Three Designs

Eduardo Fé

Introduction

The idea of Neyman Orthogonality applies to popular designs.

Here we discuss:

- Estimation under ignorability.
- Instrumental Variable estimation with binary treatment and assignment.
- Difference in Differences with panel data (briefly)

Introduction.

Throughout, we assume:

Data are available, $\{Y_i, D_i, W_i\}_{i=1}^{n}$ where

$$Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i).$$

W includes p explanatory variables (potentially many more than n). We are interested in the causal effect of D_i,

$$\tau = E[Y_i(1)] - E[Y_i(0)].$$

Assumption 1, Stable Unit Treatment Value Assumption (SUTVA). There is no interference across units, so that

$$Y_i(D_1, D_2, ..., D_n) = Y_i(D_i)$$

and there is no hidden variation in treatment.

Design 1: Ignorability.

D_i is not randomly allocated: D and $Y(d)$ are not independent.

However:

Assumption 2, Ignorability. For $d \in \{0, 1\}$, i.e. $D \perp Y(d)| \; W$

- Variation in assigned treatment is a good as random given W.

Assumption 3, Overlap / Full support. The propensity score, $e(W) = P(D = 1|W)$, is such that $P(0 < e(W) < 1) = 1$.

- For any w we will be able to find obsevations with both $D = 1$ and $D = 0$ (so that comparisons of outcomes across treatment groups, at each w, are feasible/defined).

Design 1: Ignorability.

Implication 1: We could identify the effect of D by comparing outcomes across treated and control units, given $W = w$.

$$\begin{aligned} & E(Y|D=1,W) - E(Y|D=0,W) \\ =& m(1,W) - m(0,W) \\ =& E(Y(1)|W) - E(Y(0)|W) = \tau(W) \end{aligned} \tag{1}$$

which is the **Conditional Average Treatment Effect**. Then we can retrieve the Average Treatment Effect,

$$\tau = E\Big[\tau(W)\Big] = E\Big[m(1,W) - m(0,W)\Big] \tag{2}$$

Design 1: Ignorability.

Implication 2: We could identify the effect of D using a Horvitz - Thompson (Inverse Propensity Weighting, IPW) estimator:

$$\tau(W) = E\left[Y \cdot \left(\frac{\mathbb{I}(D=1)}{e(W)} - \frac{\mathbb{I}(D=0)}{1-e(W)}\right) \middle| W\right] \tag{3}$$

Define

$$H = \frac{\mathbb{I}(D=1)}{e(W)} - \frac{\mathbb{I}(D=0)}{1-e(W)} \tag{4}$$

Clearly

$$\tau = E\left[E(Y \cdot H|W)\right]$$

Design 1: Ignorability.

Equations 1 and 3 could be used to construct estimators, replacing $E(Y|D = d, W), e(W)$ with ML estimators.

That will not work because, individually, these estimators solve moment equations which are NOT Neyman Orthogonal.

However the combination of the two estimators results in a Neyman Orthogonal moment[1]:

$$\eta(W) = [m(1, W_i) - m(0, W_i)] + [Y_i - m(D, W)] \cdot H_i \tag{5}$$

where

$$H_i = \frac{\mathbb{I}(D_i = 1)}{e(W_i)} - \frac{\mathbb{I}(D_i = 0)}{1 - e(W_i)} \tag{6}$$

has $E(\eta(W)) = ATE$ and the "derivative" of $\eta(W)$ does not depend on $m()$ or $p()$.

[1] This results in the popular "Double Robust" estimator.

Design 1: Implementation.

- Step 1: Split the sample in $k = 1, ..., K$ folds.
- Step 2: For k=1,2,...K,
  - Step 2.1. Estimate, using all but fold k, $m(.)$ and $p(.)$ using a ML procedure, denoted $\hat{m}_{-k}(.)$, $\hat{p}_{-k}(.)$
  - Step 2.2. Obtain, for each i in fold k the *residuals* $\hat{\eta}_i = \hat{\eta}_i(W_i)$,

$$\hat{\eta}_i = [\hat{m}_{-k}(1, W_i) - \hat{m}_{-k}(0, W_i)] + [Y_i - \hat{m}_{-k}(D_i, W_i)] \cdot \hat{H}_i$$

    where

$$\hat{H}_i = \frac{\mathbb{I}(D_i = 1)}{\hat{p}_{-k}(W_i)} - \frac{\mathbb{I}(D_i = 0)}{1 - \hat{p}_{-k}(W_i)}$$

- Step 3: The estimator of ATE is $\hat{\tau} = n^{-1} \sum_{i=1}^{n} \hat{\eta}_i$
- Step 4: Base inference on the following estimator of the variance of $\hat{\tau}$, $\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (\hat{\eta}_i - \hat{\tau})^2$

Design 1: Properties.

The estimator $\hat{\tau}$ is such that, under certain conditions (including that the ML provides a *decent* fit)

$$\sqrt{n}(\hat{\tau} - \tau) \sim N(0, \Sigma) \tag{7}$$

as $n \to \infty$, where $\Sigma = E(\eta(W) - \tau)^2$.

Design 1: Example

Generated 1000 samples of 500 observations from:

$$Y_i = D_i\tau + \cos(W_i'\beta)^2 + \varepsilon_i \text{ with } \varepsilon_i \sim N(0,1) \tag{8}$$

$$D_i = \mathbb{I}\left(\sin(W_i\gamma) + \cos(W_i'\gamma) + \nu_i > 0\right) \text{ with } \nu_i \sim N(0,1) \tag{9}$$

Where W_i includes 3 covariates with correlation $0.7^{|j-k|}$, $j,k = 1,2,3$

We set $\tau = 0$.

Design 1: Example.

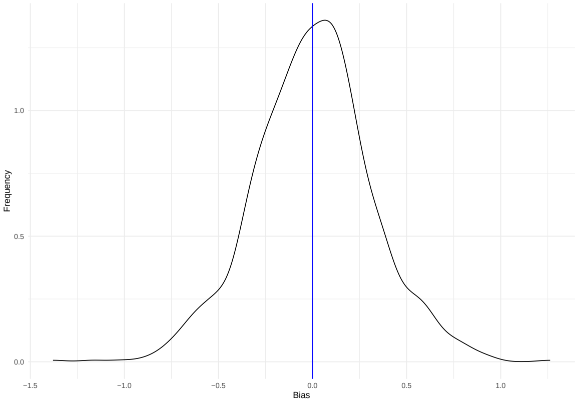


Figure 1: Distribution of the bias.

The average value of the estimate was 0.002123. The standard t-test for $H_o : \tau = 0$ rejected the null 0.028 times.

Design 2: Instrumental Variables.

As in Design 1, D is assumed endogenous, but we do not assume Conditional Independence.

We assume that there are (many) pre-treatment (exogenous) covariates W .

We now have a binary Instrumental Variable, $Z \in \{0, 1\}$ such that

Assumption 2', Exclusion: In the traditional setting Z is either randomly allocated or plausibly random in the sense that:

$$Y(Z, D(Z)) = Y(D(Z)) = Y(D)$$

Assignment only affects outcomes through its effect on uptake.

Design 2: Instrumental Variables.

If, in addition, there are no *defiers*,

Assumption 4, Monotonicity $D(1) \geq D(0)$

we can estimate a Local Average Treatment Effect (LATE)

$$\begin{aligned} \theta = LATE =& E(Y(1) - Y(0)|D(1) > D(0)) \\ =& \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} \end{aligned} \tag{10}$$

LATE is identified from data: each $E(.|Z=z)$ can be estimated with a sample mean.

That will results in the **Wald estimator**.

Design 2: Instrumental Variables.

It turns out that LATE estimation via IV is **Neyman Orthogonal**.

Recall we can estimate LATE as the solution to a two-stage least squares regression problem:

$$
\begin{aligned}
Y &= \beta_0 + \theta D + \varepsilon \\
D &= \pi_0 + \pi_1 Z + \nu
\end{aligned}
$$

- Stage 1: Estimate the second equation by OLS; obtain predicted values $\hat{D}$
- Stage 2: Regress Y on the predictions $\hat{D}$ from the first stage. The estimator of θ is the estimator of LATE and equals the **Wald Estimator** in this simple case.

Design 2: Instrumental Variables.

Looking at the 2-stage model

- θ is a single parameter and there is no regularisation in the equation for Y
- Regularisation is limited to the first stage, which is a pure **prediction** problem
- Therefore regularisation errors in the first stage are unlikely to spill over onto stage 2.

Design 2: Instrumental Variables.

The exclusion and monotonicity assumption will often be more credible conditional on explanatory variables, W.

Consider the extended LATE model

$$Y = m(D, X, \varepsilon) \tag{11}$$

$$D = p(Z, X, \nu) \in \{0, 1\} \tag{12}$$

$$Z = h(X, \epsilon) \in \{0, 1\} \tag{13}$$

where $D = p(z, X, \nu)$ is weakly increasing in z (monotonicity).

In this setting, the potential outcome is $Y(d) = m(d, X, \varepsilon)$ and

$$\theta = LATE = \frac{E\Big[E(Y|Z=1, W) - E(Y|Z=0, W)\Big]}{E\Big[E(D|Z=1, W) - E(D|Z=0, W)\Big]} \tag{14}$$

Design 2: Instrumental Variables

LATE remains identifiable provided that

- Z is relevant (it predicts D) monotonically,
- and the **support** condition $0 < P(Z = 1|W) < 1$ holds.
- The denominator

$$E\Big[E(D|Z = 1, W) - E(D|Z = 0, W)\Big]$$

  still identifies the proportion of compliers.

Design 2: Estimation

$$\theta = LATE = \frac{E\Big[E(Y|Z=1,W) - E(Y|Z=0,W)\Big]}{E\Big[E(D|Z=1,W) - E(D|Z=0,W)\Big]} \quad (15)$$

Z is, by definition, randomly allocated or independent of the potential outcomes.

The LATE thus is the ratio of two *regressions* under ignorability:

- A *regression* of Y on Z, given W
- A *regression* of D on Z, given W

As before, regularisation errors when modelling D do not affect modelling of Y.

Based estimation on the methods for Causal Machine Learning under Ignorability.

Design 2: Estimation

The LATE now depends on the *nuisance* parameters

- $m(Z, W) = E(Y|Z, W)$,
- $g(Z, W) = E(D|Z, W)$,
- $p(W) = E(Z|W)$.

Let

$$\kappa_1 = m(1, W) - m(0, W) + \left[Y - m(Z, W)\right] \cdot H(Z, W)$$
$$\kappa_2 = g(1, W) - g(0, W) + \left[D - g(Z, W)\right] \cdot H(Z, W)$$

where

$$H(Z, W) = \frac{Z}{p(W)} - \frac{1 - Z}{1 - p(W)}$$

Estimation will proceed by applying methods for conditional Ignorability to κ_1, κ_2.

Design 2: Implementation.

- Step 1: Split the sample in $k = 1, ..., K$ folds.
- Step 2: For k=1,2,...K,
  - Step 2.1. Estimate, using all but fold k, $m(.)$, $g(.)$ and $p(.)$ using a ML procedure, denoted $\hat{m}_{-k}(.)$, $\hat{g}_{-k}(.)$, $\hat{p}_{-k}(.)$
  - Step 2.2. Obtain, for each i in fold k the *residuals* $\hat{\kappa}_{1i} = \hat{\kappa}_{1i}(W_i)$, $\hat{\kappa}_{2i} = \hat{\kappa}_{2i}(W_i)$,

    $$\hat{\kappa}_{1i} = [\hat{m}_{-k}(1, W_i) - \hat{m}_{-k}(0, W_i)] + [Y_i - \hat{m}_{-k}(Z_i, W_i)] \cdot \hat{H}_i$$

    where

    $$\hat{H}_i = \frac{\mathbb{I}(Z_i = 1)}{\hat{p}_{-k}(W_i)} - \frac{\mathbb{I}(Z_i = 0)}{1 - \hat{p}_{-k}(W_i)}$$

    (same for κ_2).
- Step 3: The estimator of ATE is $\hat{\theta} = n^{-1} \sum_{i=1}^{n} \hat{\kappa}_{1i} / n^{-1} \sum_{i=1}^{n} \hat{\kappa}_{2i}$

Design 2: Implementation

Let,

$$\hat{\kappa}_{1i} = [\hat{m}_{-k}(1, W_i) - \hat{m}_{-k}(0, W_i)] + [Y_i - \hat{m}_{-k}(Z_i, W_i)] \cdot \hat{H}_i$$
$$\hat{\kappa}_{2i} = [\hat{g}_{-k}(1, W_i) - \hat{g}_{-k}(0, W_i)] + [Y_i - \hat{g}_{-k}(Z_i, W_i)] \cdot \hat{H}_i$$

and

$$\hat{\eta}(W_i) = \hat{\kappa}_{1i} + \hat{\theta} \cdot \hat{\kappa}_{2i}. \tag{16}$$

Then,

$$\hat{\Sigma} = \frac{n^{-1} \sum_{i=1}^{n} \hat{\eta}(W_i)^2}{n^{-1} \sum_{i=1}^{n} \hat{\kappa}_{2i}^2} \tag{17}$$

is a valid estimator of the variance of $\hat{\theta}$.

Design 2: Implementation

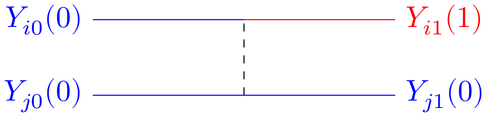
It turns out that $\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \Sigma)$ under certain weak assumptions.

The procedure implicitly relies on a Neyman Orthogonal moment. Details in:

- Chernozhukov, Hansen, Wuthrich (2020). 'Instrumental variable quantile regression', arXiv:2009.00436
- Okui, Small, Tan, and Robins (2012) 'Doubly robust instrumental variable regression'. Statistica Sinica 22.1, 173–205.

Design 3: Difference-in-Difference.

The setting is



where i belongs to a "treated" group and "j" does not.

The effect of the *intervention* is identified if

$$E(Y_{i1}(0) - Y_{i0}(0)|D = 1) = E(Y_{i1}(0) - Y_{i0}(0)|D = 0)$$

in which case τ is identified:

$$\begin{aligned}\tau = & \Big[E(Y_{i1}|D = 1) - E(Y_{i0}|D = 1)\Big] \\ & - \Big[E(Y_{i1}|D = 0) - E(Y_{i0}|D = 0)\Big] \end{aligned} \tag{18}$$

Difference-in-Differences

The common trend assumption will often be palatable only within subgroups (defined by W):

$$E(Y_{i1}(0) - Y_{i0}(0)|D = 1, W) = E(Y_{i1}(0) - Y_{i0}(0)|D = 0, W)$$

Several Neyman Orthogonal moments:

- **Chang (2020)** 'Double/Debiased Machine Learning for Difference-in-Differences Models'. Econometrics Journal 23 (2), pp. 177–191
- Sant'Anna, Zhao (2020) 'Doubly Robust Difference-in-Differences Estimators'. Journal of Econometrics 219 (1), pp. 101–122
- Zimmert (2020) Efficient Difference-in-Differences Estimation with High-Dimensional Common Trend Confounding," arXiv:1809.01643

Difference-in-Differences

One needs to distinguish the cases of panel data vs cross-sectional data. We focus on the former (required modifications are minimal).

Estimation to follow k-fold Cross-fitting, based on the estimator (Chang, 2020)

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \hat{g}_k(W_i)}{\hat{p}_k(1-\hat{g}_k(W_i))}(Y_i(1) - Y_i(0) - \hat{\Delta}(W_i)) \qquad (19)$$

where $g = E(D|W)$, $p = P(D = 1)$, $\hat{\Delta}$ is a ML estimator of

$$\Delta = E(Y(1) - Y(0)|W, D = 0)$$