Semiparametric Causal Inference

Continous Treatments, Panel Data, Heterogeneity and sundries

Apoorva Lal

August 24, 2022

Stanford

Discrete and Continous Treatments

Setup

- For i.i.d. observations $i \in \{1, .., N\}$, we observe $\{Y_i, \mathbf{X}_i, W_i\}_i^N$ where:
  - $Y_i \in \mathbb{R}$ is the **outcome**
  - $W_i \in \{0, \ldots, K\}$ is the **treatment assignment**
  - $\mathbf{X}_i \in \mathbb{R}^k$ is the **covariate vector**
- We posit the existence of **potential outcomes** $Y^0, \ldots, Y^k$ for each unit. Vertically concat them into a 'science table' that is $N \times K$.

Setup

- For i.i.d. observations $i \in \{1, .., N\}$, we observe $\{Y_i, \mathbf{X}_i, W_i\}_i^N$ where:
  - $Y_i \in \mathbb{R}$ is the **outcome**
  - $W_i \in \{0, \ldots, K\}$ is the **treatment assignment**
  - $\mathbf{X}_i \in \mathbb{R}^k$ is the **covariate vector**
- We posit the existence of **potential outcomes** $Y^0, \ldots, Y^k$ for each unit. Vertically concat them into a 'science table' that is $N \times K$.

- Treatment effects (*estimands*) are defined as functions of *potential outcomes*, and since $(K-1)/K$ of them are unobserved, we need assumptions to use *estimators* to compute them using data.
  - Extent of missingness is increasing in the number of treatments: 1/2 for binary, $(K-1)/K$ for discrete, ≈ 1 for continuous

Identifying counterfactual means and friends : Discrete Treatments

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i = k} Y_i^k$. What does this assume?

Identifying counterfactual means and friends : Discrete Treatments

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.

Identifying counterfactual means and friends : Discrete Treatments

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
    - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
    - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Identifying counterfactual means and friends : Discrete Treatments

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Then, the *Counterfactual mean* is non-parametrically identified, as are causal contrasts. Augmented IPW (Robins et al 1994)

$$\widehat{\Gamma}_i^{(w)} = \underbrace{\widehat{\mu}^w(\mathbf{X})}_{\text{Outcome Model}} + \underbrace{\frac{\mathbb{1}_{W_i=w}}{\widehat{\pi}^w(\mathbf{X})}}_{\text{(Inv) Propensity score}} \underbrace{(Y_i - \widehat{\mu}^w(\mathbf{X}))}_{\text{(Residual}}$$

Identifying counterfactual means and friends : Discrete Treatments

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Then, the *Counterfactual mean* is non-parametrically identified, as are causal contrasts. Augmented IPW (Robins et al 1994)

$$\widehat{\Gamma}_i^{(w)} = \underbrace{\widehat{\mu}^w(\mathbf{X})}_{\text{Outcome Model}} + \underbrace{\frac{\mathbb{1}_{W_i=w}}{\widehat{\pi}^w(\mathbf{X})}}_{\text{(Inv) Propensity score}} \underbrace{(Y_i - \widehat{\mu}^w(\mathbf{X}))}_{\text{(Residual}}$$

- $\widehat{\mu}^w(\cdot), \widehat{\pi}^w(\cdot)$ are *nuisance functions* (potentially) high-dim quantities incidental to low-dim target (marginal mean, causal contrast).
- All nuisance functions are henceforth *cross-fit* [allows any good ML for curve fitting - Chernozhukov et al 2018]

Identifying counterfactual means and friends : Discrete Treatments

- *Causal Consistency / SUTVA* : $Y_i = \sum_k \mathbb{1}_{W_i=k} Y_i^k$. What does this assume?
  - i's outcome is only affected by i's treatment status. This may not be the case in many settings, e.g. with peer effects/interference/spillovers/contagion.
  - In such settings, the potential outcomes are indexed by $Y^{\mathbf{W}}$. In the extreme case of unrestricted interference, the 'science table' has width K^n. Need new assumptions / different estimands.
- *Unconfoundedness*: $Y^1, Y^0 \perp\!\!\!\perp W_i | \mathbf{X}_i$. Treatment is as good as random given covariates.
- *Overlap*: $0 < \pi^w(\mathbf{X}) < 1$. Each unit has positive probability of treatment.

Then, the *Counterfactual mean* is non-parametrically identified, as are causal contrasts. Augmented IPW (Robins et al 1994)

$$\widehat{\Gamma}_i^{(w)} = \underbrace{\widehat{\mu}^w(\mathbf{X})}_{\text{Outcome Model}} + \underbrace{\frac{\mathbb{1}_{W_i=w}}{\widehat{\pi}^w(\mathbf{X})}}_{\text{(Inv) Propensity score}} \underbrace{(Y_i - \widehat{\mu}^w(\mathbf{X}))}_{\text{(Residual}}$$

- $\widehat{\mu}^w(\cdot), \widehat{\pi}^w(\cdot)$ are *nuisance functions* (potentially) high-dim quantities incidental to low-dim target (marginal mean, causal contrast).
- All nuisance functions are henceforth *cross-fit* [allows any good ML for curve fitting - Chernozhukov et al 2018]
- Implementations: `grf::causal_forest`, `npcausal::ate`, `poirot::aipw`, and `DoubleML::IIRM` in R, `econML`, `DoubleML`, `causalML` in Python

Inference

- Augmented IPW estimators attain the semiparametric efficiency bound
    - $\approx$ CRLB for semiparametric models - see Hahn (1998)
- We want estimators with a familiar 'parametric' behaviour that satisfy a CLT of the $\sqrt{n}(\widehat{\tau} - \tau) \to \mathcal{N}\left(0, V\right)$

Inference

- Augmented IPW estimators attain the semiparametric efficiency bound
  - $\approx$ CRLB for semiparametric models - see Hahn (1998)
- We want estimators with a familiar 'parametric' behaviour that satisfy a CLT of the $\sqrt{n}(\widehat{\tau} - \tau) \rightarrow \mathcal{N}\left(0, V\right)$

$$\sqrt{\mathbb{E}\left[(\widehat{\mu}_{(w)}(\mathbf{X}_i) - \mu_{(w)}(\mathbf{X}_i))\right]^2}, \sqrt{\mathbb{E}\left[(\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i))\right]^2} << n^{-1/4}$$

- $\sqrt{n}$ either by function class of $\widehat{\mu}, \widehat{\pi}$ is not too complex ('Donsker') or sample splitting (paper, tutorial)
- Variance of influence function can be used to construct CIs for marginal means or causal contrasts:
  - $\widehat{\sigma}^2_w = \widehat{\mathbb{V}}(\widehat{\Gamma}_i^{(w)})$
  - SE = $\sqrt{\widehat{\sigma}^2_w / n}$

Continuous Treatments

- Now consider a case when $w \in \mathcal{W} \subseteq \mathbb{R}$, with corresponding potential outcomes $Y_i^{(w)}$.
  - Unconfoundedness': $\mathbb{E}\left[Y^w | w, \mathbf{X}\right] = \mathbb{E}\left[Y^w | \mathbf{X}\right]$
  - Positivity': $\mathsf{f}(w|\mathbf{X}) > 0$
  - **Generalised Propensity Score** (Propensity Density) $r(w, x) := \mathsf{f}_{w|x}(w|\mathbf{X})$
    - Estimating conditional densities is hard. Recent progress

Continuous Treatments

- Now consider a case when $w \in \mathcal{W} \subseteq \mathbb{R}$, with corresponding potential outcomes $Y_i^{(w)}$.
  - Unconfoundedness': $\mathbb{E}\left[Y^w | w, \mathbf{X}\right] = \mathbb{E}\left[Y^w | \mathbf{X}\right]$
  - Positivity': $\mathsf{f}(w|\mathbf{X}) > 0$
  - **Generalised Propensity Score** (Propensity Density) $r(w,x) := \mathsf{f}_{w|x}(w|\mathbf{X})$
    - Estimating conditional densities is hard. Recent progress
- The quantity of interest is the *does response curve* $\theta(w) := \mathbb{E}\left[Y^w\right]$: expected value of the potential outcome across observations when treatment is set at w. This uses the following construction for the DR score (Kennedy et al 2017, Colangelo and Lee 2022, Klosin 2022)

$$\Gamma_i = \mu^w(\mathbf{X}_i) + \frac{K_h(w_i - w)}{r(w,x)}(Y_i - \mu^w(\mathbf{X}_i))$$

Continuous Treatments

- Now consider a case when $w \in \mathcal{W} \subseteq \mathbb{R}$, with corresponding potential outcomes $Y_i^{(w)}$.
  - Unconfoundedness’: $\mathbb{E}\left[Y^w | w, \mathbf{X}\right] = \mathbb{E}\left[Y^w | \mathbf{X}\right]$
  - Positivity’: $\mathsf{f}(w|\mathbf{X}) > 0$
  - **Generalised Propensity Score** (Propensity Density) $r(w, x) := \mathsf{f}_{w|x}(w|\mathbf{X})$
    - Estimating conditional densities is hard. Recent progress
- The quantity of interest is the *does response curve* $\theta(w) := \mathbb{E}\left[Y^w\right]$: expected value of the potential outcome across observations when treatment is set at w. This uses the following construction for the DR score (Kennedy et al 2017, Colangelo and Lee 2022, Klosin 2022)

$$\Gamma_i = \mu^w(\mathbf{X}_i) + \frac{K_h(w_i - w)}{r(w, x)}(Y_i - \mu^w(\mathbf{X}_i))$$

- Its *average derivative* can be estimated using residuals-on-residuals regression (Robinson (1988), Powell, Stock, Stoker (1989))

$$Y_i - \widehat{\mu}_i(\mathbf{X}_i) = \tau(W_i - \widehat{\pi}(\mathbf{X}_i)) + \varepsilon_i$$

Continuous Treatments

- Now consider a case when $w \in \mathcal{W} \subseteq \mathbb{R}$, with corresponding potential outcomes $Y_i^{(w)}$.
    - Unconfoundedness': $\mathbb{E}\left[Y^w | w, \mathbf{X}\right] = \mathbb{E}\left[Y^w | \mathbf{X}\right]$
    - Positivity': $\mathsf{f}(w|\mathbf{X}) > 0$
    - **Generalised Propensity Score** (Propensity Density) $r(w, x) := \mathsf{f}_{w|x}(w|\mathbf{X})$
        - Estimating conditional densities is hard. Recent progress
- The quantity of interest is the *does response curve* $\theta(w) := \mathbb{E}\left[Y^w\right]$: expected value of the potential outcome across observations when treatment is set at w. This uses the following construction for the DR score (Kennedy et al 2017, Colangelo and Lee 2022, Klosin 2022)

$$\Gamma_i = \mu^w(\mathbf{X}_i) + \frac{K_h(w_i - w)}{r(w, x)}(Y_i - \mu^w(\mathbf{X}_i))$$

- Its *average derivative* can be estimated using residuals-on-residuals regression (Robinson (1988), Powell, Stock, Stoker (1989))

$$Y_i - \widehat{\mu}_i(\mathbf{X}_i) = \tau(W_i - \widehat{\pi}(\mathbf{X}_i)) + \varepsilon_i$$

- implementation: `npcausal::ctseff`, `DoubleML::PLM`

Panel Data

Difference in Differences

- Now, write $Y^w(t)$ to denote the potential outcomes Y^1, Y^0 at time $t \in \{0, 1\}$ and $Y(t)$ to denote the realised outcome. The estimand is the ATT in the 2nd period $\mathbb{E}\left[Y^1(1) - Y^0(1) | D = 1\right]$.
- The *conditional* parallel trends assumption is written as

$$E\left[Y^0(1) - Y^0(0) \mid \mathbf{X}, D = 1\right] = E\left[Y^0(1) - Y^0(0) \mid \mathbf{X}, D = 0\right]$$

Difference in Differences

- Now, write $Y^w(t)$ to denote the potential outcomes Y^1, Y^0 at time $t \in \{0,1\}$ and $Y(t)$ to denote the realised outcome. The estimand is the ATT in the 2nd period $\mathbb{E}\left[Y^1(1) - Y^0(1) | D = 1\right]$.
- The *conditional* parallel trends assumption is written as

$$E\left[Y^0(1) - Y^0(0) \mid \mathbf{X}, D = 1\right] = E\left[Y^0(1) - Y^0(0) \mid \mathbf{X}, D = 0\right]$$

- Under (c)PT, there are multiple candidate estimators

$$\begin{aligned}
\hat{\tau}^{\text{OM}} &= \{\widehat{\mathbb{E}}[Y(1) \mid \mathbf{X}, D = 1] - \widehat{\mathbb{E}}[Y(1) \mid \mathbf{X}, D = 0]\} - \{\widehat{\mathbb{E}}[Y(0) \mid \mathbf{X}, D = 1] - \widehat{\mathbb{E}}[Y(0) \mid \mathbf{X}, D = 0]\} \\
\hat{\tau}^{\text{IPW}} &= \frac{1}{N}\sum_i Y_i(1) - Y_i(0) \frac{D - \hat{e}(\mathbf{X}_i)}{P(D = 1)(1 - \hat{e}(\mathbf{X}_i))} \\
\hat{\tau}^{\text{AIPW}} &= \frac{1}{N}\sum_i (Y_i(1) - Y_i(0) - d(\mathbf{X}_i, D = 0)) \cdot \left[\frac{D - \hat{e}(\mathbf{X}_i)}{P(D = 1)(1 - \hat{e}(\mathbf{X}_i))}\right]
\end{aligned}$$

- where $\widehat{e}$ is a propensity score and $d(\cdot)$ is an outcome model for the trend $Y(1) - Y(0)$ in untreated obs $D = 0$. code

Panel Data

- For one-way panel data: $Y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$, one idea is to partial out FEs and work with $\ddot{y}_{it}, \ddot{x}_{it}$ with clustered ML (e.g. clustered LASSO)

Panel Data

- For one-way panel data: $Y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}$, one idea is to partial out FEs and work with $\ddot{y}_{it}, \ddot{x}_{it}$ with clustered ML (e.g. clustered LASSO)
- However, most panel data typically stipulates the following (two-way fixed effects) outcome model

$$Y^0_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\beta + \varepsilon_{it}; \quad Y^1_{it} = \tau_{it} W_{it} + Y^0_{it}$$

- Want ATT, or at least convex averages of τ_{it}; not guaranteed with TWFE under staggered adoption [cf TWFE lit]

Panel Data

- For one-way panel data: $Y_{it} = \alpha_i + \mathbf{x}_{it}'\beta + \varepsilon_{it}$, one idea is to partial out FEs and work with $\ddot{y}_{it}, \ddot{x}_{it}$ with clustered ML (e.g. clustered LASSO)
- However, most panel data typically stipulates the following (two-way fixed effects) outcome model

$$Y_{it}^0 = \alpha_i + \gamma_t + \mathbf{x}_{it}'\beta + \varepsilon_{it}; \quad Y_{it}^1 = \tau_{it} W_{it} + Y_{it}^0$$

- Want ATT, or at least convex averages of τ_{it}; not guaranteed with TWFE under staggered adoption [cf TWFE lit]
- Panel regression is betting the house on a functional form. Better make it flexible, say with a factor model

$$Y_{it} = \delta_{it} W_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{it}$$

$\boldsymbol{f}_t = [f_{1t}, \ldots, f_{rt}]'$ is $k \times 1$ **common factors**, $\boldsymbol{\lambda}_i = [\lambda_{i1}, \ldots, \lambda_{ir}]'$ is $r \times 1$ **factor loadings**.

Panel Data

- For one-way panel data: $Y_{it} = \alpha_i + \mathbf{x}_{it}'\beta + \varepsilon_{it}$, one idea is to partial out FEs and work with $\ddot{y}_{it}, \ddot{x}_{it}$ with clustered ML (e.g. clustered LASSO)

- However, most panel data typically stipulates the following (two-way fixed effects) outcome model

$$Y_{it}^0 = \alpha_i + \gamma_t + \mathbf{x}_{it}'\beta + \varepsilon_{it}; \quad Y_{it}^1 = \tau_{it} W_{it} + Y_{it}^0$$

- Want ATT, or at least convex averages of τ_{it}; not guaranteed with TWFE under staggered adoption [cf TWFE lit]

- Panel regression is betting the house on a functional form. Better make it flexible, say with a factor model

$$Y_{it} = \delta_{it} W_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{it}$$

$\boldsymbol{f}_t = [f_{1t}, \ldots, f_{rt}]'$ is $k \times 1$ **common factors**, $\boldsymbol{\lambda}_i = [\lambda_{i1}, \ldots, \lambda_{ir}]'$ is $r \times 1$ **factor loadings**.

- $f_t = 1 \implies \lambda_i \times 1 = \lambda_i$ **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**
- Extends naturally to Matrix Completion

Trying to make PT hold using reweighting (Synth and friends)

- 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}
- Following Abadie, Diamond, Hainmueller (2010), a whole family of methods to **try to make parallel trends hold** using balancing methods. Comprehensive intro : Yiqing's course materials

Trying to make PT hold using reweighting (Synth and friends)

- 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}
- Following Abadie, Diamond, Hainmueller (2010), a whole family of methods to **try to make parallel trends hold** using balancing methods. Comprehensive intro : Yiqing's course materials
- The following approach is due to Arkhankelsky et al (2021) [Implemented in synthdid]
- unit weights $\widehat{\omega}^{\text{sdid}}$ *align pre-exposure trends in outcomes of unexposed units with those for exposed units* $\sum_{i=1}^{N_{co}} \widehat{\omega}^{\text{sdid}} Y_{it} \approx N_{tr}^{-1} \sum_{i=N_{co}+1}^{N} Y_{it}$

$$
\left(\hat{\omega}_0, \hat{\omega}^{\text{sdid}}\right) = \underset{\omega_0 \in \mathbb{R}, \omega \in \Omega}{\arg\min}\, \ell_{\text{unit}}\left(\omega_0, \omega\right) \quad \text{where}
$$

$$
\ell_{\text{unit}}\left(\omega_0, \omega\right) = \sum_{t=1}^{T_{\text{pre}}} \left(\omega_0 + \sum_{i=1}^{N_{co}} \omega_i Y_{it} - \frac{1}{N_{\text{tr}}} \sum_{i=N_{\text{co}}+1}^{N} Y_{it} \right)^2 + \zeta^2 T_{\text{pre}} \|\omega\|_2^2,
$$

$$
\Omega = \left\{ \omega \in \mathbb{R}_+^N : \sum_{i=1}^{N_{\text{co}}} \omega_i = 1, \omega_i = N_{\text{tr}}^{-1} \text{ for all } i = N_{\text{co}} + 1, \ldots, N \right\},
$$

SDID : contd

- time weights $\widehat{\lambda}_t^{\text{sdid}}$ that *balance pre-exposure time periods with post-exposure time periods for unexposed units*.

$$\left(\hat{\lambda}_0, \hat{\lambda}^{\text{sdid}}\right) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\ \ell_{\text{time}}\left(\lambda_0, \lambda\right) \quad \text{where}$$

$$\ell_{\text{time}}\left(\lambda_0, \lambda\right) = \sum_{i=1}^{N_{\text{co}}} \left(\lambda_0 + \sum_{t=1}^{T_{\text{pre}}} \lambda_t Y_{it} - \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it}\right)^2$$

$$\Lambda = \left\{\lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T_{\text{pre}}} \lambda_t = 1, \lambda_t = T_{\text{post}}^{-1} \text{ for all } t = T_{\text{pre}}+1, \ldots, T\right\}$$

SDID : contd

- time weights $\widehat{\lambda}_t^{\text{sdid}}$ that *balance pre-exposure time periods with post-exposure time periods for unexposed units*.

$$\left(\hat{\lambda}_0, \hat{\lambda}^{\text{sdid}}\right) = \underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg\min}\, \ell_{\text{time}}\left(\lambda_0, \lambda\right) \quad \text{where}$$

$$\ell_{\text{time}}\left(\lambda_0, \lambda\right) = \sum_{i=1}^{N_{\mathrm{co}}} \left(\lambda_0 + \sum_{t=1}^{T_{\mathrm{pre}}} \lambda_t Y_{it} - \frac{1}{T_{\mathrm{post}}} \sum_{t=T_{\mathrm{pre}}+1}^{T} Y_{it}\right)^2$$

$$\Lambda = \left\{\lambda \in \mathbb{R}_+^T : \sum_{t=1}^{T_{\mathrm{pre}}} \lambda_t = 1, \lambda_t = T_{\text{post}}^{-1} \text{ for all } t = T_{\mathrm{pre}} + 1, \ldots, T\right\}$$

- Finally: Regression

$$\left(\widehat{\tau}^{\text{sdid}}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}\right) = \underset{\tau, \mu, \alpha, \beta}{\operatorname{argmin}} \left\{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau\right)^2 \widehat{\omega}_i^{\text{sdid}} \widehat{\lambda}_t^{\text{sdid}}\right\}$$

Heterogeneous Effects

The problem

- We are interested in the **Conditional Average Treatment Effect (CATE)**:
  $\tau(\mathbf{X}) = E[Y^{(1)} - Y^{(0)}|\mathbf{X} = \mathbf{x}]$
  - This is a *function*, not a number, so we may want to summarise
    - projecting imputed effects linearly on covariates (BLP)
    - binning estimates (GATE)

The problem

- We are interested in the **Conditional Average Treatment Effect (CATE)**:

  $\tau(\mathbf{X}) = E[Y^{(1)} - Y^{(0)} | \mathbf{X} = \mathbf{x}]$
  - This is a *function*, not a number, so we may want to summarise
    - projecting imputed effects linearly on covariates (BLP)
    - binning estimates (GATE)

Parametric Outcome Modeling: Estimate OLS with interactions

- $Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \beta_3 W_i X_i + \epsilon_i$
  - Implicit outcome models: $Y_i^0 = \beta_2 X_i, Y_i^1 = Y_i^0 + \beta_1 + \beta_3 X_i$
- $\widehat{\text{CATE}}_X = \hat{\beta}_1 + \hat{\beta}_3 X_i$
- Why do we need machine learning / regularization to do this?

The problem

- We are interested in the **Conditional Average Treatment Effect (CATE)**:

$$\tau(\mathbf{X}) = E[Y^{(1)} - Y^{(0)} | \mathbf{X} = \mathbf{x}]$$

  - This is a *function*, not a number, so we may want to summarise
    - projecting imputed effects linearly on covariates (BLP)
    - binning estimates (GATE)

Parametric Outcome Modeling: Estimate OLS with interactions

- $Y_i = \beta_0 + \beta_1 W_i + \beta_2 X_i + \beta_3 W_i X_i + \epsilon_i$
  - Implicit outcome models: $Y_i^0 = \beta_2 X_i, Y_i^1 = Y_i^0 + \beta_1 + \beta_3 X_i$
- $\widehat{\text{CATE}}_X = \hat{\beta}_1 + \hat{\beta}_3 X_i$
- Why do we need machine learning / regularization to do this?
  - **Overfitting**: We know that in general, when $k \approx N$, traditional OLS methods will badly overfit
  - **Unknown Functional Form**: The analyst does not know what the underlying heterogeneity looks like
  - **fishing**: Why should the reader believe that this specification fell from the sky?

Pure ML Ideas

T-Learner

- fits separate models on the treated and controls.
- Learn $\hat{\mu}_{(0)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 0$.
- Learn $\hat{\mu}_{(1)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 1$.
- Report $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$.

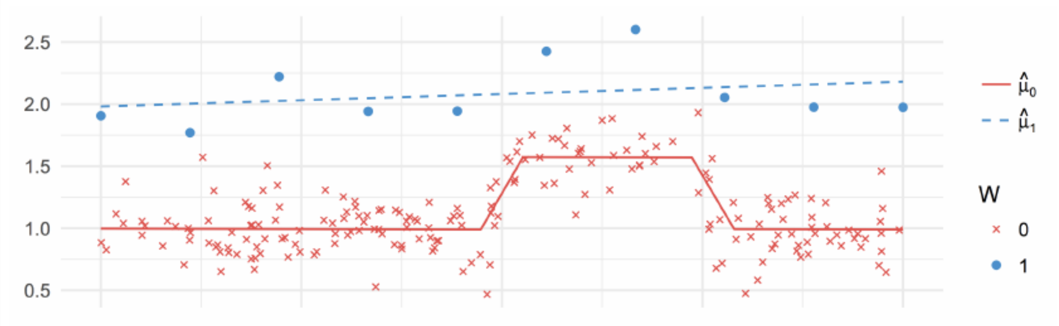
T-Learner

- fits separate models on the treated and controls.
- Learn $\hat{\mu}_{(0)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 0$.
- Learn $\hat{\mu}_{(1)}(x)$ by predicting Y_i from X_i on the subset of observations with $W_i = 1$.
- Report $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$.

S-Learner

- fits a single model to all the data.
- Learn $\hat{\mu}(z)$ by predicting Y_i from $Z_i := (X_i, W_i)$ on all the data.
- Report $\hat{\tau}(x) = \hat{\mu}((x, 1)) - \hat{\mu}((x, 0))$.

They were bad: Regularization Bias



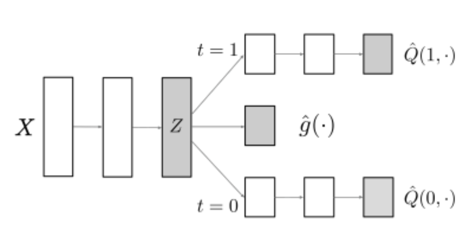
- Differential shrinkage across treatment levels leads to ‘hallucinated’ heterogeneity
- Problem is generic for any regression learner. Need some kind of ‘joint’ modelling for potential outcomes.

Sidestepping Regularisation Bias: Tailored Neural-net achitecture

Dragonnet, Tarnet, etc.

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \widehat{R}(\theta; \mathbf{X}) \text{ where}$$

$$\widehat{R}(\theta; \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} ((Q^{nn}(w_i, \mathbf{X}_i, \theta) - y_i)^2 + \alpha \text{CrossEntropy}(g^{nn}(\mathbf{X}_i; \theta), w_i))$$



https://arxiv.org/pdf/1906.02120.pdf

Sidestepping Regularisation Bias: X, R Learners

X-Learner

- Fit $\hat{\mu}^{(0)}(x), \hat{\mu}^{(1)}(x)$ using nonparametric regression
- Define pseudo-effects $\widetilde{D}_i^1 := Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)$ and use them to fit $\widehat{\tau}^1(\mathbf{X}_i)$ on $\{i : W_i = 1\}$
- Define pseudo-effects $\widetilde{D}_i^0 := \widehat{\mu}^{(1)}(\mathbf{X}_i) - Y_i$ and use them to fit $\widehat{\tau}^0(\mathbf{X}_i)$ on $\{i : W_i = 0\}$
- Aggregate them as $\widehat{\tau}(x) = (1 - \widehat{\pi}(x))\widehat{\tau}^1(\mathbf{x}) + \widehat{\pi}(x)\widehat{\tau}^0(\mathbf{x})$

https://arxiv.org/abs/1706.03461

R-Learner

- Minimise Robinson (R) Loss

$$\widehat{\tau} = \underset{\tau}{\text{argmin}} \left\{ \widehat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\}$$

$$\widehat{L}(\tau(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} ((Y_i - \widehat{\mu}(\mathbf{X}_i)) - (W_i - \widehat{\pi}(\mathbf{X}_i))\, \tau(\mathbf{X}_i))^2$$

- IOW, Regress pseudo outcome $\frac{Y - \mu(\mathbf{X})}{W - \widehat{\pi}(X)}$ on covariates $\psi(\mathbf{X}_i)$
- weights $(W - \widehat{\pi}(\mathbf{X}))^2$

https://arxiv.org/abs/1712.04912

DR-Learner

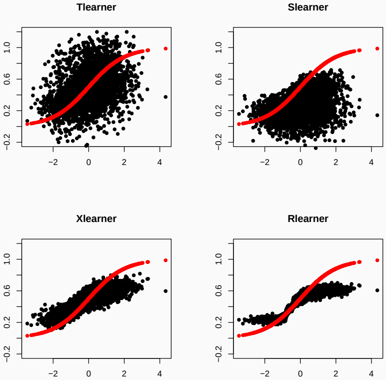
- Construct pseudo-outcomes $\widehat{\varphi}(Z) := \widehat{\Gamma}_i^1 - \widehat{\Gamma}_i^0$ using AIPW score function
- Regress it on covariates $\psi(\mathbf{X}_i)$

https://arxiv.org/abs/2004.14497

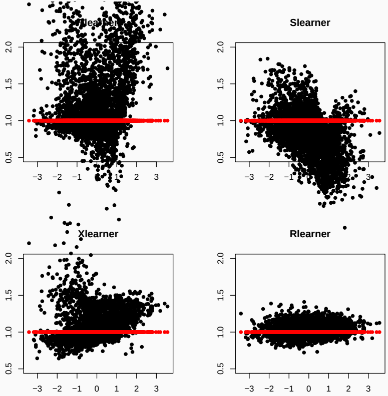
In action: RCT, Confounding

- Simulation + Implementation

Experiment



Observational



Summary of Generic Approaches [Knaus et al 2021]

Table 1. Summary of generic approaches to estimate IATEs.

Approach	w_i	Y_i^*
MOM IPW	1	$Y_{i,IPW}^*$
MOM DR	1	$Y_{i,DR}^*$
MCM	$T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}$	$2T_iY_i$
MCM with EA	$T_i \frac{D_i - p(X_i)}{4p(X_i)(1-p(X_i))}$	$2T_i(Y_i - \mu(X_i))$
Orthogonal Learning	$(D_i - p(X_i))^2$	$\frac{Y_i - \mu(X_i)}{D_i - p(X_i)}$

https://arxiv.org/abs/1810.13237

- $D_i = W_i \in \{0, 1\}$
- $T_i = 2D_i - 1 \in \{-1, 1\}$
- $Y^*_{\text{IPW}} = \frac{W_i - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)(1-\pi(\mathbf{X}_i))}$
- $Y^*_{DR} = \widehat{\Gamma}_i^1 - \widehat{\Gamma}_i^0$
- All problems solve weighted least squares

$$\min_{\tau} \left(\frac{1}{n} \sum_{i=1}^{n} w_i (Y_i^* - \tau(\mathbf{X}_i))^2 \right)$$

Stratification

- Since Het-FX estimators produce estimates of $\widehat{\tau}_i$, a gut-check for how well this works is to then stratify on $\widehat{\tau}_i$ (say, J bins), and compute $\widehat{\text{ATE}}^j$ in each bin using say AIPW
- If $\widehat{\text{ATE}}^j$s are sorted along their bin indices, this increases confidence that $\widehat{\tau}_i$s aren't all noise

Best linear predictor method

- Create synthetic predictors $C_i = \overline{\tau}(W_i - \widehat{\pi}^{-i}(\mathbf{X}_i))$ and $D = (\widehat{\tau}^{-i}(\mathbf{X}_i) - \overline{\tau})(W_i - \widehat{\pi}(\mathbf{X}_i))$
- Regress $Y_i - \widehat{\mu}^{-i}(\mathbf{X}_i) \sim \alpha C_i + \beta D_i$
- $\alpha \approx 1$ indicates quality of ATE
- $\beta \approx 1$ indicates quality of CATE estimates (p.value is an omnibus test of heterogeneity fit by $\widehat{\tau}_i$)

- https://datascience.quantecon.org/applications/heterogeneity.html
- https://grf-labs.github.io/grf/articles/diagnostics.html