Selection on Observables: Implementation

ECON 31720 Applied Microeconometrics

Francesco Ruggieri

The University of Chicago

October 14, 2020

Framework for Selection on Observables

- $D \in \{0,1\}$ is a **binary treatment**, $Y \in \mathbb{R}$ is an **outcome** of interest
- D and Y are linked by **potential outcomes** $Y(0), Y(1)$
- $X \in \mathbb{R}^{d_x}$ is a vector of predetermined, **observable** characteristics with support $\mathcal{X}$
- The treatment is **as-good-as randomly assigned conditional on observables**:

$$(Y(0), Y(1)) \perp\!\!\!\perp D | X = x \qquad \text{for all } x \in \mathcal{X}$$

- No realization of X deterministically implies a treatment state (**overlap condition**):

$$0 < p(x) < 1 \quad \forall x \in \mathcal{X}$$

where $p(x) \equiv \mathbb{P}(D = 1 | X = x)$ is the **propensity score**

Shortcomings of Imputation with Linear Regression

Last week we discussed the shortcomings of **regression-based imputation estimators**:

1. Linear regression **with additive separability** between D and X:

$$Y = \alpha^* + \beta^* D + X'\gamma^* + U \quad \text{with} \quad \mathbb{E}[U] = \mathbb{E}[DU] = 0 \text{ and } \mathbb{E}[XU] = 0_{d_x}$$

   - This approach is undesirable because it imposes $\mathbf{ATE} = \mathbf{ATT} = \mathbf{ATU} \approx \boldsymbol{\beta}^*$

2. Linear regression **without additive separability** between D and X:

$$Y = \alpha^* + \beta^* D + X'\gamma^* + XD'\delta^* + U \quad \text{with} \quad \mathbb{E}[U] = \mathbb{E}[DU] = 0 \text{ and } \mathbb{E}[XDU] = \mathbb{E}[XU] = 0_{d_x}$$

   - This approach allows for target parameters to **differ** based on the distribution of X
   - But **linear extrapolation** makes it sensitive to settings in which observables are **unequally distributed** across treatment states (**negative weights** to some outcome contrasts)

Selection on Observables à la Imbens (2015)

1. **Design** stage: assess **overlap** in covariate distributions
   - Estimate the propensity score
   - Drop units with extreme values for the estimated propensity score
2. **Supplementary analysis** stage: assess the plausibility of **selection on observables**
   - Partition the set of observables into pseudo-outcomes and other covariates
   - Estimate the average treatment effect on pseudo-outcomes, controlling for other covariates
   - Suggestive evidence in favor of unconfoundedness if estimated average treatment effect is ≈ 0
3. **Analysis** stage: estimate the **target parameter(s)** of interest
   - Blocking on the estimated propensity score or one-to-one (or -$\overline{k}$) matching with replacement

Design Stage: Assessing Overlap in Covariate Distributions

- Let $k \in \{1, \ldots, \overline{k}\}$ **index** elements of the vector of observables X

- Define the following sample **means** and sample **variances** for each k:

$$\overline{X}_{1,k} \equiv \frac{1}{n_1} \sum_{i:D_i=1} X_{i,k} \qquad \overline{X}_{0,k} \equiv \frac{1}{n_0} \sum_{i:D_i=0} X_{i,k}$$

$$S^2_{1,k} \equiv \frac{1}{n_1 - 1} \sum_{i:D_i=1} \left(X_{i,k} - \overline{X}_{1,k}\right)^2 \qquad S^2_{0,k} \equiv \frac{1}{n_0 - 1} \sum_{i:D_i=0} \left(X_{i,k} - \overline{X}_{0,k}\right)^2$$

- Use **normalized differences** in average covariates as opposed to **t-statistics**:

$$\Delta_k \equiv \frac{\overline{X}_{1,k} - \overline{X}_{0,k}}{\sqrt{\frac{S^2_{1,k}}{2} + \frac{S^2_{0,k}}{2}}} \quad \text{vs.} \quad t_k \equiv \frac{\overline{X}_{1,k} - \overline{X}_{0,k}}{\sqrt{\frac{S^2_{1,k}}{n_1} + \frac{S^2_{0,k}}{n_0}}}$$

- Large normalized differences point to an **unequal distribution of $\boldsymbol{X}$** for $d = 0, 1$

1. Framework for Selection on Observables

2. Selection on Observables à la Imbens (2015)
   - Assessing Overlap in Covariate Distributions
   - Estimating the Propensity Score
   - Ensuring Overlap in Covariate Distributions
   - Assessing Selection on Observables
   - Estimating Target Parameters

3. Summary

Design Stage: Estimating the Propensity Score

- Imbens and Rubin (2015) proposes that $p(X)$ be modeled based on **logistic regression**:

$$p\left(x\right)=\frac{\exp\left(h(x)'\gamma\right)}{1+\exp\left(h(x)'\gamma\right)} \quad \text{with } h:\mathcal{X}\rightarrow\mathbb{R}^m \text{ and } \gamma\in\mathbb{R}^m$$

  They recommend checking for **robustness** using a probit model, $p\left(x\right)=\Phi\left(x'\gamma\right)$

- The propensity score does **not** have a **structural or causal interpretation** in this setting
  - It will have such an interpretation in the context of IV with heterogeneous treatment effects

- The goal is to provide the **"best" approximation** to the conditional expectation $\mathbb{E}\left[D|X\right]$

- A key **choice** is the vector of functions of the observables, $\boldsymbol{h}\left(\cdot\right)$
  - The most common choice is $\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{x}$, but this may **not** be **flexible** enough

Design Stage: Estimating the Propensity Score

- Imbens and Rubin (2015) proposes a data-driven approach based on **stepwise regression**
  - **LASSO** seems an attractive alternative (see Belloni, Chernozhukov, and Hansen, 2012)

- This approach limits the components of $h(x)$ to be **second-order polynomials**
  - $h(x)$ contains either components of x or the product of two components of x

- How to choose among the $\overline{k}\left(\overline{k}+1\right)/2-1$ first- and second-order terms?
  1. Choose a **subset** of the covariates to be included in the **linear** part of the specification, X_B
  2. Choose a **threshold value** to include **more linear terms** based on **likelihood ratio** tests
     - Include if the null hypothesis that the coefficient on the additional covariate is 0 is rejected
  3. Analogously, choose a **threshold value** to include **quadratic terms**
     - Include if the null hypothesis that the coefficient on the second-order term is 0 is rejected

Design Stage: Estimating the Propensity Score

Once $h(\cdot)$ is chosen, the propensity score can be estimated with **maximum likelihood**:

$$\widehat{p}(x) = \frac{\exp\left(h(x)'\widehat{\gamma}\right)}{1 + \exp\left(h(x)'\widehat{\gamma}\right)}$$

General **computation tips** for any likelihood maximization problem:

- Ensure that explanatory variables have roughly the **same order of magnitude**
- When feasible, supply an **analytical gradient** and an **analytical Hessian**

Design Stage: Ensuring Overlap in Covariate Distributions

Two approaches to ensure **covariates** are **similarly distributed** across treatment states:

1. **Matching without replacement** on the propensity score
   - This approach is suitable for settings in which $n_0 \gg n_1$ and the focus is to estimate the ATT

2. **Dropping** observations with **extreme values** of the propensity score
   - Goal: reduce the sensitivity to minor specification changes by eliminating hard-to-match units

Design Stage: Ensuring Overlap in Covariate Distributions

Matching without replacement on the propensity score to create a **balanced sample**:

1. Estimate the **propensity score** and compute the estimated **log-odds ratio**:

$$\widehat{\ell}(x) \equiv \ln\left(\frac{\widehat{p}(x)}{1-\widehat{p}(x)}\right)$$

   Recall: the log-odds ratio is the inverse of the standard logistic function and is linear in x

2. **Sort treated units** based on their log-odds ratio in descending order

3. **Match** each treated unit with the **closest** control unit in terms of $\widehat{\ell}(x)$
   - Start with the first treated unit (i.e., highest log-odds ratio) and proceed without replacement

The outcome is a sample of $2 \times n_1$ units, half of them treated and half of them controls.

Design Stage: Ensuring Overlap in Covariate Distributions

Dropping observations with extreme values of the propensity score:

- This approach is based on Crump, Hotz, Imbens, and Mitnik (CHIM, 2008)
- Choose a **subset** of the covariate space, $\mathcal{A} \subset \mathcal{X}$, such that the **overlap** condition holds
  - Consider the average treatment effect $\tau(\mathcal{A}) = \mathbb{E}[Y(1) - Y(0)|X \in \mathcal{A}]$
  - Intuition: if, for some $x \in \mathcal{X}$, $n_1(x) \gg n_0(x)$ or $n_1(x) \ll n_0(x)$, then $\widehat{\mathrm{Var}}[\hat{\tau}(\mathcal{A})]$ is **large**
  - Solution: **exclude units** with covariate values such that $n_1(x) \gg n_0(x)$ or $n_1(x) \ll n_0(x)$

Design Stage: Ensuring Overlap in Covariate Distributions

- In the simple case of **homoscedasticity**, $\sigma_1^2 = \sigma_0^2 = \sigma^2$, the **asymptotic variance** is

$$\mathrm{Var}\left[\widehat{\tau}\left(\mathcal{A}\right)\right] = \frac{\sigma^2}{\mathbb{E}\left[X \in \mathcal{A}\right]}\mathbb{E}\left[\frac{1}{p\left(X\right)\left(1 - p\left(X\right)\right)}\middle| X \in \mathcal{A}\right]$$

- The set $\mathcal{A}$ that **minimizes** the asymptotic variance of this estimator is

$$\mathcal{A}^* = \{x \in \mathcal{X} | \alpha \leq p\left(x\right) \leq 1 - \alpha\}$$

where

$$\frac{1}{\alpha(1-\alpha)} = 2 \times \mathbb{E}\left[\frac{1}{p\left(X\right)\left(1 - p\left(X\right)\right)}\middle| \frac{1}{p\left(X\right)\left(1 - p\left(X\right)\right)} \leq \frac{1}{\alpha\left(1-\alpha\right)}\right]$$

Design Stage: Ensuring Overlap in Covariate Distributions

1. Estimate the **propensity score**, $\widehat{p}(X)$, and define the function

$$g : \mathcal{X} \to \mathbb{R} \quad \text{with} \quad g(x) = \frac{1}{\widehat{p}(x)(1 - \widehat{p}(x))}$$

2. Define an **objective function** by taking the sample analog of the variance above:

$$h : \mathbb{R} \to \mathbb{R} \quad \text{with} \quad h(\lambda) = \frac{1}{\left(\sum_{i=1}^{n} \mathbb{I}[g(X) \leq \lambda]\right)^2} \sum_{i=1}^{n} \mathbb{I}[g(X) \leq \lambda]\, g(X)$$

3. Compute $\widehat{\boldsymbol{\lambda}} \equiv \arg\min_\lambda h(\lambda)$ by evaluating $h(\lambda)$ at $\lambda = g(X_i)\ \forall i \in \{1, \ldots, n\}$

4. Use $\widehat{\lambda}$ to compute $\widehat{\alpha} = \frac{1}{2} - \sqrt{\frac{1}{4} - \widehat{\lambda}^{-1}}$, which **solves the equation** in α above

Supplementary Analysis Stage: Assessing Selection on Observables

- The assumption $Y(0), Y(1) \perp\!\!\!\perp D|X$ is **untestable**
- However, its **plausibility** can be assessed with **suggestive evidence**
- Intuition: estimate the treatment effect on a **pseudo-outcome**, i.e.,
  - A variable known to be **unaffected** by the treatment (typically a lagged outcome)
- Lagged outcomes are usually available in datasets for the evaluation of training programs

Supplementary Analysis Stage: Assessing Selection on Observables

1. Partition X into **lagged outcomes** and **time-invariant characteristics**:

$$X = \left(Y_{-1}, Y_{-2}, \ldots, Y_{-\bar{t}}, Z'\right)'$$

2. Assume **unconfoundedness** given only **$\bar{t} - 1$ lags** of the outcome:

$$Y(0), Y(1) \perp\!\!\!\perp D| \left(Y_{-1}, Y_{-2}, \ldots, Y_{-(\bar{t}-1)}, Z'\right)'$$

3. Assume **stationarity and exchangeability**:

$$f_{Y_{i,s}(0)|Y_{i,s-1}(0),\ldots,Y_{i,s-(\bar{t}-1)}(0),Z_i,D_i}\left(y_s|y_{s-1}, \ldots, y_{s-(\bar{t}-1)}, z, d\right) \quad \text{does not depend on } i \text{ and } s$$

4. **Test** the **implied independence** of Y_{-1} and D given controls:

$$Y_{-1} \perp\!\!\!\perp D| \left(Y_{-2}, \ldots, Y_{-\bar{t}}, Z'\right)'$$

1. Framework for Selection on Observables
2. Selection on Observables à la Imbens (2015)
   - Assessing Overlap in Covariate Distributions
   - Estimating the Propensity Score
   - Ensuring Overlap in Covariate Distributions
   - Assessing Selection on Observables
   - **Estimating Target Parameters**
3. Summary

Analysis Stage: Estimate Target Parameters

Imbens (2015) proposes two alternative approaches to estimate target parameters:

1. **Blocking** on the estimated **propensity score** and within-block **regression adjustment**
2. **Covariate matching with replacement** and within-match **regression adjustment**

Analysis Stage: Estimate Target Parameters

Blocking on the estimated propensity score and within-block regression adjustment:

1. **Partition** the range of the propensity score, the interval $[0,1]$, into $\bar{j}$ **intervals**
   - Each interval is $[b_{j-1}, b_j)$ for $j \in \{1, \ldots, \bar{j}\}$, with $b_0 = 0$ and $b_1 = 1$

2. **Estimate** the **within-block average treatment effect** with linear regression:

$$\left(\widehat{\alpha}_j, \widehat{\tau}_j, \widehat{\beta}_j'\right)' = \arg\min_{(\alpha,\tau,\beta')'} \sum_{i=1}^{n} B_i(j)\left(Y_i - \alpha - \tau D_i - X_i'\beta\right)^2 \quad \text{with} \quad B_i(j) \equiv \mathbb{I}\left[b_{j-1} < p(x_i) \leq b_j\right]$$

   - **Regression extrapolation** not salient because $f_{X|D=1}(x|d=1) \approx f_{X|D=0}(x|d=0)$ within each block

3. **Aggregate** within-block average treatment effects. For instance:

$$\widehat{\tau}^{\text{ATE}}_{\text{block}} \equiv \sum_{j=1}^{\bar{j}} \frac{n_{1j} + n_{0j}}{n} \times \widehat{\tau}_j$$

Analysis Stage: Estimate Target Parameters

A data-dependent **algorithm** to choose the **number** of blocks and their **boundaries**:

1. Estimate the **log-odds ratio**, $\widehat{\ell}(x) \equiv \ln\left(\frac{\widehat{p}(x)}{1-\widehat{p}(x)}\right)$

2. Consider block j and **perform a t-test** using the log-odds ratio:

$$t = \frac{\overline{\widehat{\ell}_{1j}} - \overline{\widehat{\ell}_{0j}}}{\sqrt{\frac{S^2_{\ell,1j}}{n_{1j}} + \frac{S^2_{\ell,1j}}{n_{0j}}}}$$, where means and variances are computed within block j

3. **Split** block j at the **median** of the values of the within-block propensity scores **unless**
   1. the t-statistic above is **smaller** than a predetermined **critical value**, or
   2. the number of **units** in any of the new potential blocks is $\overline{\boldsymbol{k}} + \mathbf{2}$ **or less**, or
   3. the number of **treated/control units** in any of the new potential blocks is **3 or less**

Analysis Stage: Estimate Target Parameters

One-to-one covariate matching with replacement and within-match regression adjustment:

1. Match all units **with replacement** (so the order does not matter)
   - Let $x, x' \in \mathcal{X}$. The **Mahalanobis metric** is

     $$||x, x'|| = (x - x')' \, \widehat{\Omega}_X^{-1} \, (x - x')$$

     where $\widehat{\Omega}_X$ is the sample variance-covariance matrix of the covariates
   - **Match** each treated unit to its **closest** untreated unit, and viceversa

2. Consider the two samples of (**observed and matched**) treated and control units
   - Within each of the **two** n-dimensional samples, estimate the **linear regression**

     $$Y_d = \alpha_d + X_d' \beta_d + U_d \quad \text{for } d \in \{0, 1\}$$

Analysis Stage: Estimate Target Parameters

3. Following Abadie and Imbens (2006, 2010), **impute potential outcomes** as

$$\widehat{Y}_i^{\text{adj}}(0) = \begin{cases} Y_i & \text{if } D_i = 0 \\ Y_i^m + (X_i - X_i^m)' \, \widehat{\beta}_0 & \text{if } D_i = 1 \end{cases} \qquad \widehat{Y}_i^{\text{adj}}(1) = \begin{cases} Y_i^m + (X_i - X_i^m)' \, \widehat{\beta}_1 & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

where Y_i^m and X_i^m denote, respectively, the **outcome and covariates** of unit i's **match**

4. Construct the **bias-adjusted** matching estimator

$$\widehat{\tau}^{\text{ATE}}_{\text{match,adj}} \equiv \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{Y}_i^{\text{adj}}(1) - \widehat{Y}_i^{\text{adj}}(0) \right)$$

As above, linear regression in this setting is "harmless" due to covariate balance.

Summary

1. Assess the **overlap** condition using **normalized differences**
2. Estimate the **propensity score** with **logistic regression**
3. Ensure the **overlap** condition holds by either
   - **Matching** without replacement on the **propensity score**, or
   - **Dropping** observations with **extreme** values of the **propensity score**
4. Assess **selection on observables** by estimating treatment effects on **pseudo-outcomes**
5. Estimate target parameters using either
   - **Blocking** on the estimated **propensity score** and within-block **regression adjustment**, or
   - **Covariate matching** with replacement and within-match **regression adjustment**