Ignorability.

Eduardo Fé

Introduction

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

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

However D_i is not randomly allocated; in particular D and $Y(d)$ are not independent.

SUTVA

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.

Ignorability

Assumption 2, Ignorability. For $d \in \{0, 1\}$:

$$D \perp Y(d)|W$$

- Variation in assigned treatment is a good as random given W.
- If we look at units with the same W, say $W = w$, then variation in D is *as if* independent of $Y(d)$.
- W contains **all** the relevant information required to explain assignment to treatment (and once that information is taken into account, any variation left in D across individuals is random).

Ignorability implies that we will be able to identify the effect of D by comparing outcomes acorss treated and control units, given $W = w$.

Overlap / Full support

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

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

Identification.

Consider the moment $E(Y|D=d,W)$. This quantity can be approximated using data[1].

Then, for any $d \in \{0,1\}$

$$\begin{aligned} E(Y|D=d,W) &= E(Y(d)|D=d,W) \text{ by SUTVA} \\ &= E(Y(d)|W) \text{ by ignorability} \end{aligned} \tag{1}$$

[1] Depending on the context, by a regression, a Machine Learning method or even by a sample mean of Y for those observations with $D=d$ and and given value of W.

Identification.

The implication is that,

$$\begin{aligned} & E(Y|D=1,W) - E(Y|D=0,W) \\ =& E(Y(1)|W) - E(Y(0)|W) = \tau(W) \end{aligned} \tag{2}$$

which is the **Conditional Average Treatment Effect**. Then we can retrieve the Average Treatment Effect[2],

$$\begin{aligned} & E\Big[E(Y|D=1,W) - E(Y|D=0,W)\Big] \\ =& E\Big[E(Y(1)|W) - E(Y(0)|W)\Big] \\ =& E\Big[\tau(W)\Big] = \tau \end{aligned} \tag{3}$$

[2] This follows from the Law of Iterated Expectations.

The role of the propensity score.

Estimation of CATE and ATE requires a good fit for $E(Y|D,W)$.

This can, in general, be complicated. Machine Learning can potentially help, but direct application of ML will not work (as explained before).

Arguably, one can often have a better understanding of the propensity score, $P(D=1|W)=p(W)$,

The role of the propensity score.

Indeed, Rosenbaum and Rubin showed that under assumptions 1 to 3,

$$Y(d) \perp D | p(W) \tag{4}$$

Implying that $E(Y|D = d, p(W)) = E(Y(d)|p(W))$.

It would then be enough to estimate this moment to identify the treatment effect.

This would be particularly convenient if $p(W)$ is known (as in randomized experiments).

The role of the propensity score.

Horvitz - Thompson: Under Assumptions 1 to 3,

$$E\left[Y \cdot \frac{\mathbb{I}(D=d)}{P(D=d|W)}\middle|W\right] = E(Y(d)|W) \tag{5}$$

Note, then,

$$\begin{aligned} &E\left[Y \cdot \left(\frac{\mathbb{I}(D=1)}{P(D=1|W)} - \frac{\mathbb{I}(D=0)}{1-P(D=d|W)}\right)\middle|W\right] \\ =&E(Y(1)-Y(0)|W) = \tau(W) \end{aligned} \tag{6}$$

Hereafter, let

$$H = \frac{\mathbb{I}(D=1)}{P(D=1|W)} - \frac{\mathbb{I}(D=0)}{1-P(D=1|W)} \tag{7}$$

Clearly

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

Operationalisation

Directly application of ML to estimate the propensity score will fail to deliver good inference.

The reason is that $E(Y \cdot H|W)$ will not satisfy Neyman Orthogonality.

Operationalisation

Suppose data come from

$$Y_i = m(D_i, W_i) + \varepsilon_i \qquad \text{where } E(\varepsilon_i | W_i, D_i) = 0 \tag{8}$$

$$D_i = p(W_i) + \nu_i \qquad \text{where } E(\nu_i | W_i) = 0 \tag{9}$$

where the functions $m(.)$ and $p(.)$ are unknown.

- Assumptions 1-3 are implicit above
- Very general model: no linearity, additivity (except for the error term)

The ATE is

$$\tau = E\left[m(1, W_i) - m(0, W_i)\right] \tag{10}$$

Operationalsiation

The following function is Neyman Orthogonal:

$$\eta(W) = [m(1, W_i) - m(0, W_i)] + [Y_i - m(D, W)] \cdot H_i \quad (11)$$

where

$$H_i = \frac{\mathbb{I}(D_i = 1)}{P(D_i = 1|W_i)} - \frac{\mathbb{I}(D_i = 0)}{1 - P(D_i = 1|W_i)} \quad (12)$$

and, critically, $E(\eta(W)) = ATE$. Combines:

- Direct estimation, via conditional mean of the outcome...
- Indirect estimation, via propensity score

Estimation with ML methods will follow the Orthogonalisation + Cross fitting procedure described in a previous lesson.

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 $\hat{\tau}$, $\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (\hat{\eta}_i - \hat{\tau})^2$

Properties.

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

$$\sqrt{n}(\hat{\tau} - \tau) \sim N(0, \Sigma) \tag{13}$$

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

Implementation.

Cross fitting structure

```
library(caret)
library(randomForest)

# Number of folds for cross fitting.
K <- 5
#Caret function to create folds
flds <- createFolds(Y, K)

# We store results for each fold in this vector:
eta <- rep(NA, n)

# Loop (from k=1 to number of folds)
for(k in 1:length(flds)){

}
```

Implementation.

Inside the 'for' loop, first get subsamples

```
    Xin<-X[flds[[k]],]
    Yin<-Y[flds[[k]]]
    Din<-D[flds[[k]]]

    Xout<-X[-flds[[k]],]
    Yout<-Y[-flds[[k]]]
    Dout<-D[-flds[[k]]]
```

Implementation.

Inside the 'for' loop, first get subsamples

```
    Xin1 <- Xin[Din==1]
    Xin0 <- Xin[Din==0]

    Xout1 <- Xout[Dout==1,]
    Xout0 <- Xout[Dout==0,]
    Yout1 <- Yout[Dout==1]
    Yout0 <- Yout[Dout==0]
    Dout1 <- Dout[Dout==1]
    Dout0 <- Dout[Dout==0]
```

Implementation.

Inside the 'for' loop, training to get $\hat{m}_{-k}(0, W_i)$ and $\hat{m}_{-k}(1, W_i)$

```
# Train on observations with D=0, and predict
# with observations in the fold

mhatD0 <- randomForest(Xout0, Yout0)
yhat0 <- predict(mhatD0, newdata = Xin)

# Train on observations with D=1, and predict
# with observations in the fold

mhatD1<-randomForest(Xout1, Yout1)
yhat1 <- predict(mhatD1, newdata = Xin)
```

Implementation.

Inside the 'for' loop, compute the propensity score $\widehat{p}_{-k}(W_i)$ and "trim"

```
# Propensity score
ghat <- randomForest(Xout, Dout)
pScore <- predict(ghat, newdata =Xin)

# Trim (value of trim is 0.01, chosen arbitrarily)
pScore<-pmax(pmin(pScore, 1-0.01),0.01)
```

Implementation.

Inside the 'for' loop, finally compute $\hat{H}_i$

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

```
Hi <- (Din/pScore) - (1-Din)/(1-pScore)
```

and store

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

```
Hi <- (Din/pScore) - (1-Din)/(1-pScore)

etai <- (yhat1 - yhat0) +
  (Yin - (yhat1*Din + yhat0*(1-Din)) ) * Hi

eta[flds[[k]]]<-etai
```

Example

Generated 1000 samples of 500 observations from:

$$Y_i = D_i\tau + \cos(W_i'\beta)^2 + \varepsilon_i \text{ with } \varepsilon_i \sim N(0,1) \quad (14)$$

$$D_i = \mathbb{I}\left(\sin(X_i\gamma) + \cos(X_i'\gamma) + \nu_i > 0\right) \text{ with } \nu_i \sim N(0,1) \quad (15)$$

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

We set $\tau = 0$.

Example.

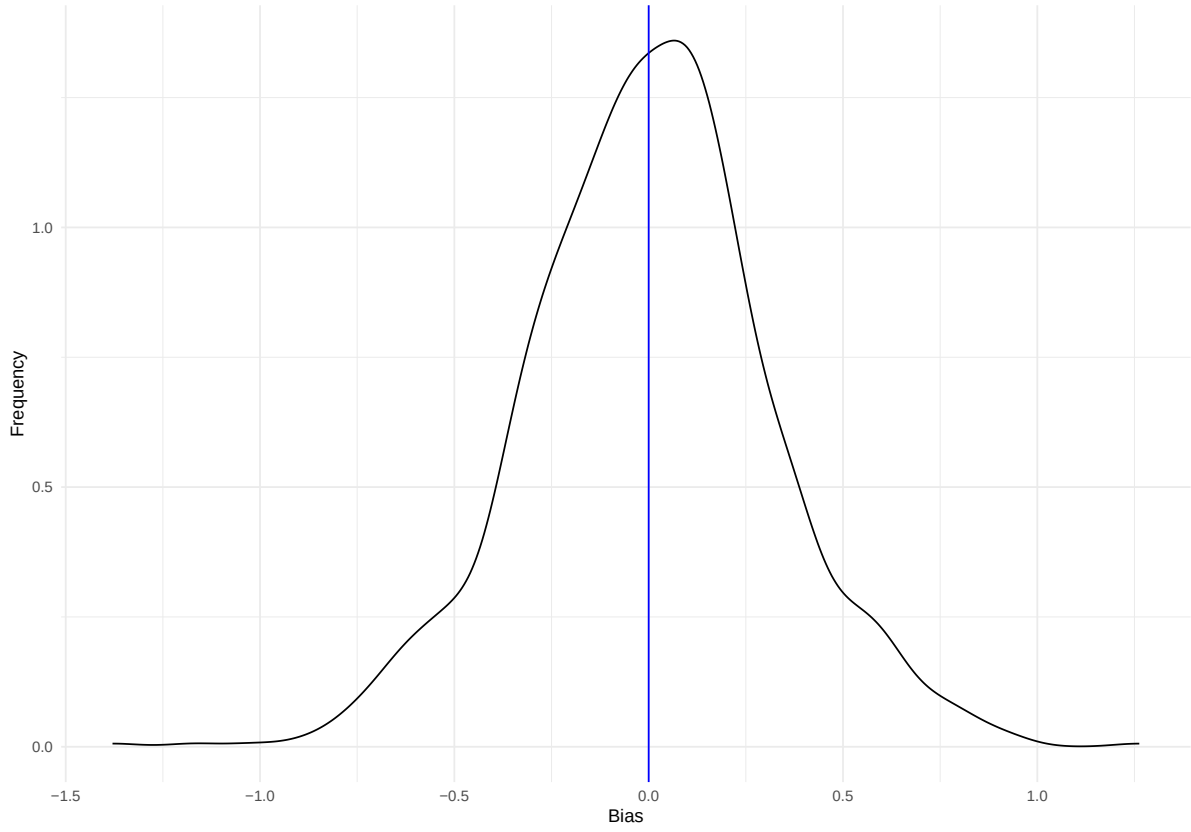

Figure 1: Distribution of the bias.

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