Questions to answer in preparing for January 11 class

# The Dynarski paper

* 1. Research Questions and Answers
     1. What are the author’s research questions? (Hint: think about the underlying big research questions.)
     2. What are the author’s answers to her research questions? Be prepared to describe magnitudes of effects.
  2. Data and Sample
     1. What data set does Dynarski use? What are its properties? What properties of the data set constrain the way Dynarski could conduct her analyses?
     2. What were the eligibility requirements?
     3. Which eligible children were excluded from the analysis? Why?
     4. There are some members of the data set that were high school seniors prior to 1979 and were eligible for the college-going benefit. Why do you think Dynarski chose not to include these individuals in the analysis?
     5. How did Dynarski deal with missing data on family income and AFQT score? What are some benefits of this approach? Drawbacks?
  3. Methodology
     1. What is the identification strategy used to make causal inferences?
     2. What is the key assumption this identification strategy makes to support the claim that it identifies causal effects? Try to use the term *secular trend* in your answer.
     3. Examine Eq. 2. What effect does the coefficient delta (*δ*) represent? What effect does the coefficient theta (*θ*) represent?
     4. Why did Dynarski include interaction terms between the covariates and the dummy variable *BEFORE* and between the covariates and the dummy variable *DECEASED*? What effect does the coefficient beta (*β*) represent?
     5. Which coefficient is “the causal parameter of interest”?
     6. Can you substitute 1s and 0s for the values that the variables in the regression equation take (Eq. 2) to fill in this table? This should give you the same answer as (v) and show you how to recapture the difference-in-difference subtraction from the regression equation. We’ve completed the first row:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Father Deceased | Not Deceased | Difference |
| 1979-1981 | α + β + δ + θ | α + θ | β + δ |
| 1982-1983 |  |  |  |
| Difference |  |  |  |

* + 1. Dynarski writes, “Standard errors in the tables are corrected for within-household correlation in error terms due to the presence of siblings in the data, as well as for heteroskedasticity due to the dichotomous dependent variable” (p. 282). What does this mean?
    2. Although she doesn’t explicitly state it, Column 2 of Table 2 and all of Table 3 represent multi-level models. Can you tell what the unit of clustering is? What type of multi-level model has she fit?
  1. Results
     1. What function does the last column in Table 1 (Column 5, labeled Difference-in-differences) serve? One of the outcome coefficients appears in another table in the paper. Which one and why only this one?
     2. In what way do the similarities and/or differences between Column 1 and 2 of Table 2 support the identifying assumptions of the paper? Try to use the terms from Week 1: *residual variance*, *treatment variable* and *bias/unbiased*.
     3. Despite the inclusion of multiple covariates in Column 2 of Table 2, the standard error on *FATHERDECxBEFORE* has increased. Why?
  2. Threats to Validity
     1. What are the threats to the internal validity of the research described in the Dynarski paper?
     2. How does Dynarski deal with these threats?
     3. Dynarski’s paper makes an untested assumption about the college enrollment and attainment trends of 18-22 year-olds with a deceased father and 18-22 year-olds without a deceased father in the years 1979-1981. What is that assumption?
     4. Explain the two types of misclassification (p. 285). How would each type of misclassification affect the estimate of the parameter of interest?
  3. External Validity
     1. What characteristics of the Social Security Program make it different from traditional need-based financial aid?
     2. How did the beneficiaries of the Social Security Program differ from the general population of 18-22 year-olds at the time period? From the college-going population?
     3. Do the differences in the characteristics of SS Program beneficiaries and other groups of college-aged students you’ve documented in (i) pose the same risks to drawing externally valid conclusions about higher education policy as the differences you’ve documented in (ii)?
     4. Under what circumstances would the differences between the SS program and need-based financial aid make it inappropriate to use Dynarski’s estimates as the basis for inferring how changes in the availability of need-based financial aid would affect college enrollment rates and educational attainments? (p. 286).

# Chapter 8 of Methods Matter

* 1. Natural experiments:
     1. Explain the term
     2. Describe common categories of natural experiments
     3. Provide an example of a natural experiment from each category that is different from those discussed in the chapter.
  2. The Angrist draft lottery paper
     1. What question(s) did Angrist ask in his draft lottery paper?
     2. What identification strategy did Angrist use to identify the effects of draft eligibility? Why was this a particularly effective strategy (self check: can you use the terms: *exogenous, equal in expectation, counterfactuals* and *omitted variable bias* in your response?)
     3. What were Angrist’s answers to his research question(s)?
     4. What do you see as threats to the internal validity of the Angrist draft lottery study?
  3. Causal analyses based on natural experiments with a discontinuity design
     1. What do you trade off in deciding on the window-size to use in estimating causal effects from natural experiments? (Use the Dynarski paper to illustrate your answer.)
     2. What are common threats to validity of analyses based on natural experiments with a discontinuity design?
     3. In the specific context of difference-in-difference natural experiments, what is the central threat to validity?
     4. The standard difference-in-difference approach (either through subtraction or the regression model presented in *MM* and Dynarski) operates through the comparison of “before” and “after” means (p. 162). What “mathematical” risk do Murnane and Willett identify about this assumption? How might you address this (and what might be some risks to that)?