Announcements -- Research Proposal

- You can find my detailed comments about your rough draft on Canvas
- Try to come see me before starting final draft if you have questions about comments
- See six example proposals on Canvas
- Final proposal due on May 3
Announcements – Exercise #4

Exercise #4 is due next week

- Please upload to Canvas, Thanks!
Background readings for today

- Roberts-Whited, *Section 6*
- Angrist-Pischke, *Sections 3.3.1-3.3.3*
- Wooldridge, *Section 21.3.5*
Quick review of last lecture on “errors”

Discuss matching

- What it does…
- And what it doesn’t do

Discuss Heckman selection model

Student presentations of “Error” papers
Quick Review [Part 1]

What are 3 data limitations to keep in mind?

- **#1 – Measurement error**; some variables may be measured with error \([e.g.\text{ industry concentration using Compustat}]\) leading to incorrect inferences
- **#2 – Survivorship bias**; entry and exit of obs. isn’t random and this can affect inference
- **#3 – External validity**; our data often only covers certain types of firms and need to keep this in mind when making inferences
Quick Review [Part 2]

- What is $AdjY$ estimator, and why is it inconsistent with unobserved heterogeneity?
  - **Answer** = $AdjY$ demeans $y$ with respect to group; it is inconsistent because it fails to account for how group mean of $X$’s affect adjusted-$Y$
    - E.g. “industry-adjust”
    - Diversification discount lit. has similar problem
    - Asset pricing has examples of this [What?]
Quick Review \([Part 3]\)

- Comparing characteristically-adjusted stock returns across portfolios sorted on some other \(X\) is example of \(AdjY\) in AP
  
  - What is proper way to control for unobserved characteristic-linked risk factors?
  
  - \textbf{Answer} = Add benchmark portfolio-period FE

\([See\ Gormley\ &\ Matsa\ (2014)]\)
Quick Review [Part 4]

- What is \( \text{AvgE} \) estimator; why is it biased?
  - \textbf{Answer} = Uses group mean of \( y \) as control for unobserved group-level heterogeneity; biased because of measurement error problem
Quick Review \textit{[Part 5]}

- What are two ways to estimate model with two, high-dimensional FE [e.g. firm and industry-year FE]?
  - \textbf{Answer #1}: Create interacted FE and sweep it away with usual within transformation
  - \textbf{Answer #2}: Use iterations to solve FE estimates
Matching – Outline

- Introduction to matching
  - Comparison to OLS regression
  - Key limitations and uses
- How to do matching
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
Matching Methods – Basic Idea [Part 1]

- Matching approach to estimate treatment effect is very intuitive and simple
  - For each treated observation, you find a “matching” untreated observation that serves as the de facto counterfactual
  - Then, compare outcome, $y$, of treated observations to outcome of matched obs.
Matching Methods – *Basic Idea* [Part 2]

- A bit more formally…
  - For each value of $X$, where there is both a treated and untreated observation…
    - Match treated observations with $X=X'$ to untreated observations with same $X=X'$
    - Take difference in their outcomes, $y$
  - Then, use average difference across all the $X$’s as estimate of treatment effect
Matching Methods – Intuition

What two things is matching approach basically assuming about the treatment?

- **Answer #1** = Treatment isn’t random; if it were, would **not** need to match on $X$ before taking average difference in outcomes

- **Answer #2** = Treatment is random *conditional* on $X$; i.e. controlling for $X$, untreated outcome captures the unobserved treated counterfactual
Matching is a “Control Strategy”

- Can think of matching as just a way to control for necessary $X$’s to ensure CMI strategy necessary for causality holds

What is another control strategy we could use to estimate treatment effect?
Matching and OLS; **not** that different

- **Answer = Regression!**
  - I.e. could just regress $y$ onto indicator for treatment with necessary controls for $X$ to ensure CMI assumption holds
  - E.g. to mirror matching estimator, you could just put in indicators for each value of $X$ as the set of controls in the regression

So, how are matching & regression different?
Matching *versus* Regression

- Basically, can think of OLS estimate as particular weighted matching estimator

  - Demonstrating this difference in weighting can be a bit technical…

  - See Angrist-Pischke Section 3.3.1 for more details on this issue, but following example will help illustrate this…
Matching vs Regression – Example [P1]

Example of difference in weighting...

- First, do simple matching estimate
- Then, do OLS where regress $y$ on treatment indicator and you control for $X$’s by adding indicators for each value of $X$

- This is very nonparametric and general way to control for covariates $X$
- If think about it, this is very similar to matching; OLS will be comparing outcomes for treated and untreated with same $X$’s
But, even in this example, you’ll get different estimates from OLS and matching.

- Matching gives more weight to obs. with $X=X'$ when there are more treated with that $X'$.
- OLS gives more weight to obs. with $X=X'$ when there is more variation in treatment [i.e. we observe a more equal ratio of treated & untreated]
Matching vs Regression – Bottom Line

- Angrist-Pischke argue that, in general, differences between matching and OLS are not of much empirical importance

- Moreover, similar to OLS, matching has a serious limitation…
Matching – *Key Limitation [Part 1]*

- What sets matching estimator apart from other estimators like IV, natural experiments, and regression discontinuity?

  - **Answer** = It does not rely on any clear source of exogenous variation!

    - I.e. If OLS estimate of treatment effect is biased, so is a matching estimator of treatment effect!
Matching – *Key Limitation [Part 2]*

- And, we abandoned OLS for a reason...
  - If original treatment isn’t random (i.e. exogenous), it is often difficult to believe that controlling for some X’s will somehow restore randomness
    - E.g. there could be problematic, *unobserved* heterogeneity
    - **Note:** regression discontinuity design is exception
  - Matching estimator suffers same problem!
Matching – *Key Limitation* [Part 3]

- Please remember this!
- Matching does **NOT** and **cannot** be used…
  - To fix simultaneity bias problem
  - To eliminate measurement error bias…
  - To fix omitted variable bias from unobservable variables [*can’t match on what you can’t observe!*]
Matching – So, *what good is it?* [Part 1]

- Prior slides would seem to suggest matching isn’t that useful…
  - Basically just another control strategy that is less dependent on functional form of $X$
  - Doesn’t resolve identification concerns

- But, there are some uses…
Matching – So, what good is it? [Part 2]

- Can be used…
  - To do robustness check on OLS estimate
  - To better screen the data used in OLS
- Can sometimes have better finite-sample properties than OLS

More about these later…
Matching – Outline

- Introduction to matching
- How to do matching
  - Notation & assumptions
  - Matching on covariates
  - Matching on propensity score
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
First some notation...

- Suppose want to know effect of treatment, $d$, where $d = 1$ if treated, $d = 0$ if not treated.

- Outcome $y$ is given by…
  - $y(1) = \text{outcome if } d = 1$
  - $y(0) = \text{outcome if } d = 0$

- Observable covariates are $X = (x_1, \ldots, x_k)$
Identification Assumptions

- Matching requires two assumptions in order to estimate treatment effect
  - “Unconfoundedness”
  - “Overlap”
Outcomes $y(0)$ and $y(1)$ are statistically independent of treatment, $d$, conditional on the observable covariates, $X$

I.e. you can think of assignment to treatment as random once you control for $X$
This assumption is **stronger** version of typical CMI assumption that we make. It is equivalent to saying treatment, \( d \), is independent of error \( u \), in following regression

\[
y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \gamma d + u
\]

**Note:** This stronger assumption needed in certain matching estimators, like propensity score.
Assumption #2 – Overlap

- For each value of covariates, there is a positive probability of being in the treatment group and in the control group
  - I.e. There will be both treatment and control observations available when match on $X$
  - Why do we need this assumption?
    - Answer = It would be problematic to do a matching estimator if we didn’t have both treated and untreated observations with the same $X$!
“Overlap” in practice

- In reality, we don’t have “overlap”
  - E.g. think about continuous variables; observations won’t have exact same $X$
  - As we’ll see shortly, we end instead use observations with “similar” $X$ in matching

  - This actually causes matching estimator to be biased and inconsistent; but there are ways to correct for this [see Abadie and Imbens (2008)]
Average Treatment Effect (ATE)

- With both assumptions, easy to show that ATE for subsample with $X = X'$ is equal to difference in outcome between treated and control observations with $X = X'$
  - See Roberts and Whited page 68 for proof
  - To get ATE for population, just integrate over distribution $X$ (i.e. take average ATE over all the $X$'s weighting based on probability of $X$)
Difficulty with **exact** matching

- In practice, difficult to use exact matches when matching on # of $X$’s (i.e. $k$) is large
  - May not have both treated and control for each possible combination of $X$’s
  - This is surely true when any $x$ is continuous (i.e. it doesn’t just take on discrete values)
Matching – *Outline*

- Introduction to matching
- How to do matching
  - Notation & assumptions
  - Matching on covariates
  - Matching on propensity score
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
Matching on Covariates – Step #1

- Select a distance metric, \( ||X_i - X_j|| \)
  - It tells us how far apart the vector of \( X \)'s for observation \( i \) are from \( X \)'s for observation \( j \)
  - One example would be Euclidean distance

\[
||X_i - X_j|| = \sqrt{(X_i - X_j) \cdot (X_i - X_j)}
\]
Matching on Covariates – Step #2

- For each observation, \( i \), find \( M \) closest matches (based on chosen distance metric) among observations where \( d \neq d_i \)
  - I.e. for a treated observation (i.e. \( d = 1 \)) find the \( M \) closest matches among untreated observations
  - For an untreated observation (i.e. \( d = 0 \)), find the \( M \) closest matches among treated observations
Before Step #3... some notation

- Define $l_m(i)$ as $m^{th}$ closest match to observation $i$ among obs. where $d \neq d_i$
  - E.g. suppose obs. $i = 4$ is treated [i.e. $d = 1$]
    - $l_1(4)$ would represent the closest untreated observation to observation $i = 4$
    - $l_2(4)$ would be the second closest, and so on

- Define $L_M(i) = \{l_m(i), \ldots, l_M(i)\}$

Just way of labeling $M$ closest obs. to obs. $i$
Create imputed untreated outcome, $\hat{y}_i(0)$, and treated outcome, $\hat{y}_i(1)$, for each obs. $i$

\[
\hat{y}_i(0) = \begin{cases} 
y_i & \text{if } d_i = 0 \\
\frac{1}{M} \sum_{j \in L_M(i)} y_j & \text{if } d_i = 1
\end{cases}
\]

\[
\hat{y}_i(1) = \begin{cases} 
y_i & \text{if } d_i = 0 \\
\frac{1}{M} \sum_{j \in L_M(i)} y_j & \text{if } d_i = 1
\end{cases}
\]

In words, what is this doing?
Interpretation…

If obs. $i$ was treated, we observe the actual outcome, $y(1)$.

But, we don’t observe the counterfactual, $y(0)$; so, we estimate it using average outcome of $M$ closest untreated observations!

$$
\hat{y}_i(0) = \begin{cases} 
y_i & \text{if } d_i = 0 \\
\frac{1}{M} \sum_{j \in L_M(i)} y_j & \text{if } d_i = 1
\end{cases}
$$

$$
\hat{y}_i(1) = \begin{cases} 
\frac{1}{M} \sum_{j \in L_M(i)} y_j & \text{if } d_i = 0 \\
y_i & \text{if } d_i = 1
\end{cases}
$$
Interpretation…

\[
\hat{y}_i(0) = \begin{cases} 
  y_i & \text{if } d_i = 0 \\
  \frac{1}{M} \sum_{j \in M(i)} y_j & \text{if } d_i = 1
\end{cases}
\]

\[
\hat{y}_i(1) = \begin{cases} 
  \frac{1}{M} \sum_{j \in M(i)} y_j & \text{if } d_i = 0 \\
  y_i & \text{if } d_i = 1
\end{cases}
\]

And vice versa, if obs. \( i \) had been untreated; we impute unobserved counterfactual using average outcome of \( M \) closest treated obs.
Matching on Covariates – Step #4

- With assumptions #1 and #2, average treatment effect (ATE) is given by:

\[
\frac{1}{N} \sum_{i=1}^{N} [\hat{y}_i(1) - \hat{y}_i(0)]
\]

In words, what is this doing?

**Answer** = Taking simple average of difference between observed outcome and constructed counterfactual for each observation
Matching – Outline

- Introduction to matching
- How to do matching
  - Notation & assumptions
  - Matching on covariates
  - Matching on propensity score
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
Another way to do matching is to first estimate a propensity score using covariates, $X$, and then match on it…
Propensity Score, $ps(x)$ [Part 1]

- Propensity score, $ps(x)$, is probability of treatment given $X$ [i.e. $Pr(d = 1 | X)$, which is equal to CEF $E[d|X]$]
  - Intuitive measure…
    - Basically collapses your k-dimensional vector $X$ into a 1-dimensional measure of the probability of treatment i.e. given the $X$’s
    - Can estimate this in many ways including discrete choice models like Probit and Logit
With unconfoundedness assumption, conditioning on $ps(X)$ is sufficient to identify average treatment effect; i.e.

- I.e. controlling for probability of treatment (as predicted by $X$) is sufficient

- Can do matching using just $ps(X)$
- Or, can regress $y$ on treatment indicator, $d$, and add propensity score as control
Matching on $ps(X)$ – Step #1

- Estimate propensity score, $ps(X)$, for each observation $i$
  
  - For example, estimate $d = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u_i$ using OLS, Probit, or Logit
  
  - Common practice is to use Logit with few polynomial terms for any continuous covariates
  
  - Predicted value for observation $i$ is its propensity score, $ps(X_i)$
Note: You only need to include $X$’s that predict treatment, $d$

- This may be less than full set of $X$’s
- In fact, being able to exclude some $X$’s (because economic logic suggests they shouldn’t predict $d$) can improve finite sample properties of the matching estimate
Matching on $ps(X) – Remaining Steps…$

- Now, use same steps as before, but choose $M$ closest matches using observations with closest propensity score
  - E.g. if obs. $i$ is untreated, choose $M$ treated observations with closest propensity scores
Propensity score – *Advantage # 1*

- Propensity score helps avoid concerns about subjective choices we make with matching

  - As we’ll see next, there are a lot of subjective choices you need to make [e.g. *distance metric*, *matching method*, etc.] when matching on covariates
Propensity score – *Advantage # 2*

- Can skip matching entirely, and estimate ATE using sample analog of

\[
E \left[ \frac{(d_i - ps(X_i)) y_i}{ps(X_i)(1 - ps(X_i))} \right]
\]

- See Angrist-Pischke, Section 3.3.2 for more details about why this works
Can get lower standard errors by instead matching on covariates if add more variables that explain \( y \), but don’t necessarily explain \( d \)

- Same as with OLS; more covariates can increase precision even if not needed for identification
- **But**, Angrist and Hahn (2004) show that using \( ps(X) \) and ignoring these covariates can actually result in better finite sample properties
Matching – Outline

- Introduction to matching
- How to do matching
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
Practical Considerations

- There are a lot of practical considerations and choices to make with matching; e.g.,
  - Which distance metric to use?
  - How many matches for each observation?
  - Match with or without replacement?
  - Which covariates $X$ should be used?
  - Use propensity score, and if so, how measure it?
Choice of distance metric [*Part 1*]

- What is downside to simple Euclidean distance metric from earlier?

\[ \| X_i - X_j \| = \sqrt{(X_i - X_j)'(X_i - X_j)} \]

- **Answer** = It ignores the potentially different scales of each variable [*which is why it typically isn’t used in practice*]

- Which variables will have more effect in determining best matches with this metric?
Choice of distance metric [*Part 2*]

- Two other possible distance metrics standardize distances using inverse of covariates’ variances and covariances
  - Abadie and Imbens (2006)
    \[
    \|X_i - X_j\| = \sqrt{(X_i - X_j)' \text{diag}(\Sigma^{-1}_X)(X_i - X_j)}
    \]
  - Mahalanobis [*probably most popular*]
    \[
    \|X_i - X_j\| = \sqrt{(X_i - X_j)' (\Sigma^{-1}_X)(X_i - X_j)}
    \]
Should you match based on covariates, or instead match using a propensity score?

- And, if use propensity score, should you use Probit, Logit, OLS, or nonparametric approach?

Unfortunately, no clear answer

- Want whichever is going to be most accurate…
- But, probably should show robustness to several different approaches
And, how many matches? \textit{[Part 1]}

- Again, no clear answer…
- Tradeoff is between bias and precision
  - Using single best match will be least biased estimate of counterfactual, \textit{but} least precise
  - Using more matches increases precision, \textit{but} worsens quality of match and potential bias
Two ways used to choose matches

- “Nearest neighbor matching”
  - This is what we saw earlier; you choose the $m$ matches that are closest using your distance metric

- “Caliper matching”
  - Choose all matches that fall within some radius
  - E.g. if using propensity score, could choose all matches within 1% of observation’s propensity score

**Question:** What is intuitive advantage of caliper approach?
And, how many matches? [Part 3]

- **Bottom line advice**
  - Best to try multiple approaches to ensure robustness of the findings
    - If adding more matches (or expanding radius in caliper approach) changes estimates, then bias is potential issue and should probably stick to smaller number of potential matches
    - If not, and only precision increases, then okay to use a larger set of matches
Matching with replacement

- Each observation can serve as a match for multiple observations
- Produces better matches, reducing potential bias, but at loss of precision

Matching without replacement
With or without replacement? [Part 2]

- Bottom line advice...
  - Roberts-Whited recommend to do matching with replacement...
    - Our goal should be to reduce bias
    - In matching without replacement, the order in which you match can affect estimates
Which covariates?

- Need all X’s that affect outcome, $y$, and are correlated with treatment, $d$ [Why?]
  - Otherwise, you’ll have omitted variables!

- But, do not include any covariates that might be affected by treatment
  - Again, same “bad control” problem

**Question:** What might be way to control for X that could be a “bad control”?  

**Answer:** Use lagged X
Matches for whom?

- If use matches for all observations (as done earlier), you estimate ATE
  - But, if only use and find matches for treated observations, you estimate average treatment effect on treated (ATT)
  - If only use and find matches for untreated, you estimate average treatment effect on untreated (ATU)
Matching – Outline

- Introduction to matching
- How to do matching
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
Testing “Overlap” Assumption

- If only one $X$ or using $ps(X)$, can just plot distribution for treated & untreated
- If using multiple $X$, identify and inspect worst matches for each $x$ in $X$
  - If difference between match and observation is large relative to standard deviation of $x$, might have problem
If there is lack of “Overlap”

- Approach is very subjective…
  - Could try discarding observations with bad matches to ensure robustness
  - Could try switching to caliper matching with propensity score
How might you try to test unconfoundedness assumption?

- Answer = Trick question; you can’t! We do not observe error, $u$, and therefore can’t know if treatment, $d$, is independent of it!

- Again, we cannot test whether the equations we estimate are causal!
But, there are other things to try...

- Similar to natural experiment, can do various robustness checks; e.g.
  - Test to make sure timing of observed treatment effect is correct
  - Test to make sure treatment doesn’t affect other outcomes that should, theoretically, be unaffected
- Or, look at subsamples where treatment effect should either be larger or smaller
Matching – Outline

- Introduction to matching
- How to do matching
- Practical considerations
- Testing the assumptions
- Key weaknesses and uses of matching
Weaknesses Reiterated [Part 1]

- As we’ve just seen, there isn’t clear guidance on how to do matching
  - Choices on distance metric, matching approach, # of matches, etc. are subjective
  - Or, what is best way to estimate propensity score? Logit, probit, nonparametric?

- Different researchers, using different methods might get different answers!
Weaknesses Reiterated \([Part 2]\)

- And, as noted earlier, matching is not a way to deal with identification problem
  - Does **NOT** help with simultaneity, unobserved omitted variables, or measurement error
  - Original OLS estimate of regressing \(y\) on treatment, \(d\), and \(X\)'s is similar but weighting observations in particular way
Often see a researcher estimate:

\[ y = \beta_0 + \beta_1 d + ps(X) + u \]

- \( d \) = indicator for some non-random event
- \( ps(X) \) = prop. score for likelihood of treatment estimated using some fancy, complicated Logit

Then, researcher will claim:

“Because \( ps(X) \) controls for any selection bias, I estimate causal effect of treatment”
Tangent – Related Problem \textit{[Part 2]}

- Researcher assumes that observable $X$ captures \textbf{ALL} relevant omitted variables
  - I.e. there aren’t any \underline{unobserved} variables that affect $y$ and are correlated with $d$
  - This is often not true… Remember long list of unobserved omitted factors discussed in lecture on panel data

- \textbf{Just because it seems fancy or complicated doesn’t mean it’s identified!}
Another Weakness – Inference

- There isn’t always consensus or formal method for calculating SE and doing inference based on estimates

- So, what good is it, and when should we bother using it?
Use as a robustness check

- Can use as robustness check to OLS estimation of treatment effect
  - It avoids functional form assumptions imposed by the regression; so, provides a nice sanity check on OLS estimates
  - Angrist-Pischke argue, however, that it won’t find much difference in practice if have right covariates, particularly if researcher uses regression with flexible controls for $X$
Can use matching to screen sample used in later regression

**Ex. #1** – Could estimate propensity score; then do estimation using only sample where the score lies between 10% and 90%

- Helps ensure estimation is done only using obs. with sufficient # of controls and treated
- Think of it as ensuring sufficient overlap
Use as precursor to regression \textit{[Part 2]}

- \textbf{Ex. #2} – Could estimate effect of treatment using only control observations that match characteristics of treated obs.
  
  \begin{itemize}
  \item E.g. If industry $X$ is hit by shock, select control sample to firms matched to similar industry
  \end{itemize}
Matching – *Practical Advice*

- User-written program, “*psmatch2,*” in Stata can be used to do matching and obtain estimates of standard errors
  - Program is flexible and can do variety of different matching techniques
Summary of Today [Part 1]

- “Matching” is another control method
  - Use to estimate treatment effect in cases where treatment is random after controlling for $X$
  - Comparable to OLS estimation of treatment effect, just without functional form assumptions
- Besides controlling for $X$, matching does **NOT** resolve or fix identification problems
Many different ways to do matching; e.g.

- Match on covariates or propensity scores
- Nearest neighbor or caliper matching

Primarily used as robustness test

- If have right covariates, $X$, and relatively flexible OLS model, matching estimate of ATE will typically be quite similar to OLS
In First Half of Next Class

- Standard errors & clustering
  - Should you use “robust” or “classic” SE?
  - “Clustering” and when to use it

- Limited dependent variables… are Probit, Logit, or Tobit needed?

- Related readings… see syllabus
Assign papers for next week…

- Morse (JFE 2011)
  - Payday lenders

- Colak and Whited (RFS 2007)
  - Spin-offs, divestitures, and investment

  - Credit ratings & sovereign credit ceiling
Break Time

- Let’s take our 10 minute break
- We’ll quickly cover Heckman selection models and then do presentations when we get back
Heckman selection models

- Motivation
- How to implement
- Limitations [i.e., why I don’t like them]
Motivation [Part 1]

- You want to estimate something like...

\[ Y_i = \beta X_i + \varepsilon_i \]

- \( Y_i \) = post-IPO outcome for firm \( i \)
- \( X_i \) = vector of covariates that explain \( Y \)
- \( \varepsilon_{i,t} \) = error term
- **Sample** = all firms that did IPO in that year

- What is a potential concern?
**Motivation [Part 2]**

- **Answer** = certain firms ‘self-select’ to do an IPO, and the factors that drive that choice might cause $X$ to be correlated with $\varepsilon_{i,t}$
  - It’s basically an omitted variable problem!
  - If willing to make some assumptions, can use Heckman two-step selection model to control for this selection bias
How to implement [Part 1]

- Assume to choice to ‘self-select’ [in this case, do an IPO] has following form...

\[
IPO_i = \begin{cases} 
1 & \text{if } \gamma Z_i + \eta_i > 0 \\
0 & \text{if } \gamma Z_i + \eta_i \leq 0 
\end{cases}
\]

- \(Z_i\) = factors that drive choice [i.e., IPO]
- \(\eta_{i,t}\) = error term for this choice
How to implement [Part 2]

- Regress choice variable (i.e., IPO) onto Z using a Probit model.
- Then, use predicted values to calculate the Inverse Mills Ratio for each observation, \( \lambda_i = \frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)} \).
- Then, estimate original regression of \( Y_i \) onto \( X_i \), but add \( \lambda_i \) as a control!

Basically, controls directly for omitted variable; e.g. choice to do IPO.
Limitations [Part 1]

- Model for choice [i.e., first step of the estimation] must be correct; otherwise inconsistent!
- Requires assumption that the errors, $\varepsilon$ and $\eta$, have a **bivariate normal distribution**
  - Can’t test, and no reason to believe this is true
    [i.e., what is the economic story behind this?]
  - And, if wrong… estimates are inconsistent!
Limitations \textit{[Part 2]}

- Can technically work if $Z$ is just a subset of the $X$ variables \textit{[which is commonly what people seem to do]}, but...

  - But, in this case, all identification relies on non-linearity of the inverse mills ratio \textit{[otherwise, it would be collinear with the $X$ in the second step]}

  - \textbf{But again, this is entirely dependent on the bivariate normality assumption and lacks any economic intuition!}
Limitations [Part 3]

- When $Z$ has variables not in $X$ [i.e., excluded instruments], then could just do IV instead!
  - I.e., estimate $Y_i = \mathbf{bX}_i + IPO_i + \varepsilon_i$ on full sample using excluded IVs as instruments for IPO
  - Avoids unintuitive, untestable assumption of bivariate normal error distribution!