EC 709: Discussion 1

Liang Zhong[1]

Boston University

samzl@bu.edu

September 2023

[1]Thanks to Rubaiyat Alam for kindly sharing his lecture notes

Overview

1. Weighting
2. Omitted Variable Bias
3. Preview: Multiple Hypotheses Testing
4. Preview: Potential Outcome framework

Table of Contents

1. Weighting
2. Omitted Variable Bias
3. Preview: Multiple Hypotheses Testing
4. Preview: Potential Outcome framework

Review: Weighting

- Problem: Non-representative sample, i.e. certain strata oversampled. When should we use the weight provided to correct for oversampling?
- Four cases in Kevin's videos:
  - Stratified sampling on an exogenous variable (e.g. Survey data)
  - Grouped data (e.g. state-level data)
  - Endogenous sampling: the probability of selection varies with the dependent variable even after conditioning on the explanatory variables
  - Differing slope coefficient (e.g. heterogeneous effect)
- Reference: Solon, Haider and Wooldridge (2015), "What Are We Weighting For?"

Need weighting – Heteroskedasticity

1. Stratified sampling on an exogenous variable + Individual-level Heteroskedasticity
    - Same reason for any Heteroskedasticity – GLS is more efficient
    - Weighting is unrelated to stratified sampling
    - Assume individual-level homoskedastic for the rest of the discussion

2. Grouped data + NO group-level FE
    - v_s: group-level error term; u_{is}: individual-level error term; N_s: number of observations in the group
    - $\sigma^2_{v_s} = \sigma^2_u / N_s \rightarrow$ Heteroskedasticity
    - In STATA, you need [$aweight = N_s$] to correct it

Need weighting – Endogenous sampling

Examples regarding PSID, which deliberately oversampled low-income households:

1. Estimating the earnings return to an additional year of schooling
   - Run IV regression of log earnings on years of schooling to handle endogeneity, with other control variables
   - Without correction for the oversampled low-income population $\Rightarrow$ inconsistent estimation of the parameters of interest

   $\Leftarrow$ Sampling criterion (family income) is related to the error term in the regression for log earnings
   - To perform IV estimation, need to weight the IV orthogonality conditions by the inverse probabilities of selection

2. "Special" Example: Estimating descriptives statistics of population
   - Actual value poverty rate in US $= 13\%$
   - Unweighted mean from PSID $= 26\% > 13\%$
   - Weighted mean using inverse probabilities of selection from PSID $= 12\% \approx 13\%$

Don't weight – Homoskedastic

1. Stratified sampling on an exogenous variable + Individual-level Homoskedastic
    - OLS is BLUE. Weighting creates heteroskedasticity and less efficient

2. Grouped data + group-level FE
    - v_s: group-level error term; u_{is}: individual-level error term; N_s: number of observations in the group; α_s: group-level fixed effect
    - $\sigma^2_{v_s} = \sigma^2_\alpha + \sigma^2_u / N_s$

$\rightarrow$ If N_s and σ^2_α large enough, then almost homoskedastic

$\rightarrow$ If N_s or σ^2_α is small then not homoskedastic, might need to weight

Don't weight – Differing slope coefficient

- Erroneous view: With unmodeled heterogeneous effects, weighting to reflect population shares identifies the population average partial effect
- In reality: if unmodeled heterogeneous effects, both OLS and WLS are generally inconsistent. Neither is better than other
- Suggestion by Solon et al:
  - Run both OLS and WLS: if very different estimates could mean heterogeneous effects
  - Study heterogeneity rather than trying to average it out; the latter is hard to extrapolate to different settings

Practical Guide and Wrap up

- Before weighting, think about What are we weighting for?
  - Heteroskedasticity: Test it before deciding whether to weight
  - $\Leftarrow$ For Grouped data: Regress residual square on a constant and N_s^{-1} (Modified Breusch-Pagan)
  - Endogenous-Sampling: Always weight
  - $\Leftarrow$ For M-estimation, weight criterion function by inverse probability of selection (Wooldridge, 1999)

- General suggestion by Solon et al: Run both OLS and WLS and compare estimates
  - Under exogenous sampling and correct model specification, should be similar
  - If different, indicates one or the other violated (e.g. heterogeneous effects not included)
  - It is advisable to use robust standard errors in most cases

Table of Contents

1. Weighting
2. Omitted Variable Bias
3. Preview: Multiple Hypotheses Testing
4. Preview: Potential Outcome framework

Review: Omitted Variable Bias

Problem: How to handle Omitted Variable Bias (OVB)?

- Find IV might not be feasible in some settings
- A "rough" sensitivity analysis: Add more covariates and show the stability of the coefficients
  - Oster(2019): Coefficient Stability $\not\equiv$ No OVB
- A "better" sensitivity analysis: Derive a bound for the coefficient
  - Oster (2019) provided the bound assuming values on R_{max} and δ_{Oster}
  - The bounding approach is useful in a varies of other settings (e.g. missing value)

  $\Leftarrow$ Relate to Partial identification which EC711 will cover in detail

Reference: Oster (2019), "Unobservable Selection and Coefficient Stability: Theory and Evidence" and STATA package **psacalc**

Oster (2019): Setup

$$Y = \beta X + \Phi\omega^o + W_2 + \epsilon$$

- β is the coefficient of interest
- X is scalar treatment
- ω^o is a matrix of observed covariates
- $W_2 = \gamma Z$ an unobserved index, where Z a matrix of unobserved covariates
- Define $W_1 = \Phi\omega^o$, assume WLOG W_1 and W_2 are orthogonal (everything holds if they are correlated, just need to transform W2)
- Let $\sigma_{iX} = cov(W_i, X), \sigma_i^2 = var(W_i)$ for $i \in \{1, 2\}$

Oster (2019): Notation

- δ_{Oster} satisfied:

$$\delta_{Oster} \frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}$$

  - Oster (2019): "proportion of selection on unobserved controls with selection on observed controls"
- Practically, Regressing Y on X gives coefficient $\mathring{\beta}$ and R-squared $\mathring{R}$
- Practically, Regressing Y on X and ω^o gives coefficient $\tilde{\beta}$ and R-squared $\tilde{R}$
- Theoretically, Regressing Y on X, ω^o and W_2 gives coefficient $\beta^\star$ and R-squared R_{max}

Oster (2019): Bias-adjusted estimator

- Under strong conditions:

$$\beta^{\star} \approx \tilde{\beta} - \delta_{Oster}[\mathring{\beta} - \tilde{\beta}]\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}}$$

  - Coefficient Stability $\Rightarrow |\mathring{\beta} - \tilde{\beta}| \to 0$
  - However, a small $\tilde{R} - \mathring{R}$ would magnify the bias

  $\Rightarrow$ Bias is proportional to coefficient change scaled by change in R-squared

- Without strong conditions, the bias estimate might not be unique, and give three different values

$\Rightarrow$ Might want to assume the bias is fairly small, then:

  - $\beta^{\star}$ should be close to $\tilde{\beta}$
  - The OVB doesn't change the direction of the covariance between the observable index and the treatment

$\Rightarrow$ **psacalc** choose the estimation with both conditions hold

Implementation - Derive Boundary

- To gain some intuition, assume strong conditions hold:

$$\beta^\star \approx \tilde{\beta} - \delta_{Oster}[\mathring{\beta} - \tilde{\beta}]\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}}$$

  - WLOG assume $\mathring{\beta} - \tilde{\beta} > 0$ and $\delta_{Oster} > 0$

  $\Rightarrow$ $\beta^\star$ is decreasing with δ_{Oster} and R_{max}

  - If further assume $R_{max} \leq \hat{R}_{max}$, $\delta_{Oster} \leq 1$

  $\Rightarrow$ $\Delta = [\beta^\star(\hat{R}_{max}, 1), \tilde{\beta}]$

- How to implement in STATA?
  - Lower Bound: psacalc beta varname, delta(1) rmax($\hat{R}_{max}$)
  - Upper Bound: Regressing Y on X and ω^o
- With the bound, we can argue whether the coefficient is significant or not by comparing if 0 is in Δ

Implementation - Evaluate Robustness

- δ_{Oster} is a function of $\beta^{\star}$ and R_{max}

$\Rightarrow$ Assume value for R_{max} , find $\delta_{Oster} = \delta$ for which $\beta^{\star} = 0$

  - Interpretation: Degree of selection on unobservables relative to observables required to explain away the result
  - $\delta = 2$ suggests unobservables need to be twice as important as observables to produce zero treatment effect
  - the larger the δ the "better", $\delta = 1$ suggested as a good cutoff

- How to implement in STATA: psacalc delta varname
  - The default assumed $\beta^{\star} = 0$ and $R_{max} = 1$

Discussion on δ_{Oster}

- Cinelli & Hazlett (2020): "constructing indices W_1 and W_2 based on relationships to the outcome is not innocuous"

$\Rightarrow$ δ_{Oster} captures not only the relative influence of ω^o and Z over the treatment but also their association with the outcome!

  - $\delta_{Oster} = \frac{cov(\gamma Z, X)}{var(\gamma Z)} \frac{var(\Phi \omega^o)}{cov(\Phi \omega^o, X)} = \frac{\lambda}{\gamma} \frac{\Phi}{\theta}$
  - where λ and θ are the coefficients of the regression $X = \theta \omega^o + \lambda Z + \epsilon_X$

  $\Rightarrow$ With $\Phi = \theta = 1$, and any $p = \gamma = \lambda$, $\delta_{Oster} = 1$

- Oster(2019) trying to argue $\delta_{Oster} = 1$ as equal selection between observables and unobservables, but we might view $p > 1$ as the unobservable has more explanatory power

Framewrok of Cinelli & Hazlett (2020)

- Denote partial R^2s:

$$R^2_{Y\sim Z|X,\omega^o} = \frac{R^2_{Y\sim Z+X+\omega^o} - R^2_{Y\sim X+\omega^o}}{1 - R^2_{Y\sim X+\omega^o}}$$

$$R^2_{X\sim Z|\omega^o} = \frac{R^2_{X\sim Z+\omega^o} - R^2_{X\sim \omega^o}}{1 - R^2_{X\sim \omega^o}}$$

  - Added explanatory power when including the unobserved confounder Z to outcome/treatment regressions
  - $R^2_{Y\sim Z|X,\omega^o} = 1$: Z explaining all residual variance of the outcome

- A different procedure using partial R^2s:

$$|\beta^\star - \tilde{\beta}| = se(\tilde{\beta})\sqrt{\frac{R^2_{Y\sim Z|X,\omega^o} R^2_{X\sim Z|\omega^o}}{1 - R^2_{X\sim Z|\omega^o}} df}$$

  - $se(\tilde{\beta})$: standard error of $\tilde{\beta}$; df: degree of freedom for $\tilde{\beta}$'s regression
  - Need further assumption on the direction of bias, but the magnitude is already helpful

About Sensitivity Analysis

- For those who are interested:
  - R package **sensemakr** available: `https://cran.r-project.org/web/packages/sensemakr/index.html`
  - A Shiny web application: `https://carloscinelli.shinyapps.io/robustness_value/`
- Both Oster(2019) and Cinelli & Hazlett (2020) tell the researcher how strong unobserved confounding would have to be to change meaningfully the treatment effect estimate beyond some level we are interested in and employ observed covariates to argue for bounds on unobserved confounding where possible
- Many other papers use sensitivity analysis on OVB or other empirical issue
  - Tell us what we would have to be willing to believe to accept the substantive claims that were initially made (Rosenbaum, 2005, 2010, 2017)
  - Oster (2019) serves as a good example of how these papers look like

Table of Contents

1. Weighting
2. Omitted Variable Bias
3. Preview: Multiple Hypotheses Testing
4. Preview: Potential Outcome framework

P-hacking and Multiple Hypotheses Testing

P-hacking is a severe issue in social sciences and occurs when:

1. Doing multiple tests on a single data set
    - Involves running lots of regressions with different specifications until we get $p < 0.05$ for some variable
    - the "significant" results are picking up some spurious correlation in the data that is endemic to our sample and not the population
    - Extreme Example: Use a data set consisting of **independent random numbers**, and running a large number of hypothesis testing, 5% of those would be "statistically significant" at the 0.05 level in the long run (definition of the significance level)
2. Doing the same test (or different ones) on different data sets
    - See videos 7&8 for simulation and the jelly bean comic

- Both count as Multiple Hypotheses Testing (MHT)

Problem of MHT

- Intuition: Test the null hypothesis of a fair coin
- Suppose flip it 10 times, and it comes out heads 9 times
  - Under the null, $P = 10 \times (.5)^{10} \approx 0.01$, reject at level $\alpha = 0.05$
- Now, suppose I have 100 fair coins (a fact that I do not know) and flip them all 10 times to see if they are fair
  - For a pre-selected coin, P(9 heads out of 10 toss)≈ 0.01
  - P(some coin in the 100 will get 9 heads) $\approx 1 - (1 - 0.01)^{100} \approx 0.66$

  $\Rightarrow$ a very good chance of finding one "unfair" coin if we are searching and not pre-selecting.
- Formally: When testing a single null (which is true):

$$P(\text{making a type I error}) = \alpha$$

- When testing multiple hypotheses (all of which are true):

$$P(\text{making at least one type I error}) = 1 - (1 - \alpha)^n > \alpha$$

$\Rightarrow$ The more we 'look' at the data/test more hypotheses, the more likely to falsely reject

Pre-analysis plan

- Would pre-selecting solve the problem?
  - Known as "pre-analysis plan": Pre-specify exactly how the analysis will be done before looking at data. This plan is then submitted to an online registry and adhered to for RCT
  - Also applied to empirical works other than RCT to prevent the researcher from running many specifications in search of a significant result
- Maybe or maybe not:
  - As mentioned by Kevin in video 8, it might encourage researchers to list all the specifications in the pre-analysis plan

  $\Rightarrow$ unlikely to solve this problem and may exacerbate it
- If we strictly limit the specifications for the researchers, there might be other practical problem

Problem of Pre-analysis plan

1. Specify a *single* outcome metric, circumventing MHT
   - Sometimes difficult to settle on a metric

   $\Rightarrow$ Want to see the effect of treatment on educational outcomes: Do you use i) test scores and ii) school dropout rates?
   - In practice, the significance of either outcome variable would be interesting to the policy-maker
2. Subgroup analysis: Pre-specified to prevent the regressions on various subgroups
   - Limited amount of mechanisms you can check that lie behind the results
   - Hinders exploratory analysis that points out important hypotheses
   - Impossible to find "surprising" patterns in data

- Overall, Pre-analysis plans are suited to papers that see the effect of treatment on outcomes (e.g., RCT)

So how to handle MHT?

- Key Intuition: Adjust the size so that the more tests are carried out, the more difficult it gets to reject a null

$\Rightarrow$ When testing more and more nulls, you will be penalized by facing lower and lower p-values to go under to reject the null

- How do we determine the p-value we need?
  - Econometricians target some criteria for MHT (e.g., FWER and FDR) parallel to the "size" in the single testing case
  - However, we don't want to be too conservative and worsen the issue of Type II error
  - Same size power trade-off applies here
- I will leave you to video 9&10 for details

Table of Contents

1. Weighting
2. Omitted Variable Bias
3. Preview: Multiple Hypotheses Testing
4. Preview: Potential Outcome framework

Beyond OLS

- Previous three lectures focused on practical problems for empirical research
  - All of them implicitly assume you are running regressions
- Run regression is not the solution to everything
  1. Wrong functional form (misspecification) $\rightarrow$ endogeneity $\rightarrow$ inconsistency
  2. No information on the underlying mechanisms

- So three languages appear in the STATs world:
  - Rubin causality model (RCM): minimum assumption on the functional form
  - Structural equation modeling (SEM): modeling the mechanisms.
    - Not the same as the structural model, and outdated nowadays
  - Directed acyclic graph (DAG): A modern version of SEM, uses the graph to illustrate the association between key covariates
    - Seldomly used in Econ, won't cover in this course

Some History on the RCM

- Parameter of interest: $Y_i(\text{Treated}) - Y_i(\text{Control})$
  - But we cannot observe both in practice

  $\Rightarrow$ Causal inference is a missing data problem! (Rubin, 1975)

- How to solve it?
  - Neyman (1923) proposed the potential outcome framework

  $\Rightarrow$ First ever formalize the problem, but only for Experiments, proposed RCT
  - Fisher (1925) proposed randomization inference (Fisher Randomization testing)

  $\Rightarrow$ Another clever way to handle the missing data problem; I'll cover it next time
  - Rubin (1975) extends the framework to observational studies

  $\Rightarrow$ Proposed the key assumption: Selection on observables (unconfoundedness)
- I'll leave to videos 11-14 for details

Thank You!