli). A predictor $m : \mathcal{X} \to \mathcal{Y}$, where the labels are produced by an (unknown) classifier f. Let $\mathbb{P}$ be an (unknown) distribution on $\mathcal{X}$. The error of m w.r.t. f is defined by

$$\mathcal{R}_{\mathbb{P},f}(m) = \mathbb{P}\left[m(\boldsymbol{X}) \neq f(\boldsymbol{X})\right] = \mathbb{P}\left[\{\boldsymbol{x} \in \mathcal{X} : m(\boldsymbol{x} \neq f(\boldsymbol{x}))\}\right] \text{ where } \boldsymbol{X} \sim \mathbb{P}$$

The empirical risk is defined as

$$\widehat{R}(m) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{m(\boldsymbol{x_i}) \neq y_i}$$

A perfect classifier (in the sense that $R_{\mathbb{P},f}(m) = 0$) does not exist, so we aim for **Probably Approximately Correct (PAC)** learners that have $R_{\mathbb{P},f}(m) \leq \varepsilon$ w.p. $1-\delta$. The space of models m is restricted to be in finite set $\mathcal{M}$. It can be shown that $\forall \varepsilon, \delta, \mathbb{P}, f$, if $n \geq \varepsilon^{-1}\log[(\delta)^{-1}|\mathcal{M}|]$, then $R_{\mathbb{P},f}(m^*) \leq \varepsilon \; w.p. \geq 1-\delta$ where

$$m^* \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{m(\boldsymbol{x_i})=y_i}\right]$$

Defn 7.7 (Vapnik-Chervonenkis Dimension).
loosely represents the expressive capacity of a set of functions.
Consider k points $\{x_1, \ldots, x_k\}$ and the set

$$E_k = \{m(x_1), \ldots, m(x_k) : \text{for } m \in \mathcal{M}\} \equiv \{-1, +1\}^k$$

we say that m shatters all the points if $|E_k| = 2^k$, i.e. all combinations are possible. Linear functions can shatter 2 points.
The VC dimension of $\mathcal{M}$ is

$$\text{VC}(\mathcal{M}) := \sup\{k \text{ s.t. } \mathcal{M} \text{ shatters } \{x_1, \ldots, x_k\}\}$$

Defn 7.8 (Support Vector Machines).
Let $y \in \{-1, 1\}$. A linear classifier can then be written as $h(x) = \text{sgn}(H(x))$ where

$$H(x) = a_0 + \sum_{i=1}^{d} a_i x_i$$

Suppose $\exists$ a hyperplane $H(x)$ s.t. $Y_i H(x_i) \geq 1 \; \forall i$.
The hyperplane $\hat{H}(x) = \hat{a}_0 + \sum_{i=1}^{N} \hat{a}_i x_i$ that separates the data and maximises the 'margin' is given by minimising $1/2\sum_{j=1}^{d} a_j^2$ subject to $Y_i H(x_i) \geq 1$.

Defn 7.9 (Boosting and Sequential Learning).
Typically, the function space $\mathcal{M}$ is large and complex, so a natural idea is to learn iteratively. Loosely, estimate a model m_1 for y from $\mathbf{X}$, which produces error ε_1. Next, estimate m_2 for ε_1 from $\mathbf{X}$, which produces ε_2, and so on. So, after k steps,

$$m^k(\cdot) = \underbrace{m_1(\cdot)}_{\sim y} + \underbrace{m_2(\cdot)}_{\sim \varepsilon_1} + \ldots \underbrace{m_k(\cdot)}_{\sim \varepsilon_{k-1}}$$

where the first error is $y - m(x)$ and so on, and can also be seen as the gradient associated with the quadratic loss function, $\varepsilon = \nabla\ell$. So, an equivalent representation is

$$m^{(k)} = m^{(k-1)} + \underset{h\in\mathcal{H}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell \left(\underbrace{y_i - m^{(k-1)}(\boldsymbol{x_i})}_{\varepsilon_{k,i}}, h(\boldsymbol{x_i}) \right) \right\}$$

where $\mathcal{H}$ is a space of 'weak learners' (typically step functions). To ensure 'slow' learning, one typically applies a shrinkage parameter $\varepsilon_1 = y - \alpha m_1(\boldsymbol{x}_1)$ $\alpha \in (0,1)$.

Arthur Charpantier's series on the probabilistic foundations of econometrics and machine learning cover this and more and have an excellent bibliography

- Econometrics
  - https://freakonometrics.hypotheses.org/57649
  - https://freakonometrics.hypotheses.org/57674
  - https://freakonometrics.hypotheses.org/57693
  - https://freakonometrics.hypotheses.org/57703
- ML
  - https://freakonometrics.hypotheses.org/57705
  - https://freakonometrics.hypotheses.org/57745
  - https://freakonometrics.hypotheses.org/57782
  - https://freakonometrics.hypotheses.org/57790
  - https://freakonometrics.hypotheses.org/57813
- Bibliography https://freakonometrics.hypotheses.org/57737

7.1.3 Goodness of Fit for Classification

Defn 7.10 (Calibration and Discrimination).

- **calibration**: Bin predicted probabilities $\widehat{y}$ into bins $\{g_k\}$, and within each compute $\widehat{\overline{Y}}_{g_k}$ (average predicted probability) and $\overline{Y}_{g_k}$. Plot the two average against each other. In a well calibrated model, the binned averages trace the identity line.
- **discrimination**: Discrimination is a measure of whether $Y = 1$ observations have high $\widehat{Y}$, and correspondingly $Y = 0$ values have low $\widehat{Y}$. Many measures; listed below

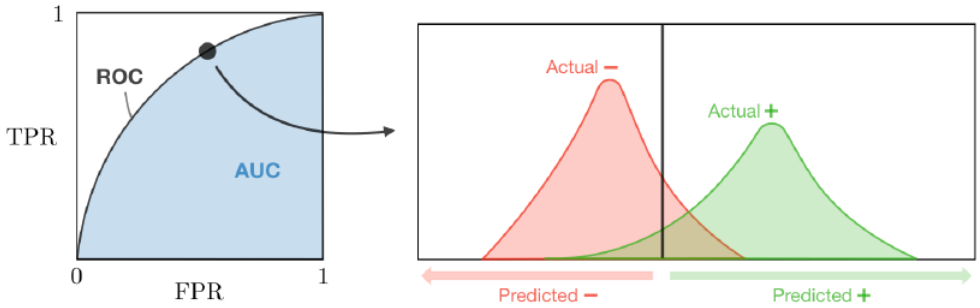


Figure 8: AUC

Defn 7.11 (Confusion Matrix).

	Observed $Y = 1$	$Y = 0$
Predicted positive $(\widehat{Y} > c)$	True Positive (TP)	False Positive (FP)
Predicted negative $(\widehat{Y} < c)$	False Negative (FN)	True Negative (TN)
	Total Positive(P)	Total Negative(N)

- **Accuracy** = $(TP + TN)/(P + N)$ - Overall performance
- **Precision** = $TP/(TP + FP)$ - How accurate positive predictions are
- **Sensitivity = Recall = True positive Rate** = TP/P - Coverage of actual positive sample
- **Specificity = True Negative Rate** = TN/N - Coverage of actual negative sample
- **Brier Score** =

$$\frac{1}{N}\sum_i \left(\hat{Y}_i - Y_i\right)^2 = \overbrace{\frac{1}{N}\sum_k^K \left(\hat{Y}_k - \bar{Y}_k\right)^2}^{\text{Calibration}} + \overbrace{\frac{1}{N}\sum_k^K n_k \left(\bar{Y}_k\left(1-\bar{Y}_k\right)\right)}^{\text{Refinement}}$$

- **F1 Score** = $\frac{2TP}{2TP+FP+FN}$: hybrid metric for unbalanced classes

Fact 7.4 (Receiver Operating Curve (ROC) / Area Under the Curve (AUC)).
Is the plot of TPR vs FPR by varying the threshold c.

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Figure 9: Wikipedia table for confusion matrix

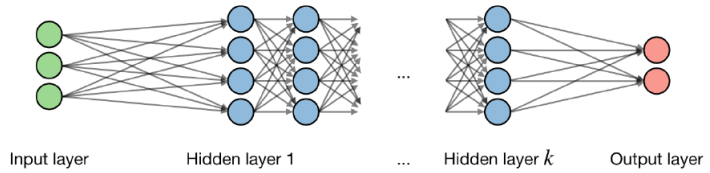


Figure 10: Neural Network Components

Defn 7.12 (Random Forests).
Suppose we have a training set $\{(\boldsymbol{X}_i, Y_i, D_i)\}_{i=1}^N$, a test point x, and a tree predictor

$$\widehat{\mu}(x) = T(x; \{(\boldsymbol{X}_i, Y_i, D_i)\}_{i=1}^N)$$

Equivalently,

$$\widehat{\mu}(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i(\boldsymbol{x}) Y_i \text{ where } \alpha_i(\boldsymbol{x}) = \frac{\mathbb{1}_{\boldsymbol{x}_i \in \mathcal{L}(\boldsymbol{X})}}{|i : \boldsymbol{x}_i \in \mathcal{L}(\boldsymbol{x})|}$$

where $\mathcal{X}$ is partitioned into leaves $\mathcal{L}(x)$, where leaves are constructed to maximise heterogeneity between nodes . Do this until all leaves have $2\times$ `minimum leaf size` observations. Regression trees overfit, so we need to use cross-validation + other tricks.
Random forests build and average many different trees T^* by

- Bagging / subsampling training set (Breiman)
- Selecting the splitting variable at each step from m out of p randomly drawn features (Amit and Geman)

$$\widehat{\tau}(x) \frac{1}{B} \sum_{b=1}^B T_b^*(x; \{(\boldsymbol{X}_i, Y_i, D_i)\}_{i=1}^N)$$

Defn 7.13 (Neural Network).
Generalised nonparametric regression with many 'layers', with components outline in 10. For the i^{th} layer of the network and j^{th} hidden layer of the unit, we have

$$z_j^{[i]} = \boldsymbol{w}_j^{[i]T} \boldsymbol{x} + b_j^{[i]}$$

where w, b, z are the weight (coefficient), bias (intercept) and output respectively.

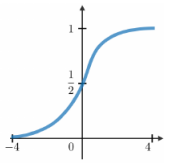
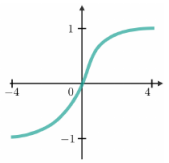
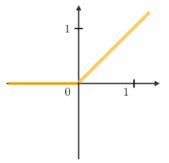
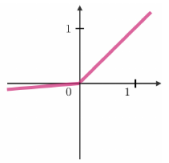
Sigmoid	Tanh	ReLU	Leaky ReLU
$g(z) = \frac{1}{1+e^{-z}}$	$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$g(z) = \max(0,z)$	$g(z) = \max(\epsilon z, z)$ with $\epsilon \ll 1$
			

Figure 11: Activation Functions

Defn 7.14 (Activation Function).
Activation functions are used at the end of a hidden layer to introduce non-linearities into the model. Common ones are

Neural networks frequently use the **cross-entropy** loss function.

Fact 7.5 (Fitting Neural Networks).
Learning rate is denoted by η, which is the pace at which the weights get updated. This can be fixed or adaptively changed using **ADAM**.
Back-propagation is a method to update the weights in the neural net by taking into account the actual output and desired output. The derivative with respect to weight $\boldsymbol{w}$ is computed using the chain rule and is of the following form

$$\frac{\partial L(z,y)}{\partial w} = \frac{\partial L(z,y)}{\partial a} \frac{\partial a}{\partial z} \frac{\partial z}{\partial w}$$

So the weight is updated

$$w \leftarrow w - \eta \frac{\partial L(z,y)}{\partial w}$$

1. Take a batch of training data
2. Perform forward propagation to compute corresponding loss
3. Perform back propagation to compute gradients
4. Use the gradients to update the weights over the network

7.2 Unsupervised Learning

There is no distinction between a label/outcome y_i and predictor $\mathbf{X}_i$ in a wide variety of problems. The goal of unsupervised methods is to characterise the joint

distriution of the data $\mathbf{X}$ using latent factors, clusters, etc.

Defn 7.15 (Principal Components Analysis).
Original data x_i in $\mathbb{R}^k$. We approximate orthogonal unit vectors $w_l \in \mathbb{R}^k$ and associated scores $[L \leq k$ weights $z_{il}]$ to minimise reconstruction error

$$J(\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i - \sum_{l=1}^{L} z_{il}\boldsymbol{w}_l\right\|^2$$

where $\widehat{\boldsymbol{x}}_i = \mathbf{W}\boldsymbol{z_i}$ subject to the constraint that the smoother matrix $\mathbf{W}$ is orthonormal. Equivalently, the objective function can be written as
$J(\mathbf{W}, \mathbf{Z}) = \left\|\mathbf{X} - \mathbf{W}\mathbf{Z}^T\right\|_{\text{Fr}}$ where $\mathbf{Z}$ is $N \times L$ with z_i in its rows.
The optimal solution sets each $\boldsymbol{w}_l$ to be the l-th eigenvector of the empirical covariance matrix. Equivalently, $\widehat{\mathbf{W}} = \mathbf{V}_L$, which contains the L eigenvectors with the largest eigenvalues of empirical covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$.

Defn 7.16 (Truncated SVD).
If we rank singular values of the data matrix $\mathbf{X}$, we can construct a rank L approximation, the **truncated SVD**

$$\mathbf{X} \approx \mathbf{U}_{:,1:L}\mathbf{S}_{1:L,1:L}\mathbf{V}'_{:,1:L}$$

This is **identical** to the optimal reconstruction $\widehat{X} = \mathbf{Z}\widehat{\mathbf{W}}'$.

Fact 7.6 (Dimension selection for PCA).

$$\sum_{j=L+1}^{J} \lambda_l = \textbf{error}(L)$$

The error is the sum of remaining eigenvalues of the covariance matrix. Total variance explained = (sum of included eigenvalues)/(sum of all eigenvalues)

8 Bayesian Statistics

8.1 Setup

Notation: per the Murphy textbook, some statements use notation $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ as shorthand for **data**.

Theorem 8.1 (Bayes Theorem).

$$\underbrace{f(\boldsymbol{\theta}|\boldsymbol{X})}_{\text{posterior}} = \frac{f(\boldsymbol{X}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\boldsymbol{X}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto \underbrace{\mathcal{L}(\boldsymbol{\theta})}_{\text{likelihood}}\ \underbrace{f(\boldsymbol{\theta})}_{\text{prior}}$$

Example 8.2 (Bayesian Updating Steps).

1. Use Bayes Rule to come up with a posterior probability of some hypothesis H_u given event E. Your prior is $P(H_u)$.

$$P(H_u|E) = \frac{P(E|H_u)P(H_u)}{P(E|H_u)P(H_u) + P(E|H_u^c)P(H_u^c)}$$

   Call the posterior probability $P(H_u')$.

2. Given second event E', use posterior probability from step 1 as your prior in the second update step.

$$P(H_u|E') = \frac{P(E'|H_u)P(H_u')}{P(E'|H_u)P(H_u') + P(E'|H_u^c)P((H_u')^c)}$$

Defn 8.1 (Exchangeability).
A sequence of random variables $y_1, \ldots, y_n$ is finitely exchangeable if their joint density remains the same under any re-ordering or re-labeling of the indices of the data.

$$p(y_1, \ldots, y_n) = p(y_{z(1)}, \ldots, y_{z(n)})$$

Exchangeability justifies use of the prior: If the data are exchangeable, then there is a parameter θ that drive the stochastic model generating the data and there exists a density over θ that does not depend on the data itself. The data are conditionally i.i.d., given the prior θ.

Independence vs. Exchangeability: Independence is a stronger condition than exchangeability (it is a special case of exchangeability). Exchangeability only requires that the marginal distribution of each random variable is the same, i.e. $p(y_1) = p(y_2)$. Independence requires that $p(y_1|y_2) = p(y_1)$. As a result, you can

have exchangeability in situations where you do not have independence, most notably sampling without replacement. If the marginal probabilities are unknown, then we only have exchangeability (not independence) even if the samples are drawn with replacement, due to the possibility that there is only one unit with a particular value of y.

Fact 8.3 (Posterior Quantities of Interest).
With the full posterior, one can compute **Posterior Mean, median, and mode** (the latter is sometimes called the Maximum A Posteriori estimate).
One can also compute

Defn 8.2 (Highest Posterior Density Region $R(\theta)$).
, which is a region such that the the parameter lies in the region with probability $1-\alpha$

$$1-\alpha = \mathbf{Pr}\left(\theta \in R(\theta)|y\right) = \int_{R(\theta)} p(\theta|y)d\theta$$

Defn 8.3 (Posterior Predictive Density).
Consider out-of-sample prediction for a single observation $\tilde{y}$. The posterior predictive density is

$$p(\tilde{y}|y_1,\ldots,y_n) = \int_{-\infty}^{\infty} p(\tilde{y}|\theta,y_1,\ldots,y_n)p(\theta|y_1,\ldots,y_n)d\theta$$

Because $\tilde{y}$ is independent of y conditional on θ (exchangeability), we can simplify this as

$$\begin{aligned} p(\tilde{y}|y_1,\ldots,y_n) &= \int_{-\infty}^{\infty} p(\tilde{y}|\theta,y_1,\ldots,y_n)p(\theta|y_1,\ldots,y_n)d\theta \\ &= \int_{-\infty}^{\infty} p(\tilde{y}|\theta)p(\theta|y_1,\ldots,y_n)d\theta \end{aligned}$$

This is just the data density for y multiplied by the posterior density for θ.

Example 8.4 (Posterior predictive density for Bernoulli trial).
Consider $\tilde{y} \sim$ Bernoulli(θ). The posterior predictive density is

$$\begin{aligned} p(\tilde{y}|y) &= \int_0^1 p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &= \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}}p(\theta|y)d\theta \end{aligned}$$

So if we want to know the posterior predictive probability $p(\tilde{y}=1|\theta)$, we can compute it as

$$\begin{aligned} p(\tilde{y}=1|\theta) &= \int_0^1 \theta p(\theta|y)d\theta \\ &= E[\theta|y] \end{aligned}$$

which is the posterior mean.

Defn 8.4 (Uninformative Prior vs. Informative Prior).
An uninformative prior on θ produces a posterior density that is proportional to the likelihood (differing only by the constant of proportionality). This implies that the mode of the posterior density is the θ that maximizes the likelihood function. An informative prior on θ yields a posterior mean that is a precision-weighted average of the prior mean and the MLE.

Stan dev team recommendations: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Theorem 8.5 (Bernstein-Von Mises Theorem / Bayesian CLT).

$$\theta|y \sim^a \mathcal{N}\left(\hat{\theta}, \mathcal{I}(\hat{\theta})^{-1}\right)$$

As $N \to \infty$, the likelihood component of the posterior becomes dominant and as a result frequentist and bayesian inferences will be based on the same limiting multivariate normal distribution.

Defn 8.5 (Bayesian Model Selection).
To choose between Bayesian models, we compute the posterior over models

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m,\mathcal{D})}$$

which allows us to pick the MAP model $\widehat{m} = \arg\max p(m|\mathcal{D})$. If we use a uniform prior over models $p(m) \propto 1$, this amounts to picking the model wich maximises

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

which is called the **marginal likelihood / integrated likelihood / evidence** for model m.

8.2 Conjugate Priors and Updating

Defn 8.6 (Improper Priors).
Priors of the form $f(\theta) \propto c\ c > 0$ are improper because $\int f(\theta)d\theta = \infty$. Improper

priors generally not a problem as long as resulting posterior is well defined.

Fact 8.6 (Flat priors are not invariant).
Suppose $X \sim \text{Bernoulli}(p)$, and we choose prior $f(p) = 1$. Define transformation $\psi = \log(p/(1-p))$. Resulting distribution of ψ is $f_\psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}$

Defn 8.7 (Jeffreys' Prior).
Method of constructing invariant priors.
$f(\theta) \propto I(\theta)^{1/2}$. For multiparameter model, $f(\theta) \propto |I(\theta)|^{1/2}$

Example 8.7 (Jeffreys' Prior).
Given $\gamma = h(\theta)$, $\frac{\partial \mathcal{L}}{\partial \gamma} = \frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \theta}{\partial \gamma}$ and

$$\frac{\partial^2 \mathcal{L}}{\partial \gamma^2} = \frac{\partial^2 \mathcal{L}}{\partial \theta^2}\left(\frac{\partial \theta}{\partial \gamma}\right)^2 + \frac{\partial \mathcal{L}}{\partial \theta}\frac{\partial^2 \theta}{\partial \gamma^2}$$

Taking expectations wrt sample density sends second piece to zero (since $\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial \theta}\right] = 0$), so

$$\mathcal{I}(\gamma) = \mathcal{I}(\theta)\left(\frac{\partial \theta}{\partial \gamma}\right)^2 \implies |\mathcal{I}(\gamma)|^{1/2} = |\mathcal{I}(\theta)|^{1/2}\left|\frac{\partial \theta}{\partial \gamma}\right|$$

Defn 8.8 (Conjugate Prior).
Analytically tractable expressions for the posterior are derived when sample and prior densities form a **natural conjugate pair**, defined as having the property that sample, prior, and posterior densities all lie in the same class of densities.
Exponential family is essentially the only class of densities to have natural conjugate priors.
A one parameter member of the exponential family has density for N obs that can be expressed as

$$\mathcal{L}(y|\theta) = \prod \exp\left((a(\theta)) + b(y) + c\left(\theta\right)u\left(y\right)\right) \propto \exp\left(Na(\theta) + c(\theta)\sum_i u(y)\right)$$

Example 8.8 (Beta-Binomial Updating).
Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, and we take prior $f(p) = 1$. By Bayes thm, the posterior is of the form

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s(1-p)^{n-s} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Instead we take $f(p) = \text{Beta}(\alpha, \beta)$. Uniform prior is a special case with $\alpha = \beta = 1$.
In general, the posterior is of the form

$$\begin{aligned} f(p|x^n) &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)}p^{\sum x_i}(1-p)^{n-\sum x_i} \\ p|x^n &\sim \text{Beta}\left(\sum x_i + 1, n - \sum x_i + 1\right) \\ &\sim \text{Beta}\left(\alpha', \beta'\right) \end{aligned}$$

```
rbeta(n, shape1, shape2)
```

Quantity	Formula
Posterior Mean	$\frac{\alpha'}{\alpha'+\beta'} = \frac{\sum x_i + 1}{\sum x_i + 1 + n - \sum x_i + 1} = \frac{\sum x_i + 1}{n+2}$
Posterior mode	$\frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{\sum x_i}{n}$
Posterior variance	$\frac{\alpha'\beta'}{(\alpha'+\beta')^2(\alpha'+\beta'+1)}$

Posterior predictive distribution	$\sim$ Beta-Binomial(n, a, b)
	$\sim$ Beta-Binomial$(n, \alpha + \sum x_i, \beta + n - \sum x_i)$
	`library(extraDistr)`
	`rbbinom(n, size, alpha, beta)`

Example 8.9 (How to find Beta hyper-parameters using a prior proportion and variance).
Suppose we have a proportion from a previous study θ_0, with variance $V(\theta_0; \alpha, \beta)$. Then we can create a constant

$$\gamma = \frac{\theta_0(1-\theta_0)}{V(\theta_0; \alpha, \beta)} - 1$$

And compute the hyper-parameters α and β for our prior distribution as

$$\begin{aligned} \alpha &= \gamma\theta_0 \\ \beta &= \gamma(1-\theta_0) \end{aligned}$$

Surprisingly this works! See Jackman p.55 for a worked out example.

Example 8.10 (Gamma-Poisson Updating).
Let $Y_1, \ldots, Y_n \sim \text{Poisson}(\lambda)$. This means that

$$\begin{aligned} p(\boldsymbol{y}|\lambda) &= \prod_{i=1}^{N} \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \\ &\propto \lambda^{\sum y_i} \exp(-n\lambda) \end{aligned}$$

We specifiy a Gamma prior on λ, which has density

$$p(\lambda; a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$$

So then the posterior for λ is

$$\begin{aligned} p(\lambda|\boldsymbol{y}) &\propto p(\lambda)p(\boldsymbol{y}|\lambda) \\ &\propto \lambda^{a-1} \exp(-b\lambda)\lambda^{\sum y_i} \exp(-n\lambda) \\ &= \lambda^{\sum y_i + a - 1} \exp(-\lambda(b+n)) \\ &\sim \text{Gamma}(\sum y_i + a, b+n) \\ &\sim \text{Gamma}(a', b') \\ &\texttt{rgamma(n, shape, rate)} \end{aligned}$$

A flat prior is $a = b = 0$.

Quantity	Formula
Posterior Mean	$\frac{a'}{b'} = \frac{\sum y_i + a}{b+n}$
Posterior mode	$\frac{a'-1}{b'} = \frac{\sum y_i + a - 1}{b+n}$
Posterior variance	$\frac{a'}{(b')^2} = \frac{\sum y_i + a}{(b+n)^2}$
Posterior predictive distribution	$\sim$ Negative Binomial(y, θ) $\sim$ Negative Binomial$(a, 1 - \frac{1}{b+1})$ `rnbinom(rnbinom(n, size, prob)`

Example 8.11 (Dirichlet-Multinomial Updating).

$$\begin{aligned} p(\theta|\alpha_1, \ldots \alpha_k) &\propto \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \\ p(y|\theta) &\propto \prod_{j=1}^{K} \theta_j^{y_j} \end{aligned}$$

where y_j is the count of observations in category j. For 3 categories, the posterior is:

$$\begin{aligned} p(\theta_1, \theta_2, 1-\theta_1-\theta_2|y) &\propto \theta_1^{\alpha_1 + y_1 - 1} \theta_2^{\alpha_2 + y_2 - 1} (1-\theta_1-\theta_2)^{\alpha_3 + y_3 - 1} \\ &\sim \text{Dirichlet}(\alpha_1 + y_1, \alpha_2 + y_2, \alpha_3 + y_3) \end{aligned}$$

Example 8.12 (Normal-Normal updating).
$y \sim \mathcal{N}(\mu, \sigma^2)$., where σ^2 is known but mean μ is not known. The joint density of y is

$$\mathcal{L}(y|\theta) = \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{N}{2\sigma^2}(\bar{y} - \theta)^2\right)$$

Given a normal prior $\theta \sim \mathcal{N}(\mu, \tau^2) \implies f(\mu) \propto \exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right)$, we can write the posterior density of the form

$$f(\theta|y) \propto \exp\left(-\frac{N}{2\sigma^2}(\theta - \bar{y})^2\right) \exp\left(\frac{(\theta-\mu)^2}{2\tau^2}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{(\theta-\mu_1)^2}{\tau_1^2}\right)\right)$$

where $\mu_1 = \tau_1^2(N\bar{y}/\sigma^2 + \mu\tau^2)$ and $\tau_1^2 = (N/\sigma^2 + 1/\tau^2)^{-1}$. Posterior mean is a weighted sum of prior mean μ and sample mean $\bar{y}$ with weights that reflect the precision of the likelihood via $N\sigma^2$ and prior τ^2.

Three cases (ref. Jackman p.80-94):

1. Variance known, mean unknown. Model:

$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ y &\sim \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

Quantity	Formula
Posterior Mean	$\frac{\mu_0(\frac{1}{\sigma_0^2}) + \bar{y}(\frac{n}{\sigma^2})}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \bar{y}\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}$
Posterior variance	$\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} = \frac{\sigma_0^2\sigma^2}{\sigma^2 + n\sigma_0^2}$
Posterior predictive distribution	$\sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ $\tilde{\mu} = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n}$ $\tilde{\sigma}^2 = \sigma^2 + \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$

2. Variance and mean both unknown. Prior densities:

$$
\begin{aligned}
p(\mu, \sigma^2) &= p(\mu|\sigma^2)p(\sigma^2) \\
p(\mu|\sigma^2) &\sim \text{Normal}(\mu_0, \sigma^2/n_0) \\
p(\sigma^2) &\sim \text{Scaled-Invese-}\chi^2(\nu_0/2, \nu_0\sigma_0^2/2)
\end{aligned}
$$

Conditional posterior densities:

$$
\begin{aligned}
\mu|\sigma^2, y &\sim \mathcal{N}(\mu_1, \sigma^2/n_1) \\
\sigma^2|y &\sim \text{Scaled-Invese-}\chi^2(\nu_1/2, \nu_1\sigma_1^2/2)
\end{aligned}
$$

where

$$
\begin{aligned}
n_1 &= n_0 + n \\
\mu_1 &= \frac{n_0\mu_0 + n\bar{y}}{n_1} \\
\nu_1 &= \nu_0 + n \\
\nu_1\sigma_1^2 &= v_0\sigma_0^2 + \sum_{i=1}^{N}(y_i - \bar{y})^2 + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{y})^2
\end{aligned}
$$

Marginal posterior density of μ:

$$
\begin{aligned}
p(\mu) &\sim \text{Student-T}(\mu_1, \sqrt{\sigma_1^2/n_1}, v_1) \\
&\sim \texttt{brms::rstudent\_t(n, df, mu = 0, sigma = 1)}
\end{aligned}
$$

where

$$
\begin{aligned}
n_1 &= n_0 + n \\
\mu_1 &= \frac{n_0\mu_0 + n\bar{y}}{n_1} \\
\nu_1 &= \nu_0 + n \\
\sigma_1^2 &= S_1/\nu_1 \\
S_1 &= \nu_0\sigma_0^2 + (n-1)\sum_{i=1}^{N}(y_i - \bar{y})^2 + \frac{n_0 n}{n_1}(\bar{y} - \mu_0)^2
\end{aligned}
$$

Posterior predictive distribution for $\tilde{y}$:

$$
p(\tilde{y}|y) \sim \text{Student-T}(\mu_1, \sigma_1\sqrt{(n_1+1)/n_1}, v_1)
$$

where

$$
\begin{aligned}
n_1 &= n_0 + n \\
\mu_1 &= \frac{n_0\mu_0 + n\bar{y}}{n_1} \\
\nu_1 &= \nu_0 + n \\
\sigma_1^2 &= S_1/\nu_1 \\
S_1 &= \nu_0\sigma_0^2 + (n-1)\sum_{i=1}^{N}(y_i - \bar{y})^2 + \frac{n_0 n}{n_1}(\bar{y} - \mu_0)^2
\end{aligned}
$$

3. Improper reference prior. Prior densities:

$$
p(\mu, \sigma^2) \propto 1/\sigma^2
$$

Posterior densities

$$\mu|\sigma^2, y \sim \mathcal{N}(\bar{y}, \sigma^2/n)$$

$$\sigma^2|y \sim \text{Scaled-Invese-}\chi^2\left(\frac{n-1}{2}, \frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{2}\right)$$

which implies

$$\frac{\mu-\bar{y}}{\sqrt{S/((n-1)n)}} \sim t_{n-1}$$

Posterior predictive distribution

$$p(\tilde{y}|y) \sim \text{Student-T}(\bar{y}, s\sqrt{\frac{n+1}{n}}, n-1)$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y-\bar{y})^2$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Fact 8.13 (Simulation from Posterior).
Posterior can often be approximated by simulation.

- Draw $\theta_1, \ldots, \theta_B \sim p(\theta|x^n)$
- Histogram of $\theta_1, \ldots, \theta_B$ approximates posterior density $p(\theta|x^n)$

Methods for this: Markov-Chain Monte-Carlo, Metropolis-Hastings, Hamiltonian Monte-Carlo

Fact 8.14 (Conjugacy for Discrete Distributions).

Likelihood	Conjugate prior	Posterior hyperparameters
Bern (p)	Beta (α, β)	$\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i$
Bin (p)	Beta (α, β)	$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$
NBin (p)	Beta (α, β)	$\alpha + rn, \beta + \sum_{i=1}^{n} x_i$
Po (λ)	Gamma (α, β)	$\alpha + \sum_{i=1}^{n} x_i, \beta + n$
Multinomial(p)	Dir (α)	$\alpha + \sum_{i=1}^{n} x^{(i)}$

Fact 8.15 (Conjugacy for Continuous Distributions).

Likelihood	Conjugate prior	Posterior hyperparameters
Unif $(0, \theta)$	Pareto(x_m, k)	$\max\{x_{(n)}, x_m\}, k+n$
Exp (λ)	Gamma (α, β)	$\alpha + n, \beta + \sum_{i=1}^{n} x_i$
$\mathcal{N}(\mu, \sigma_c^2)$	$\mathcal{N}(\mu_0, \sigma_0^2)$	$\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma_c^2}\right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_c^2}\right)$, $\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_c^2}\right)^{-1}$
$\mathcal{N}(\mu_c, \sigma^2)$	Scaled Inverse Chi-square(ν, σ_0^2)	$\nu + n, \frac{\nu\sigma_0^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\nu+n}$
$\mathcal{N}(\mu, \sigma^2)$	Normal-scaled Inverse Gamma$(\lambda, \nu, \alpha, \beta)$	$\frac{\nu\lambda + n\bar{x}}{\nu+n}$, $\nu + n$, $\alpha + \frac{n}{2}$, $\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{\gamma(\bar{x}-\lambda)^2}{2(n+\gamma)}$

8.3 Computation / Markov Chains

Defn 8.9 (Stochastic Process $\{X_t : t \in T\}$).
is a collection of random variables.

Defn 8.10 (Markov Chains).
The process $\{X_n : n \in T\}$ is a Markov Chain if

$$\mathbf{Pr}\left(X_n = X | X_0, \ldots, X_{n-1}\right) = \mathbf{Pr}\left(X_n = x | X_{n-1}\right)$$

Defn 8.11 (Monte Carlo Integration).
First rewrite the integral to be evaluated $I = \int_a^b h(x)dx$ as follows

$$I = \int_a^b h(x)dx = \int_a^b w(x)f(x)dx$$

where $w(x) = h(x)(b-a)$ and $f(x) = \frac{1}{b-1}$. Since f is the probability density for a uniform r.v. over (a, b), we can write $I = \mathbb{E}_f\left[w(X)\right]$ where $X \sim \mathsf{U}\left[a, b\right]$. If we generate $X_1, \ldots, X_N \sim \mathsf{U}\left[a, b\right]$, by LLN,

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} w(X_i) \xrightarrow{p} \mathbb{E}\left[w(X)\right] = I$$

Defn 8.12 (Reversibility / Detailed Balance).
The goal is the generate sequences $\boldsymbol{\theta}_1, \boldsymbol{\theta_2}, \ldots$ from $f(\boldsymbol{\theta}|\boldsymbol{y})$. An MCMC scheme will generate samples from the conditional if

$$P(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})f(\boldsymbol{\theta}_{j-1}|\boldsymbol{y}) = P(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}_j)f(\boldsymbol{\theta}_j|\boldsymbol{y})$$

where $P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)$ is the pdf of $\boldsymbol{\theta}_i$ given $\boldsymbol{\theta}_j$. The LHS is the joint pdf of $\boldsymbol{\theta}_j, \boldsymbol{\theta}_{j-1}$ from the chain, **if $\boldsymbol{\theta}_{j-1}$ is from** $f(\boldsymbol{\theta}|\boldsymbol{y})$. Integrating RHS over $d\boldsymbol{\theta}_{j-1}$ yields $f(\boldsymbol{\theta}_j|\boldsymbol{y})$, so the result states that given $\boldsymbol{\theta}_{j-1}$ is from the correct posterior distribution, the chain generates $\boldsymbol{\theta}_j$ also from the posterior $f(\boldsymbol{\theta}|\boldsymbol{y})$.

Defn 8.13 (Gibbs Sampling).
Basic idea - turn high dimensional problem into several one-dimensional problems. Suppose (X, Y) has joint density $f_{X,Y}(x, y)$. Suppose it is possible to simulate from conditional distributions $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Let (X_0, Y_0) be starting values. Assuming we have drawn $(X_0, Y_0), \ldots, (X_i, Y_i)$, we generate (X_{i+1}, Y_{i+1}) as follows

- $X_{n+1} \sim f_{X|Y}(x|Y_n)$
- $Y_{n+1} \sim f_{Y|X}(y|X_{n+1})$
- $\ldots$ for multiple parameters

Example 8.16 (Gibbs for Univariate Normal).
. Let $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$. Define precision $\tau = 1/\sigma^2$

- **Likelihood**: $\mathsf{f}\left(y|\mu, \tau\right) \sim \tau^{n/2}\exp\left(\frac{1}{2}\tau\sum_{i=1}^{n}(y_i - \mu)^2\right)$
- **(Noninformative) Prior**: $\pi\mu, \tau \sim \tau$

Posterior Distribution

$$\pi\left(\mu, \tau|y\right) \sim \tau^{(n/2)+1}\exp\left(-\frac{1}{2}\tau\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

full conditionals:

- $\pi(\mu|\tau, y) = \mathcal{N}\left(\bar{y}, (n\tau)^{-1}\right)$
- $\pi(\tau|\mu, y) = \Gamma\left(\frac{n}{2}, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2\right)$

However, it is typically impossible to write out or sample from full-conditionals.

Defn 8.14 (Metropolis Algorithm).
Let $q(y|x)$ be an arbitrary, friendly distribution we can sample from. The conditional density $q(y|x)$ is called the proposal distribution. MH creates a sequence of observations $X_0, \ldots$ as follows Choose X_0 arbitrarily. Suppose we have generated $X_0, X_1, \ldots, X_i$. Generate X_{i+1} as follows

- Generate **proposal** $Y \sim q(y|X_i)$
- Evaluate $r := r(X_i, Y)$ where

$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$$

- Set

$$X_{i+1} = \begin{cases} Y & \text{w.p } r \\ X_i & \text{w.p } 1 - r \end{cases}$$

Defn 8.15 (Expectation Maximisation).
Let $\boldsymbol{x}_i$ be observed and $\boldsymbol{z}_i$ missing. The goal is to maximise the log-likelihood of the observed data

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log p(\boldsymbol{x}_i|\theta) = \sum_{i=1}^{n}\left[\sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})\right]$$

Cannot push log inside the sum because of unobserved variables. EM tackles the problem as follows. Define **complete data log likelihood** as $\ell_c(\boldsymbol{\theta}) := \sum_{i=1}^{n}\log p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})$. This cannot be computed, since $\boldsymbol{z}_i$ is unknown.
Instead, define $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \mathbb{E}\left[\ell_c(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\theta}^{t-1})\right]$ where t is the iteration number and Q is called the **auxiliary function**.

- **Expectation (E) Step**: Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$, which is an expectation wrt old params $\boldsymbol{\theta}^{t-1}$.
- **Maximisation (M) Step**: Optimise the Q function wrt $\boldsymbol{\theta}$.
  Compute $\boldsymbol{\theta}^t = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$. For MAP estimation, $\boldsymbol{\theta}^t = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) + \log p(\boldsymbol{\theta})$

Example 8.17 (EM for probit regression).
Probit has the form $p(y_i = 1|z_i) = \mathbb{1}_{z_i>0}$ where $z_i \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ is the latent variable. The complete data log likelihood, assuming a $\mathcal{N}(0, \boldsymbol{\Sigma}_0)$ prior on $\boldsymbol{\beta}$.

$$\begin{aligned}\ell(\boldsymbol{z}, \boldsymbol{\beta}|\boldsymbol{\Sigma}_0) &= \log p(\boldsymbol{y}|\boldsymbol{z}) + \log \mathcal{N}(\boldsymbol{z}|\mathbf{X}\boldsymbol{\beta}, \mathbf{I}) + \log \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\Sigma}_0) \\ &= \sum_{i=1}^{n} \log p(y_i|z_i) - \frac{1}{2}(\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'(\boldsymbol{\Sigma}_0)^{-1}\boldsymbol{\beta} + \text{const}\end{aligned}$$

The posterior in the E step is a **truncated Gaussian**

$$p(z_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \begin{cases} \mu_i + \frac{\phi(\mu_i)}{\Phi(\mu_i)} & \text{if } y_i = 1 \\ \mu_i - \frac{\phi(\mu_i)}{\Phi(\mu_i)} & \text{if } y_i = 0 \end{cases}$$

where $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. In the M step, we estimate $\boldsymbol{\beta}$ using ridge, where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{z}]$.

$$\hat{\beta} = \left((\boldsymbol{\Sigma}_0)^{-1} + \mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\mu}$$

8.4 Hierarchical Models

Defn 8.16 (Heirarchical Priors).
Parameters in a prior are modeled as having a distribution that depends on **hyperparameters**. This results in joint posteriors of the form

$$f(\theta, \tau|y) \propto \underbrace{\mathcal{L}(y|\theta)}_{\text{likelihood}} \underbrace{f(\theta|\tau)}_{\text{parameter prior}} \underbrace{f(\tau)}_{\text{hyperparameter prior}}$$

Represented by the graphical model $\boldsymbol{\tau} \to \boldsymbol{\theta} \to \mathcal{D}$.
We are typically interested in the marginal posterior of θ, which is obtained by integrating the joint posterior w.r.t τ.

By treating $\boldsymbol{\tau}$ as a latent variable, we allow data-poor observations to borrow strength from data rich ones.

8.4.1 Empirical Bayes

In hierarchical models, we need to compute the posterior on multiple layers of latent variables. For example, for a two-level model, we need

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})$$

We can employ a computational shortcut by approximating the posterior on the hyper-parameters with a point-estimate $p(\boldsymbol{\eta}|\mathcal{D}) \approx \delta_{\hat{\boldsymbol{\eta}}}(\boldsymbol{\eta})$, where $\hat{\boldsymbol{\eta}} = \arg\max p(\boldsymbol{\eta}|\mathcal{D})$. Since $\boldsymbol{\eta}$ is usually much smaller than $\boldsymbol{\theta}$ in dimensionality, we can safely use a uniform prior on $\boldsymbol{\eta}$. Then, the estimate becomes

$$\hat{\eta} = \arg\max p(\mathcal{D}|\boldsymbol{\eta}) = \arg\max \underbrace{\left[\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}\right]}_{\text{marginal likelihood}}$$

This violates the principle that the prior should be chosen independently of the data, but is a cheap computational trick. This produces a hierarchy of Bayesian methods in increasing order of the number of integrals performed.

Example 8.18 (Cancer Rates across cities).
Suppose we measure the number of people in various cities N_i and the number of people who died of cancer x_i. We assume $x_i \sim \text{Bin}(N_i, \theta_i)$ and want to estimate the cancer rates θ_i. The MLE solution would be to either estimate them all separately, or estimate a single θ for all cities.
The hierarchical approach is to model $\theta_i \sim \text{Beta}(a, b)$, and write a joint distribution

$$\begin{aligned}p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{N}) &= p(\boldsymbol{\eta})\prod_{i=1}^{N}\text{Bin}(x_i|N_i, \theta_i)\,\text{Beta}(\theta_i, \boldsymbol{\eta}) \\ &\text{analytically integrate out } \theta_i \\ &= \prod_{i=1}^{N}\int \text{Bin}(x_i|N_i, \theta_i)\,\text{Beta}(\theta_i, \boldsymbol{\eta})\,d\theta_i \\ &= \prod_{i=1}^{N}\frac{\text{Beta}(a + x_i, b + N_i - x_i)}{\text{Beta}(a, b)}\end{aligned}$$

where $\eta := (a, b)$. We can also put covariates on $\theta_i = f(\boldsymbol{x}_i'\beta)$.

8.4.2 Hierarchy of Bayesianity

Method	Definition
Maximum Likelihood	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}\|\boldsymbol{\theta})$
MAP Estimation	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})$
Empirical Bayes	$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})d\boldsymbol{\theta} = \arg\max_{\boldsymbol{\eta}} p(\mathcal{D}\|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} = \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})d\boldsymbol{\theta} = \arg\max_{\boldsymbol{\eta}} p(\mathcal{D}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}\|\mathcal{D}) \propto p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$

8.5 Graphical Models

Defn 8.17 (Chain rule of probability).
Any joint distribution can be represented as follows

$$p(x_{1:v}) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)\dots p(x_v|x_{1:V-1})$$

where V is the number of variables [and we have dropped the parameter vector $\boldsymbol{\theta}$]. It follows that

The joint distribution $p(\boldsymbol{x}) = p(x_1, \dots, x_K)$ can be written as

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|\text{Pa}_k)$$

where Pa_k denotes the *parent nodes* of x_k, which are nodes that have arrows pointing to x_k.

Defn 8.18 (Conditional Independence).
X and Y are said to be conditionally independent iff the conditional joint can be written as the product of the conditional marginal.

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow p(X, Y|Z) = p(X|Z)p(Y|Z)$$

Defn 8.19 (d-separation).
If the condition $\mathcal{A} \perp\!\!\!\perp \mathcal{B}|\mathcal{C}$, it must be the case that **all paths are blocked**. All paths are blocked iff

- Arrows on the path meet either head to tail or tail to tail at the node, and the note is in the set $\mathcal{C}$
- Arrows meet head to head at the node, and neither the node nor any of its descendents is in the set $\mathcal{C}$

8.5.1 Empirical Bayes

Example 8.19 (James-Stein Estimator for batting averages (Efron and Morris)).
We suppose that each player's MLE value p_i (his batting average in the first 90 tries) is a binomial proportion,

$$p_i \sim \text{Bi}\left(90, P_i\right)/90$$

Here P_i is his true average, how he would perform over an infinite number of tries; TRUTH $_i$ is itself a binomial proportion, taken over an average of 370 more tries per player.
At this point there are two ways to proceed. The simplest uses a normal approximation to (7.17)

$$p_i \dot{\sim} \mathcal{N}\left(P_i, \sigma_0^2\right)$$

where σ_0^2 is the binomial variance

$$\sigma_0^2 = \bar{p}(1-\bar{p})/90$$

with $\bar{p} = 0.254$ the average of the p_i 's. Letting $x_i = p_i/\sigma_0$, applying (7.13), and transforming back to $\hat{p}_i^{\text{JS}} = \sigma_0 \hat{\mu}_i^{\text{JS}}$, gives James-Stein estimates

$$\hat{p}_i^{\text{JS}} = \bar{p} + \left[1 - \frac{(N-3)\sigma_0^2}{\sum\left(p_i - \bar{p}\right)^2}\right](p_i - \bar{p})$$

A second approach begins with the arcsin transformation

$$x_i = 2(n+0.5)^{1/2} \sin^{-1}\left[\left(\frac{np_i + 0.375}{n+0.75}\right)^{1/2}\right]$$

9 Dependent Data: Time series and spatial statistics

9.1 Time Series

A time series is a sequence of data points $\{w_t\}_{t=1}^{T}$ observed over time. In a random sample, points are iid, so the joint distribution $f(w_1, \dots, w_T) = \prod_{t=1}^{T} f(w_t)$. In time series, this is clearly violated, since observations that are temporally close to each other tend to be more similar.

Defn 9.1 (Stochastic Process).
is a sequence of random variables $\{\dots, Y_{-1}, Y_0, Y_1, \dots\}$ that are indexed w.r.t the elements in a set of indices $\{Y_t : t \in \mathcal{T}\}$. Hypothetical repeated realisations of a stochastic process look like

$$\left\{w_t^{(1)}, w_t^{(2)}, \dots, w_t^{(n)}\right\}_{t=-\infty}^{\infty}$$

The index set $\mathcal{T}$ may be either countable, in which case we get a *discrete time process* or an uncountable, in which case we get a *continous time process*.

State Space We assume $\exists$ a set $\mathcal{Y} \in \mathbb{R}$ s.t. $\forall t \in \mathcal{T}, Y_t \in \mathcal{Y}$. Then, $\mathcal{Y}$ is called the **State Space** of the stochastic process.

Defn 9.2 (Martingales).
Consider a random process $\{Y_t\}_{t=1}^{\infty}$ and an increasing sequence of information sets $\{\mathcal{F}_t\}_{t=1}^{\infty}$ i.e. collection of $\sigma-$fields s.t. $\mathcal{F}_0 \subset \mathcal{F}_1 \dots \mathcal{F}_\infty \subset \mathcal{F}$. If Y_t belongs to the information set $\mathcal{F}_t$ and is absolutely integrable [i.e. $y_t \in L_0(\mathcal{F}_t) \cap L_1(\mathcal{F})$], and $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] = Y_t \ \forall t < \infty$ then $\{Y_t\}_{t=0}^{\infty}$ is called a martingale. In words, the conditional expected value of the next observation, given all the past observations, is equal to the most recent observation.

Defn 9.3 (Autocovariance).
The autocovariance of Y_t is the covariance between Y_t and its j^{th} lagged value

$$\gamma_{jt} := \mathbb{E}[Y_t - \mu_t][Y_{t-j} - \mu_{t-j}]$$

the variance covariance matrix of $y = \{y_t\}$ is given by

$$\mathbb{V}[y] = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \cdots & \cdots \\ \vdots & \ddots & \dots & \vdots \\ \gamma_{T-1} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}$$

the j^{th} order correlation coefficient $\rho_j := \gamma_j/\gamma_0$.

Defn 9.4 (Stationarity $\equiv I(0)$).
A random process is said to be stationary if the distribution functions of $(X_{t_1}, X_{t_2} \dots)$ and $(X_{t_1+j}, X_{t_2+j} \dots)$ are the same $\forall t_1, \dots, t_k, h \in \mathbb{Z}$.
A process is said to be covariance (or *weakly*) stationary if

1. $\mathbb{E}[Y_t] = \mu \ \forall \ t \in T$
2. $\gamma_{jt} = \mathbb{E}[Y_t - \mu_t][Y_{t-j} - \mu_{t-j}] = \gamma_j \ \forall t \in T$

i.e. neither the mean nor the autocovariances depend on the date t; stationary expectation, variance, and covariance. Most relevant variables aren't stationary, but their *detrended* or *first-differenced* versions may be.

Defn 9.5 (Markov Process).
If $X_0, X_1, \dots$ is a **Markov Process**,

$$\mathbf{Pr}(X_{n+1} \le x | X_1, \dots, X_n) = \mathbf{Pr}(X_{n+1} \le x | X_n)$$

that is, the conditional distribution of X_{n+1} given $X_1, \dots, X_n$ does not depend on $X_1, \dots, X_{n-1}$.

Markov Chain A Markov chain is simply a **Markov process in which the state-space is a countable set**. Since a Markov chain is a markov process, the conditional distribution of $X_{t+1}|X_1, \dots, X_t$ depends only on X_t. The conditional distribution is often represented by a **Transition matrix** where

$$\mathbf{P}_{ij}^{t,t+1} = \mathbf{Pr}(X_{t+1} = j | X_t = i) \quad ; i, j = 1, \dots J$$

If $\mathbf{P}$ is the same $\forall t$, we say the Markov chain has *stationary* transition probabilities.

Defn 9.6 (Ergodic Processes).
A stationary process is *ergodic* if any two variables positioned far apart in the sequence are *almost independently distributed*.
$\{x_t\}$ is ergodic if, for any two bounded functions $f(.)$ in $k+1$ variables and $g(.)$ in $l+1$ variables,

$$\lim_{N\to\infty} |\mathbb{E}[f(x_t, \dots, x_{t+k}) g(x_{t+N}, \dots, x_{t+l+N})]| - |\mathbb{E}[f(x_t, \dots, x_{t+k})]| \, |\mathbb{E}[g(x_{t+N}, \dots, x_{t+l+N})]| = 0$$

i.e. $\lim_{j\to\infty} \gamma_j = 0$
Sufficient condition for ergodicity is x_t be **covariance stationary** and $\sum_{j=0}^{\infty} |\gamma_j| < \infty$
Ergodic processes have the following property

$$\mathbb{V}\left[\sum_{t=1}^{T} x_t\right] = \sum_{j=1-T}^{T-1} (T - |j|)\gamma_j$$

this result implies that

$$\lim_{T\to\infty} \mathbb{V}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right] = \sum_{j=-\infty}^{\infty} \gamma_j < \infty$$

This permits us to swap is for ts and derive Asymptotic theory with dependent observations, such as LLN and CLT.

Defn 9.7 (Brownian Motion).
A family of r.v.s $\{X_t\}$ indexed by a *continuous* variable t over $[0, \infty)$ is a Brownian Motion iff

1. $X(0) = 0$
2. $\{X(s_i + t_i) - X_(S_i)\}$ over an arbitrary collection of disjoint intervals $(s_i, s_i + t_i$ are independent r.v.s
3. $\forall\ s, t \geq 0, X(s+t) - X(s) \sim \mathcal{N}(0, t)$

Defn 9.8 (White Noise).
White noise is a sequence $\{\varepsilon_t\}$ whose elements have mean zero and variance σ^2, and for which ε_t's are uncorrelated over time

1. $\mathbb{E}[\varepsilon_t] = 0$
2. $\mathbb{E}\left[\varepsilon_t^2\right] = \sigma^2$
3. $\mathbb{E}[\varepsilon_t \varepsilon_{t-j}] = 0\ \forall\ j \neq 0$

Defn 9.9 (Moving average : MA(q)).
A moving average of order q, $MA(q)$ is a weighted average of the q most recent values of a white noise defined as

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q + \varepsilon_{t-q}$$

Defn 9.10 (Autoregressive : AR(p)).
An autoregressive process of order p, $AR(p)$ is given by Y_t as a linear combination of p lags of itself and one white noise

$$Y_t = \mu + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

Defn 9.11 (Autoregressive Moving Average: ARMA(p, q)).
ARMA(p, q) combines AR(p) and MA(q)

$$Y_t = \mu + \underbrace{\varepsilon_t + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q + \varepsilon_{t-q}}_{\text{MA(q)}} + \underbrace{\phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}}_{\text{AR(p)}}$$

Theorem 9.1 (Wold Theorem).
Consider AR(1): $Y_t = \rho Y_{t-1} + \varepsilon_t$. Since this holds at t, it holds at $t-1 \implies Y_{t-1} = \rho Y_{t-2} + \varepsilon_{t-1}$. Substitute into original to get $Y_t = \rho(\rho Y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$. Repeat ad infinitum to obtain, as long as $\rho < 1$

$$Y_t = \sum_{s=0}^{\infty} \rho^s \varepsilon_{t-s}$$

In other words, AR(1) $\equiv$ MA(∞) ; they are different representations of the same underlying stochastic process.
Wold Representation: All covariance-stationary time series processes can be represented by / decomposed into a deterministic component and a $MA(\infty)$

Defn 9.12 (Time Trends).
In a stationary process, $\mathbb{E}[x_t] = \mu$, which is seldom true. A less restrictive assumption that allows for nonstationarity is to specify the mean as a function of time.

$$\mathbb{E}[x_t] = \alpha + \beta t \ \text{ specify } x_t = \alpha + \beta t + \varepsilon_t; \varepsilon \text{ stationary}$$

Defn 9.13 (Random walk $\equiv I(1)$).
is a a process such that $\mathbb{E}[x_t | x_{t-1}, x_{t-2}, \dots] = x_{t-1}$.
$x_t = x_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ = AR(1) process with $\phi = 1 =:$ **Unit Root**. Rewrite as

$$x_t = x_{t-1} + \varepsilon_t = x_0 + \sum_{j}^{t} \varepsilon_j$$

Random walk with drift

$$x_t = x_{t-1} + \delta + \varepsilon_t = \delta t + \varepsilon_t + \varepsilon_{t-1} \dots \varepsilon_1 + x_0 = x_0 + \delta t + \sum_{j}^{t} \varepsilon_j$$

Defn 9.14 (Unit Root Tests).
For the following model $x_t = \rho x_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ = AR(1)
test $\rho = 1$. Distribution of $\widehat{\rho}$ under the null $\rho = 1$ is **non-standard**: CLT not valid.
test to use: Dickey Fuller, Augmented Dickey Fuller, Phillips-Perron.

Defn 9.15 (Cointigration).
Let $y_t \sim I(1) \wedge x_t \sim I(1)$. y_t and x_t are said to be cointegrated if $\exists \psi$ s.t. $y_t - \psi x_t \sim I(0)$. For example, let

$$y_t = \beta x_t + \varepsilon_t$$
$$x_t = x_{t-1} + \upsilon_t$$

where $(\varepsilon_t, \upsilon_t)$ is white noise. Then, $y_t, x_t \sim I(1)$, but $y_t - \beta x_t \sim I(0)$, with cointegration vector $a = (1, -\beta)$.

Defn 9.16 (Hodrick-Prescott (HP) filter).
decomposes an observed time series $X_t, t = 1, 2, \ldots, n$ into a trend $\overline{X}_t$ and a stationary component $\widetilde{X}_t = X_t - \overline{X}_t$ so that the trend $\left\{\overline{X}_t\right\}_{t=1}^n$ minimises

$$\frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X}_t)^2 + \underbrace{w\frac{1}{n}\sum_{t=2}^{n-1}\left((\overline{X}_{t+1} - \overline{X}_t) - (\overline{X}_t - \overline{X}_{t-1})\right)^2}_{\text{Penalty for incorporating fluctuations}}$$

w is a tuning parameter. In quarterly data, $w = 1600$.

9.1.1 Regression with time series

Basic assumption in conventional OLS with time series is $\mathbb{E}[y_t|x_1, \ldots, x_T] = \mathbb{E}[y_t|x_t] = x_t'\beta$. Equivalently, $y_t = x_t'\beta + \varepsilon_t$ $\mathbb{E}[\varepsilon_t|X] = 0$ where $X = (x_1, \ldots, x_T)'$. The second classical assumption is $\mathbb{E}\left[\varepsilon_t^2|x\right] = \sigma^2$ $\forall t; \mathbb{E}[\varepsilon_t\varepsilon_{t-j}] = 0$ $\forall$ t, j.
$\mathbb{E}[u_t, u_{t-j}|X] \neq 0$ is called autocorrelation. Fix: **Newey-West HAC consistent variance estimator** 'meat'

$$\widehat{V} = \widehat{\Gamma}_0 + \sum_{i=1}^{m}\left(1 - \frac{j}{m+1}\right)(\widehat{\Gamma}_j + \widehat{\Gamma}_j') \text{ where } \widehat{\Gamma}_j = \frac{1}{T-j}\sum_{t=j+1}^{T}\widehat{\varepsilon}_j\widehat{\varepsilon}_{t-j}x_t x_{t-j}'$$

with variance estimated the normal way

$$\widehat{W} = \left(\frac{1}{T}\sum_{t=1}^{T}x_t x_t'\right)^{-1}\widehat{V}\left(\frac{1}{T}\sum_{t=1}^{T}x_t x_t'\right)^{-1}$$

Defn 9.17 (Error Correction Mechanism).
Consider the model

$$y_t = \delta + \alpha y_{t-1}\beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t$$

Subtract y_{t-1} and add $-\beta_0 x_{t-1} + \beta_0 x_{t-1}$ to the l.h.s. we get

$$y_t - y_{t-1} = \delta - (1-\alpha)y_{t-1} + \beta_0(x_t - x_{t-1}) + (\beta_0 + \beta_1)x_{t-1} + \varepsilon_t$$

and

$$\Delta y_t = \delta + \beta_0\Delta x_t - (1-\alpha)(y_{t-1} - \gamma x_{t-1}) + \varepsilon_t$$

where γ is the long run effect

$$\gamma = \frac{\beta_0 + \beta_1}{1-\alpha}$$

Defn 9.18 (Testing for trend-breaks - sequential Chow Test).
A Quandt Likelihood ratio test begins with no knowledge of when the trend break occurs [although researchers typically know of the timing for substantive reasons], and sequentially estimates the following model

$$\Delta Y_t = \log Y_t - \log Y_{t-1} = \alpha + \delta_0 D_t(\tau) + \varepsilon_t$$

where ΔY_t is the first difference of the outcome, and $D_t(\tau)$ is an indicator variable equal to zero for all years before τ and one for all subsequent years. The researcher varies τ and tests the null that $\delta_0 = 0$, and the largest F-statistic is used to determine the best possible break point. Use Andrews (2003) critical values to account for multiple-testing.

9.2 Spatial Statistics

Defn 9.19 (Spatial Stochastic Process, Autocorrelation).
A spatial stochastic process is a collection of random variables $y(u)$ indexed by location u: $\left\{y_i, i \in \mathcal{D} \subset \mathbb{R}^d\right\}$, where $\mathcal{D}$ is either a continuous surface of a finite set of discrete locations.
For each location u, $y(u)$ is a random variable, and thus needs to be modeled. Basic approach is to assume $\mathbb{E}[y(u)], \mathbb{V}[y(u)]$ exist, and decompose

$$y(u) = \underbrace{m(u)}_{\text{mean function}} + \underbrace{e(u)}_{\text{error}}$$

mean function $m(u) = \mathbb{E}[y(u)]$ and stochastic error process $e(u)$ s.t. $\mathbb{E}[e(u)] = 0$.

9.2.1 Kriging - modeling $m(u)$

Main reference: Christensen (2019, ch 8)

Defn 9.20 (Universal Kriging).
Assume linear structure for $m(u)$. p known functions of u, $x_1(u), \dots, x_p(u)$ s.t.

$$m(u) = \sum_{j=1}^{p} \beta_j x_j(u)$$

A special case of this is the **Ordinary Kriging** model where

$$m(u) = \mu$$

for unknown μ. The most basic model is **Simple Kriging** where

$$m(u) = \mu_0$$

with known μ_0.

Fact 9.2 (BLP of spatial data: Kriging).
Assume the universal kriging model $m(u) = \sum_j^p \beta_j x_j(u)$ holds, we have data on locations $u_1, \dots, u_n$, and that we wish to predict the value of $y(u_0)$. The model can be written

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \mathbb{E}[e] &= 0 \\ \text{Cov}[e] &= \boldsymbol{\Sigma} = [\sigma_{ij}] = \sigma(u_i, u_j) \; i, j = 1, \dots, n \end{aligned}$$

Let $\Sigma_{Y0} := \begin{bmatrix} \sigma_{10} \\ \vdots \\ \sigma_{n0} \end{bmatrix}$

The best linear unbiased predictor of y_0 is

$$\widehat{y}_0 = x_0' \widehat{\beta} + \delta'(Y - X\widehat{\beta})$$

where $\widehat{\beta} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y$ and $\delta = \Sigma^{-1}\Sigma_{Y0}$.

9.2.2 Spatial Autocorrelation: Modelling $e(u)$

Spatial Autocorrelation is expressed as

$$\begin{aligned} \sigma(u, w) &:= \text{Cov}[y(u), y(w)] = \text{Cov}[e(u), e(w)] = \sigma(w, u) \\ &= \mathbb{E}[y(u)y(w)] - \mathbb{E}[y(u)]\,\mathbb{E}[y(w)] \neq 0 \;\; \forall i \neq j \end{aligned}$$

Covariance is often modelled in terms of an unknown parameter $\boldsymbol{\theta}$, in which case we write $\sigma(u, w; \boldsymbol{\theta})$. Assumptions made about $e(u)$ include

1. second-order stationary
2. strictly stationary
3. intrinsically stationary
4. increment stationary
5. isotropic

Covariance functions can be modelled 3 basic ways:

1. Specify a particular functional form on the stochastic process generating the random variables $\{y_i, i \in \mathcal{D}\}$, from where covariance structure follows
2. Model the covariance structure directly, typically as a function of a small number of parameters
3. Leave covariance unspecified and estimate it nonparametrically

Defn 9.21 (Stationarity).
A process $y(u)$ is said to be *strictly stationary* if $\forall k$, locations $u_1, \dots, u_k$, and Borel sets $C_1, \dots, C_k$, and any vector $h \in \mathbb{R}^d$,

$$\mathbf{Pr}\left(y(u_1) \in C_1, \dots, y(u_k) \in C_k\right) = \mathbf{Pr}\left(y(u_1 + h) \in C_1, \dots, y(u_k + h) \in C_k\right) \quad (12)$$

i.e. the joint density is translation invariant. In particular, $m(u) = m(u + h)$, so

$$m(u) = \mu \quad (13)$$

Also, $\sigma(u, w) = \sigma(u + h, w + h)$. Let $h = -w$, so $\sigma(u, w) = \sigma(u - w, 0)$, and so the **covariance function is a function of $u - w$ alone**. To make this explicit, we write

$$\sigma(u, w) = \sigma(u - w) = \sigma(h) \quad (14)$$

If $y(u)$ is strictly stationary and the joint distribution of all the random variables in 12 is multivariate gaussian, the process is called a **Gaussian Process**.

A second-order (weak) stationary process satisfies 13 and 14, but may or may not satisfy 12.
An increment-stationary process satisfies 13 and

$$\mathbf{Pr}\left(y(u_2)-y(u_1)\in C_1,\ldots,y(u_k)-y(u_{k-1})\in C_k\right)= \tag{15}$$
$$\mathbf{Pr}\left(y(u_2+h)-y(u_1+h)\in C_1,\ldots,y(u_k+h)-y(u_{k-1}+h)\in C_k\right) \tag{16}$$

Brownian motion is increment-stationary but not stationary.

Defn 9.22 ((Semi-)Variogram).
These are defined directly on increment-stationary processes. For a process satisfying 13, the semivariogram is defined

$$\begin{aligned}\gamma(u,w)&=\frac{1}{2}\mathbb{E}\left[y(u)-y(w)\right]^2=\frac{1}{2}\mathbb{V}\left[y(u)-y(w)\right]\\&=\left\{\mathbb{V}\left[y(u)\right]+\mathbb{V}\left[y(w)\right]-2\text{Cov}\left[y(w),y(u)\right]\right\}\\&=\left\{\sigma(u,u)+\sigma(w,w)-2\sigma(w,u)\right\}\end{aligned}$$

The variogram is $2\gamma(u,w)$. For an increment-stationary process, $\gamma(u,w)=\gamma(u+h,w+h)\ \forall\ h$, and we write

$$\gamma(u,w)=\gamma(u-w,0)=\gamma(u-w) \tag{17}$$

An **intrinsically-stationary** process satisfies 13 and 17. All second-order stationary processes are intrinsically stationary, but not vice versa.

Fact 9.3 (Semivariogram estimation).
For a linear model, stipulate a nonnegative definate weighting matrix, and fit
$Y=X\beta+e\ \ ,\mathbb{E}\left[e\right]=0\ \ ;\text{Cov}\left[e\right]=\Sigma_0$
to obtain residuals $\hat{e}_0=Y-X\widehat{\beta}$. For any vector h, there is a finite number N_h of pairs of observations y_i,y_j for which $u_i-u_j=h$. For each of these pairs, list the corresponding residual pairs, $(\widehat{e}_{0i},\widehat{e}_{0i(h)}),i=1,\ldots,N_h$. If $N_h\geq 1$, the traditional empirical covariance estimator is

$$\widehat{\sigma}(h)=\widehat{\sigma}(-h)=\frac{1}{N_h}\sum_{i=1}^{N_h}\widehat{e}_i\widehat{e}_{i(h)}$$

The traditional empirical semivariogram estimator in ordinary kriging (no covariates) is

$$\widehat{\gamma}(h)=\frac{1}{2N_h}\sum_{i=1}^{N_h}(y_i-y_{i(h)})^2$$

Defn 9.23 (Isotropy).
A second-order stationary process is said to be *isotropic if*

$$\sigma(u-w)=\sigma(||u-w||)$$

An intrinsically stationary process is isotropic if

$$\gamma(u-w)=\gamma(||u-w||)$$

Defn 9.24 (Spatial Autoregressive Processes).

$$\mathbf{y}-\mu\boldsymbol{\iota}=\rho\mathbf{W}(\mathbf{y}-\mu\boldsymbol{\iota})+\boldsymbol{\varepsilon}=\left(\mathbf{I}-\rho\mathbf{W}\right)^{-1}+\boldsymbol{\varepsilon}$$

where $\mathbf{W}$ is a $N\times N$ weight matrix. a spatial lag for y_i

$$\mathbf{W}y_i=\sum_j w_{ij}y_j$$

Defn 9.25 (Direct Representation of Spatial Autocorrelation).
A parsimonious specification of a small number of parameters for the covariance matrix is typically presumed.

$$\text{Cov}\left[\varepsilon_i,\varepsilon_j\right]=\sigma^2 f(d_{ij},\varphi)$$

where $\varepsilon_i,\varepsilon_j$ are residuals, σ^2 is the error variance, d_{ij} is the distance between i,j, and f is a distance decay function such that $\frac{\partial f}{\partial d}<0$ and $|f(d_{ij},\varphi)|\leq 1$, with $\varphi\in\Phi$ being a $p\times 1$ vector.

Defn 9.26 (Moran's I).
The generalised Moran's I is a weighted, scaled cross-product

$$\mathcal{I}:=\frac{n\sum_{i=1}^n\sum_{j\neq i}w_{ij}(y_i-\overline{y})(y_j-\overline{y})}{\sum_{i=1}^n\sum_{j\neq i}w_{ij}\sum_i(y_i-\overline{y})^2}$$

Its expected value is $\frac{-1}{n-1}$.
A test for Moran's I involves *shuffling the locations of points* and computing $\mathcal{I}$ S times. This produces a randomization distribution under H_0.
A Monte-carlo P-value is

$$\widehat{p}=\frac{1+\sum_{s=1}^S\mathbb{1}_{I_s^*\geq I_{obs}}}{S+1}$$

9.2.3 Spatial Linear Regression

A simple spatial regression is

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

the solution is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}(^\top\mathbf{I} - \rho\mathbf{W})\mathbf{y}$$

Its reduced form is

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$$

The spatial lag term induces correlation between the error and explanatory variables, and thus must be treated as an endogenous variable.
A spatial error model is simply an linear model with a non-spherical but typically parametric structure in the error covariance matrix.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \underbrace{\lambda\mathbf{W}\boldsymbol{\xi} + \boldsymbol{\eta}}_{\text{Composite error } \boldsymbol{\varepsilon}}, \ \eta \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$$

$$\mathbb{E}\left[\varepsilon\varepsilon'\right] = \boldsymbol{\Omega}(\boldsymbol{\theta})$$

Example 9.4 (Kelly (2020)'s 'Direct' standard errors).
A covariance function decomposes into a systematic part and idiosyncratic noise as follows

$$\boldsymbol{\Sigma}_{ij} = \sigma^2 C(\|i-j\|, \boldsymbol{\pi}) + \tau^2\mathbb{1}_{ij} \equiv \sigma^2\mathbf{C}_{UU} + \tau^2\mathbf{I}$$

where C is a correlation function, $\|i-j\|$ is the distance between points i, j.
Kelly recommends using a Whittle-Matern function defined next. These parameters can be fitted on the error distribution to estimate the covariance matrix.

Defn 9.27 (Covariance Function).
A covariance function $\operatorname{Cov}\left(Y(\mathbf{x}), Y\left(\mathbf{x}'\right)\right)$ describes the joint variability between a stochastic process $Y(\cdot)$ at two locations $\mathbf{x}$ and $\mathbf{x}'$. This covariance function is vital in spatial prediction. The fields package includes common parametric covariance families (e.g. exponential and Matern) as well as nonparametric models (e.g. radial and tensor basis functions).
When modeling $\operatorname{Cov}\left(Y(\mathbf{x}), Y\left(\mathbf{x}'\right)\right)$, we are often forced make simplifying assumptions.

- Stationarity assumes we can represent the covariance function as

  $$\operatorname{Cov}(Y(\mathbf{x}+\mathbf{h}), Y(\mathbf{x})) = C(\mathbf{h})$$

  for some function $C : \mathbb{R}^d \to \mathbb{R}$ where $\dim(\mathbf{x}) = d$.

- - Isotropy assumes we can represent the covariance function as

  $$\operatorname{Cov}(Y(\mathbf{x}+\mathbf{h}), Y(\mathbf{x})) = C(\|\mathbf{h}\|)$$

  for some function $C : \mathbb{R} \to \mathbb{R}$, where $\|\cdot\|$ is a vector norm.

Exponential :

$$\operatorname{Cov}\left(Y(\mathbf{x}), Y\left(\mathbf{x}'\right)\right) = C(r) = \rho e^{-r/\theta} + \sigma^2\mathbf{1}_{\mathbf{x}=\mathbf{x}'}$$

Matern:

$$\operatorname{Cov}\left(Y(\mathbf{x}), Y\left(\mathbf{x}'\right)\right) = C(r) = \rho\left(\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{r}{\theta}\right)^\nu K_\nu\left(\frac{r}{\theta}\right)\right) + \sigma^2\mathbf{1}_{\mathbf{x}=\mathbf{x}'}$$

where K_ν is a modified Bessel function of the second kind, of order ν
Matern covariance depends on $(\rho, \theta, \nu, \sigma^2)$, while exponential depends on $\rho, \theta, \sigma^2)$, where

- θ: is the range of the process at which observations become uncorrelated
- ρ : marginal variance / 'sil'
- σ^2 : small scale variation such as measurement error
- ν : smoothness

Fact 9.5 (Workhorse Spatial Regression).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \overbrace{\beta\mathbf{W}\mathbf{y}}^{\text{AC in Y}} + \overbrace{\mathbf{W}\mathbf{X}\boldsymbol{\theta}}^{\text{AC in X}} + \overbrace{\mathbf{W}\boldsymbol{\nu}\boldsymbol{\lambda}}^{\text{AC in errors}} + \boldsymbol{\varepsilon}$$

Here, $\mathbf{W}$ is a weight matrix (typically row-standardised), so $\mathbf{WM}$ is a spatial lag.
In spatial econometrics, the above form nests many popular regressions

- **Spatially Autoregressive** (SAR) Model : $\boldsymbol{\lambda} = \boldsymbol{\theta} = \mathbf{0}$
- **Spatially lagged** x : $\boldsymbol{\beta} = \boldsymbol{\lambda} = \mathbf{0}$
- **Spatial Durbin Model** : $\boldsymbol{\lambda} = \mathbf{0}$

- **Spatial Error model** : $\boldsymbol{\beta = \lambda = 0}$

Fact 9.6 (The Reflection Problem).
In the Social Interactions Literature (e.g. Manski (1993)), the above expression is written in the form of conditional expectations

$$y_i = \mathbf{x}_i'\boldsymbol{\gamma} + \mathbb{E}\left[y|\mathbf{w}_i\right]' \overbrace{\beta}^{\text{Endogenous}} + \mathbb{E}\left[\mathbf{x}_i|\mathbf{w}_i\right]' \overbrace{\boldsymbol{\theta}}^{\text{Contextual}} + \mathbb{E}\left[\nu|\mathbf{w}_i\right]' \underbrace{\lambda}_{\text{Correlated}} + \varepsilon_i$$

in practice, the expectations are replaced with empirical counterparts $\widehat{\mathbb{E}}(y|\mathbf{w}_i) = \mathbf{W}y$ and so on, so the estimation steps are isomorphic.
Define unobservables as $\boldsymbol{\upsilon} = \mathbf{W}\nu + \boldsymbol{\varepsilon}$, and assume they are uncorrelated with observables $\mathbf{x}$; that is, there is no sorting and no omitted spatial variables. Then, we can write

$$\mathbf{y} = \mathbf{X}\gamma + \mathbf{W}\mathbf{y}\beta + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\upsilon}$$

Premultiplying by $\mathbf{W}\mathbf{y}$ gives

$$\mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{X}\gamma + \mathbf{W}\mathbf{W}\mathbf{y}\beta + \mathbf{W}\mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{W}\boldsymbol{\upsilon}$$

This shows that $\mathbf{G}\mathbf{y}$ is correlated with $\boldsymbol{\upsilon}$, i.e. $\mathbb{E}\left[\boldsymbol{\upsilon}|\mathbf{W}\mathbf{y}\right] \neq 0$, and least square estimates of the above regression are biased.
If we assume $\mathbf{W}$ is idempotent (by constructing a block-diagonal, transitive matrix), we can simplify the above expression to

$$\begin{aligned}
\mathbf{W}y &= \mathbf{W}\mathbf{X}\frac{\boldsymbol{\gamma}+\boldsymbol{\theta}}{\mathbf{1}-\boldsymbol{\beta}} + \mathbf{W}\upsilon/(1-\beta) \quad \text{Plugging in definition for Wy} \\
y &= \mathbf{X}\underbrace{\gamma/(1-\beta)}_{\widetilde{\gamma}} + \mathbf{W}\mathbf{X}\underbrace{(\gamma\beta+\theta)/(1-\beta)}_{\widetilde{\theta}} + \underbrace{\upsilon + \mathbf{W}\upsilon\beta/(1-\beta)}_{\widetilde{\upsilon}}
\end{aligned}$$

In summary, β, θ cannot be separately identified from the composite parameters $\widetilde{\beta}, \widetilde{\theta}$. This is the *reflection problem* of (Manski, 1993).

9.2.4 Spatial Modelling

Based on Rue and Held (2005) and various lecture notes.

Defn 9.28 (Conditional / Markov Independence).
x_1, x_2 are conditionally independent given x_3 if, for a given value of x_3, learning x_2 gives one no additional information about x_1. The density representation is therefore

$$\mathsf{f}\left(\boldsymbol{x}\right) = \mathsf{f}\left(x_1|x_3\right)\mathsf{f}\left(x_2|x_3\right)\mathsf{f}\left(x_3\right)$$

which is a simplification of the general representation.

$$\mathsf{f}\left(\boldsymbol{x}\right) = \mathsf{f}\left(x_1|x_2,x_3\right)\mathsf{f}\left(x_2|x_3\right)\mathsf{f}\left(x_3\right)$$

Theorem 9.7 (Factorisation Criterion for Conditional Independence).

$$x \perp\!\!\!\perp y|z \Leftrightarrow \mathsf{f}\left(x,y,z\right) = g(x,z)h(y,z)$$

for some functions f, g , and $\forall z$ with $\mathsf{f}\left(z\right) > 0$

Example 9.8 (AR1 GMRF).

$$x_t = \phi x_{t-1} + \varepsilon_t \; ; \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}\left(0,1\right) \; , |\phi| < 1$$

This can be re-expressed as

$$x_t|x_1,\ldots,x_{t-1} \; \sim \mathcal{N}\left(\phi x_{t-1}, 1\right) \; \forall t = 2,\ldots,n$$

So, for $x_s \; , x_t \; , 1 \leq s < t \leq n$,

$$x_s \perp\!\!\!\perp x_t|\left\{x_{s+1},\ldots,x_{t-1}\right\} \;\; \text{if} \;\; t-s>1$$

In addition to the conditional distribution, also assume the marginal distribution of $x_1 \overset{\text{iid}}{\sim} \mathcal{N}\left(0, 1/(1-\phi)^2\right)$, which is the stationary distribution of this process. Then, the join distribution of $\mathbf{x}$ is

$$\begin{aligned}
\mathsf{f}\left(\boldsymbol{x}\right) &= \mathsf{f}\left(x_1\right)\mathsf{f}\left(x_2|x_1\right)\ldots,\mathsf{f}\left(x_n|x_{n-1}\right) \\
&= \frac{1}{(2\pi)^{n/2}}\left|\mathbf{Q}\right|^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{Q}\mathbf{x}\right)
\end{aligned}$$

where $\mathbf{Q}$ is a precision matrix of the form

$$\boldsymbol{Q} = \begin{pmatrix}
1 & -\phi & & & & & \\
-\phi & 1+\phi^2 & -\phi & & & & \\
 & \ddots & \ddots & \ddots & & & \\
 & & & -\phi & 1+\phi^2 & -\phi & \\
 & & & & -\phi & 1 &
\end{pmatrix}$$

This tridiagonal form is due to the fact that $x_i \perp\!\!\!\perp x_j$ if $|i-j| > 1$ given the rest of the sequence. This is generally true for any GMRF: $Q_{ij} = 0 \; , i \neq j \implies x_i \perp\!\!\!\perp x_j|\left\{x_k : k \neq i,j\right\}$.

While the conditional independence structure is readily apparent from the precision matrix, it isn't evident in the covariance matrix $\mathbf{\Sigma} = \mathbf{Q}^{-1}$, which is completely dense with entries

$$\sigma_{ij} = \frac{1}{1-\phi^2}\phi^{|i-j|}$$

Entries of the covariance matrix $\mathbf{\Sigma}$ only give direct information about the *marginal* dependence structure, not the conditional one.

Defn 9.29 (Spatial Gaussian Process (GP)).
A spatial process $Y(\mathbf{s})$ $s \in \mathcal{D} \subset \mathbb{R}^2$ is said to follow a **Gaussian Process** if *any realisation* $\boldsymbol{Y} = (Y(\mathbf{s})_1, \ldots, Y(\mathbf{s})_n)'$ at the finite number of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ follows an $N-$ variate Gaussian. More precisely, let $\mu(\mathbf{s} : \mathcal{D} \to \mathbb{R}$ denote a mean function returning a mean at location $\mathbf{s}$ (typically assumed to be linear in covariates $\mathbf{X}(\mathbf{s}) = (1, X_1(\mathbf{s}), \ldots, X_p(\mathbf{s}))'$) and $\mathbb{C}(\mathbf{s}_1, \mathbf{s}_2) : \mathcal{D}^2 \to \mathbb{R}^+$ denote a covariance function. Then, $Y(\mathbf{s})$ follows a spatial Gaussian process, and $\mathbf{Y}$ has a density

$$\mathrm{f}_{\mathbf{Y}}(\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)|\mathbf{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'(\mathbf{\Sigma})^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}$$

Where $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_N))'$ is the mean vector and $\mathbf{\Sigma} = \{\mathbb{C}(\mathbf{s}_i, \mathbf{s}_j)\}_{ij}$ is the $N \times N$ covariance matrix. Evaluating this density requires $\mathcal{O}(N^3)$ operations and $\mathcal{O}(N^2)$ memory, which means it does not scale well with large datasets. See Heaton et al. (2019) for overview of alternatives.

Defn 9.30 (Conditional Autoregressions (Besag 1974)).
Let $\boldsymbol{x}$ be associated with some property of points (typically location), with no natural ordering of the indices. The joint density of a zero-mean GMRF is specified by each of the n full-conditionals

$$x_i|\boldsymbol{x}_{-i} \sim \mathcal{N}\left(\sum_{j:j\neq i}\beta_{ij}x_j, (\kappa)_i^{-1}\right)$$

these are called **CAR** models. The associated precision matrix is

$$\mathbf{Q} = Q_{ij} = \begin{cases} \kappa_i & i = j \\ -\kappa_i\beta_{ij} & i \neq j \end{cases}$$

which is symmetric and positive-definite.

Defn 9.31 (Gaussian Markov Random Field (GMRF)).
A random vector $\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ is called a GMRF wrt a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$ iff its density has the form

$$\mathrm{f}(\boldsymbol{x}) = (2\pi)^{-n/2}|\mathbf{Q}|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

and $Q_{ij} \neq 0 \Leftrightarrow \{i,j\} \in \mathcal{E} \forall i \neq j$. If $\mathbf{Q}$ is completely dense, $\mathcal{G}$ is completely connected. In spatial settings, $\mathbf{Q}$ is typically sparse [depending on how neighbours are defined.]
Key summary quantities

- $$\mathbb{E}[x_i|\boldsymbol{x}_{-i}] = \mu_i - \frac{1}{Q_{ii}}\sum_{j:j\sim i}Q_{ij}(x_j - \mu_j)$$
- $\mathrm{Prec}(x_i, \boldsymbol{x}_{-i}) = Q_{ii}$ and
$$\mathrm{Corr}(x_i, x_j|\boldsymbol{x}_{-ij}) = \frac{-Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, i \neq j$$

Fact 9.9 (Markov Properties of GMRFs).
Let $\boldsymbol{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The following are equivalent

1. **Pairwise Markov Property**: $x_i \perp x_j|\boldsymbol{x}_{-ij}$ if $\{i,j\} \notin \mathcal{E} \wedge i \neq j$
2. **Local Markov Property**; $x_i \perp \boldsymbol{x}_{\{i,ne(i)\}}|\boldsymbol{x}_{ne(i)}$ $\forall i \in \mathcal{V}$
3. **Global Markov**: $\boldsymbol{x}_A \perp \boldsymbol{x}_B|\boldsymbol{x}_C$ for disjoint sets A, B, C where C separates A, B and A and B are nonempty.

Defn 9.32 (Linear Gaussian Process Models).
let the spatial process at location $\mathbf{s} \in \mathcal{D}$ be

$$Z(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) \quad , \forall \mathbf{s} \in \mathcal{D}$$

where $\mathbf{X}(\mathbf{s})$ collects a $p-$ vectors of covariates for site $\mathbf{s}$, and $\boldsymbol{\beta}$ is a p-vector of coefficients. Spatial dependence can be imposed by modelling $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ as a zero-mean stationary Gaussian Process. Distributionally, this implies that for any $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$, if we let $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))'$, and $\mathbf{\Theta}$ be the parameters of the model

$$\mathbf{w}|\mathbf{\Theta} \sim \mathcal{N}(0, \mathbf{\Sigma}(\mathbf{\Theta}))$$

where $\mathbf{\Sigma}(\mathbf{\Theta})$ is the covariance matrix of a n-dimensional normal density. We need $\mathbf{\Sigma}(\mathbf{\Theta})$ to be Symmetric, PD for this distribution to be proper.
Special cases:

- **Exponential Covariance Matrix**: $\boldsymbol{\Theta} = (\psi, \phi, \kappa)$ $\boldsymbol{\Sigma}(\boldsymbol{\Theta}) = \psi\mathbf{I} + \kappa\mathbf{H}(\phi)$, where the i, jth element of $\mathbf{H}(\phi) = \exp(-\|\mathbf{s}_i - \mathbf{s}_j\| / \phi)$. The 'nugget' ψ is the variance of the non-spatial error, κ dictates the scale, and ϕ dictates the range of the spatial dependence.
- **Matern Covariance**: $\boldsymbol{\Theta} = (\psi, \kappa, \phi, \nu) > 0$ for distance $x := \|\mathbf{s}_i - \mathbf{s}_j\|$.

$$\text{Cov}\left[x; \phi, \psi, \kappa, \nu\right] = \begin{cases} \frac{\kappa}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}x/\phi)^{\nu} K_{\nu}(2\sqrt{\nu}x/\phi) & \text{if } x > 0 \\ \psi + \kappa & \text{if } x = 0 \end{cases}$$

where $K_{\nu}(x)$ is a modified Bessel function of order ν.

Defn 9.33 (Linear GMRF Models).
Specifying Σ directly can be awkward when dealing with irregular spatial data [i.e. every real use case].
So, random effects $\mathbf{w}$ are modelled *conditionally*. Let $\mathbf{w}_{-i}$ denote the vector of $\mathbf{w}$ excluding $\mathbf{w}(\mathbf{s}_i)$. Model $w(\mathbf{s}_i)$ in terms of its full-conditional.

$$w(\mathbf{s}_i)|\mathbf{w}_{-i}, \boldsymbol{\Theta} \sim \mathcal{N}\left(\sum_{j=1}^{n} c_{ij} w(\mathbf{s}_j), \kappa_i^{-1}\right), \; i = 1, \ldots, n$$

where c_{ij} describes the neighbourhood structure.

1. Besag (1974) proved that if $\mathbf{Q}$ is symmetric PD, with κ_i in the diagonals and $-\kappa_i c_{ij}$ in the off-diagonals. $\mathbf{w}|\boldsymbol{\Theta} \sim \mathcal{N}\left(0, \mathbf{Q}^{-1}\right)$. Simplest version assumes common precision parameter $\kappa_i = \tau$.
2. **Intrinsic GMRF**: $f(\mathbf{w}|\boldsymbol{\Theta}) \sim \tau^{(N-1)/2} \exp(-\mathbf{w}'\mathbf{Q}(\tau)\mathbf{w})$. When $c_{ij} = 1$ for neighbours (i.e. adjacency matrix instead of distances), it simplifies further to

$$(\mathbf{w}|\boldsymbol{\Theta}) \sim \tau^{(N-1)/2} \exp\left(-\frac{1}{2}\sum_{i \sim j}(w(\mathbf{s}_i) - w(\mathbf{s}_j))^2\right)$$

Defn 9.34 (Gaussian Process Spatial GLMs).
Let $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ and $\{w(\mathbf{s} : \mathbf{s} \in D)\}$ be two spatial processes on $D \subset \mathbb{R}^d (d \in \mathbb{Z}^+)$. Assume $Z(\boldsymbol{s}_i)$s are conditional independent given random effects $w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n)$, and that $Z(\mathbf{s}_i)$ follow some common distributional form, and

$$\mathbb{E}\left[Z(\mathbf{s}_i)|\mathbf{w}\right] = \mu(\mathbf{s}_i) \; \forall i = 1, \ldots n$$

Let $\eta(\mathbf{s}) = h(\mu(\mathbf{s}))$ for some known link function $h(\cdot)$ e.g. $h(x) = \log\left(\frac{x}{1-x}\right)$ for logit. Assume linear form for projection $\eta(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s})$. Spatial dependence via $\mathbf{w}|\boldsymbol{\Theta} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}(\boldsymbol{\Theta})\right)$, where Σ is often Matern.

A Mathematical Background

A.1 Proof Techniques

1. **Direct Proof / *modus ponens***: If R is a true statement and $R \implies S$ is a true conditional statement, then S is a true statement. Direct proofs typically involve *backwards-forwards reasoning* - take all statements that follow from R that might relate to S, list them in $\mathcal{R}$. Then, take all statements that follow from S, list them in $\mathcal{S}$. Then, look for statements $r, s \in \mathcal{R} \times \mathcal{S}$ that have a straightforward proof, and write proof of the form $R \implies r \implies s \implies S$.
2. **Contrapositive**: Since every conditional statement is equivalent to its contrapositive, proving $\neg Q \implies \neg P$ is equivalent to proving $P \implies Q$.
3. **Proof by contradiction**: Assume P is true, and assume Q is false [i.e. $\neg Q$ is true], and show that $\neg Q \implies S$ (using P and other possible intermediate results) where S is known to be false. Conclude that $\neg Q$ must be false, so Q must be true, and we have proved that $P \implies Q$.
4. **Induction** [only applies to statements pertaining to well ordered sets] $\mathcal{N}$
   - Assume a base case - $P(0)$ is a true statement
   - Prove whenever $P(k)$ is true, $P(k+1)$ is true
   - Therefore $P(n)$ is true for every $n \in \mathcal{N}$

A.2 Set Theory

A set is a collection of objects. E.g. $\mathbb{R}, \mathbb{Q}, \mathbb{Z}, \mathbb{N}$.
Set operations:

- Intersection : $A \cap B$
- Union : $A \cup B$
- Difference: $A \setminus B := \{x : x \in A \wedge x \notin B\}$
- cartesian product: $A \times B := \{(a, b) : a \in A \wedge b \in B\}$

Defn A.1 (Power set).
Set of all subsets of S is itself a set. Denoted as $\mathfrak{P}(S)$

$$|\mathbb{C}| > |\mathbb{R}| > |\mathbb{Q}| > |\mathbb{Z}| > |\mathbb{N}|$$

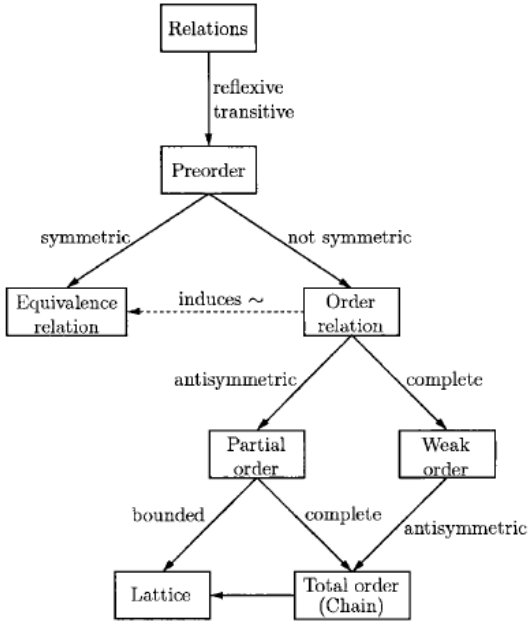


Figure 12: Types of Relations

A.2.1 Relations

Given two sets X and Y, any subset of their Cartesian product $X \times Y$ is called a binary relation. For any pair of elements $(x, y) \in R \subseteq RX \times Y \implies xRy$.
Properties of binary relations:

- **reflexive** $xRx \; \forall \; x \in X$
- **transitive** if $xRy \wedge yRz \implies xRz$
- **symmetric** If $xRy \implies yRx$
- **antisymmetric** If $xRy \wedge yRx \implies x = y$
- **asymmetric** if $xRy \implies \neg(yRx)$
- **complete** if either xRy or yRx or both $\forall x, y, z \in X$

Defn A.2 (Equivalence Relations).
An equivalence relation R on a set X is a relation that is reflexive, transitive, and

symmetric. Given an equivalence relation $\sim$, the set of elements that are related to a given element a :

$$\sim(a) := \{x \in X : x \sim a\}$$

is called the equivalence class of a. e.g. Indifference $\sim$ preference relation is an equivalence relation, but the preference relation $\succeq$ is not because it isn't symmetric.

Defn A.3 (Order Relations).
A relation that is reflexive and transitive but not symmetric is called an order relation: $x \succeq y$. This is also called a *weak order*. $\succ$ is not an order relation because it is not reflexive (and is called a *strong* order). Every order relation also induces an equivalence relation: $x \sim y \Leftrightarrow x \succeq y \wedge y \succeq x$

An ordered set $(X, \succeq)$ consists of a set X together with an order relation $\succeq$ defined on X.

A.2.2 Intervals and Contour Sets

Given an ordered set and two elements $a, b \in X$ s.t. $b \succeq a$, we can define

- The open interval (a, b) : set of all elements strictly between a and b.
- The closed interval $[a, b]$: set of all elements between a and b s.t. $[a, b]\{x \in X : a \preccurlyeq x \preccurlyeq b\}$

Analogously, for arbitrary ordered sets, $(X, \succeq)$ we can define

- **Upper contour set** $\succeq(a) := \{x \in X : x \succeq a\}$: set of all elements that follow or dominate a
- **Lower contour set** $\preccurlyeq(a) := \{x \in X : x \preccurlyeq a\}$: set of all elements that preced a in the order $\succeq$

A partial order is a relation that is reflexive, transitive, and antisymmetric.

Defn A.4 (Meet and Join).
The **join** of a partially ordered set $\mathcal{S}$ is the **supremum** and is denoted $\bigvee \mathcal{S}$. $\max(a, b)$ is sometimes written $a \vee b$.
The **meet** of a poset is the **infimum** and is denoted $\bigwedge \mathcal{S}$. $\min(a, b)$ is sometimes written $a \wedge b$.

A.2.3 Algebra

Defn A.5 (Groups).
A set $\mathcal{G}$ and an operation $\otimes : \mathcal{G} \times \mathcal{G} \to G$ defined on $\mathcal{G}$. Then $G := (\mathcal{G}, \otimes)$ is called a group if the following conditions hold:

1. Closure of $\mathcal{G}$ under $\otimes$: $\forall x, y \in \mathcal{G}, x \otimes y \in \mathcal{G}$
2. Associativity: $\forall x, y, z \in \mathcal{G}, (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. Neutral element: $\exists e \in \mathcal{G} \forall x \in G$ s.t. $x \otimes e = \ e \otimes x = x$
4. Inverse element: $\forall x \in \mathcal{G}, \exists y \in \mathcal{G} : x \otimes y = e \wedge y \otimes x = e$.
5. **if additionally** $\forall x, y \in \mathcal{G} : x \otimes y = y \otimes x$, then G is an *Abelian/Commutative* group

$(\mathbb{Z}, +), (\mathbb{R}^{m \times n}, +)(\mathbb{R} \backslash \{0\}, .)$ are all groups

Defn A.6 (Vector Spaces).
A real valued vector space $V = (\mathcal{V}, +, \cdot)$ is a vector space with two operations

$$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$$
$$\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$$

Where

1. $(\mathcal{V}, +)$ is an Abelian Group
2. Distributivity
   - $\forall \lambda \in \mathbb{R}, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V} : \lambda \cdot (\boldsymbol{x} + \boldsymbol{y}) = \lambda \cdot \boldsymbol{x} + \lambda \cdot \boldsymbol{y}$
   - $\forall \lambda, \psi \in \mathbb{R}, \boldsymbol{x} \in \mathcal{V} : (\lambda + \psi) \cdot \boldsymbol{x} = \lambda \cdot \boldsymbol{x} + \psi \cdot \boldsymbol{x}$
3. Associativity : $\forall \lambda, \psi \in \mathbb{R}, \boldsymbol{x} \in \mathcal{V} : \lambda \cdot (\psi \cdot \boldsymbol{x}) = (\lambda\psi) \cdot \boldsymbol{x}$
4. Neutral Item wrt outer operation: $\forall \boldsymbol{x} \in \mathcal{V} : \mathbf{1} \cdot \boldsymbol{x} = \boldsymbol{x}$

A.3 Analysis and Topology

Preliminaries:
Vectors : $x := (x_1, \cdots, x_k)$, where $x_i \in \mathbb{R}$

A.3.1 Metric Spaces

Defn A.7 (Euclidian Distance).

$$d_2(x,y) := \|x-y\|_2 := \left(\sum_{i=1}^{k}(x_i-y_i)^2\right)^{1/2}$$

Requirements for a metric(e.g. $d_2 : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}\ \forall x,y,v \in \mathbb{R}^k$):

- $d_2(x,y) = 0 \Leftrightarrow x = y$: a point is at zero distance from itself
- $d_2(x,y) = d_2(y,x)$: distance is symmetric
- $d_2(x,y) \leq d_2(x,v) + d_2(v,y)$: triangle inequality

We can generalise this definition to arbitrary nonempty sets S.

Defn A.8 (Metric Spaces).
A metric space is a nonempty set S and a metric of distance $\rho : S \times S \to \mathbb{R}\, \forall x,y,v \in S$ s.t.

- $\rho(x,y) = 0 \Leftrightarrow x = y$
- $\rho(x,y) = \rho(y,x)$
- $\rho(x,y) \leq \rho(x,v) + \rho(v,y)$

For example, $(\mathbb{R}^k, d_2)$ is a metric space. Many additional metric spaces in $\mathbb{R}^k$ are generated by a norm.

Defn A.9 (Norm).
A norm on $X \subseteq \mathbb{R}^k$ is a mapping $X \ni x \mapsto \|x\| \in \mathbb{R}$ s.t. $\forall x,y \in \mathbb{R}^k$ and $\gamma \in \mathbb{R}$ satisfying

- **Nonnegativity**: $\|x\| \geq 0\ \forall x\ in X$
- **Non degeneracy**: $\|x\| = 0 \Leftrightarrow x = 0$
- **Homogeneity**: $\|\gamma x\| = |\gamma|\,\|x\|$
- **Triangle Inequality**: $\|x+y\| \leq \|x\| + \|y\|$

Each norm $\|.\|$ on $\mathbb{R}^k$ generates a metric ρ on $\mathbb{R}^k$ via $\rho(x,y) := \|x-y\|$.
E.g. $\|x\|_2 := (\sum_{i=1}^{k} x_i^2)^{1/2}$ generates euclidian distance d_2.
The pair $(X, \|\cdot\|)$ consisting of a vector space X together with a norm $\|\cdot\|$ is called a **normed linear space**.

Defn A.10 (Banach Space).
A banach space$(X, \|\cdot\|)$ is a **normed linear space** that is a **complete**(in the Cauchy-convergence sense) metric space with respect to the metric derived from its norm.

Defn A.11 (p-norm).
Also known as **Minkowski Norm**
A class of norms that includes $\|.\|_2$ as a special case is the $\|.\|_p$ defined by

$$\|x\|_p := \left(\sum_{i=1}^{k}|x_i|^p\right)^{1/p} \quad x \in \mathbb{R}^k$$

$\|.\|_p$s give rise to a class of metric spaces $(\mathbb{R}^k, d_p)$ where $d_p(x,y) := \|x-y\|_p\ \forall x,y \in \mathbb{R}^k$.

Examples:

- 1: Taxicab
- 2: Euclidian
- ∞: Chebychev

Defn A.12 (Frobinius Norm).
Frobinius Norm of a matrix $\mathbf{A}$ is

$$\|\mathbf{A}\|_{\mathrm{Fr}} = \sqrt{\sum_{i=1}^{M}\sum_{j}^{N} a_{ij}^2} = \sqrt{\mathrm{trace}(\mathbf{A}'\mathbf{A})}$$

Defn A.13 (Sup, Inf).
$a,b \in \mathbb{R}$, $[a,b]$ denotes the set of real numbers satisfying $a \leq x \leq b$. (or) denotes a strict inequality (i.e. closed from above or below).
If $\mathcal{S} \subset \mathbb{R}$ is bounded from above, $\exists y$ s.t. $x \leq y\ \forall x \in \mathcal{S}$.. Then y is the *least upper bound* or **supremum** of $\sup\{x : x \in \mathcal{S}\}$. If $\mathcal{S}$ is not bounded from above, we write $\sup_{x\in\mathcal{S}} = \infty$.
Similarly, the *greatest lower bound* of a set or *infimum* is denoted $\inf_{x\in\mathcal{S}}(x) \vee \inf\{x : x \in \mathcal{S}\}$

Defn A.14 (Sequences, Liminf, Limsup).
A sequence $x_1, x_2, \ldots x_n$ is denoted by $\{x_i\}_{i=1}^{\infty}$ or $\{x_i\}$ when the range of the indices is clear.
Let $\{x_i\}$ be an infinite sequence of real numbers and $\exists S$ s.t. $(1)\forall \varepsilon > 0, \exists N$ s.t. $\forall n > N, x_n < S + \varepsilon$ and $(2)\forall \varepsilon > 0$ and $M > 0, \exists n > M$ s.t. $x_n > S - \varepsilon$. Then, S is the $\limsup\{x_n\}$.
If $\{x_n\}$ is not Bounded from above, $\limsup x_n = \infty$.

Defn A.15 (Cauchy Criterion).
A sequence (x_n) in a metric space (S, ρ) is said to be a Cauchy sequence if, $\forall \epsilon > 0, \exists N \in \mathbb{N}$ s.t. $\rho(x_j, x_k) < \epsilon$ whenever $j, k \geq N$ (intuitively, points in a Cauchy sequence get tighter together).

Let (x_n) be a sequence of vectors in $\mathbb{R}^k$. Suppose for any $\epsilon > 0$, $\exists n \in \mathbb{N}$ s.t. $\forall p, q > n, \rho(x^p, x^q) < \epsilon$. Then, (x_n) has a limit.
More basic definition: $\{a_n\}_{n=1}^{\infty} \to A$ if $\forall \epsilon > 0, \exists N$ s.t. $\forall n \geq N, |a_n - A| < \epsilon$

- $\{a_n b_n\} \to AB$
- $\{a_n + b_n\} \to A + B$

Sequences
Let $S = (S, \rho)$ be a metric space. A sequence $(x_n) \subset S$ is said to converge to $x \in S > 0, \exists N \in \mathbb{N}$ s.t. $n \geq N \implies \rho(x_n, x) < \epsilon$.

Theorem A.1.
A sequence in (S, ρ) can have at most one limit

Defn A.16 (ϵ ball).
centered on $x \in S$ with radius $\epsilon > 0$ is the set

$$B(\epsilon, x) := \{z \in S : \rho(z, x) < \epsilon\}$$

Set Definitions

Defn A.17 (Bounded Set).
A subset E of S is called *bounded* if $E \subset B(n, x)$ for some $x \in S$ and some suitably large $n \in \mathbb{N}$ (intuition - some arbitrarily large ϵ ball can fit E inside it).
A sequence (x_n) in S is called bounded if its range $\{x_n : n \in \mathbb{N}\}$ is a bounded set.

Defn A.18 (Closed Set).
A set $F \subset S$ is closed IFF for every convergent sequence contained in F, the limit of the sequence is also in F.
A closed set contains all its limit points. That is , if (x_k) is a convergent sequence of points in S, then $\lim_{k\to\infty} x_k$ is in S as well.

Defn A.19 (Open Set).
A subset of an arbitrary metric space S is open iff its complement is closed, and closed iff its complement is open.
A set $S \in \mathbb{R}^k$ is called open if, $\forall x \in S \exists \epsilon > 0$ s.t. $y \in B(\epsilon, x), \rho(x, y) < \epsilon$ is in S.

If F is a closed, bounded subset of $(\mathbb{R}, \|.\|)$, then $\sup F \in F$.

- A set $S \subset \mathbb{R}^k$ is open iff its complement is closed.
- the union of *any* number of **open** sets is open
- the intersection of a *finite* number of **open** sets is open.
- the intersection of *any* number of **closed** sets is closed
- the union of a *finite* number of **closed** sets is closed.

Defn A.20 (Boundary and Closure).
A point $x \in S$ is called an **interior point** of S if the set $\{y : \rho(y, x) < \epsilon\}$ is contained in S for all $\epsilon > 0$ sufficiently small. A point is called a **boundary point** if $\{y : \rho(y, x) < \epsilon\} \cap S^c$ is non-empty for all $\epsilon > 0$ sufficiently small. The set of all boundary points in $\mathcal{A}$ is denoted by $\partial\mathcal{A}$.
The **closure** of a set S is the set S combined with all points that are the limits of sequence of points in S.

Defn A.21 (Complete Set).
A subset $A \subset S$ is said to be complete iff every cauchy sequence in A converges to some point in A.

Defn A.22 (Compact Set).
The set $K \subset S$ is called compact if every sequence contained in K has a subsequence that converges to a point in K.

Defn A.23 (Convex Set).
A set $S \subset \mathbb{R}^k$ is called convex if, $\forall \lambda \in [0, 1]$ and $a, a' \in S$, we have $\lambda + (1 - \lambda)a' \in S$.
(i.e. all convex combinations of two points in a set are also in the set).

Theorem A.2 (Bolzano-Weierstrass).
Every bounded sequence in euclidian space $(\mathbb{R}^k, d_2)$ has at least one convergent subsequence.

Theorem A.3 (Heine-Borel).
A subset $(\mathbb{R}^k, d_2)$ is precompact in the same iff it is bounded and compact.
IOW : Compact $\Leftrightarrow$ Closed $\wedge$ Bounded

Theorem A.4.
All metrics on $\mathbb{R}^k$ induced by a norm are equivalent.

A.4 Functions

A function f from set A to B, written as $A \ni x \mapsto f(x) \in B$ or $f : A \to B$ is a rule associating every element in A to one and only one element in B. The point b is also written as $f(a)$, and is called the *image* of a under f. For $D \subset B$, the set $f^{-1}(D)$ is the set of all points in A that map into D under F, and is called the *preimage* of D under F. $f^{-1}(D) := \{a \in A : f(a) \in D\}$
a function $f : A \to B$ is called

- *injective / one-to-one* if distinct elements of A are always mapped into distinct elements of B
- *surjective / onto* if every element of B is the image under f of at least one point in A
- *bijective* if a function is both injective and surjective

Defn A.24 (Continuous functions).
A real valued function f on $\mathbb{R}^k$ is **continous at point** a if $\forall \epsilon > 0, \exists \delta > 0$ s.t. $\rho(x,a) < \delta \implies |f(x) - f(a)| < \epsilon$.
Equivalently, $\lim_{x \to a} f(x) = f(a)$
A function is said to be continuous on the set $S \subset \mathbb{R}^k$ if, $\forall a \in S \wedge \forall \epsilon > 0, \exists \delta > 0$ s.t. $\forall \{x : \rho(x,a) < \delta\}, |f(x) - f(a)| < \epsilon$. Equivalently, in $\lim_{x \to a} f(x)$, we require the sequence of points that converge to a to be entirely in S.

- The sum of two continuous functions is continuous
- The product of two continuous functions is continuous
- The quotient of two continuous functions is continuous at any point where the denominator is nonzero

Defn A.25 (ϵ, δ definition of limit).

$$\lim_{x \to c} f(x) = L \iff \forall \varepsilon > 0, \exists \delta > 0, \text{ s.t. } 0 < |x - c| < \delta \Rightarrow |f(x) - L| < \varepsilon$$

Defn A.26 (Lipshitz Continuity).
Given two metric spaces $(\mathcal{X}, \rho_X)$, $(\mathcal{Y}, \rho_Y)$, a function $f : \mathcal{X} \to \mathcal{Y}$ is called Lipshitz continuous if $\exists K \in \mathbb{R}$ s.t. $\forall x_1, x_2 \in \mathcal{X}$,

$$\rho_Y(f(x_1), f(x_2)) \leq K \rho_X(x_1, x_2)$$

such a K is referred to as a **Lipshitz constant** for the function f.
A Real valued function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz if $\exists K > 0$ such that

$$|f(x_1) - f(x_2)| \leq K |x_1 - x_2|$$

This limits how fast a function can change. Every function that has bounded first-derivatives is Lipshitz continuous. **A differentiable function is Lipshitz if and only if it has a bounded derivative**.

Defn A.27 (Holder Continuity).
A function defined on $\mathcal{X}$ is said to be Holder of order $\alpha > 0$ if $\exists M \geq 0$ such that

$$\rho_Y(f(x), f(y)) \leq M \rho_X(x,y)^\alpha \ \forall x, y \in \mathcal{X}$$

this is also called **Uniform Lipshitz**.

Theorem A.5 (Continuity).
A function f $S \to Y$ is continuous iff the preimage $f^{-1}(G)$ of every open set $G \subset Y$ is open in S.

Defn A.28 (Continuous function).
f is continuous if $\forall \epsilon > 0, \exists \delta > 0$ s.t. $|x - x_0| < \delta \ \forall \ x \implies |f(x) - f(x_0)| < \epsilon$.

Theorem A.6.
Let function f $S \to Y$, where S, Y are metric spaces and f is continuous. If $K \subset S$ is compact, then so is $f(K)$, the image of K under f.

Example A.7 (Gamma Function).

$$\Gamma[\alpha] := \int_0^\infty t^{\alpha-1} e^{-t} dt = \int_0^1 (\log(1/t))^{\alpha-1} \, dt$$

Beta function : $B(\alpha, \beta) = \Gamma(\alpha) \cdot \Gamma(\beta) / \Gamma(\alpha + \beta)$

Theorem A.8 (Weirstrass Maximum Theorem).
Let $f : k \to \mathbb{R}$, where $K \subset (S, \rho)$ (an arbitrary metric space). If f is continuous and K is compact, then f attains its supremum and infimum on K.
In case of continuous functions on compact domains, optima always exist.

Defn A.29 (Differentiability).
The function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at x_0 if

$$\exists \lim R(x) = \frac{(f(x) - f(x_0))}{(x - x_0)} = f'(x_0)$$

, i.e. $(f(x) - f(x_0))/(x - x_0)$ has a limit as $x \to x_0$.
The derivative of f at x_0 is this limit and is denoted $f'(x_0)$ or $\frac{\partial f}{\partial x}|_{x=x_0}$

Fact A.9 (Restrictiveness of Function Classes).
Differentiability $\subset$ **Continuity** $\subset \exists$ **Limit** i.e. not all functions with limits are continuous, not all continuous functions are differentiable.
More generally,
Continuously Differentiable $\subset$ **Lipshitz Continuous** $\subset \alpha-$ **Holder Continuous** $\subset$ **Uniformly Continuous** $\subset$ **Continuous**

Fact A.10 (Properties of differentiable functions).

- **Linearity**: $f, g : X \to Y$ are differentiable at $\boldsymbol{x}$, then $f + g$ and αf are differentiable at x with $\nabla (f + g)(x) = \nabla f(x) + \nabla g(x)$; $\nabla \alpha f(x) = \alpha \nabla f(x)$

- **Chain Rule**: $g \cdot f$ differentiable with $\nabla g \cdot f(x) = \nabla g(f(x)) \cdot \nabla f(x)$

Theorem A.11 (Rolle's theorem).
Let $f : [a,b] \to \mathbb{R}$, f is continuous and differentiable. $f(a) = f(b) \implies \exists c \in [a,b]$ s.t. $f'(c) = 0$.

Theorem A.12 (Mean Value theorem).
$f : [a,b] \to \mathbb{R}$, f is continuous and differentiable. Then,

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Defn A.30 (Epigraph of a function).
The epigraph of a function f is epi$f := \{(x,t) : \} f(x) \leq t$ (i.e. area above the function).

Defn A.31 (Concavity and Convexity for Real-valued functions).
Let $f : [a,b] \to \mathbb{R}, x, y \in [a,b]; t \in (0,1)$. Then,

- F is **convex** if $f((1-t)x + ty) \leq (1-t)f(x) + tf(x)$. $f'' \geq 0$: i.e. the epigraph of f is a convex set.
- F is **concave** if $f((1-t)x + ty) \geq (1-t)f(x) + tf(x)$. $f'' \leq 0$

A.4.1 Fixed Points

Defn A.32 (Fixed Point).
Let $T : S \to S$, where S is any set. An $x^* \in S$ is called a fixed point of T on S if $Tx^* = x^*$.

If $S \subset \mathbb{R}$, then fixed points of T are those points in S where T meets the 45 degree line.

Theorem A.13 (Brouwer's Fixed Point Theorem).
Consider the space $(\mathbb{R}^k, d)$, where d is the metric induced by any norm. Let $S \subset \mathbb{R}^k$, and let $T : S \to S$. If T is continuous and S is both compact and convex, then T has at least one fixed point in S.

Defn A.33 (Mapping Categories).
Let (S, ρ) be a metric space. $T : S \to S$ is a map. it is called

- *nonexpansive* if $\rho(Tx, Ty) \leq \rho(x,y)\ \forall x, y \in S$
- *contracting* if $\rho(Tx, Ty) < \rho(x,y)\ \forall x, y \in S, x \neq y$
- *uniformly contracting* with modulus $\lambda \in [0,1)$ if $\rho(Tx, Ty) < \lambda\rho(x,y)\ \forall x, y \in S, x \neq y$

Theorem A.14 (Hahn-Banach Fixed Point Theorem).
Let $T : S \to S$, where (S, ρ) is a complete metric space. If T is a uniform contraction on S with modulus λ, then T has a unique fixed point $x^* \in S$. Moreover for every $x \in S$ and $n \in \mathbb{N}$, we have $\rho(T^n x, x^*) \leq \lambda^n \rho(x, x^*) \implies T^n x \to x^*$ as $n \to \infty$

A.5 Measure

Defn A.34 ($\sigma-$ field / Event Space).
A σ-algebra (also σ-field) is a collection $\mathcal{F}$ of subsets of Ω that

- $\Omega \in \mathcal{F} \wedge \emptyset \in \mathcal{F}$: includes Ω itself and the null set
- $A \in \mathcal{F} \implies \Omega - A =: A^C \in \mathcal{F}$: is closed under complement
- $A_1, A_2, \cdots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$: is closed under countable unions.

This is effectively the definition of the *event space* $\mathcal{S}$ for a sample space Ω.

Defn A.35 (Measure).
A measure μ on a set $\mathcal{X}$ assigns a nonnegative value $\mu(A)$ to many subsets of $\mathcal{X}$. For a collection $\mathcal{F}$ subsets of Ω, a measure is a map

$$\mu : \mathcal{F} \to [0, \infty]$$

Given $A \in \mathcal{F}$, $\mu(A)$ is a measure of the 'size' of set A.
A function μ on a $\sigma-$ field $\mathcal{A}$ of $\mathcal{X}$ is a measure of

- **Null empty-set**: $\mu(\emptyset) = 0$
- **Non-Negativity**: $\forall A \in \mathcal{A}, 0 \leq \mu(A) \leq \infty \implies \mu : \mathcal{A} \to [0, \infty]$
- **Countable Additivity:** If $A_1, A_2, \ldots$ are disjoint elements of $\mathcal{A}$ (i.e. $A_i \cap A_j = \emptyset\ \forall i \neq j$),

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

Existence ensured by Caratheodory's Extension Theorem.

Examples of measures μ:

- If $\mathcal{X}$ is countable, let $\mu(A) = \#A =$ number of points in A. This counting measure can be defined for any subset $A \subset \mathcal{X}$, then the $\sigma-$ field $\mathcal{A}$ is the collection of all subsets of $\mathcal{X} =: \mathfrak{X} = 2^{\mathcal{X}}$, the *power set* of $\mathcal{X}$.
- If $\mathcal{X} = \mathbb{R}^k$, define $\mu(A) = \int \ldots_A \int dx_1 \ldots dx_k$

Defn A.36 (Borel Set).
Given a topology on Ω, a Borel σ-field is a σ field generated by the family of open subsets of Ω, i.e. the smallest σ field that contains all the open sets.
The Lebesgue measure of a set A can be defined implicitly for any set B in a $\sigma-$ field $\mathcal{B}$ called the **Borel sets** of $\mathbb{R}^n$. $\mathcal{B}$ is the smallest $\sigma-$ field that contains all 'rectangles'

$$(a_1, b_1) \times \cdots \times (a_n, b_n) := \{x \in \mathbb{R}^n : a_i < x_i < b_i, i = 1, \ldots, n\}$$

Example A.15 (Uncountable Sample Spaces).
Suppose $\Omega = \mathbb{R}$. We say that $I \subset \mathbb{R}$ is a bounded interval if $\exists a, b, a < b$ s.t. $I \in \{[a, b], (a, b), (a, b], [a, b)\}$. Define $\mathcal{C}^1 := \{I \subset \mathbb{R}, I \text{ is a bounded interval}\}$, the smallest $\sigma-$ algebra that contains $\mathcal{C}^1$ is denoted by $\mathcal{B}^1$ and is called the Borelian $\sigma-$algebra Thus, a countable union of open intervals belongs to the Borelean σ-algebra. Since every open subset of $\mathbb{R}$ can be written as a countable union of open intervals, it is therefore also a Borelean set. The closed subsets are Borelean since a closed set is the complement of an open set.

Defn A.37 (Lebesgue Measure).
Basic problem: how to assign each subset of $\mathbb{R}^k$, i.e. each element of $\mathfrak{P}(\mathbb{R}^k)$ a real number that will represent its 'size'.
With $\mathbb{R}^n, n = 1, 2, 3$, $\mu(A)$ is the length, area, or volume of A, respectively. μ is a *Lebesgue measure* on $\mathbb{R}^k$.

Defn A.38 (Measure Space).
If $\mathcal{A}$ is a $\sigma-$field of subsets of $\mathcal{X}$, the pair $(\mathcal{X}, \mathcal{A})$ is called a measurable space, and if μ is a measure on $\mathcal{A}$, the triple $(\mathcal{X}, \mathcal{A}, \mu)$ is called a measure space.

Defn A.39 (Probability Space).
A measure μ is called a *probability measure* if $\mu(\mathcal{X}) = 1$, and then the triple $(\mathcal{X}, \mathcal{A}, \mu)$ is called a *probability space*.

Defn A.40 (Measurable functions).
If $(\mathcal{X}, \mathcal{A})$ is a measurable space and f is a real-valued function on $\mathcal{X}$, f if measureable if

$$f^{-1}(B) := \{x \in \mathcal{X} : f(x) \in B\} \in \mathcal{A}$$

for every Borel set B.

A.6 Integration

An integral is a map assigning a number to a function, where the number is viewed as the area/volume 'under' the function. Given a measure space $(\Omega, \mathcal{F}, \mu)$ and a measureable function $f : \Omega \to \mathbb{R}$, an integral $\int f d\mu$ is a map from f to number such that the following three properties hold

- If $f \geq 0$, then $\int f d\mu \geq 0$
- $\forall a \in \mathbb{R}, \int a\phi d\mu = a \int \phi d\mu$
- $\int (f + g) d\mu = \int f d\mu + \int g d\mu$

Defn A.41 (Riemann Sums).
Suppose f is a bounded function defined on $[a, b]$. An increasing sequence $P := \{a = x_0 < x_1 < x_2 < \cdots < x_n = b\}$ defines a *partition* of the interval. The **mesh size** of the partition is defined to be

$$|P| = \max\{|x_i - x_{i-1}| : i = 1, \ldots, N\}$$

To each partition we associate two approximations of the area under the graph of f, by the rules

$$U(f, P) := \sum_{j=1}^{N} \sup_{x \in [x_{j-1}, x_j]} f(x)(x_j - x_{j-1})$$

$$L(f, P) := \sum_{j=1}^{N} \inf_{x \in [x_{j-1}, x_j]} f(x)(x_j - x_{j-1})$$

these are called the *upper* and *lower* **Riemann Sums** For any partition, $U(f, P) \geq L(f, P)$.

Defn A.42 (Riemann Integrability).
A bounded function f defined on an interval $[a, b]$ is said to be Riemann integrable if

$$\inf_P U(f, P) = \sup_P L(f, P); \int_a^b f(x)dx =: \text{Reimann Integral}$$

Suppose f is piecewise continuous, defined on $[a, b]$. Then, f is Reimann integrable and

$$\int_a^b f(x)dx = \lim_{N \to \infty} \sum_{j=1}^{N} f\left(a + \frac{j}{N}(b - a)\right) \frac{b - a}{N}$$

Theorem A.16 (Fundamental Theorem of Calculus).
If f is continuous on $[a, b]$ then $F(x) := \int_a^x f(t)dt$ is differentiable on the open interval (a, b) and $F'(x) = f(x)\ \forall x \in (a, b)$. F is called the **anti-derivative**.

$$\int_a^b f(t)dt = F(b) - F(a)$$

Defn A.43 (Indicator Function).
Given a measurable space $(\Omega, \mathcal{F})$ and a set $E \in \mathcal{F}$, we define an indicator function $f_e : \Omega \to \mathbb{R}$ defined by

$$f_E(\omega) = \mathbb{1}_{\omega \in E}$$

This function is measurable.

Defn A.44 (Integration of Simple Functions).
Any function f of the form $f(\omega) \sum_{i=1}^{n} a_i \mathbb{1}_{\omega \in E_i}$ $\forall a_i \in \mathbb{R} \wedge E_1, \dots \mathbb{E}_n \in \mathcal{F}$ constitutes a finite partition of Ω. A countable sum of measurable functions is measurable, which implies that f is measurable. Then we define

$$\int f d\mu := \sum_{i=1}^{n} a_i \mu(E_i)$$

Defn A.45 (Lebesgue Integral / Integration of Measurable Functions).
For any measurable function f, define $f^+ := \max\{f, 0\}$ and $f^- := \max\{-f, 0\}$, which are also measurable. We also have $f = f^+ - f^-$ and $|f| = f^+ - f^-$. When either $\int f^- d\mu$ or $\int f^- d\mu$ is finite, we define the integral

$$\int f d\mu := \int f^+ d\mu - \int f^- d\mu$$

When both $\int f^- d\mu$ and $\int f^- d\mu$ are finite, we say f is integrable w.r.t. μ. Lebesgue integrals intuitively slice the function horizontally, while Reimann integrals slice vertically.

A.7 Probability Theory

Defn A.46 (Kolmogorov Axioms).
The triple $(\Omega, \mathcal{S}, P)$ is a probability space if it satisfies the following

- **Unitarity**: $\mathbf{Pr}(\Omega) = 1$
- **Non Negativity**: $\forall s \in S, \mathbf{Pr}(a) \geq 0$ $\mathbf{Pr}(a) \in \mathbb{R} \wedge \mathbf{Pr}(a) < \infty$
- **Countable Additivity**: If $A_1, A_2, \dots, \in S$ are *pairwise disjoint*[*i.e.* $\forall i \neq j, A_i \cap A_j = \emptyset$], Then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

Other properties for any event A, B

- $A \subset B \implies \mathbf{Pr}(A) \leq \mathbf{Pr}(B)$
- $\mathbf{Pr}(A) \leq 1$
- $\mathbf{Pr}(A) = 1 - \mathbf{Pr}(A^c)$
- $\mathbf{Pr}(\emptyset) = 0$

Stated differently:

Defn A.47 (Probability).
Let P be a probability measure on a measurable space $(\mathcal{E}, \mathcal{B})$, so $(\mathcal{E}, \mathcal{B}, P)$ is a probability space. Sets $B \in \mathcal{B}$ are called events, points $e \in \mathcal{E}$ are called outcomes, and $P(B)$ is called the probability of B.
Let $(\mathcal{E}, \mathcal{B})$ be a measurable space. Let $\mathbb{P} : \mathcal{B} \to [0, 1]$ be a 'set' function mapping the $\sigma-$ algebra of subsets of $\mathcal{E}$ into the real line. We say $\mathbb{P}$ is a probability measure if, for events $A, B \in \mathcal{E}$,

1. $0 \leq \mathbf{Pr}(A) \leq 1$: Events range from never happening to always happening
2. $\mathbf{Pr}(\mathcal{E}) = 1$: Something must happen
3. $\mathbf{Pr}(\emptyset) = 0$: Nothing never happens
4. $\mathbf{Pr}(A) + \mathbf{Pr}(A^c) = 1$: A must either happen or not happen
5. $\mathbf{Pr}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbf{Pr}(A_n)$: $\sigma-$ additivity for countable disjoint events
   - **Boole's Inequality** $\mathbf{Pr}(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty}$ for any sequence of events
6. **Monotonicity**: for events $A, B; A \subseteq B \implies \mathbf{Pr}(A) \leq \mathbf{Pr}(B)$

Defn A.48 (Random Variable).
A measurable function $X : \Omega \to \mathbb{R}$ s.t. $\forall r \in \mathbb{R}, \{\omega \in \Omega : X(\omega) \leq r\} \in \mathcal{E}$ (event space) is called a random variable. In other words, a random variable is a function from the sample space to the real line $\mathbb{R}$, and the probability of its value being in a given interval is well defined.

Defn A.49 (Probability Distribution).
The probability measure P_X living on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for any $A \in \mathcal{B}(\mathbb{R})$,

$$P_X(A) = P(\{e \in \mathcal{E} : X(e) \in A\}) =: P(X \in A)$$

for Borel sets A is called the *distribution* of X. The notation $X \sim Q$ is used to indicate that X has distribution $Q \implies P_X = Q$.

Defn A.50 (Cumulative Distribution Function).
Map $\mathbb{F} : \mathbb{R} \to [0,1]$ such that

$$F_X(x) = P(X \leq x) = P(\{e \in \mathcal{E} : X(e) \leq x\}) =: P_x((-\infty, x])$$

for $x \in \mathbb{R}$
Properties of CDF $\mathbb{F}(u)$

- **Boundary property**: $\lim_{u \to -\infty} \mathbb{F}(u) = 0, \lim_{u \to \infty} \mathbb{F}(u) = 1$
- **Nondecreasing**: $\mathbb{F}(x) \leq \mathbb{F}(y)$ if $x \leq y$
- **Right continuous**: $\lim_{u \downarrow x} \mathbb{F}(u) = \mathbb{F}(x)$

A.7.1 Densities

Theorem A.17 (Radon-Nikodym).
If a finite measure P is absolutely continuous wrt a $\sigma-$ finite measure μ, then $\exists$ a nonnegative measurable function f s.t.

$$P(A) = \int_A f d\mu =: \int f \mathbb{1}_A d\mu$$

The function f in this theorem is called the Radon-Nikodym derivative of P with respect to μ, or the density of P with respect to μ, denoted

$$f = \frac{dP}{d\mu}$$

Defn A.51 (Absolutely Continuous Random Variables).
If a random variable has density p wrt Lebesgue measure on $\mathbb{R}$, then X or its distribution P_X is called *absolutely continuous* with density p. Then, from R-N ,

$$F_X(x) = P(X \leq x) = P_X((-\infty, x]) = \int_{-\infty}^{x} p(u) du$$

Using the fundamental theorem of calculus, p can be found from the CDF F_X by differentiation, $p(x) = F'_X(x)$.

Defn A.52 (Discrete Random Variables:).
Let $\mathcal{X}_0$ be a countable subset of $\mathbb{R}$. The measure $\mu := \mu(B) = \#(\mathcal{X} \cap B)$ for borel sets B is also called *counting measure* on $\mathcal{X}_0$. Then,

$$\int f d\mu = \sum_{x \in \mathcal{X}_0} f(x)$$

Suppose X is a random variable s.t. $P(X \in \mathcal{X}_0) = P_X(\mathcal{X}_0 = 1)$. Then, X is called a *discrete random variable*.
The density p of P_X w.r.t. μ satisfies

$$P(X \in A) = P_X(A) = \int_A p d\mu = \sum_{x \in \mathcal{X}_0} p(x) \mathbb{1}_A(x)$$

In particular, if $A = \{y\}$ s.t. $y \in \mathcal{X}_0$, then $X \in A \Leftrightarrow X = y$, and so

$$P(X = y) = \sum_{x \in \mathcal{X}_0} p(x) \mathbb{1}_{\{y\}}(x) = p(y)$$

The density p is called the *mass function* for X.

A.7.2 Moments

Defn A.53 (Expectation:).
If X is a random variable on a probability space $(\mathcal{E}, \mathcal{B}, P)$, then the expectation of $X \sim P_X$ (i.e. density p), is defined as

$$\mathbb{E}[X] := \int X dP = \int x dP_X(x) = \int x p(x) d(x)$$

For discrete RV X with $P(X \in \mathcal{X}_0) = 1$ for a countable set $\mathcal{X}_0$, if μ is counting measure on $\mathcal{X}_0$, and p is the mass function given by $p(x) = P(X = x)$,

$$\mathbb{E}[X] := \int x dP_X(x) = \int x p(x) d\mu(x) = \sum_{x \in \mathcal{X}_0} x p(x)$$

Defn A.54 (Variance).
The *variance* of a random variable X with finite expectation is defined as

$$\mathbb{V}[X] = \mathbb{E}[X - \mathbb{E}[X]]^2$$

If X is absolutely continuous with density p,

$$\mathbb{V}[X] = \int (x - \mathbb{E}[X])^2 p(x) dx$$

If X is discrete with mass function p,

$$\mathbb{V}[X] = \sum_{x \in \mathcal{X}_0} (x - \mathbb{E}[X])^2 p(x)$$

A.7.3 Random vectors

If $X_1, \ldots, X_n$ are random variables, then the function $X : \mathcal{E} \to \mathbb{R}^n$ defined by

$$X(e) = \begin{pmatrix} X_1(e) \\ \vdots \\ X_n(e) \end{pmatrix}, e \in \mathcal{E}$$

is called a *random vector*. The definitions above extend naturally to random vectors, e.g. the distribution P_X of X is

$$P_X(B) = P(X \in B) := P(\{e \in \mathcal{E} : X(e) \in B\})$$

for Borel sets $B \in \mathbb{R}^n$. The expectation of a random vector X is the vector of expectations

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

A random vector is said to be absolutely continuous if the CDF can be written as

$$\mathbb{F}(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(z_1, \ldots, z_n) dz_1 \ldots dz_n$$

Defn A.55 (Random Matrices).
A matrix W is called a *random matrix* if the entries W_{ij} are random variables.

Defn A.56 (Covariance).
The *covariance* of a random vector X is the matrix of covariances of the variables in X

$$[\mathrm{Cov}(X)]_{ij} = \mathrm{Cov}(X_i, X_j)$$

If $\mu = \mathbb{E}[X]$ and $(X - \mu)'$ is the transpose of the mean deviation, then

$$\mathrm{Cov}(X_i, X_j) := \mathbb{E}[X_i - \mu_i](X_j \mu_j) = \mathbb{E}[(X - \mu)(X - \mu)']_{ij}$$

so

$$\mathrm{Cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)'] = \mathbb{E}[XX'] - \mu\mu'$$

A.7.4 Product Measures and Independence

Let $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$ be measure spaces. Then $\exists$ a unique product measure $\mu \times \nu$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \vee \mathcal{B})$ such that $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)\ \forall A \in \mathcal{A}, B \in \mathcal{B}$. The σ-field $\mathcal{A} \vee \mathcal{B}$ is defined formally as the smallest $\sigma-$ field containing all sets $A \times B$ with $A \in \mathcal{A}$, $B \in \mathcal{B}$.

Theorem A.18 (Fubini).
Integration against the product measure $\mu \times \nu$ can be accomplished by iterated integration against μ and ν, in either order

$$\int d(\mu \times \nu) = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y)$$

Defn A.57 (Independence of random variables).
Suppose $X_i : \Omega \to \mathbb{R}, 1 \leq i \leq m$ are random variables.
$X_1, X_2, \ldots, X_m$ are indepepndent for all $B_1, B_2, \ldots, B_m$ Borel subsets of $\mathbb{R}$, it is true that

$$\mathbf{Pr}(X_i \in B_i, \forall i\ 1 \leq i \leq m) = \mathbf{Pr}(X_1 \in B_1) \ldots \mathbf{Pr}(X_m \in B_m)$$

A.7.5 Conditional Expectations

Defn A.58 (Conditional Probability).
Given two r.v.s X, Y with finite second moments, $\mathbb{E}[Y|X]$ is defined as a $\sigma(X)-$ measurable function $m(X)$ such that

$$m(.) \operatorname*{argmin}_h \mathbb{E}_{\mathbb{P}}[Y - m(X)]^2$$

For continuous $X, Y \in \mathbb{R}^2$ with pdf $f(x, y)$, For any $y^* \in \mathbb{R}$, the conditional probability of the event $\{Y \leq y^*\} := \mathbb{P}_{Y|X}(Y \leq y^*|x)$ is defined as a function satisfying

$$\int_{-\infty}^{x^*} \mathbb{P}_{Y|X}(Y \leq y^*|x) f_X(dx) = \mathbb{P}_{XY}(X \leq x^*, Y \leq y^*)\ \forall x^* \in \mathbb{R}$$

The conditional CDF is

$$\mathbb{F}_{Y|X}(y^*) = \mathbb{P}_{Y|X}(Y \leq y^*|x) = \int_{-\infty}^{y^*} \underbrace{\frac{f(x, y)}{f_X(x)}}_{\text{conditional density } f(y|x)} dy$$

Defn A.59 (Bounded Lipshitz Distance).
Let $X \sim F_X; Y \sim F_Y$. Define the the class of 'Bounded Lipshitz functions' with Lipshitz constant 1 as

$$BL(1) := \left\{ h : \mathbb{R} \to \mathbb{R} \text{ s.t. } |h(x) - h(x)| \leq |x - y| \wedge \sup_{x \in \mathbb{R}} |h(x)| \leq 1 \right\}$$

Then the Bounded Lipshitz Distance is

$$d_{BL}(F_X, F_Y) := \sup_{h \in BL(1)} |\mathbb{E}_{F_X}[h(X)] - \mathbb{E}_{F_Y}[h(Y)]|$$

F_X and F_Y are said to be 'close' if $d_{BL}(F_X, F_Y)$ is small.

A.7.6 Order Statistics

Fact A.19 (Distribution of Order Statistics).
Suppose $X_1, X_2, \ldots, X_n$ are IID r.v.s with distribution $\mathbb{F}_x()$. To each $\omega \in \Omega$ define $\max\{X_1, \ldots, X_n\}(\omega) = \max\{X_1(\omega), \ldots, X_n(\omega)\}$. We want the distribution F of $\max\{X_1, \ldots, X_n\}$

$$\begin{aligned} G(r) &= \mathbf{Pr}(\{\omega \in \Omega; \max\{X_1, \ldots, X_n\} \leq r\}) \\ &= \mathbf{Pr}(\cap_{i=1}^n [X_i \leq r]) = \prod_{i=1}^N \mathbf{Pr}(X_i \leq r) \\ &= \prod_{i=1}^N F(r) = \mathbb{F}^n(r) \end{aligned}$$

If F has a density, G has a density too

$$g(r) = G'(r) = n\mathbb{F}^{n-1}(r)\mathsf{f}(r)$$

More generally, the distribution function of $X_{(m)}$ is given by $\mathbb{F}_{(m)}$

$$\begin{pmatrix} \mathbb{F}_{(m)}(t) = \sum_{i=m}^n \\ ni\mathbb{F}(t)^i(1-\mathbb{F}(t))^{n-i} \end{pmatrix}, -\infty < t < \infty$$

Severini (2005, chap 7).

Example A.20 (Max of n iid $\mathsf{U}[0,1]$ vars).
has a distribution with denisty $g(r) = nr^{n-1}$

Example A.21 (Distribution of second highest value Y^2).
(Useful for Vickrey auctions)

$$\begin{aligned} \mathbb{F}_{Y^2}(r) &= \mathbb{F}^m(r) + m(1-\mathbb{F}(r))\mathbb{F}^{m-1}(r) \\ \mathsf{f}_{Y^2}(r) &= m(m-1)(1-\mathbb{F}(r))\mathbb{F}^{m-2}(r)\mathsf{f}(r) \end{aligned}$$

A.8 Linear Functions and Linear Algebra

A.8.1 Linear Functions

Defn A.60 (Linear Function / Homomorphism).
A function $f : X \to Y$ between two linear spaces X, Y is linear if it preserves the linearity of sets X and Y through the following properties ($\forall\, \boldsymbol{x}_1, \boldsymbol{x}_2 \in X$)

- **Additivity** $f(\boldsymbol{x}_1 + \boldsymbol{x}_2) = f(\boldsymbol{x}_1) + f(\boldsymbol{x}_2)$
- **Homogeneity** $f(\alpha \boldsymbol{x}_1) = \alpha f(\boldsymbol{x}_1)$
- $f : V \to W$ linear and bijective is called an *isomorphism*
- A linear function that maps onto $\mathbb{R}$ is called a *linear functional*.
- A linear function that maps onto *itself* is called a *linear operator / automorphism*.

Every linear function mapping from a finite-dimensional domain X can be represented by a matrix.

Fact A.22 (Typology of Linear functions).
Linear Functions $f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x_2}) = \alpha f(\boldsymbol{x}_1 + (1-\alpha) f(\boldsymbol{x}_2)$ Imply

- **Additivity** : $f(\boldsymbol{x}_1 + \boldsymbol{x}_2 = f(\boldsymbol{x}_1) + f(\boldsymbol{x}_2)$ generalises to
  - **Convex Functions** $f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2)$ generalises to
  - **Quasiconvex functions** $f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \min(f(\boldsymbol{x}_1, \boldsymbol{x}_2)$
- **Homogeneity**: $f(\alpha \boldsymbol{x}) = \alpha f(\boldsymbol{x})$ generalises to
  - **Homogeneous functions** $f(\alpha \boldsymbol{x}) = \alpha^k f(\boldsymbol{x})$, $\alpha > 0$ generalises to
  - **Homothetic functions** $f(\boldsymbol{x}_1) = f(\boldsymbol{x}_2) \implies f(\alpha \boldsymbol{x}_1) = f(\alpha \boldsymbol{x}_2)$ $\alpha > 0$

Defn A.61 (Inner Product / Dot Product).
An inner product on a vector space V is a mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies, $\forall x, y, z \in V \wedge a\ in \mathbb{R}$

- $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$
- $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$
- $\langle x, ay \rangle = a \langle x, y \rangle$
- $\langle x, y \rangle = \langle y, x \rangle$

$$\boldsymbol{x}'\boldsymbol{y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_{i=1}^{n} x_i y_i$$

This generalises to an inner product of two functionals $u, v : \mathbb{R} \to \mathbb{R}$ where

$$\langle u, v \rangle = \int_a^b u(x)v(x)dx$$

An inner product defines a norm $\|v\| = \sqrt{\langle v, v \rangle}$. This gives us a restatement of the Cauchy-Schwartz inequality $|\langle x, y \rangle| \leq \|x\| \, \|y\|$

Defn A.62 (Orthogonal / Orthonormal Vectors).

1. $\langle x, y \rangle = 0 \implies x \perp y$ (**Orthogonality**). Furthermore, if $\|\boldsymbol{x}\| = 1 = \|\boldsymbol{y}\|$, then they are said to be *orthonormal*.
2. $\langle x, y \rangle = \pm 1 \implies$ x parallel to y

Defn A.63 (Symmetric, Positive Definite Matrices).
$\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, positive definite if $\forall \boldsymbol{x} \in V : \boldsymbol{x}'\mathbf{A}\boldsymbol{x} > 0$.

Defn A.64 (Hadamard Product).
For conformable matrices $\mathbf{A}$ and $\mathbf{B}$ with identical dimensions, the element-wise product

$$c_{ij} = a_{ij} \times b_{ij} \; \forall i, j \in \dim(A) \equiv \mathbf{A} \odot \mathbf{B}$$

is called the hadamard product.

Defn A.65 (Euclidian Norm).
Euclidian norm of a vector $\boldsymbol{x} \in \mathbb{R}^N$ is defined as

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

Fact A.23 (Angle between two vectors).
Angle between $\boldsymbol{u}$ and $\boldsymbol{v}$ is given by

$$\cos\theta = \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\| \, \|\boldsymbol{v}\|}$$

Theorem A.24 (Cauchy-Schwarz Inequality).
$\|\langle x, y \rangle\| \leq \|x\| \, \|y\|$

Defn A.66 (Trace).

$$\text{Trace}(A) = \sum_{n=1}^{N} a_{nn}$$

For conformable matrices $A, B; \text{tr}(AB) = \text{tr}(BA)$

Defn A.67 (Eigenvalues and Eigenvectors).
For a square matrix A, scalar λ and vector x that satisfies $Ax = \lambda x$ constitute an **eigenvalue** and **eigenvector** respectively.

Defn A.68 (Orthogonal / Orthonormal Matrix).
A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is *orthogonal* iff its columns are orthonormal so that

$$\mathbf{A}\mathbf{A}' = I = \mathbf{A}'\mathbf{A} \implies \mathbf{A}^{-1} = \mathbf{A}'$$

Defn A.69 (Kernel, Rank Nullity).

For $\mathbf{\Phi} : V \to W$, we define the *kernel/null space* as

$$\ker(\mathbf{\Phi}) := \mathbf{\Phi}^{-1}(\mathbf{0}_W) = \{\boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \mathbf{0}_W\}$$

and the *image/range*

$$\text{Im}(\mathbf{\Phi}) := \Phi(V) = \{\boldsymbol{w} \in W | \exists \boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \boldsymbol{w}\}$$

The dimension of the image is called the **rank** of $\mathbf{\Phi}$. The dimension of the kernel is called the **nullity** of $\mathbf{\Phi}$. If $\mathbf{X}$ is finite dimensional,

$$\text{rank } \mathbf{\Phi} + \text{nullity } \mathbf{\Phi} = \dim \mathbf{X}$$

This is the **rank-nullity result**
A linear function Φ has **full rank** if $\text{rank}\Phi(X) = \min\{\text{rank}X, \text{rank}Y\}$

Defn A.70.
Nonsingular Matrices
A matrix $\mathbf{A} \in \mathbb{R}_{n \times n}$ with columns $\mathbf{a}_1, \ldots \mathbf{a}_n$ is *non-singular* or *one-to-one* if

$$\mathbf{A} \text{ is one-to-one} \Leftrightarrow \mathbf{a_1}, \ldots \mathbf{a_n} \text{ is a basis} \Leftrightarrow \ker\mathbf{A} = \{\mathbf{0}\}$$

A.8.2 Projection

Defn A.71 (Projection).
Let V be a vector space and $U \subseteq V$ is a subspace of V. A linear mapping $\pi : V \to U$

is called a projection if $\pi^2 = \pi \circ \pi = \pi$. Since homeomorphisms can be expressed by a transformation matrix, projections can be represented as a **projection matrix** $\mathbf{P}_\pi$ with the property $\mathbf{P}_\pi^2 = \mathbf{P}$. *Projection matrices are always symmetric*.

Example A.25 (Projection onto general subspaces).
We look at orthogonal projections of vectors $\boldsymbol{y} \in \mathbb{R}^n$ onto lower dimensional subspaces $X \subseteq \mathbb{R}^n$ with $\dim(X) = m \geq 1$. Assume $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m)$ is an ordered basis of X. Therefore, any projection $\pi_U(\boldsymbol{x}) = \sum_{i=1}^m \lambda_i x_i = \mathbf{X}\boldsymbol{\lambda}$. The problem, then, is to find $\lambda_1, \dots, \lambda_m$ coordinates of the projection (with respect to basis X) where $\pi_U(\boldsymbol{x}) = \mathbf{B}\boldsymbol{\lambda}$ given $\mathbf{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_m] \in \mathbb{R}^{n \times m}$ and $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_m]^T \in \mathbb{R}^m$. The solution is the familiar OLS coef vector

$$\boldsymbol{\lambda} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$$

where $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is also called the **pseudo-inverse** of $\mathbf{B}$, which can be computed as long as $(\mathbf{X}'\mathbf{X})^{-1}$ is full rank.
The projection matrix is therefore

$$\mathbf{P}_\pi = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Defn A.72 (Gram-Schmidt Orthogonalisation).
Any basis $(\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ of an $n-$ dimensional vector space V can be transformed into an orthogonal/orthonormal basis $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ where $\text{span}[\boldsymbol{b}_1, \dots, \boldsymbol{b}_n] = \text{span}[\boldsymbol{u}_1, \dots, \boldsymbol{u}_n]$ as follows

$$\begin{aligned} \boldsymbol{u}_1 &:= \boldsymbol{b}_1 \\ \boldsymbol{u}_k &:= \boldsymbol{b}_k - \pi_{\text{span}[\boldsymbol{u}_1, \dots, \boldsymbol{u_{k-1}}]}(\boldsymbol{b}_k) \qquad\qquad k = 2, \dots, n \end{aligned}$$

where the kth basis vector $\boldsymbol{b}_k$ is projected onto the subspace spanned by the first $k-1$ constructed orthogonal vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{k-1}$.
This is the same as FWL Theorem, but older.

A.8.3 Matrix Decompositions

Defn A.73 (Spectral / Eigenvalue Decomposition).
A square matrix $\mathbf{A}$ admits to an eigen-decomposition if it can be factorised as $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$ where

- $\mathbf{Q}$ is a $n \times n$ matrix whose ith column is the eigenvector $\boldsymbol{q}_i$ of A (**orthogonal matrix**)
- $\boldsymbol{\Lambda}$ is a diagonal matrix with corresponding eigenvalues $\Lambda_{ii} = \lambda_i$

Fact A.26 (Orthogonal Matrices).
If $\mathbf{Q}, \mathbf{N}$ are $N \times N$ orthogonal matrices

- $\mathbf{Q}^T = \mathbf{Q}^{-1}$ is also orthogonal
- $\mathbf{QN}$ is orthogonal
- $det(\mathbf{Q}) \in \{-1, 1\}$

Defn A.74 (Cholesky Decomposition).
If $\mathbf{A}$ is positive definite, then it admits to

- $\mathbf{A} = \mathbf{R}^T\mathbf{R}$ where $\mathbf{R}$ is non-singular **upper triangular**
- $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{L}$ is non-singular **lower triangular**

Defn A.75 (QR Decomposition).
If $\mathbf{A}$ is a $N \times K$ matrix with full column rank, $\exists$ a factorisation $\mathbf{A} = \mathbf{QR}$ where

- $\mathbf{Q}$ is an orthogonal matrix
- $\mathbf{R}$ is $K \times K$ upper triangular and nonsingular (invertible)

Example A.27 (QR Decomposition for OLS).
$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ is often numerically unstable, so we can define $\mathbf{X} = \mathbf{QR}$. The, the OLS estimate can be written as $\hat{\boldsymbol{\beta}} = (\mathbf{R})^{-1}\mathbf{Q}'\boldsymbol{y}$. The homoscedastic variance is $\mathbb{V}\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{R}'\mathbf{R})^{-1}\sigma^2$.
Using the same decomposition, $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{QQ}'\boldsymbol{y}$.

Defn A.76 (Singular Value Decomposition).
Any $n \times p$ matrix $\mathbf{Z}$ may be written as

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'$$

where $\mathbf{U}$ is a $n \times n$ orthogonal matrix, $\mathbf{V}$ is a $p \times p$ orthogonal matrix, and $\boldsymbol{\Sigma}$ is a $n \times p$ diagnonal matrix with non-negative elements.

Example A.28 (SVD of covariance matrix equivalence with spectral decomposition).
For a square covariance matrix $\mathbf{X}'\mathbf{X}$, if $\mathbf{X} = \mathbf{USV}'$, then

$$\mathbf{X}'\mathbf{X} = \mathbf{VS}^T\mathbf{U}^T\mathbf{USV}^T = \mathbf{VDV}'$$

where $\mathbf{D} = \mathbf{S}^2$ contains the square singular values. In other words,

$$\mathbf{U} = \text{evec}(\mathbf{XX}'),\ \mathbf{V} = \text{evec}(\mathbf{X}'\mathbf{X}),\ \mathbf{S}^2 = \text{eval}(\mathbf{X}'\mathbf{X}) = \text{eval}(\mathbf{XX}')$$

A.8.4 Matrix Identities

For conformable matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$,

$$\begin{aligned}
\mathbf{A}(\mathbf{B}+\mathbf{C}) &= \mathbf{AB}+\mathbf{AC} \\
\mathbf{A}+\mathbf{B}^\top &= \mathbf{A}^\top+\mathbf{B}^\top \\
\mathbf{AB}^\top &= \mathbf{B}^\top\mathbf{A}^\top \\
(\mathbf{AB})^{-1} &= (\mathbf{B})^{-1}(\mathbf{A})^{-1} \\
\operatorname{trace}(\mathbf{ABC}) &= \operatorname{trace}(\mathbf{CBA}) = \operatorname{trace}(\mathbf{BCA})
\end{aligned}$$

A.8.5 Partitioned Matrices

Defn A.77 (Partitioned Matrices).
It can be useful to partition a matrix as follows

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

Multiplying a partitioned matrix with a stacked vector $\mathbf{c}$

$$\mathbf{Xc} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} X_{11}c_1 + X_{12}c_2 \\ X_{21}c_1 + X_{22}c_2 \end{bmatrix}$$

Fact A.29 (Inverse of 2×2 partitioned matrix).

$$\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}^{-1} = \begin{bmatrix} X_{11}^{-1} + X_{11}^{-1}X_{12}F_2X_{21}(X_{11})^{-1} & -(X_{11})^{-1}X_{12}F_2 \\ -F_2X_{21}(X_{11})^{-1} & F_2 \end{bmatrix}$$

where $F_2 = \left(X_{22} - X_{21}(X_{11})^{-1}X_{12}\right)^{-1}$

A.9 Function Spaces

Almost all of this is based on / stolen from Larry Wasserman
http://www.stat.cmu.edu/ larry/=sml/functionspaces.pdf and Racine, Su, and Ullah (2013).

Defn A.78 (Function spaces).
Let U be any set, let bU be the collection of all bounded functions s.t. $f : U \to \mathbb{R}$ (i.e. $\sup_{x\in U} |f(x)| < \infty$ and let

$$d_\infty(f,g) := \|f-g\|_\infty := \sup_{x\in U} |f(x)-g(x)|$$

Spaces of functions can be treated as linear vector spaces

Defn A.79 (Inner Product and Norm in function spaces).

$$\langle f, g\rangle = \int f(x)g(x)dx$$

which leads to a norm for functions

$$\|f\|_2^2 = \int f^2(x)dx$$

Defn A.80 (Eigenvalues and Eigenfunctions).
An operator $\mathcal{O}$ is a higher-order function that maps from one function to another. A derivative and integral are both operators.
Operators can have eigenvalues and eigenfunctions such that

$$\mathcal{O}f = \lambda f$$

$\exp ax$ is an eigenfunction for both differentiation and integration.

Defn A.81 (Hilbert Space $\mathcal{H}$).
is a complete ($:=$ every Cauchy sequence in the space converges to a point in it), inner product space. Equivalently, it is a *vector space endowed with an inner product and an associated norm and metric such that every Cauchy sequence has a limit in* $\mathcal{H}$.
Intuitively, it means it doesn't have any 'holes' in it ($\mathbb{Q}$ is *not* a complete space because $\sqrt{2}$ is missing from it).
Every Hilbert space is a Banach space but the reverse is not true in general. In a hilbert space, $\|f_n - f\| \to 0$ as $n \to \infty$.
If V is a hilbert space and L is a closed subspace then $\forall v, \exists y \in L$ called a projection of v onto L that minimises $\|v-z\|$ over $z \in L$. The set of elements orthogonal every $z \in L$ is denoted $L^\perp$. Every $v \in L$ can be written as $v = w + z$ where z is the projection of v onto L and $w \in L^\perp$.

Example A.30 ($\mathcal{R}$).
, the set of random variables defined on a common probability space $\{\Sigma, \mathcal{F}, \mu\}$ is a Hilbert space with inner product $\langle X, Y\rangle = \mathbb{E}[XY]$, associated norm $\|X\| = \sqrt{\mathbb{E}[X^2]}$ and metric $\|X - Y\|$.

Example A.31 ($L^2(w)$).
the space of Borel-measurable real functions f on $\mathbb{R}$ given density $w(x)$ satisfying $\int_{-\infty}^{\infty} f(x)^2 w(x)\mathrm{d}x < \infty$ and associated norm $\|f\| = \sqrt{\langle f, f\rangle}$ and metric $\|f - g\|$ is a hilbert space.

Defn A.82 (Orthogonal / Direct Sum).
If L and M are spaces such that every $\ell \in L$ is orthogonal to every $m \in M$, then we define the orthogonal sum as

$$L \oplus M = \{l + m : l \in L, m \in M\}$$

A set of vectors $\{e_t, t \in T\}$ is *orthonormal* if $\langle e_s, e_t \rangle = 0$ when $s \neq t$ and $\|e_t\| = 1 \forall T$. This is also called an **orthonormal basis**. **Every hilbert space has an orthonormal basis. A Hilbert space is said to be *separable* if there exists a countable orthonormal basis**.

A.9.1 L_p spaces

Let $\mathcal{F}$ be a collection of functions $[a, b] \mapsto \mathbb{R}$. The L_p norm on $\mathcal{F}$ is defined by

$$\|f\|_p = \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}$$

For $p = \infty$, we define the sup norm $\|f\|_\infty = \sum_x |f(x)|$.
The space $L_p(a, b)$ is defined as

$$L_p(a, b) := \left\{ f : [a, b] \to \mathbb{R} : \|f\|_p < \infty \right\}$$

Every L_p space is a Banach Space.

- Cauchy Schwartz: $\left(\int f(x)g(x)dx \right)^2 \leq \int f^2(x)dx \int g^2(x)dx$
- Minkowski : $\|f + g\|_p \leq \|f\|_p + \|g\|_p$ where $p > 1$
- Holder: $\|fg\|_1 \leq \|f\|_p \|g\|_q$ where $(1/p) + (1/q) = 1$

Example A.32 (L_2 space).
Functions where $\|f\|_2^2 < \infty$ are said to be *square-integrable*, and the space of square-integrable functions is called L_2. Many familiar results from vector spaces carries over into L_2.
$L_2(a, b)$ is a Hilbertspace. The *inner product* between two functions $f, g \in L_2(a, b)$ is $\int_a^b f(x)g(x)dx$ and the norm of f is $\|f\|^2 = \int_a^b f^2(x)dx$. With this inner product, $L_2(a, b)$ is a separable Hilbert space; that is, we can find a countable orthonormal basis $\phi_1, \phi_2, \ldots;$, that is $\|\phi_j\| = 1 \; \forall j$, and $\int_a^b \phi_i(x)\phi_j(x) = 0 \; \forall i \neq j$. It follows that if $f \in L_2(a, b)$,

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x) \;\; \text{where } \theta_j = \int_a^b f(x)\phi_j(x)dx$$

are the coefficients. Parseval's identity $\int_a^b f^2(x)dx = \sum_{j=1}^{\infty} \theta_j^2$.
The **span** of L_2 is

$$\left\{ \sum_{j=1}^{\infty} a_j \phi_j(x) : a_1, \ldots, a_n \in \mathbb{R} \right\}$$

The projection of $f = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$ onto the span $\{\phi_1, \ldots, \phi_n\}$ is $f_n = \sum_{j=1}^{n} \theta_j \phi_j(x)$, which we call the **n-term linear approximation of** f.

Defn A.83 (Bases in function space).
A sequence of functions $\psi_1, \ldots$ can be considered a basis. An orthonormal basis is one that admits to

$$f = \sum_{j=1}^{\infty} \langle f, \psi_j \rangle \, \psi_j$$

Mononomials $1, x, x^2, \ldots$ are a basis for L_2 on $[0, 1]$ and $\mathbb{R}$, but they aren't orthogonal.

Example A.33 (Famous Bases).
A popular basis for L_2 on $[0, 1]$ are the sines and cosines, which may be written as $\phi_1 = 1$, $\phi_{2k} = \sin 2k\pi x$, $\phi_{2k+1} = \cos 2k\pi x$. Coefficients in this expansion are referred to as the *Fourier transform* of the original function.
A **cosine basis** on $[0, 1]$ is

$$\phi_0(x) = 1, \;\; \phi_j(x) = \sqrt{2} \cos(2\pi j x) \;\; , j = 1, 2, \ldots$$

Legendre basis on $(-1, 1)$ is

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}\left(3x^2 - 1\right), \quad P_3(x) = \frac{1}{2}\left(5x^3 - 3x\right), \ldots$$

The Haar basis on [0,1] consists of functions

$$\left\{ \phi(x), \psi_{jk}(x) : j = 0, 1, \ldots, k = 0, 1, \ldots, 2^j - 1 \right\}$$

where

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$\psi_{jk}(x) = 2^{j/2}\psi\left(2^j x - k\right)$ and

$$\psi(x) = \begin{cases} -1 \text{ if } 0 \le x \le \frac{1}{2} \\ 1 \text{ if } \frac{1}{2} < x \le 1 \end{cases}$$

This is a doubly indexed set of functions so when f is expanded in this basis we write

$$f(x) = \alpha\phi(x) + \sum_{j=1}^{\infty}\sum_{k=1}^{2^j-1}\beta_{jk}\psi_{jk}(x)$$

where $\alpha = \int_0^1 f(x)\phi(x)dx$ and $\beta_{jk} = \int_0^1 f(x)\psi_{jk}(x)dx$. The Haar basis is an example of a wavelet basis.

Defn A.84 (Holder Spaces).
Let β be a positive integer. Let $T \subset \mathbb{R}$. The Holder space $H(\beta, L)$ is the set of functions $g : T \to \mathbb{R}$ such that

$$\left|g^{(\beta-1)}(y) - g^{(\beta-1)}(x)\right| \le L|x-y|, \quad \text{for all } x, y \in T$$

The special case $\beta = 1$ is sometimes called the Lipschitz space. If $\beta = 2$ then we have

$$|g'(x) - g'(y)| \le L|x-y|, \quad \text{for all } x, y$$

Roughly speaking, this means that the functions have bounded second derivatives.

Multivariate version There is also a multivariate version of Holder spaces. Let $T \subset \mathbb{R}^d$. Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d, s! = s_1! \cdots s_d!, x^s = x_1^{s_1} \cdots x_d^{s_d}$ and

$$D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$$

The Hölder class $H(\beta, L)$ is the set of functions $g : T \to \mathbb{R}$ such that

$$|D^s g(x) - D^s g(y)| \le L\|x-y\|^{\beta-|s|}$$

for all x, y and all s such that $|s| = \beta - 1$

Defn A.85 (Sobolev Space).
A Sobolev space is a space of functions possessing sufficiently many derivatives for some application domain. Formally,

Let f be integrable on every bounded interval. Then f is weakly differentiable if there exists a function f' that is integrable on every bounded interval, such that $\int_x^y f'(s)ds = f(y) - f(x)$ whenever $x \le y$. We call f' the weak derivative of f. Let $D^j f$ denote the j^{th} weak derivative of f
The Sobolev space of order m is defined by

$$W_{m,p} = \{f \in L_p(0,1) : \|D^m f\| \in L_p(0,1)\}$$

The Sobolev ball of order m and radius c is defined by

$$W_{m,p}(c) = \left\{f : f \in W_{m,p}, \|D^m f\|_p \le c\right\}$$

Defn A.86 (Mercer Kernel and Theorem).
A Mercer kernel is a continuous function $K : [a,b] \times [a,b] \to \mathbb{R}$ such that $K(x,y) = K(y,x)$, and such that K is positive semidefinite, meaning that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j)\, c_i c_j \ge 0$$

for all finite sets of points $x_1, \ldots, x_n \in [a,b]$ and all real numbers $c_1, \ldots, c_n$. The function

$$K(x,y) = \sum_{k=1}^{m-1}\frac{1}{k!}x^k y^k + \int_0^{x\wedge y}\frac{(x-u)^{m-1}(y-u)^{m-1}}{(m-1)!^2}du$$

is an example of a Mercer kernel. The most commonly used kernel is the Gaussian kernel

$$K(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$$

Defn A.87 (Reproducing Kernel Hilbert Spaces).
Given a kernel K, let $K_x(\cdot)$ be the function obtained by fixing the first coordinate. That is, $K_x(y) = K(x,y)$. For the Gaussian kernel, K_x is a Normal, centered at x. We can create functions by taking liner combinations of the kernel:

$$f(x) = \sum_{j=1}^{k}\alpha_j K_{x_j}(x)$$

Let $\mathcal{H}_0$ denote all such functions:

$$\mathcal{H}_0 = \left\{ f : \sum_{j=1}^{k} \alpha_j K_{x_j}(x) \right\}$$

Given two such functions $f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x)$ and $g(x) = \sum_{j=1}^{m} \beta_j K_{y_j}(x)$ we define an inner product

$$\langle f, g \rangle = \langle f, g \rangle_K = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j)$$

In general, $f($ and $g)$ might be representable in more than one way. You can check that $\langle f, g \rangle_K$ is independent of how $f($ or $g)$ is represented. The inner product defines a norm:

$$\|f\|_K = \sqrt{\langle f, f, \rangle} = \sqrt{\sum_j \sum_k \alpha_j \alpha_k K(x_j, x_k)} = \sqrt{\alpha^T \mathbb{K} \alpha}$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)^T$ and $\mathbb{K}$ is the $k \times k$ matrix with $\mathbb{K}_{jk} = K(x_j, x_k)$

The Reproducing Property .
Let $f(x) = \sum_i \alpha_i K_{x_i}(x)$. Note the following crucial property:

$$\langle f, K_x \rangle = \sum_i \alpha_i K(x_i, x) = f(x)$$

This follows from the definition of $\langle f, g \rangle$ where we take $g = K_x$. This implies that

$$\langle K_x, K_x \rangle = K(x, x)$$

This is called the reproducing property. It also implies that K_x is the representer of the evaluation functional.
The completion of $\mathcal{H}_0$ with respect to $\| \cdot \|_K$ is denoted by $\mathcal{H}_K$ and is called the RKHS generated by K.
Evaluation Functionals. A key property of RKHS's is the behavior of the evaluation functional. The evaluation functional δ_x assigns a real number to each function. It is defined by $\delta_x f = f(x)$. In general, the evaluation functional is not continuous. This means we can have $f_n \to f$ but $\delta_x f_n$ does not converge to $\delta_x f$. For example, let $f(x) = 0$ and $f_n(x) = \sqrt{n} I\left(x < 1/n^2\right)$. Then $\|f_n - f\| = 1/\sqrt{n} \to 0$. But $\delta_0 f_n = \sqrt{n}$ which does not converge to $\delta_0 f = 0$. Intuitively, this is because Hilbert spaces can contain very unsmooth functions.

But in an RKHS, the evaluation functional is continuous. Intuitively, this means that the functions in the space are well-behaved. To see this, suppose that $f_n \to f$. Then

$$\delta_x f_n = \langle f_n K_x \rangle \to \langle f K_x \rangle = f(x) = \delta_x f$$

so the evaluation functional is continuous.
A Hilbert space is a RKHS if and only if the evaluation functionals are continuous.

Theorem A.34 (Representer Theorem).
Let ℓ be a loss function depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$ and on $f(X_1), \ldots, f(X_n)$. Let $\widehat{f}$ minimize

$$\ell + g\left(\|f\|_K^2\right)$$

where g is any monotone increasing function. Then $\widehat{f}$ has the form

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

for some $\alpha_1, \ldots, \alpha_n$

A.10 Calculus and Optimisation

A.10.1 Calculus

Defn A.88 (Derivative).
The derivative of a function f at point x, when defined, is the tangent to the function at x.

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Defn A.89 (Gradient, Jacobian, Hessian).
For function $f : \mathbb{R}^m \to \mathbb{R}$, we define the

$$\nabla_{\boldsymbol{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

which collects the partial derivatives in a column vector.
The matrix of partial derivatives of f is called the *Hessian*, denoted by $\mathsf{H}(\mathbf{x})$

$$\mathsf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

For a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, we can construct the **Jacobian**, which collects all $m \times n$ partial derivatives.

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \nabla f_1(\boldsymbol{x}) \\ \nabla f_2(\boldsymbol{x}) \\ \vdots \\ \nabla f_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\boldsymbol{x}) & \frac{\partial}{\partial x_2} f_1(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\boldsymbol{x}) \\ \frac{\partial}{\partial x_1} f_2(\boldsymbol{x}) & \frac{\partial}{\partial x_2} f_2(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_n} f_2(\boldsymbol{x}) \\ & \cdots & & \\ \frac{\partial}{\partial x_1} f_m(\boldsymbol{x}) & \frac{\partial}{\partial x_2} f_m(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\boldsymbol{x}) \end{bmatrix}$$

Theorem A.35 (Taylor's theorem).
$f : \mathbb{R} \to \mathbb{R}$ admits to Taylor expansion around a such that

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 \cdots = \sum_{n=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n$$

For a function with multiple arguments $f : \mathbb{R}^k \to \mathbb{R}$, the second-order Taylor expansion around the point $\boldsymbol{x}_0$ is

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)\mathsf{H}(\mathbf{x})(\mathbf{x} - \mathbf{x}_0)$$

Fact A.36 (Sufficient Conditions for Local Maxima and Minima).
Let $f(\mathbf{x})$ have continuous first and second order partial derivatives in the $\varepsilon-$ neighbourhood of the optimum $\mathbf{x}_0$.

- If $\nabla f(\mathbf{x}_0) = 0$ and $\mathsf{H}(\mathbf{x}_0)$ is *positive definite*, then $\mathbf{x_0}$ is a **local minimum**.
- If $\nabla f(\mathbf{x}_0) = 0$ and $\mathsf{H}(\mathbf{x}_0)$ is *negative definite*, then $\mathbf{x_0}$ is a **local maximum**.

Theorem A.37 (Generalised (Everett) Lagrange Multiplier Theorem).
Let $\lambda_1, \ldots, \lambda_m$ be nonnegative real numbers, and suppose $\mathbf{x}_0$ maximises the *Lagrangian* $M(\mathbf{x}, \boldsymbol{\lambda})$

$$M(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{j=1}^{m} \lambda_j g_j(\mathbf{x})$$

Then, $\mathbf{x}_0$ maximises $f(\mathbf{x})$ subject to constraints ($x \in \mathcal{S}$)

$$g_j(\mathbf{x}) \leq g_j(\mathbf{x}_0) \quad , j = 1, \ldots m$$

Theorem A.38 (Inverse Function Theorem).
If $\varphi : U \to \mathbb{R}^d$ is differentiable at a and D_{φ_a} is invertible, then $\exists U', V'$ such that $a \in U' \subseteq U, \varphi(a) \in V' \wedge \varphi : U' \to V'$ is bijective. Further, the inverse function $\psi : V' \to U'$ is differentiable.

Theorem A.39 (Implicit function theorem).
Let $U \subseteq \mathbb{R}^{d+1}$ be a domain and $f : U \to \mathbb{R}$ be a differentiable function. If $x \in \mathbb{R}^d \wedge y \in \mathbb{R}$, we'll concatenate the two vectors and write $(x, y) \in \mathbb{R}^{d+1}$.
Suppose $c = f(a, b)$, and $\partial_y f(a, b) \neq 0$. Then, $\exists U' \ni a \wedge$ differentiable function $g : U' \to \mathbb{R}$ s.t. $g(a) = b \wedge f(x, g(x)) = c \; \forall \, x \in U'$.
Further, $\exists V' \ni b$ s.t. $\{(x, y) | x \in U', y \in V', f(x, y) = c\} = \{(g, g(x)) | x \in U'\}$. IoW, $\forall x \in U', f(x, y) = c$ has a unique solution $y = g(x) \in V'$.

Fact A.40 (Differentiating implicit fns using tangent planes).
Let $f : \mathbb{R}^2 \to \mathbb{R}$ be differentiable and consider the implicitly defined curve

$$\Gamma := \left\{(x, y) \in \mathbb{R}^2 | f(x, y) = c\right\}$$

(i.e. a level set of f). Pick $(a, b) \in \Gamma$, suppose $\partial_y f(a, b) \neq 0$. By IFT, we know $y-$coordinate of this curve can locally be expressed as a differentiable function of x. Directly differentiating $f(x, y) = c$ w.r.t. x gives

$$\partial_x f + \partial_y f \frac{dy}{dx} = 0 \Leftrightarrow \frac{dy}{dx} = \frac{-\partial_x f(a, b)}{\partial_y f(a, b)}$$

Differentiation Rules

Function Rules	$f(x)$	$f'(x)$
	x^a	ax^{a-1}
	e^x	e^x
	$\log x$	$\frac{1}{x}$
Linear Rule	$(af+bg)$	$a\frac{\partial f}{\partial x}+b\frac{\partial g}{\partial x}$
Product Rule	$(f\cdot g)'$	$f'(x)g(x)+f(x)g'(x)$
Quotient Rule	f/g	$\frac{f'(x)g(x)-f(x)g'(x)}{(g(x))^2}$
Chain Rule	$f(g(x))$	$\frac{\partial f}{\partial g}\frac{\partial g}{\partial x}$

Matrix Derivatives Let $\boldsymbol{a}, \boldsymbol{x} \in \mathbb{R}^n$, and $\mathbf{A}$ be a conformable matrix

- $\frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$
- $\frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}'} = \boldsymbol{a}'$
- —
- $\frac{\partial}{\partial \boldsymbol{x}'}\mathbf{A}\boldsymbol{x} = \mathbf{A}$
- $\frac{\partial}{\partial \boldsymbol{x}}\mathbf{A}\boldsymbol{x} = \mathbf{A}'$
- $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}'\mathbf{A}\boldsymbol{x} = (\mathbf{A}+\mathbf{A}')\boldsymbol{x}$
- $\frac{\partial}{\partial \mathbf{A}}\boldsymbol{x}'\mathbf{A}\boldsymbol{x} = \boldsymbol{x}\boldsymbol{x}'$
- $\frac{\partial}{\partial \mathbf{A}}\log|\mathbf{A}| = (\mathbf{A}')^{-1}$

General Results from Optimisation Theory Luenberger (1997) and Rustagi (2014)

- **Projection Theorem** - In $\mathbb{R}^k$, the shortest line from a point to the plane is funished by the *perpendicular* from the point to the plane. Core idea carries through to higher dimensions and infinite-dimensional Hilbert Space
- **Hahn-Banach Theorem**: given a sphere and a point not in the sphere, there exists a hyperplane separating the point and the sphere.
- **Duality**: The **shortest** distance from a point to a convex set is equal to the **maximum** of the distances from the point to a hyperplane separating the point from the convex set.
- **Differentials**: Set derivative of the objective function to zero.

A.10.2 Linear Programming

Maximise

$$\max_{\mathbf{x}} \quad Z = \mathbf{c}^\top \mathbf{x}$$

subject to

$$\begin{aligned} \mathbf{A}\mathbf{x} &\leq \mathbf{b} \\ \mathbf{x} &\geq 0 \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the choice vector, $\mathbf{c} \in \mathbb{R}^n$ is a given vector, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a known matrix of constants, and $\mathbf{b} \in \mathbb{R}^m$ is a vector of constants.

Defn A.90 (Standard form of Linear programs).
By introducing m *slack* variables $y_1, \ldots, y_m$, $\mathbf{y} = (y_1, \ldots, y_m)^\top$ for every inequality with $\mathbf{y} \geq 0$, we can convert every linear programming into its *standard form*

$$\begin{aligned} \min_{\mathbf{x}} \quad & Z = (\mathbf{c}^\top \mathbf{x}^\top + \mathbf{0}^\top \mathbf{y}) \text{ subject to} \\ \mathbf{A}\mathbf{x} + \mathbf{I}\mathbf{y} &= \mathbf{b} \\ \mathbf{x} \geq 0, \mathbf{y} &\geq 0 \end{aligned}$$

Defn A.91 (Primal and Dual).
Primal

$$\begin{aligned} \min_{x} \quad & \mathbf{c}^\top \mathbf{x} \text{ s.t.} \\ \mathbf{A}\mathbf{x} &\geq b \\ \mathbf{x} &\geq 0 \end{aligned}$$

Dual

$$\max_y \ \mathbf{b}^\top \mathbf{y} \ \text{ s.t.}$$
$$\mathbf{A}^\top \mathbf{y} \leq c$$
$$\mathbf{y} \geq 0 \ \ , \mathbf{y} \in \mathbb{R}^{n+m}$$

Theorem A.41 (Duality Theorem).
A feasible solution $\mathbf{x}_0$ to the primal is optimal IFF there exists a feasible solution $\mathbf{y}_0$ to the dual problem such that

$$\mathbf{c}^\top \mathbf{x_0} = \mathbf{b}^\top \mathbf{y_0}$$

Dantzig's Simplex method, Karmarkar's Algorithm.

A.10.3 Nonlinear Optimisation

$$\text{minimise[maximise]} \ \ f(\mathbf{x}) \ \text{ s.t.}$$
$$g_i(\mathbf{x}) \leq a_i \ \ i = 1, \ldots, k$$
$$\mathbf{x} \geq 0$$

Saddle Point Suppose we have $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\phi(\cdot)$ is a real valued function. Then, $(\mathbf{x_0}, \mathbf{y_0}), \mathbf{x_0} \geq \mathbf{0}, \mathbf{y_0} \geq \mathbf{0}$ is a *saddle-point* of $\phi(\mathbf{x}, \mathbf{y})$ if

$$\phi(\mathbf{x_0}, \mathbf{y}) \leq \phi(\mathbf{x_0}, \mathbf{y_0}) \leq \phi(\mathbf{x}, \mathbf{y_0})$$

$\forall \mathbf{x}, \mathbf{y} \geq \mathbf{0}$.

$$\phi(\mathbf{x_0}, \mathbf{y_0}) = \min_{\mathbf{x}} \max_{\mathbf{y}} \ \phi(\mathbf{x}, \mathbf{y})$$
$$\phi(\mathbf{x_0}, \mathbf{y_0}) = \max_{\mathbf{y}} \min_{\mathbf{x}} \ \phi(\mathbf{x}, \mathbf{y})$$

Defn A.92 (Quadratic Program).

$$Q = \mathbf{a}^\top \mathbf{x} - \frac{1}{2} \mathbf{x}^\top \mathbf{B} \mathbf{x} \ \text{ s.t.}$$
$$\mathbf{C}^\top \mathbf{x} \leq \mathbf{d}$$
$$\mathbf{x} \geq 0$$

where $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ is symmetric, positive definite, $\mathbf{C} \in \mathbb{R}^{n \times k}$ is a matrix of constraints, and $\mathbf{d} \in \mathbb{R}^k$.

Constrained Maximisation

Fact A.42 (General Proposition).
We want to maximise $f(x) \text{ subject to } g(x) = c$ (implicitly defined by $S := \{g = c\}$). Suppose $\nabla g \neq 0 \forall x \in S$. If f attains a constrained local maximum (or minimum) at a on the surface S, $\exists \lambda \in \mathbb{R}$ s.t. $\nabla f(a) = \lambda \nabla g(a)$.

Generic problem of the form

Defn A.93 (Lagrangian).

$$\max_{x_1, x_2 \in \mathbb{R}^n} f(x_1, x_2) \text{ s.t. } g(x_1, x_2) = b$$

First, write

$$\mathcal{L}(x_1, x_2, \lambda) = f(x_1, x_2) + \lambda \left[g(x_1, x_2) - b \right]$$

differentiating wrt x_1, x_2, λ yields FOCs

$$[x_1] : \frac{\partial \mathcal{L}}{\partial x_1} = f_1(x_1, x_2) + \lambda g_1(x_1, x_2) = 0$$
$$[x_2] : \frac{\partial \mathcal{L}}{\partial x_2} = f_2(x_1, x_2) + \lambda g_2(x_1, x_2) = 0$$
$$[\lambda] : \frac{\partial \mathcal{L}}{\partial \lambda} = g(x_1, x_2) - b = 0$$

which gives us three (potentially nonlinear) equations with three unknowns (x_1, x_2, λ), that can be solved simultaneously.
To check sufficiency, the second-order condition analogue is the determinant of the **bordered hessian matrix**

$$\text{BH}(x_1, x_2, \lambda) = \begin{bmatrix} 0 & -g_1(x_1, x_2) & -g_2(x_1, x_2) \\ -g_1(x_1, x_2) & f_{11}(x_1, x_2) & f_{12}(x_1, x_2) \\ -g_2(x_1, x_2) & f_{21}(x_1, x_2) & f_{22}(x_1, x_2) \end{bmatrix}$$

If $\det \text{BH} > 0$, then it is negative definite, which implies that the (x_1^*, x_2^*) that solves the system is indeed a local maximum.

Defn A.94 (Hessian, Definiteness).
The *Hessian* for of a C^2 [twice differentiable] function $f : \mathbb{R}^n d \to \mathbb{R}$ is defined by the matrix

$$\mathrm{H}f = \begin{bmatrix} \partial_1\partial_1 f & \partial_2\partial_1 f & \dots & \partial_d\partial_1 f \\ \partial_1\partial_2 f & \partial_2\partial_2 f & \dots & \partial_d\partial_2 f \\ \vdots & \vdots & \dots & \vdots \\ \partial_1\partial_d f & \partial_2\partial_d f & \dots & \partial_d\partial_d f \end{bmatrix}$$

- If $(\mathbf{A}\boldsymbol{v}) \cdot \boldsymbol{v} \leq 0 \; \forall v \in \mathbb{R}^d$, $\mathbf{A}$ is said to be *negative semi-definite*.
- If $(\mathbf{A}\boldsymbol{v}) \cdot \boldsymbol{v} < 0 \; \forall v \in \mathbb{R}^d$, $\mathbf{A}$ is said to be *negative definite*.
- If $(\mathbf{A}\boldsymbol{v}) \cdot \boldsymbol{v} \geq 0 \; \forall v \in \mathbb{R}^d$, $\mathbf{A}$ is said to be *positive semi-definite*.
- If $(\mathbf{A}\boldsymbol{v}) \cdot \boldsymbol{v} > 0 \; \forall v \in \mathbb{R}^d$, $\mathbf{A}$ is said to be *positive definite*.

Numerical Optimisation

Root-finding We want to evaluate the roots of the equation

$$y = f(x) = 0, \; x \in \mathbb{R}$$

Assume the inverse of f, denoted f^{-1} exists.

$$x = f^{-1}(y) = g(y)$$

Finding the root of $f(x) = 0$ is equivalent to evaluating $g(0) = x$.
Canonical newton-raphson is

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Quasi-Newton General version of update rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \lambda_k \cdot \mathbf{A}_k \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_k)$$

Step length $\lambda = 1$ for both N-R and BHHH.

Defn A.95 (Newton Raphson).
set $\mathbf{A}_k = (\mathsf{H}(\theta))^{-1}$
Update rule:

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

For Log-likelihood,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_k)\right)^{-1} \frac{\partial \ell}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_k) \equiv \boldsymbol{\theta}_k - (\mathsf{H}(\boldsymbol{\theta}_k))^{-1} \mathsf{s}(\boldsymbol{\theta}_k)$$

Defn A.96 (Berndt-Hall-Hall-Hausman (BHHH)).
Uses Information-matrix equality. Set $A_k = \frac{1}{N}(\mathsf{S}(\theta_k)\mathsf{S}(\theta_k)')$ to be outer product of scores

$$A_k := \left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell}{\partial \theta}(\theta_k)\frac{\partial \ell}{\partial \theta'}(\theta_k)\right)^{-1}$$

B Bibliography

References

Abadie, Alberto (Apr. 2003). "Semiparametric instrumental variable estimation of treatment response models". *Journal of econometrics* 113.2, pp. 231–263 (cit. on p. 55).

— (Jan. 2005). "Semiparametric Difference-in-Differences Estimators". en. *The Review of economic studies* 72.1, pp. 1–19 (cit. on p. 65).

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program". *Journal of the American statistical Association* 105.490, pp. 493–505 (cit. on pp. 69, 70).

Acharya, Avidit, Matthew Blackwell, and Maya Sen (Aug. 2016). "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects". *The American political science review* 110.3, pp. 512–529 (cit. on p. 81).

Amemiya, Takeshi (1985). "Advanced Econometrics". *Harvard University Press, Cambridge, Massachusetts* (cit. on p. 16).

Anatolyev, Stanislav and Nikolay Gospodinov (2011). *Methods for estimation and inference in modern econometrics*. CRC Press (cit. on p. 33).

Angrist, Joshua D. and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press (cit. on p. 56).

Arkhangelsky, Dmitry et al. (2021). "Synthetic Difference-in-Differences". *American Economic Review* 111.12, pp. 4088–4118 (cit. on p. 72).

Aronow, Peter M and Allison Carnegie (2013). "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable". *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* 21.4, pp. 492–506 (cit. on p. 56).

Aronow, Peter M and Benjamin T Miller (2019). *Foundations of agnostic statistics*. Cambridge University Press.

Athey, Susan, Mohsen Bayati, et al. (Oct. 2017). "Matrix Completion Methods for Causal Panel Data Models". arXiv: 1710.10251 [math.ST] (cit. on p. 69).

Athey, Susan and Guido Imbens (2006). "Identification and inference in nonlinear difference-in-differences models". *Econometrica* 74.2, pp. 431–497 (cit. on p. 69).

— (2016a). "Recursive partitioning for heterogeneous causal effects". *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360 (cit. on p. 50).

— (July 2016b). "The Econometrics of Randomized Experiments". arXiv: 1607.00698 [stat.ME] (cit. on p. 35).

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". *The Annals of Statistics* 47.2, pp. 1148–1178 (cit. on p. 49).

Austen-Smith, David and Jeffrey S Banks (2000). *Positive political theory I: collective preference*. Vol. 1. University of Michigan Press.

Bach, Philipp et al. (2021). "DoubleML–An Object-Oriented Implementation of Double Machine Learning in R". *arXiv preprint arXiv:2103.09603* (cit. on p. 47).

Bai, Jushan (2009). "Panel data models with interactive fixed effects". *Econometrica* 77.4, pp. 1229–1279 (cit. on p. 73).

Bang, Heejung and James M Robins (2005). "Doubly robust estimation in missing data and causal inference models". *Biometrics* 61.4, pp. 962–973 (cit. on p. 45).

Bazen, Stephen (2011). *Econometric methods for labour economics*. Oxford University Press (cit. on p. 77).

Bellemare, Marc F, Jeffrey R Bloem, and Noah Wexler (2020). "The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion". *Working paper* (cit. on p. 80).

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (May 2014). "High-Dimensional Methods and Inference on Structural and Treatment Effects". *The journal of economic perspectives: a journal of the American Economic Association* 28.2, pp. 29–50 (cit. on p. 46).

Ben-Michael, Eli et al. (Oct. 28, 2021). "The Balancing Act in Causal Inference". *arXiv* [*stat.ME*] (cit. on p. 45).

Billingsley, Patrick (2008). *Probability and measure*. John Wiley & Sons.

Blackwell, Matthew and Adam N Glynn (Nov. 2018). "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables". *The American political science review* 112.4, pp. 1067–1082 (cit. on p. 74).

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2022). "Quasi-experimental shift-share research designs". *The Review of Economic Studies* 89.1, pp. 181–213 (cit. on pp. 57, 58).

Boyd, Stephen and J Duchi (2012). "EE364b: Convex optimization II". *Course Notes, http://www.stanford.edu/class/ee364b*.

Callaway, Brantly and Pedro HC Sant'Anna (2020). "Difference-in-differences with multiple time periods". *Journal of Econometrics* (cit. on p. 68).

Cameron, A Colin and Pravin K Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Carter, Michael (2001). *Foundations of mathematical economics*. MIT Press.

Chaisemartin, Clément de and Xavier D'Haultfœuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". *The American economic review* (cit. on p. 67).

Chernozhukov, Victor, Denis Chetverikov, et al. (Feb. 2018). "Double/debiased machine learning for treatment and structural parameters". *The econometrics journal* 21.1, pp. C1–C68 (cit. on pp. 45, 46).

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly (2013). "Inference on counterfactual distributions". *Econometrica* 81.6, pp. 2205–2268 (cit. on p. 77).

Chernozhukov, Victor, Christian Hansen, and Martin Spindler (2015). "Post-selection and post-regularization inference in linear models with many controls and instruments". *American Economic Review* 105.5, pp. 486–90 (cit. on p. 59).
Christensen, Ronald (2019). *Advanced linear modeling*. Springer (cit. on p. 120).
Cinelli, Carlos and Chad Hazlett (Feb. 17, 2020). "Making sense of sensitivity: extending omitted variable bias". *Journal of the Royal Statistical Society. Series B, Statistical methodology* 82 (1), pp. 39–67 (cit. on p. 52).
Cornelissen, Thomas et al. (Aug. 2016). "From LATE to MTE: Alternative methods for the evaluation of policy interventions". *Labour economics* 41, pp. 47–60 (cit. on p. 58).
Cover, Thomas M (1999). *Elements of information theory*. John Wiley & Sons (cit. on p. 10).
Deisenroth, Marc Peter, A Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for machine learning*. Cambridge University Press.
Doudchenko, Nikolay and Guido Imbens (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*. Tech. rep. National Bureau of Economic Research (cit. on p. 70).
Efron, Bradley (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM (cit. on p. 30).
Frangakis, Constantine E and Donald B Rubin (2002). "Principal stratification in causal inference". *Biometrics* 58.1, pp. 21–29 (cit. on p. 60).
Frölich, Markus and Stefan Sperlich (2019). *Impact evaluation*. Cambridge University Press (cit. on p. 35).
Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020). "Bartik instruments: What, when, why, and how". *American Economic Review* 110.8, pp. 2586–2624 (cit. on p. 57).
Goodman-Bacon, Andrew (2018). *Difference-in-differences with variation in treatment timing*. Tech. rep. National Bureau of Economic Research (cit. on p. 68).
Györfi, László et al. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.
Hahn, Jinyoung (1998). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects". *Econometrica* (cit. on pp. 41, 43, 45).
Hainmueller, Jens (2012). "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies". *Political analysis*, pp. 25–46 (cit. on p. 44).
Hastie, Tibshirani, Robert Tibshirani, and H Jerome (2009). *Elements of Statistical Learning*. Springer, NY.
Heaton, Matthew J et al. (2019). "A case study competition among methods for analyzing large spatial data". *Journal of Agricultural, Biological and Environmental Statistics* 24.3, pp. 398–425 (cit. on p. 124).
Heckman, James J and Edward J Vytlacil (2007). "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". *Handbook of econometrics* 6, pp. 4779–4874 (cit. on p. 58).
Hernán, Miguel A, Babette Brumback, and James M Robins (June 2001). "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments". *Journal of the American Statistical Association* 96.454, pp. 440–448 (cit. on p. 74).
Hirano, Keisuke and Guido Imbens (2004). "The propensity score with continuous treatments". *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164, pp. 73–84 (cit. on p. 40).
Hirano, Keisuke, Guido Imbens, and Geert Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score". *Econometrica* 71.4, pp. 1161–1189 (cit. on p. 43).
Imai, Kosuke, Luke Keele, and Teppei Yamamoto (Feb. 2010). "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects". en. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1, pp. 51–71 (cit. on p. 80).
Imai, Kosuke and Marc Ratkovic (Jan. 2014). "Covariate balancing propensity score". en. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 76 (1), pp. 243–263 (cit. on p. 44).
Imbens, Guido (2000). "The role of the propensity score in estimating dose-response functions". *Biometrika* 87.3, pp. 706–710 (cit. on p. 40).
— (Feb. 2004). "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". *The review of economics and statistics* 86.1, pp. 4–29 (cit. on pp. 39, 40, 43).
Imbens, Guido and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (cit. on p. 35).
Imbens, Guido and Stefan Wager (May 2017). "Optimized Regression Discontinuity Designs". arXiv: 1705.01677 [stat.ME] (cit. on p. 63).
Jann, Ben (2008). "A Stata implementation of the Blinder-Oaxaca decomposition". *Stata journal* 8.4, pp. 453–479.
Jiang, Zhichao, Shu Yang, and Peng Ding (Dec. 3, 2020). "Multiply robust estimation of causal effects under principal ignorability". *arXiv* [*stat.ME*] (cit. on p. 60).
Keener, Robert W (2011). *Theoretical statistics: Topics for a core course*. Springer.
Kelly, Morgan (2020). "Direct Standard Errors for Regressions with Spatially Autocorrelated Residuals" (cit. on p. 122).
Kennedy, Edward H (Oct. 2015). "Semiparametric theory and empirical processes in causal inference". arXiv: 1510.04740 [math.ST] (cit. on pp. 82, 86).
Koenker (2005). *Quantile Regression* (*Econometric Society Monographs; No. 38*). Cambridge university press (cit. on p. 28).

Kosorok, Michael R (2008). *Introduction to empirical processes and semiparametric inference*. Springer (cit. on p. 85).
Lechner, Michael (2011). "The Estimation of Causal Effects by Difference-in-Difference Methods". *Foundations and Trends® in Econometrics* 4.3, pp. 165–224 (cit. on p. 64).
Lee, Myoung-jae (2016). *Matching, regression discontinuity, difference in differences, and beyond*. Oxford University Press.
Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky (2018). "Balancing covariates via propensity score weighting". *Journal of the American Statistical Association* 113.521, pp. 390–400 (cit. on p. 44).
Liu, Licheng, Ye Wang, and Yiqing Xu (July 2021). "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data". arXiv: 2107.00856 [stat.ME] (cit. on p. 69).
Luenberger, David G (1997). *Optimization by vector space methods*. John Wiley & Sons (cit. on p. 144).
Manski, Charles F (1993). "Identification of endogenous social effects: The reflection problem". *The review of economic studies* 60.3, pp. 531–542 (cit. on p. 123).
Menezes, Flavio M and Paulo Klinger Monteiro (2005). *An introduction to auction theory*. OUP Oxford.
Mogstad, Magne and Alexander Torgovitsky (Aug. 2018). "Identification and Extrapolation of Causal Effects with Instrumental Variables". *Annual review of economics* 10.1, pp. 577–613 (cit. on p. 59).
Mullainathan, Sendhil and Jann Spiess (2017). "Machine learning: an applied econometric approach". *Journal of Economic Perspectives* 31.2, pp. 87–106 (cit. on p. 102).
Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press (cit. on p. 91).
Newey, Whitney K and Daniel McFadden (1994). "Large sample estimation and hypothesis testing". *Handbook of econometrics* 4, pp. 2111–2245 (cit. on p. 32).
Oster, Emily (2019). "Unobservable selection and coefficient stability: Theory and evidence". *Journal of Business & Economic Statistics* 37.2, pp. 187–204 (cit. on p. 51).
Owen, Art B (2001). *Empirical likelihood*. Chapman and Hall/CRC (cit. on p. 33).
Racine, Jeffrey, Liangjun Su, and Aman Ullah (2013). *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics*. Oxford University Press (cit. on p. 139).
Robins, James M, Miguel Angel Hernan, and Babette Brumback (2000). *Marginal structural models and causal inference in epidemiology* (cit. on p. 74).
Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American statistical Association* 89.427, pp. 846–866 (cit. on p. 45).
Robinson, Peter M (1988). "Root-N-consistent semiparametric regression". *Econometrica: Journal of the Econometric Society*, pp. 931–954 (cit. on pp. 49, 89).
Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". *Biometrika* 70.1, pp. 41–55 (cit. on p. 42).
Rue, Havard and Leonhard Held (2005). *Gaussian Markov random fields: theory and applications*. CRC press (cit. on p. 123).
Ruppert, David, Matt P Wand, and Raymond J Carroll (2003). *Semiparametric regression*. 12. Cambridge university press.
Rustagi, Jagdish S (2014). *Optimization techniques in statistics*. Elsevier (cit. on p. 144).
Scholkopf, Bernhard and Alexander J Smola (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series (cit. on p. 91).
Schrimpf, Paul (2018). *Mathematics for Economics - Lecture Notes*.
Severini, Thomas A (2005). *Elements of distribution theory*. Cambridge University Press (cit. on p. 136).
Shalizi, Cosma (2019). *Advanced data analysis from an elementary point of view*.
Stachurski, John (2009). *Economic dynamics: theory and computation*. MIT Press.
— (2016). *A primer in econometric theory*. Mit Press.
Tsiatis, Anastasios (2007). *Semiparametric theory and missing data*. Springer Science & Business Media (cit. on pp. 82, 83).
Van Der Laan, Mark J and Daniel Rubin (2006). "Targeted maximum likelihood learning". *The international journal of biostatistics* 2.1 (cit. on p. 45).
Vohra, Rakesh V (2004). *Advanced mathematical economics*. Routledge.
Ward, Michael D and John S Ahlquist (2018). *Maximum Likelihood for Social Science: Strategies for Analysis*. Cambridge University Press.
Wasserman, Larry (2006). *All of nonparametric statistics*. Springer Science & Business Media (cit. on p. 82).
— (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.
White, Halbert (2014). *Asymptotic theory for econometricians*. Academic press (cit. on pp. 18, 22).
Williams, Christopher KI and Carl Edward Rasmussen (2006). *Gaussian processes for machine learning*. Vol. 2. 3. MIT press (cit. on p. 91).
Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data*. MIT press (cit. on p. 67).
Xu, Yiqing (Jan. 2017). "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models". *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* 25.1, pp. 57–76 (cit. on pp. 69, 73).

Zhao, Qingyuan (Jan. 22, 2016). "Covariate Balancing Propensity Score by Tailored Loss Functions". *arXiv* [*stat.ME*] (cit. on p. 45).
Zhou, Zhengyuan, Susan Athey, and Stefan Wager (Oct. 10, 2018). "Offline multi-action policy learning: Generalization and optimization". *arXiv* [*stat.ML*] (cit. on p. 50).