ECON 190 – Seminar in Economics
Drake University, Spring 2024
William M. Boal

Blackboard: http://drake.blackboard.com
Email: william.boal@drake.edu
Useful links: wmboal.com/seminar

BOAL'S ECON 190

COURSE PACKET

SPRING 2024

A. Reading guides
B. Writing assignments
C. Slideshow handouts
D. Stata labs

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Blackboard: drake.blackboard.com
Helpful links: wmboal.com/seminar
Email: william.boal@drake.edu

TENTATIVE COURSE SYLLABUS

1. Resources | 2. Requirements | 3. Schedule

1. Resources

Description from Course Catalog: Capstone course for the economics major, with three components. Students will (1) access, read, discuss and critique recent academic economic research on a subject of current interest; (2) take a comprehensive exam in economics; and (3) write an economics research paper on a topic of their choice in consultation with the instructor. May be repeated for credit with instructor permission.

Prerequisites: ECON 170 or STAT 170.

CBPA Promises: "Our graduates will have the skills and experiences to thrive in a complex, diverse, and evolving world. They will be (1) Proficient in their fields, (2) Data-driven, strategic, and innovative problem solvers, (3) Effective communicators, (4) Socially and ethically responsible leaders, and (5) Global and multicultural citizens." This course addresses all five Promises, but especially Promises (1), (2) and (3).

Who should take this course: ECON 190 is required for the Economics major and the Quantitative Economics major. ECON 190 can count as a 100-level elective course for the Economics minor.

Class meetings: CRN 8665 meets Mondays and Wednesdays, 8:00 to 9:30 AM, in room 102 Aliber Hall.

How to contact instructor:

- Electronic mail: william.boal@drake.edu
- Office: 319 Aliber Hall
- Telephone and voice mail: 271-3129

The quickest way to reach me is by email, which I check continually throughout the day. Please do *not* send messages by Blackboard, which I check infrequently.

Office hours: My office hours this semester are **TBA**. If these hours are inconvenient due to schedule conflicts, please send email to schedule a special appointment and suggest some alternate times.

Resources to purchase:

- Required: Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: the path from cause to effect. Princeton*, New Jersey: Princeton University Press. ISBN 978-0-691-15284-4 (paperback).
- Required: *Boal's Econ 190 Course Packet*. Available for purchase at **TBA**. Please bring it to class every day.
- Required: Stata software. Stata is currently the most popular software for econometric research. Stata comes in different varieties—see http://www.stata.com/order/new/edu/gradplans/student-pricing. You need Stata/BE, which costs $48 for a six-month student license.
- Recommended: A three-ring binder and highlighter for your course packet.

Online resources:

- Drake email. Course announcements will occasionally be sent to this account, so check it daily. Announcements often get diverted to "Junk" or "Clutter" folders, so check them as well as your inbox.
- Quizzes and writing assignments are posted on Blackboard. If you have difficulty accessing Blackboard, please contact the Drake ITS HelpDesk at 271-3001.
- Miscellaneous helpful links are posted at wmboal.com/seminar .

2. Requirements

Course grade: Each assignment is graded on a scale from zero to 100. Your overall course SCORE is calculated as a weighted average, using the following weights.

10% **Quizzes on readings**. Each day that a reading is due, students will complete a short quiz at the beginning of class.

20% **Midterm essay**. Students will write a short essay summarizing the readings on inequality. The essay must be written in academic style. See separate assignment sheet for details.

5% **Stata labs**. A series of Stata labs will be done mostly in class but sometimes finished outside of class. Students must submit their Stata log files on Blackboard.

50% **Research paper**. Each student will write an empirical research paper. The paper should estimate a relationship or test a hypothesis using econometric tools we discuss in class to analyze data that you collect. The paper need not be on "inequality"—it may be on any economic topic. The length should be 5 to 8 pages, plus references, tables, and graphs. See separate assignment sheet for details. A required template is posted on Blackboard.

15% **Comprehensive exam**. As this is a capstone course, an exam measuring proficiency in all areas of economics will be given during Finals Week. The content will be about 45% microeconomics, 40% macroeconomics, and 15% econometrics.

An overall course SCORE of 97 or above is required for an A+, 93 for an A, 90 for an A-, 87 for a B+, 83 for a B, 80 for a B-, 77 for a C+, 73 for a C, 70 for a C-, 67 for a D+, 63 for a D, and 60 for a D-. SCORES will not be rounded before awarding letter grades. Extra credit work is not available.

Academic writing style: This is a writing-intensive course. *All* writing should be in the impersonal, objective style of academic economics. Most economics journals put citations and bibliography in a format roughly similarly to APA style, so we will too. For a convenient guide, see http://library.albany.edu/cfox .

Writing Center: Any piece of writing longer than a few sentences can become difficult for readers to follow. The tutors at Drake's Writing Center can help you make sure your writing tells a coherent story. Students are required to visit the Writing Center at least twice: once before submitting the midterm essay and once before submitting the first draft of the research paper. Appointments are made at https://library.drake.edu/writing-center/. It is OK to cancel or reschedule an appointment, but no-shows are unacceptable because they needlessly deprive other students of appointments.

Policy on absences: Attendance is taken at every class. Students may miss up to three classes for any reason without penalty. Thereafter, one point will be deducted from the course SCORE for each absence. Athletic team trips, documented by a sheet from the Director of Athletics, will not be counted as absences.

Policy on grade corrections: Accurate grading is important. If you find an error, please let me know as soon as possible. The deadline for regrading homework, problem sets, or midterm exams is the day of the final exam.

Disability accommodation: Any student who has a disability that substantially limits their ability to perform in this course under normal circumstances should contact Student Disability Services, 271-1835, to request accommodation. All relevant information will be kept strictly confidential.

Policy on academic integrity: The CBPA's Academic Integrity Policy (www.drake.edu/cbpa/about/cbpapolicies) applies to this course. The consequences of violating this policy vary, depending on my evaluation of the severity of the dishonesty. A violation (such as cheating, plagiarism, or fabrication) can result in a grade of zero on the test or assignment, an F for the course grade, or even expulsion from the University. Please read the policy and ask for clarification if necessary.

3. Schedule

All readings marked by a box (☐) must be done BEFORE class. There will be a quick quiz at the beginning of class, followed by detailed class discussion. A guide to each reading is included in the course packet.

Monday readings on wage inequality should be downloaded from Cowles Library databases (library.drake.edu/find/article-databases) unless otherwise noted. Find them using the Library's *SUPERSEARCH* engine or its EBSCO databases *Academic Source Premier* and *EconLit,* or *JSTOR*. Articles from the *Journal of Economic Perspectives* can also be downloaded free from www.aeaweb.org/journals/jep . Take notes using the reading guide in your course packet.

Wednesday readings on econometrics are from the required book by Angrist and Pischke. You will probably use one of the methods in this book for your own research paper. Again, take notes using the reading guide in your course packet.

In addition to discussing readings, we will use any available class time to discuss and practice *doing* economic research. I have prepared six lab exercises which we will do together in class. These exercises require Stata, so make sure it is installed on your computer before the first day of class and bring your computer to every class. These exercises demonstrate how to download and analyze data, skills you will use for your own research paper.

Week	Monday: Reading research	Wednesday: Doing research	Friday: Deadlines
1 **Jan 29**	**Welcome** • Review syllabus, answer questions. • Sign up for research paper brainstorming sessions. • Review basic econometrics and measures of inequality.	**Correlation versus causality** ☐ Angrist & Pischke (2015) introduction and chapter 1, pp. 1-11. In class: Stata lab on National Health Interview Survey (NHIS).	
2 **Feb 5**	**Trends: What CPS data show** ☐ What is the CPS? View description at https://www.census.gov/programs-surveys/cps/about.html . ☐ Guzman, G., & Kollar, M. A. (2023). *Income in the United States: 2022*. https://www.census.gov/library/publications/2023/demo/p60-279.html. In class: Create time-series graphs of inequality measures from Guzman and Kollar report, using Microsoft Excel.	**Randomized trials** • Angrist & Pischke (2015) chapter 1, pp. 12-33. In class: Begin Stata lab on Current Population Survey.	
3 **Feb 12**	**Trends: What tax return data show** ☐ Atkinson, A. B., Piketty, T., & Saez, E. (2011). Top incomes in the long run of economic history. *Journal of Economic Literature, 49*(1), 3-71.	**Regression** • Angrist & Pischke (2015) chapter 2, pp. 47-78. In class: Finish Stata lab on Current Population Survey.	

Week	Monday: Reading research	Wednesday: Doing research	Friday: Deadlines
4 **Feb 19**	**Explanations: Skill-biased technical change** ☐ Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Cambridge, Massachusetts: Harvard University Press. Read chapter 3. [This book is on reserve at Cowles Library in either print or e-book format. How to access: **TBA**.]	**Instrumental variables** • Angrist & Pischke (2015) chapter 3, pp. 98-138. In class: Stata lab on instrumental variables.	Submit research paper proposal, including research question and data description, on Blackboard.
5 **Feb 26**	**Explanations: Decline of unions and falling real minimum wage** ☐ Fortin, N. M., & Lemieux, T. (1997). Institutional changes and rising wage inequality: is there a linkage? *Journal of Economic Perspectives*, 11(2), 75-96.	**Regression Discontinuity Designs** • Angrist & Pischke (2015) chapter 4, pp. 147-177. In class: Stata lab on regression discontinuity.	
6 **Mar 4**	**Explanations: Information technology** ☐ Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: an empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279-1333. ☐ Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in U.S. wage inequality: revising the revisionists. *Review of Economics and Statistics, 90*(2), 300-323.	**Difference-in-differences** ☐ Angrist & Pischke (2015) chapter 5, pp. 178-202. In class: Begin Stata lab on National Longitudinal Survey of Youth (NLSY).	
7 **Mar 17**	**Explanations: Immigration** ☐ Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review, 43*(2), 245-257. ☐ Borjas, G. J., Freeman, R. B., & Katz, L. F. (1996). Searching for the effect of immigration on the labor market. *American Economic Review*, 86(2), 246-251.	**Research paper workshop I** Each student brings draft paper copies of • Table 1: variable definitions • Table 2: descriptive statistics • Key equation(s) to be estimated In class: Finish Stata lab on National Longitudinal Survey of Youth (NLSY).	
8 **Mar 25**	**Explanations: International trade** ☐ Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The China syndrome: local labor market effects of import competition in the United States. *American Economic Review, 103*(6), 2121-2168.	**Class discussion: What should a research paper look like?** In class: Begin Stata lab on Consumer Expenditure Survey.	Submit draft data section and methodology section of research paper (use the template!) on Blackboard.

Week	Monday: Reading research	Wednesday: Doing research	Friday: Deadlines
9 **Apr 1**	**Explanations: "Superstars," CEOs, and finance professionals** ☐ Rosen, S. (1981). The economics of superstars. *American Economic Review, 71*(5), 845-858. ☐ Bivens, J., & Mishel, L. (2013). The pay of corporate executives and financial professionals as evidence of rents in top 1 percent incomes. *Journal of Economic Perspectives*, 27(3), 57-78. ☐ Kaplan, S. N., & Rauh, J. D. (2013). It's the market: the broad-based rise in the return to top talent. *Journal of Economic Perspectives*, 27(3), 35-56.	**The causal effect of schooling on wages** ☐ Angrist & Pischke (2015) chapter 6, pp. 209-234. In class: Finish Stata lab on Consumer Expenditure Survey.	Schedule appointment with Writing Center to discuss **midterm essay**. Bring a draft and the assignment sheet. Work with Writing Tutor to ensure your essay tells a coherent story.
10 **Apr 8**	**Class discussion: Improving writing** Economic ideas are complicated so clear writing is important. We will review samples of economic writing and try to improve them.	**Research paper workshop II** Each student brings a draft paper copy of Table 3: estimates.	Submit midterm essay.
11 **Apr 15**	**Continue research paper workshop II**	**Finish research paper workshop II**	Schedule appointment with Writing Center to discuss 1st draft of **research paper**. Bring a draft and the assignment sheet. Work with Writing Tutor to ensure your paper tells a coherent story.
12 **Apr 22**	**Oral presentations of research papers** Make a PowerPoint presentation, at least one slide for each section of your paper and one slide for each table or graph. Fonts should be at least 20 pt for readability.	**Continue oral presentations of research papers**.	Submit 1st draft of research paper (use the template!) and Stata log file on Blackboard.
13 **Apr 29**	**Continue oral presentations of research papers**.	**Finish oral presentations of research papers**	
14 **May 6**	**Review for comprehensive exam:** • microeconomics	**Review for comprehensive exam:** • macroeconomics • econometrics.	Submit final draft of research paper and Stata log file on Blackboard.
Finals Week		**Comprehensive Exam** Day and time **TBA**, in the regular classroom.	Have a good summer!

[end of syllabus]

SECTION A
READING GUIDES

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist & Pischke (2015) introduction and chapter 1, pp. 1-11.

Q1: Why is a simple comparison of graduation rates of students with and without student loans NOT an "other things equal" comparison? (p. xii)

Q2: What does the Latin phrase "ceteris paribus" mean?

Q3: According to Angrist and Pischke, what must hold for a correlation between two variables to yield causal understanding? (p. xiii)

Q4: Define the following terms: outcome, treatment, treatment group, control group, counterfactual (complete sentences are not necessary). (p. 3).

Q5: The authors use different notation from most econometrics textbooks. What does "Y_{1i}" mean? What does "Y_{0i}" mean? Are both values ever observed? (p. 6).

Q6: Suppose individual i receives treatment of some kind (such as being given health insurance) but individual j does not. Which of the following are observed and which are counterfactual: Y_{1i} , Y_{0i} , Y_{1j} , Y_{0j} ? (pp. 6, 7)

Q7: It is tempting (but maybe fruitless) to measure the causal effect of the treatment as $(Y_{1i} - Y_{0j})$. This expression is the difference between the following two expressions: $(Y_{1i} - Y_{0i})$ and $(Y_{0i} - Y_{0j})$. Which expression represents the true causal effect of the treatment? Which expression represents selection bias? (p. 8)

Q8: In the National Health Interview Survey, according to the first line of table 1.1 (p. 5), is $(Y_{1i} - Y_{0j})$ typically positive or negative?

Q9: Suppose better educated people are typically healthier regardless of insurance status, but they are also more likely to have insurance. Then is (Y0i- Y0j) typically positive or negative?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

U.S. Census Bureau. (2022). Current Population Survey https://www.census.gov/programs-surveys/cps/about.html

Q1: What does "CPS" stand for?

Q2: To answer the following questions, scroll down and click on "Subject definitions." In the CPS, what is the difference between a "family" and a "household"?

Q3: What types of labor income does the CPS ask about? (See the definition of "earnings.")

Q4: What kinds of capital income does the CPS ask about? (See "income measurement.")

Q5: What kinds of government transfers (benefits) does the CPS ask about? (See "income measurement.")

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Guzman, G., & Kollar, M. A. (2023). *Income in the United States: 2022*. G. P. Office.
https://www.census.gov/content/dam/Census/library/publications/2023/demo/p60-279.pdf

Q1: What does "CPS ASEC" stand for? (p. 1)

Q2: How many households were surveyed in the 2023 ASEC? When was the ASEC data collected? (p. 54)

Q3: What was the response rate to the ASEC in 2023?

Q4: Does the ASEC collect before-tax income or after-tax income? (p. 13)

Q5: Does the ASEC collect information on the Earned Income Tax Credit, SNAP ("food stamps"), health benefits, and subsidized housing? (p. 13).

Q6: Does the ASEC collect information on capital gains income? (p. 13)

Q7: What kinds of income are most likely to be underreported in the ASEC? (p. 13)

Q8: What is "equivalence-adjusted income"? (p. 7)

Q9: Look at tables A-4a and A-4b, on pages 32-35. Compute the percent increase or decrease, to the nearest tenth of a percentage point, since 1980 in the following:

	1980 value	1922 value	% increase (or decrease)
(a) the 90th/10th income ratio			
(b) the share of the lowest quintile			
(c) the share of the highest quintile			
(d) the share of the highest 5 percent			
(e) the Gini index			

Overall, has inequality of household income increased or decreased since 1980?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist & Pischke (2015) chapter 1, pp. 12-17. "Randomized Trials."

Q1: The Law of Large Numbers says that the average in a sample converges to what, as the sample size increases? (p. 13)

Q2: It is crucial to understand the notation in this book. Translate the following expressions into English:

(a) Y_{1i}

(b) Y_{0i}

(c) $E[Y_{1i}]$

(d) $E[Y_{1i} \mid D_i = 1]$

Q3: Random assignment ensures that which of the following expressions is zero? Why is that a good thing? (p. 15)

$E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$

$E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]$

$E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$

Q4: The differences in columns (3) and (6) of table 1.1 are generally large, but the differences in columns (2) through (5) of table 1.3 are generally small? Why? What is the crucial difference between the two data sets? (p. 19) [Hint: answer is a 13-letter word beginning with "r".]

Q5: In the RAND Health Insurance Experiment, did more generous health plans increase or decrease health care use? Did more generous health plans improve or worsen health outcomes? (pp. 22-23).

Q7: In the Oregon Health Plan lottery, did winners increase or decrease their health care use? Did winners enjoy better HEALTH outcomes? (pp. 26-29).

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Atkinson, A. B., Piketty, T., & Saez, E. (2011). Top incomes in the long run of economic history. *Journal of Economic Literature*, 49(1), 3-71.

Q1: Define the following terms:

vingtile (p. 5)

tax unit (p. 15)

pre-fisc income (p. 34)

post-fisc income (p. 34)

marginal tax rate (p. 34)

realized capital gains (p. 35)

tax evasion (p. 36)

tax avoidance (p. 36)

capital levy (p. 66).

Q2: Which data are better for measuring inequality at the HIGH end of the distribution: tax data or household surveys like the CPS? Why? (p. 29)

Q3: Consider figure 3, p. 8, which shows the income share and composition of the top 0.1%. What were the two biggest income components for this group in 1916? What are the two biggest income components for this group today?

Q4: Consider figures 8-10 on pp. 41-43. Summarize the trend for the top 1% share of total income for each of the following countries: United States, United Kingdom, Germany, and Sweden.

Q5: How might war affect the capital income enjoyed by those at the top? (pp. 62-63)

Q6: How might war affect the inequality of labor income? (p. 63)

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist, J. D., & Pischke, J.-S. (2015). Chapter 2. "Regression."

Q1: Why is a comparison of the earnings of Harvard graduates and University of Massachusetts graduates *not* an "other things equal" comparison? (p. 48)

Q2: What is the "first and most important step" to obtain an "other things equal" comparison? (p. 50)

Q3: Failure to control for differences (other than the treatment) between groups can result in _____ bias (fill in the blank). (p. 55)

Q4: What are the three kinds of variables in a regression equation? (p. 56)

Q5: If the dependent variable is the natural log of earnings, and the estimated coefficient of the private-school dummy variable were 0.07, how should this estimate be interpreted? (p. 60)

Q6: When is a coefficient estimate considered "statistically significant," usually? (p. 62)

Q7: A regression explaining a person's earnings seems to show a statistically significant effect of college selectivity, measured as the average test score at the college they attended (table 2.4, p. 67, row 1, columns 1-3). But the effect disappears when controls are included for the average test scores at all the colleges the same person applied to (row 1, columns 4-6). Which estimates suffer from selection bias? Why?

Q8: Suppose hypothetically we regress people's log earnings on their years of schooling alone and get a coefficient of 0.15 (the "short" equation). However, suppose that parents' earnings also affect a person's earnings (perhaps because parents help their children get jobs and start careers) but we have omitted parents' earnings as a regressor from our regression. Had we included parents' earnings in the ("long") equation, it would have had a coefficient of 0.4. Meanwhile, a regression of parent's earnings on years of schooling (of their children) would have a coefficient of 0.05. Is our original coefficient estimate of 0.15 biased up or down? How large is the omitted variable bias (OVB)? What is an unbiased estimate of the effect of schooling on earnings? [Hint: use second equation on p. 72.]

Q9: "Robust" has a specific meaning in econometrics. What are "robust" estimates of treatment effects? (p. 75).

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Harvard University Press.

Q1: What is "skill-biased technical change," also called "technology-skill complementarity"? (pp. 90-93).

Q2: How long do the authors believe the U.S. has experienced "skill-biased technical change" (SBTC)? Do the authors believe that SBTC is a recent phenomenon related to computers? (pp. 91-94, 124-125)

Q3: One hypothesis for why the relative demand for unskilled workers fell is globalization: industries and plants employing low-skilled workers have gone overseas. Why do the authors reject that hypothesis as the main explanation? (p. 98).

Q4: What is an "annual log change"? (table 3.2, p. 101)

Q5: The first column of table 3.2, p. 101, shows changes in the relative wages of skilled and unskilled workers. Skilled workers are defined as college graduates, and unskilled workers as having a high school diploma or less. (Workers with some college are divided evenly between the two skill groups.) In the decades since 1950, have the wages of skilled workers risen or fallen relative to the wages of unskilled workers? Any exceptions?

Q6: Table 3.2, p. 101, uses the elasticity of substitution to infer changes in the relative demand for skilled labor from data on changes in the relative wage and relative supply, as follows. Recall that the elasticity of substitution (σ) between skilled (S) and unskilled (U) labor is defined with calculus as

$$\sigma = - d \ln(X_S/X_U) / d \ln(MP_S/MP_U) .$$

If labor markets are competitive, workers are paid their marginal products (VMP=real W), so we have

$$\sigma = - d \ln(X_S/X_U) / d \ln(W_S/W_U) .$$

Taking reciprocals,

$$(-1/\sigma) = d \ln(W_S/W_U) / d \ln(X_S/X_U) .$$

For small changes,

$$\Delta \ln(W_S/W_U) = (-1/\sigma) \Delta \ln(X_S/X_U).$$

All this so far assumes no technical change. The authors allow skill-biased technical change to change relative *demand* by writing

$$\Delta \ln(W_S/W_U) = (-1/\sigma) [\Delta \ln(X_S/X_U)_{supply} - \Delta \ln(X_S/X_U)_{demand}] .$$

This can be rearranged to give

$$\Delta \ln(X_S/X_U)_{demand} = \sigma [\Delta \ln(W_S/W_U)] + \Delta \ln(X_S/X_U)_{supply} .$$

In words, the change in relative demand equals σ times the change in the relative wage, plus the change in relative supply.

Let's use this equation for a quick calculation. The authors' preferred estimate of σ is 1.64. Suppose hypothetically that the change in the relative wage were 0.10 and there were no change in relative supply. What would be the authors' inferred change in relative demand?

Q7: Now suppose the change in the relative wage were 0.10 and the change in the relative supply were 0.20. What would be the authors' inferred change in relative demand?

Q8: In table 3.2, p. 101, which are generally larger--changes in relative supply or the authors' inferred changes in relative demand? Which side is "winning the race" from 1950 to the present?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist, J. D., & Pischke, J.-S. (2015). Chapter 3. "Instrumental Variables."

Q1: Students at KIPP charter schools have higher test scores than students in nearby schools. So why are some people skeptical of the effectiveness of KIPP schools? (p. 100).

Q2: Table 3.1 on page 104 shows data from the KIPP school in Lynn, Massachusetts. What is the effect of winning the lottery on actually attending this KIPP school? What is the effect of winning the lottery on math score? What is the effect of winning the lottery on verbal score? [Hint: Use the numbers in the column labeled lottery "Winners versus losers."]

Q3: In the KIPP schools example, the authors want to measure the effect of KIPP *attendance* (a dummy variable) on students' test scores, but they use a KIPP *offer* of admission (another dummy variable) as an instrument. Explain why the "offer" variable meets the three requirements of an instrument given at the bottom of page 106.

(i) The *relevance* requirement is that the *instrument have a causal effect on the treatment variable.* A KIPP offer satisfies this requirement because...

(ii) The *independence* requirement is that the *instrument be uncorrelated with any omitted variables.* A KIPP offer satisfies this requirement because...

(iii) The *exclusion* requirement is that the instrument *affect the outcome only through the treatment variable*. A KIPP offer satisfies the exclusion requirement because...

Q4: In the KIPP schools example, the authors compute the instrumental-variables (IV) estimate of the effect of attendance on math score. How should the IV estimate of the effect of attendance on verbal score be computed? What is its value? (See equation 3.1 on page 107).

Q5: In the Minnesota Domestic Violence Experiment example, the authors want to measure the effect of "coddling" by police officers (a dummy variable) on recurrence of domestic assault, but they use random color-coded report forms as an instrument. Explain why the "color-code" variable meets the three requirements of an instrument given at the bottom of page 106.

(i) The *relevance* requirement is that the *instrument have a causal effect on the treatment variable.* The color code satisfies this requirement because...

(ii) The *independence* requirement is that the *instrument be uncorrelated with any omitted variables.* The color code satisfies this requirement because...

(iii) The *exclusion* requirement is that the instrument *affect the outcome only through the treatment variable*. The color code satisfies this requirement because...

Q6: In the Minnesota Domestic Violence Experiment example, what is the estimate based on LATE (IV)? What is the estimate based on treatment delivered (OLS)? Why are they different? (p. 120).

Q7: In the "Population Bomb" example, the authors want to measure the effect of family size on education of first-born children, but they use a dummy variable for first two children being the same sex as an instrument. Explain why the "same-sex" dummy variable meets the three requirements of an instrument given at the bottom of page 106.

(i) The *relevance* requirement is that the *instrument have a causal effect on the treatment variable.* The "same-sex" dummy variable meets this requirement because...

(ii) The *independence* requirement is that the *instrument be uncorrelated with any omitted variables.* The "same-sex" dummy variable meets this requirement because...

(iii) The *exclusion* requirement is that the instrument *affect the outcome only through the treatment variable*. The "same-sex" dummy variable meets this requirement because...

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Fortin, N. M., & Lemieux, T. (1997). Institutional changes and rising wage inequality: is there a linkage? *Journal of Economic Perspectives*, 11(2), 75-96.

Q1: The authors of this paper base most of their arguments on graphs of density functions for log wages. In theory, without a legal minimum wage, the density function for log wages would presumably be a smooth bell-shaped curve. With a strong legal minimum wage, the curve would presumably be squeezed or bunched at the low end. With this in mind, how did the fall in the real minimum wage from 1979 to 1988 affect wage inequality among MEN? (See figure 1, p. 83).

Q2: How did the fall in the real minimum wage from 1979 to 1988 affect wage inequality among WOMEN? (See figure 1, p. 83).

Q3: Unions have multiple effects on overall wage inequality. How might unions decrease wage inequality? How might they increase inequality? (p. 79, and figure 2 on p. 85)

Q4: How did deunionization since 1979 affect the wage INEQUALITY of men? Of women? (See figure 2 on p. 85).

Q5: According to the calculations in table 2 (p. 89), which institutional change (deunionization or the falling real minimum wage) contributed the most toward rising wage inequality among men? Among women?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist, J. D., & Pischke, J.-S. (2015). Chapter 4. "Regression Discontinuity Designs."

Q1: In a regression discontinuity (RD) design, what is a "running variable"? How is it related to the dummy treatment variable in a "sharp RD" design? (p. 151)

Q2: If the effect of the running variable is not linear, quadratic terms and interactions with the dummy treatment variable are sometimes included. In that case, to simplify interpretation of estimation results, the running variable is typically redefined to equal zero at the cutoff, as shown in equation (4.3) on page 155. In equation (4.3), find the derivative d Ma / d Da, when a = a0.

Q3: In a "fuzzy RD" design, how is the running variable related to the treatment variable? (pp. 165-166)

Q4: "Peer effects" are the effects of people around us on our own success. There has been considerable research on "peer effects" in education--that is, the effect on a student's learning of the other students at the same school. Formerly, a typical study estimating peer effects would regress an individual student's test score on the same student's past performance and the average test score of their peers, as in equation (4.6) on page 170. The coefficient of peers' average test scores, θ_1 ("theta 1") in equation (4.6), was thought to estimate the causal effect of peers. Today, it is recognized that θ_1 may not have a causal interpretation. Why not?

Q5: In a "fuzzy RD" design, coefficients are estimated using instrumental variables or two stage least-squares. (pp. 172-173).

(a) In the first stage, the dependent variable is the treatment intensity, such as the average test score of peers. What are the regressors?

(b) In the second stage, the dependent variable is the outcome of interest, such as the individual student's test score. What are the regressors?

(c) Which variable is the *instrument*, excluded from the second stage?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: an empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279-1333.

Q1: This article focuses on the distinction between “routine” and “nonroutine tasks” at work. The CPS records the occupations of individual workers, but not what kinds of tasks they do. Where did the authors find information on what kinds of tasks workers do? (p. 1281).

Q2: As computers decrease in price and increase in speed, how do they affect the demand for "routine-task" labor and "nonroutine-task" labor, according to the authors? (pp. 1284-1285).

Q3: The authors' model assumes that workers can supply either routine labor ("task input") or nonroutine labor, but their model implies that the wage of routine labor is "pinned down by the price of computer capital" (p. 1288). Why?

Q4: I suggest you skip sections IA and IB, but read carefully the “three propositions” on pages 1290-1291. These “propositions” are the key implications of the authors’ model which they will test against the data. Fill in the blanks below with either "more" or "less," according the authors' model.

(a) The more routine labor an industry or occupation used in the past, the _____ it will invest in computers.

(b) The more an industry or occupation invests in computers, the _____ it will subsequently increase its use of nonroutine labor.

Q5: Read pages 1291-1292, which explain how the authors measure changes in the amount of routine labor at the "extensive" and "intensive" margins. What do they mean by changes at the "extensive margin"? What do they mean by changes at the “intensive margin"? (p. 1292).

Q6: In the authors' data, what task codes do they classify as "routine"? What task codes do they classify as "nonroutine"? (p. 1293).

Routine	Nonroutine

Q7: Consider the regression results listed as equation (12) on page 1294. Why do they confirm proposition P1 (p. 1290)? Is the estimate statistically significant? Why or why not?

Q8: According to figure 1, entitled "Trends in Routine and Nonroutine Task Input, 1960 to 1998" (p. 1296), which task inputs rose over this period? Which fell?

Q9: Table II, panel B (p. 1298) shows the shifts in tasks performed in the economy as a whole, decomposed into shifts between industries and shifts within industries. Did the shift to nonroutine tasks occur because the economy shifted toward industries where nonroutine skills were needed ("between"), or because all industries began to use more nonroutine skills ("within")?

Q10: The estimates in table III (p. 1304) show the effects of changes in computer use on changes in tasks. Focus on the column labelled "1990-1998," and the row labelled "D Computer use 1984-1997." Do the signs (positive/negative) confirm proposition P2 (p. 1291)? Are the estimates statistically significant? Why or why not?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in U.S. wage inequality: revising the revisionists. *Review of Economics and Statistics, 90*(2), 300-323.

Q1: In the introduction, the authors say their evidence will show that the wage distribution "polarized" after 1987 (p. 301, column 2). This is more complicated than a simple rise in overall wage inequality. What do they mean by "polarization" of the wage distribution?

Q2: The authors argue that any "parallel movement in earnings and employment" (that is, movements of wages and employment in the same direction, positive or negative) is evidence that shifts in labor demand are responsible for observed changes in wages (p. 301, column 2). Explain their argument in terms of a supply-and-demand graph. Could a "parallel movement in earnings and employment" ever be caused by a shift in supply or by a change in the minimum wage? (Hint: sketch a graph before answering.)

Q3: The figures in this paper are important and revealing. Does figure 1 (p. 303) show that overall wage inequality decreased or increased from 1963 to 2005? Why?

Q4: Consider figure 3 (p. 304) which shows what happened to inequality in more detail. Describe what happened to inequality at the top half of the wage distribution and what happened to inequality at the bottom half.

Q5: The authors acknowledge that the real minimum wage is negatively correlated with inequality, as noted by Card and DiNardo (2002). But they are skeptical that that drop in the minimum wage caused the increase in inequality. Why? (See discussion on page 311, and figure 7 on p. 312).

Q6: Figure 10 (p. 318) shows what kinds of tasks are typically performed by workers at different education ("skill") levels. For example, the curve for abstract tasks shows that hardly any workers with low education perform abstract tasks while most workers with high education perform abstract tasks. Turn your attention to the curve for routine tasks, which starts at the left-hand margin at about 50. If routine tasks are replaced by computers, which workers are most vulnerable? Why? (See also discussion at the top of page 319.)

Q7: Figure 11 (p. 319) shows changes over time for occupations, sorted by skill range (education) from low skill to high skill. During the period 1980-1990, which skill range suffered the greatest employment loss and smallest wage increase? During the period 1990-2000, which skill range suffered the greatest employment loss and smallest wage increase? (See also discussion at the bottom of page 319.)

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist, J. D., & Pischke, J.-S. (2015). Chapter 5. "Difference-in-Differences."

Q1: If treatment and control groups are different in many ways, simple cross-section regression is likely to suffer from omitted variable bias or selection bias. How can the difference-in-differences method still produce a causal estimate?

Q2: Why is the "common trends" assumption important for the difference-in-differences method? (pp. 184-186)

Q3: Consider regression equation (5.3) on page 188. Which parameter (α "alpha," β "beta," γ "gamma," or δ "delta") is the causal effect--that is, the effect of the Atlanta Fed's policy of increasing lending to troubled banks?

Q4: Consider regression equation (5.5) on page 194. There are many parameters in this equation: α "alpha," δ_r, 50 β_k ("beta-k"), and 13 γ_j ("gamma-j"). Which parameter is the causal effect of the MLDA on death rates?

Q5: What are "time effects" and "state effects" (pp. 193-194)? Why are they all present in equation (5.5) (p. 194), instead of a single treatment dummy and a single post-treatment dummy as in equation (5.3) (p. 188)?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market. *Industrial and Labor Relations Review*, 43(2), 245-257.

Q1: When did Cuban immigrants on the "Mariel Boatlift" arrive in Miami? How many settled there permanently? (pp. 245-246)

Q2: What does the author mean by a "natural experiment" (p. 245)? Why was the Mariel Boatlift incident a natural experiment (pp. 245-246)?

Q3: How were the Mariel Boatlift immigrants different from other Cuban immigrants? (table 2, p. 249)

Q4: The author notes that black workers in Miami experienced a fall in wages and a rise in unemployment in 1982 (table 4, p. 251), but does not attribute these to the Mariel Boatlift. Why not?

Q5: What caused the decrease in wages of Cuban workers in Miami in 1980-1981 (table 3, p. 250), according to the author?

Q6: What caused the increase in unemployment rate of Cuban workers in Miami in 1980-1981 (table 4, p. 251), according to the author?

Q7: It is certainly surprising that so many new arrivals apparently had so little effect on the Miami low-wage labor market. According to the author, what offsetting mechanisms might have cushioned the effect of the Mariels on the Miami labor market? (pp. 255-256).

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Borjas, G. J., Freeman, R. B., & Katz, L. F. (1996). Searching for the effect of immigration on the labor market. *American Economic Review*, 86(2), 246-251.

Q1: Immigrants vary widely in their level of skill, but do they compete mostly with high skilled native workers or mostly with low-skilled native workers? (p. 246).

Q2: The paper describes *two* approaches for measuring the impact of immigration on wages. The first is the "area" approach, which compares outcomes in *different areas of the country* that have received more or fewer immigrants. (Card's Mariel Boatlift paper uses this approach.) The first column of page 247 shows a regression equation where the dependent variable is a worker's wage and the explanatory variables include the worker's age, education, and the ratio of immigrants to natives in the area. Would you expect the ratio of immigrants to natives in the area to have a positive or negative effect on the wage? Why?

Q3: Estimates of the coefficient of the ratio of immigrants to natives in table 1 (p. 247) are statistically insignificant or positive, contrary to expectations. How do the authors explain this result?

Q4: Estimates of the coefficient of the ratio of immigrants to natives in table 2 (p. 248) are for slightly different regressions based on the same underlying data source. The dependent variable is now the *change* in wages (from 1980 to 1990) for a particular group in a particular area and the explanatory variables include the *change* in the ratio of immigrants to natives in the same group. Columns (2) and (4) show estimates where area is controlled for. Do the signs of the coefficients make more sense? Why?

Q5: The discussion on pp. 248 and 249 suggests two possible explanations why immigration has no detectable effect on local labor markets, but a detectable effect on wide-area labor markets. What are those two reasons?

Q6: The second approach for measuring the impact of immigration on wages is the "factor-proportions" approach, which uses the elasticity of substitution σ ("sigma") to compute changes in education groups' relative wages for the country as a whole:

$$\Delta\ln(w_1/w_2) = (-1/\sigma)\,\Delta\ln(x_1/x_2),$$

where w_1 and w_2 are the wages of workers in groups 1 and 2, and x_1 and x_2 are quantities of workers. The last column of table 3 on p. 249 shows the authors' estimates of $\Delta\ln(x)$ caused by immigration and international trade, for different groups of workers. Given any two of the estimates in this column, $\Delta\ln(x_1/x_2)$ is calculated as $\Delta\ln(x_1) - \Delta\ln(x_2)$. Then $\Delta\ln(x_1/x_2)$ is multiplied by $(-1/\sigma)$, where σ is the elasticity of substitution, to compute the percent change in relative wages $\Delta\ln(w_1/w_2)$. Let's do this for high school versus college.
First, calculate $\Delta\ln(x_1/x_2)$ for x_1=high school equivalents and x_2 = college equivalents.

Q7: Second, compute $(-1/\sigma)$ using the authors' preferred estimate of σ ("sigma"), from a paper by Katz and Murphy (1992), which is 1.41.

Q8: Finally, multiply $\Delta\ln(x_1/x_2)$ by $(-1/\sigma)$ to get $\Delta\ln(w_1/w_2)$. Then interpret your answer--immigration and trade caused the ratio of high-school wages to college wages to fall by how much? (Remember that a change in a logarithm of a variable is a *percent change* in the variable itself.)

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The China syndrome: local labor market effects of import competition in the United States. *American Economic Review*, 103(6), 2121-2168.

Like many articles in the *American Economic Review*, this article is quite long and involved. However, the conclusion (pp. 2158-2159) is clear and succinct. I suggest you begin by reading the abstract and the conclusion carefully. Then skim the introduction (pp. 2121-2125). Then look at the crucial tables, numbered 3-8 (pp. 2137-2149).

Q1:This paper examines the effects of imported goods from China on local labor markets, which the authors call "CZs" in the paper (pp. 2122, 2132). What are "CZs"? Where did the authors get the definitions of these "CZs"?

Q2: U.S. trade with China is largely driven by comparative advantage. In what industries does China have a comparative advantage? (p. 2123)

Q3: The authors' treatment variable is "the change in Chinese import exposure per worker," denoted ΔIPW_{uit} and defined by equation (3) on page 2128. In equation (3), the subscript u means "overall U.S.," the subscript i denotes a particular local labor market (or region), the subscript j denotes a particular industry, the subscript uc means "from China to the U.S.," and the subscript t denotes a particular time period. Suppose for example j denotes the shoe industry. Then ΔM_{ucjt} denotes the rise in shoe imports from China over a particular time period. What does (L_{ijt}/L_{ujt}) mean? And why is ΔM_{ucjt} multiplied by (L_{ijt}/L_{ujt}) ?

Q4: Also in equation (3) on page 2128, what does L_{it} mean and why do the authors divide by it?

Q5: The authors worry that estimates of a simple regression of outcomes like employment on Chinese import exposure per worker will not have a causal interpretation, so they instead compute "instrumental variables" (IV) estimates. They also spend a great deal of the paper checking IV assumptions, which can make for tedious reading. For now, we will skip all this and jump to their results, which are quite interesting.

Table 3 (p. 2137) shows estimates of the effect of Chinese import exposure on local manufacturing employment. Is the effect negative, as hypothesized? Is it statistically significant, and if so why? Is it "robust"--that is, does it hold up when different control variables are included?

Q6: A reduction in local employment would not be a big problem if these workers could find jobs by moving to other local labor markets. Table 4 (p. 2142) shows estimates of the effect of Chinese import exposure on total local population. The authors emphasize the bottom row, labeled "Panel C, full controls." Is the effect negative, as hypothesized? Is it statistically significant? Do workers who lose their jobs in manufacturing indeed move to other local labor markets? (See also discussion at bottom of p. 2142).

Q7: Similarly, a reduction in manufacturing employment would not be a big problem if these workers could find jobs outside manufacturing. By definition, workers who lose jobs in manufacturing must do one of three things: take nonmanufacturing jobs, become unemployed, or leave the labor force (NILF). Panel A of Table 5, columns (2) through (4) (p. 2143) examines these possibilities. When workers lose their jobs in manufacturing due to Chinese imports, what do they do? How do we know? (See also discussion at bottom of p. 2143 and top of p. 2144).

Q8: Consider Panel B of table 5, p. 2143. Do Chinese imports have a larger effect on workers with college degrees or workers with no college? How do we know? (See also discussion at bottom of p. 2144 and top of p. 2145).

Q9: Consider again table 5, p. 2143. Some people who leave the labor force enroll in Social Security Disability Insurance (SSDI). What effect do Chinese imports have on SSDI enrollment? How do we know? (See also discussion on p. 2145.)

Q10: Tables 6 (p. 2146) and 7 (p. 2147) show estimates of the effect of Chinese imports on local wages. Do Chinese imports tend to increase or decrease local average wages? Explain the differences with respect to college versus no college, and manufacturing versus nonmanufacturing.

Q11: Table 8 (p. 2149) shows estimates of the effect of Chinese imports on government transfer (benefit) programs. Do Chinese imports tend to increase or decrease government transfers? For which programs is the effect significant at the 5 percent level?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Rosen, S. (1981). The economics of superstars. *American Economic Review, 71*(5), 845-858.

Q1: What is the "phenomenon of Superstars," as defined by Rosen? (p. 845)

Q3: In developing his model of "Superstars," what special feature of *consumer preferences* does the author assume? Are consumers willing to substitute one good dentist for two mediocre dentists? (p. 846).

Q3: What special feature of *technology* allows slightly more talented persons to gain much higher incomes?

Q4: Why can singers, actors, and athletes become "Superstars" today, but could not do so in the nineteenth century?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Bivens, J., & Mishel, L. (2013). The pay of corporate executives and financial professionals as evidence of rents in top 1 percent incomes. *Journal of Economic Perspectives*, 27(3), 57-78.

Q1: What two occupations account for the majority of the increase in income in the top 1.0 percent and the top 0.1 percent?

Q2: During what decade did CEO compensation grow most rapidly?

Q3: We have read papers that attribute the rise in relative pay for college-educated workers to an increase in relative demand for such workers. Do the authors attribute the rise in pay for CEOs to an increase in demand? Explain.

Q4: Do the authors attribute the rise in pay for finance workers to an increase in demand? Explain.

Q5: The authors cite studies finding that lower marginal tax rates on top incomes increase top incomes but do not seem to raise overall economic growth (pp. 72-73). This is a surprise, because if top managers and entrepreneurs enjoy greater rewards, one might expect they would supply more effort and thereby raise economic growth. How do the authors explain these findings?

Q6: Some authors, such as Goldin and Katz (2008), propose increasing the supply of highly-skilled workers in order to reduce inequality. Do the authors of this paper propose increasing the supply of CEOs and finance workers? What do they propose?

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Kaplan, S. N., & Rauh, J. D. (2013). It's the market: the broad-based rise in the return to top talent. *Journal of Economic Perspectives*, 27(3), 35-56.

Q1: Figure 1 (p. 37) shows average and median pay of CEOs. Why is the average always greater than the median?

Q2: What is the difference between a publicly-traded company and a privately-held company? (p. 38).

Q3: One hypothesis explaining the rise in CEO pay is that CEOs are increasingly able to extract rents from compliant boards of directors. How does a comparison of pay for CEOs in publicly-traded companies versus privately-held companies help test this hypothesis? What do researchers find? (p. 39)

Q4: How does a comparison with top hedge fund managers and law partners help test the rent-extraction hypothesis? (table 1, p. 40) What do researchers find?

Q5: How does a comparison with athletes help test the rent-extraction hypothesis? (figure 3, p. 42) What do researchers find?

Q6: What theory of inequality do the authors find most plausible in explaining the "broad-based rise in the return to top talent"? (pp. 42-43).

[end of reading guide]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

READING GUIDE

Angrist, J. D., & Pischke, J.-S. (2015). Chapter 6. "The Wages of Schooling."

Q1: The general formula for omitted-variable bias is
OVB = (regression of omitted variable on included variable) times (effect of omitted variable in long equation)
= δ ("delta") times γ ("gamma") in the second equation on page 212 (see also p. 72).
Suppose ability is omitted from the equation for wages, and determine whether the coefficient of schooling is biased up or biased down as follows.

(a) What is presumably the sign of "regression of omitted on included" (p. 212) and why?

(b) What is the sign of "effect of omitted in long equation" and why?

(c) What should we therefore expect to be the sign of OVB, positive or negative?

Q2: What is a "bad control"? (p. 217). Why might some people argue that IQ score is a bad control for ability?

Q3: How can data on twins be used to eliminate ability bias in estimating the returns to schooling? (pp. 217-219).

Q4: What happens to the estimated coefficient of a regressor if that regressor is measured with error? (pp. 219-221).

Q5: The twins studies by Ashenfelter, Krueger, and Rouse, described in section 6.2, seek to estimate the coefficient of schooling on earnings while correcting for two problems. Ability bias would tend to raise the estimated coefficient of schooling. Measurement error on schooling would tend to lower the estimated coefficient of schooling. Consider the estimates in table 6.2 (p. 220).

(a) Differencing should correct for ability bias but could exacerbate measurement error. Did differencing raise or lower the estimated coefficient of schooling? (Compare columns (1) and (2).)

(b) So differencing either fixed ability bias or exacerbated measurement error. We want to know which. Instrumental variables (IV) applied to the differenced equation should correct for measurement error if there is any. Did IV applied to the differenced equation raise or lower the estimated coefficient of schooling? (Compare columns (2) and (4).)

(c) Bottom line: was there really any ability bias in column (1)?

(d) Another way to figure out whether ability bias was present is to apply IV without differencing. By itself, IV should correct for both ability bias and measurement error. Did IV raise or lower the estimated coefficient of schooling? (Compare columns (1) and (3).) Again, was there really any ability bias in column (1)?

Q6: In the compulsory-attendance study by Angrist and Acemoglu, why should one worry about the common-trends assumption? In other words, why should state-specific trends be introduced as control variables? (pp. 226-227).

Q7: In the study by Angrist and Krueger, quarter-of-birth (QOB) is used as an instrument for schooling, in hopes of correcting for ability bias. Explain why the QOB variable meets the three requirements of an instrument given at the bottom of page 106 (pp. 233-234).

Q8: Again, IV should correct for both ability bias (which biases the coefficient up) and measurement error (which biases the coefficient down). In the quarter-of-birth study by Angrist and Krueger, the IV estimates of the coefficient of schooling (table 6.5, p. 232) are as large or larger than the OLS estimates. So which seems to be the more important problem in their data--ability bias or measurement error? Why?

[end of reading guide]

SECTION B
WRITING ASSIGNMENTS

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

ASSIGNMENT: MIDTERM ESSAY

Each student will write a midterm essay reflecting on the academic research we have read through Kaplan and Rauh (2013). The essay should use and cite evidence presented in that research. The length should be about 1000 to 1500 words, plus bibliography.

Question to be addressed: *Is wage inequality in the United States likely to continue to grow in the future?* First, summarize and synthesize the evidence in our readings. What factors have caused the increase in wage inequality in the last fifty years? *Cover all the "explanations" in the syllabus*. Next, predict whether *each* of these factors will continue, stop, or perhaps reverse. It is difficult to predict the future, but give your best guesses and defend them. Give reasons for your predictions, but acknowledge uncertainty as appropriate. Finally, summarize your predictions to answer the question to be addressed.

Audience: Assume your readers know some economics and understand terms like "elasticity of substitution," "Gini coefficient," "human capital," and "standard error." But assume your readers are not familiar with trends and explanations for wage inequality.

Organization: The top of the first page should contain the title ("Is Wage Inequality in the United States Likely to Grow in the Future?"), your name (no "by"), and the date. The body may be divided into sections if desired. An alphabetized bibliography formatted in APA style should appear on a separate page at the end. Include and refer to one or more of the figures you created in Excel from data in the Census Bureau report, *Income in the United States*.

Writing style: Academic economics is written in an impersonal, objective style. Personal feelings do not belong. Uncertainty and conflicting evidence are recognized. Suspicions and conjectures are acknowledged as such. Conclusions are never overstated. As in natural science, use of the passive tense is perfectly acceptable.

Citations and bibliography: Citations in the text must use (author date) style. List of references must be in APA format (see http://library.albany.edu/cfox). Note that Monday readings in the syllabus are already listed in APA format, so for your bibliography, you can simply copy and paste!

Writing Center: You must visit the Writing Center before submitting this essay. Appointments are made at https://library.drake.edu/writing-center/ .

Proofread your work: Points will be deducted for spelling and grammatical errors, for deviation from APA style in citations and bibliography, for omitting page numbers, or for not using the Writing Workshop.

Checkpoints:

Due dates	*What*
Friday April 5	Schedule an appointment at the Writing Workshop. Bring your essay and this assignment sheet. Work with the Writing Tutor to ensure your essay tells a coherent story.
Friday April 12	Submit midterm essay on Blackboard by 11:59 PM.

GRADING RUBRIC FOR MIDTERM ESSAY

Student name		
Component	Score	Comments
Does the paper show clear organization? Can the reader discern the author's outline and line of argument?	/20	
Are all the factors that have caused the increase in wage inequality over the last fifty years clearly explained?	/20	
Are sensible predictions offered for these factors in the future? Are those predictions defended?	/20	
Is the question in the title answered directly in a concluding paragraph?	/20	
Is proper academic writing style observed, including author-date citations, APA-style bibliography, topic sentences, objective tone, etc.?	/20	
Points deducted for spelling and grammatical errors, omitting page numbers, or not visiting Writing Workshop.		
TOTAL	/100	

[end of assignment]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

ASSIGNMENT: RESEARCH PAPER

Each student will write an empirical research paper. The paper should estimate a relationship or test a hypothesis using real data and econometric tools we discuss in class. The paper need not be on "inequality"—it may be on any economic topic. The idea need not be original—your paper can simply replicate an existing paper, perhaps with different data. The length should be 5 to 8 pages, plus references, tables, and graphs.

Getting started: Think about what general topic interests you in the first few weeks of class. Stop by during my office hours to chat. I can suggest data sources and papers to check. Allow plenty of time for finding and accessing data—this is often the most time-consuming part of any research project. Think carefully about whether you are measuring a cause-and-effect relationship.

Organization: Your paper must look professional and conform to formatting conventions of economic research, so that professionals in the field are more likely to read and understand it. To make your job easier, a template in Microsoft Word has been posted on Blackboard. ***All students must use the template*** and may deviate from it only with the instructor's prior permission. Be sure to delete all square brackets as you fill in your own material.

Writing style: Academic economics is written in an impersonal, objective style. Personal feelings do not belong. Uncertainty and conflicting evidence are recognized. Suspicions and conjectures are acknowledged as such. Conclusions are never overstated. As in natural science, use of the passive tense is perfectly acceptable.

Citations and bibliography: Citations in the text must use (author date) style. List of references must be in APA format (see http://library.albany.edu/cfox).

Writing Center: You must consult the Writing Center before submitting your first draft. Appointments can be made at https://library.drake.edu/writing-center/ .

Proofread your work: Points will be deducted for spelling and grammatical errors, for deviation from APA style in citations and bibliography, for not using the template provided on Blackboard, or for not using the Writing Workshop.

Oral presentations: In late April, each student will make a 10-minute presentation to the class and Economics faculty members. Prepare a PowerPoint presentation that summarizes your paper. Create at least one slide for each section of your paper and one slide for each table or graph. Fonts should be at least 20 pt for readability.

Deadline	*Checkpoint*	*Length*	*Points*
Friday Feb 23, 11:59 PM.	(1) Submit proposal, including research question and data description, on Blackboard.	1 page	10
Friday Mar 29, 11:59 PM.	(2) Submit draft data section and methodology section of research paper (use the template!) on Blackboard. Also submit Stata log file showing how you computed descriptive statistics.	2-4 pages	10
By Friday Apr 19	Schedule appointment with Writing Workshop to discuss 1st draft of research paper. Bring your entire paper and this assignment sheet. Work with the Writing Tutor to ensure your paper tells a coherent story.		
Friday Apr 26, 11:59 PM.	(3) Submit 1st draft of research paper (use the template!) on Blackboard. Also submit Stata log file showing how you computed the descriptive statistics and the estimates.	5-8 pages plus references, tables, and graphs	20
Apr 22-May 1	Oral presentations.	10 minutes	
Friday May 10, 11:59 PM.	(4) Submit final draft of research paper and Stata log file on Blackboard. Also submit Stata log file showing how you computed the descriptive statistics and the estimates.	5-8 pages plus references, tables, and graphs	60

[end of assignment]

[TITLE OF PAPER]

[Author]

Drake University

[Date]

ABSTRACT: [Give a summary of your paper in 100 to 200 words. Briefly state your research question, your method, and your results. Highlight in a sentence or two what is unique about your paper. Provide enough information for someone to decide whether they want to read further, but maintain an objective and impersonal style.]

KEYWORDS: [Give 3-5 specific words or phrases that should lead a search engine to your paper. Good examples: returns to schooling, female labor-force participation, Islamic banking, rent control. Bad examples: schooling, participation, banking, housing, price, quantity, elasticity, cause, effect (too general).]

1. INTRODUCTION

[What is the general topic and why is it interesting? What research question(s) are you asking in this paper? Foreshadow your method and results, but don't give them all away.]

2. LITERATURE REVIEW

[What is already known about this question from economic theory and/or from previous empirical papers? Find 3-5 previous papers using Google Scholar or using journal databases accessed through Cowles Library such as EBSCO, ECONLIT, or JSTOR. (If you download a paper on your topic in the *Journal of Economic Perspectives* from www.aeaweb.org/journals/jep, subsequent papers that cited that paper will be listed at the very end of the PDF document. This is a handy way to find the most recent published research on your topic.)

You are not expected to completely understand the methodology of the papers you review. Just give the main idea of each paper, which should be clear from reading the abstract, introduction, and conclusion.

Do not merely list the papers. Compare them to each other and tell a story about how previous research has tackled your question. How will your approach differ from previous papers? Cite all papers using APA's author (date) style.

In this and every section, the first paragraph should tell the reader where you are going, and the last paragraph should summarize key points and prepare the reader for the next section.]

3. DATA

[What data will you use to address your research question? Where were the data obtained? Who created the data? Are the data well-suited to your research question? Why? Explain the structure of the data set. Is it a cross-section, a time-series, or a panel? What is the unit of observation: a person, a county, a state, a country, etc., over a week, a month, a year, etc.? How many observations are there? Did you exclude any observations, and if so, why? If you transformed any of the data, explain how and why. Discuss your table of variable definitions and your table of descriptive statistics—see examples below. Are the mean values and the ranges sensible? If not, the reader will not take your analysis seriously. Discuss any interesting features of the data.]

4. METHODOLOGY

[What equation are you estimating? Write it out like this:

$$\ln(\text{wage}) = \beta_1 + \beta_2 \text{ schooling} + \beta_3 \text{ experience} + \beta_4 \text{ experience}^2 + \text{error}$$

Identify the outcome (or dependent) variable, the treatment (or independent) variable, and the controls. What functional form (logs, ratios, per-capita units, square terms, interactions, etc.) did you choose and why? How does the equation bear on your research question?

Will the estimates have a causal interpretation, and if so, why? In other words, what is your *identification strategy* (randomization, regression, instrumental variables, regression discontinuity, difference-in-differences, or something else)? Is there a risk of omitted variable bias or selection bias? If so, what controls have been omitted?]

5. RESULTS

[Give a table of coefficient estimates and include the standard error directly below each estimate in parentheses—see example below. Discuss the values of the point estimates. Do the signs and values of all the coefficients make sense? Why or why not?

Which coefficient estimate addresses your research question? What is its sign and value? Is it statistically significantly different from zero at 5%? (Refer to p-value or t-statistic. Remember that the R-square value is NOT a test statistic.) If not, report a 95% percent confidence interval for it. Remember that a low t-statistic can mean either that the coefficient is small or that the standard error is large.]

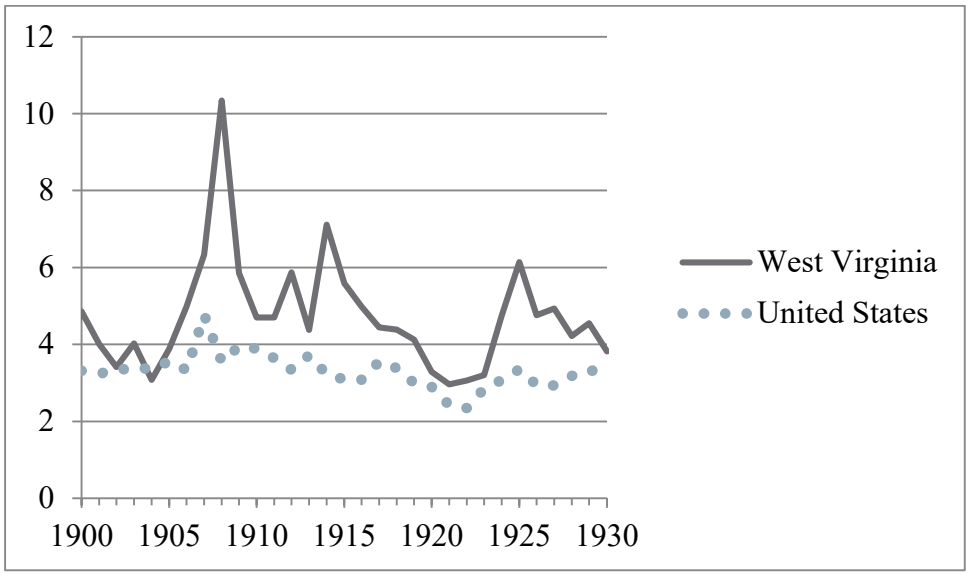
6. CONCLUSIONS

[What answers can you now give to the question(s) you posed in the introduction? What questions remain unanswered?]

REFERENCES

[The list of references must start on a new page. References must include all papers cited in your paper and all data sources including websites, but nothing else (no "suggested readings.") List everything in APA format—see http://library.albany.edu/cfox for examples, and note that readings in our syllabus are listed in APA format. References must be in alphabetical order of first author's surname. If first authors are the same, order papers by second author's surname, etc. If all authors are the same, list them chronologically by date of publication.]

FIGURE 1: COAL FATALITIES PER THOUSAND COAL WORKERS



SOURCES: West Virginia—West Virginia Office of Miners' Health, Safety and Training, http://www.wvminesafety.org/historicprod.htm, accessed June 8, 2014. United States—Mine Safety and Health Administration, http://arlweb.msha.gov/stats/centurystats/coalstats.asp, accessed March 8, 2016. West Virginia data are for fiscal years ending June 30 through 1924, and calendar years thereafter; data for 1925 cover 18 months.

[Figures can help tell your story because they often communicate more clearly than the most carefully-written prose. Figures and tables must start on a new page. Note the difference between figures and tables: figures are time plots, scatterplots, graphs, diagrams, and maps, while tables are arrays of words or numbers. If your data are time series, you are REQUIRED to include a time plot of the dependent variable and most important independent variable(s). If your methodology is difference-in-differences, you are REQUIRED to include a time plot of the dependent variable in "treated" and "control" groups so that the reader can check the "parallel-trends" assumption. If your dataset is small (under 100 observations) you are REQUIRED to include a scatterplot of the key dependent variable against the key independent variable. In other cases, figures are recommended but not required.

Figures can appear at the end of your paper, or they can be embedded in the narrative immediately after they are referenced for the first time, but not both. If you choose to place your figures at the end of the paper, put each one on a separate page. If you choose to embed your figures, be sure that each one starts on a new page.]

TABLE 1: VARIABLE DEFINITIONS

Variable	Definition	Formula
Wage	Average hourly earnings	Earnings last week / hours worked last week
Schooling	Years of schooling completed	
Experience	Years of potential labor-market experience	Age – schooling - 6
Union	Member of a labor union	1=member, 0=nonmember

SOURCE: Current Population Survey, March 2009.

[A table of variable definitions is required. The above is an example. Note that for readability, names of variables are plain English, not Stata variable names. Note the formatting: there are no vertical lines, and only two horizontal lines.

Tables can appear at the end of your paper, or they can be embedded in the narrative immediately after they are referenced for the first time, but not both. If you choose to place your tables at the end of the paper, put each one on a separate page. If you choose to embed your tables, be sure that each one starts on a new page and does not break across pages—in MS Word, choose *Paragraph>Line and page breaks>Keep with next*. Every table must be discussed, at least briefly, in your text, but include enough notes to the table that it can be understood without reading the text.]

TABLE 2: DESCRIPTIVE STATISTICS

	Number of observations	Mean	Standard deviation	Minimum	Maximum
Wage	953	20.57	4.65	5.50	88.75
Schooling	952	13.72	1.75	8.00	20.00
Experience	952	15.38	3.95	2.00	52.00
Union	921	0.092	0.289	0.00	1.00

SOURCE: Current Population Survey, March 2009. Accessed at ipums.org, February 2011.

[A table of descriptive statistics is required. All statistics necessary for this table can be computed by the Stata command `summarize` , but do *not* simply paste Stata output into your paper. Instead, use the more professional format shown above. The purpose of this table is to help the reader understand your estimates in the next table, so compute the numbers for this table *after* you have dropped unusable observations. For readability, names of variables should be plain English, not Stata variable names. Show all digits for whole numbers and at least three significant digits for fractions. Note the formatting: there are no vertical lines, and only two horizontal lines. You can add columns for additional statistics, such as the median (computed with `summarize` option `detail`), if they help the reader understand the data better.]

TABLE 3: ESTIMATES OF THE EFFECT OF SCHOOLING ON LOG HOURLY WAGE

	(1)	(2)	(3)
Schooling	0.127	0.143	0.133
	(0.304)	(0.272)	(0.265)
Experience		0.012	0.011
		(0.002)	(0.002)
Experience, squared		-0.0006	-0.0005
		(0.002)	(0.002)
Union member			0.143
			(0.062)
Adjusted R-square	0.038	0.044	0.047
Number of observations	952	952	921

NOTES: Dependent variable is log hourly wage. Standard errors robust to heteroskedasticity are shown in parentheses.

[A table of estimates is required. Do *not* simply paste Stata output into your paper. Instead, use the more professional format shown above. Note that for readability, names of variables are plain English, not Stata variable names. Note the formatting: there are no vertical lines, and only two horizontal lines. Standard errors must be in parentheses, so they are easy to distinguish from coefficient estimates, and notes at the bottom should describe how those standard errors were computed. Show three decimal places after the decimal point—more if the number would otherwise show as zero. Show at least three columns of alternative estimates to illustrate how robust your results are to alternative equation specifications. Add more regressors (X variables) as you move from left to right. If you have more *rows* of estimates than can fit on the page, and some of the coefficients are not important to your story, don't display unimportant coefficients in the table, but say in the notes "Control variables include …" If you have more *columns* than can fit on the page, create another table.]

DATA APPENDIX

[A data appendix is required. It should start on a new page. The purpose of a data appendix is to document your data well enough that another economist could find the same data and replicate your results. So your data appendix must give a detailed explanation of how you obtained the data, including any websites and dates accessed. It should also describe which observations you deliberately dropped from the analysis and why (for example, data were incomplete, data seemed to contain errors, etc.) and how you transformed the data. Finally, describe the software (Stata, Excel, etc.) used to make the calculations. In the case of Stata, indicate which commands were used to compute the estimates (`regress`, `ivregress`, `xtreg`, `xtivreg`, etc.) If you include tables in your appendix, number them A1, A2, etc., and use the same formatting as regular tables. End your appendix with the following sentence.]

Data and Stata code used to compute the estimates are available from the author on request.

[And mean it! Keep a copy of your data file and your Stata do-file.]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

GRADING RUBRIC
RESEARCH PAPER CHECKPOINT 1
PROPOSAL

Student name		
Component	Score	Comments
RESEARCH QUESTION: What is your research question? Are you trying to estimate a causal relationship?	/3	
DATA: What is the source of your data? Are the data cross-section, time series, panel, or pooled? What is the unit of observation? What is the outcome variable? What is the treatment variable(s)?	/7	
OVERALL	/10	

[end of rubric]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

GRADING RUBRIC
RESEARCH PAPER CHECKPOINT 2
DATA AND METHODOLOGY SECTIONS

Student name		
Component	Score	Comments
DATA SECTION TEMPLATE: What data will you use to address your research question? Where were the data obtained? Who created the data? Are the data well-suited to your research question? Why? Explain the structure of the data set. Is it a cross-section, a time-series, or a panel? What is the unit of observation: a person, a county, a state, a country, etc., over a week, a month, a year, etc.? How many observations are there? Did you exclude any observations, and if so, why? If you transformed any of the data, explain how and why. Discuss your table of variable definitions and your table of descriptive statistics—see examples below. Are the mean values and the ranges sensible? If not, the reader will not take your analysis seriously. Discuss any interesting features of the data.	/4	
METHODOLOGY SECTION TEMPLATE: What equation are you estimating? Write it out like this: ln(wage) = β1 + β2 schooling + β3 experience + β4 experience2 + error Identify the outcome (or dependent) variable, the treatment (or independent) variable, and the controls. What functional form (logs, ratios, per-capita units, square terms, interactions, etc.) did you choose and why? How does the equation bear on your research question? Will the estimates have a causal interpretation, and if so, why? In other words, what is your identification strategy (randomization, regression, instrumental variables, regression discontinuity, difference-in-differences, or something else)? Is there a risk of omitted variable bias or selection bias? If so, what controls have been omitted?	/4	
Table of variable definitions	/1	
Table of descriptive statistics	/1	
Points deducted for spelling or grammatical errors, for not following template, or for skipping class workshop		
TOTAL	/10	

[end of rubric]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

GRADING RUBRIC
RESEARCH PAPER CHECKPOINT 3
FIRST DRAFT

See template posted on Blackboard for expected contents of each section.

Student name		
Component	Score	Comments
Section 1: Introduction	/3	
Section 2: Literature review	/3	
Section 3: Data	/3	
Section 4: Methodology	/3	
Section 5: Results	/3	
Section 6: Conclusions	/3	
Format: title page, abstract, keywords, references, tables, figures, data appendix, and page numbers.	/2	
Points deducted for spelling or grammatical errors, for deviation from APA style in citations and bibliography, for not using the template provided on Blackboard, for not using Writing Center, for skipping class research workshop, or for not supplying Stata log file.		
OVERALL	/20	

[end of rubric]

SECTION C
SLIDESHOW HANDOUTS

REVIEW OF ORDINARY LEAST SQUARES (REGRESSION)

- How can we test and measure economic relationships in the real world?

Analyzing data

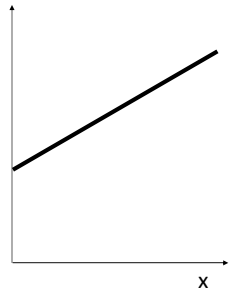
- To test and measure economic relationships, labor economists use _____ from the real world.
- We usually want to find out _____ one variable affects another, and if so, how _____ the relationship is.

Examples of economic relationships

- Does more education increase a worker's earnings? If so, how much?
- Does an increase in the minimum wage reduce employment? If so, how much?
- Do unions raise wages? If so, how much?
- Do increased unemployment benefits cause people to remain unemployed longer? If so, how much?

Linear relationships

- Suppose x and y have a linear relationship: $y = \beta_1 + \beta_2 x$.
- β_1 = _____
- β_2 = _____



Meaning of slope

- If x increases by a small amount, then y changes by β_2 times that amount.
- Example: Suppose $\beta_2 = 2$ and x increases by 0.4. Then y increases by (approximately) _____.

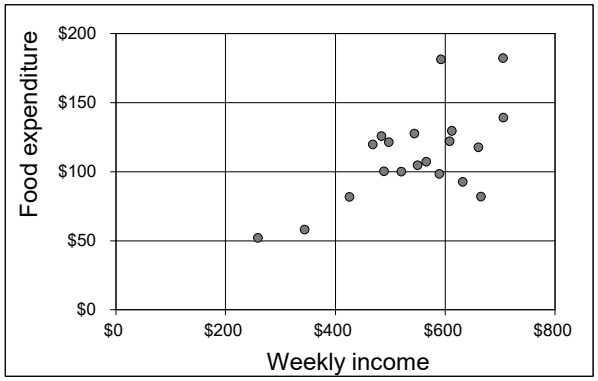
Measuring relationships

- Suppose we have data on x_i and y_i, for i = 1 through n.
- We believe that x and y have a roughly _____ relationship: $y = \beta_1 + \beta_2 x$.
- How can we estimate β_1 and β_2?

Example 1: household income and food expenditure

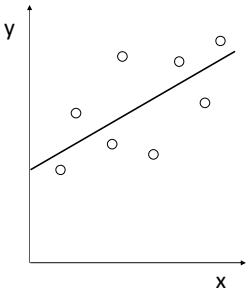
Household no.	Weekly income	Food expenditure	Household no.	Weekly income	Food expenditure
1	258.3	52.25	11	564.6	107.48
2	343.1	58.32	12	588.3	98.48
3	425	81.79	13	591.3	181.21
4	467.5	119.9	14	607.3	122.23
5	482.9	125.8	15	611.2	129.57
6	487.7	100.46	16	631	92.84
7	496.5	121.51	17	659.6	117.92
8	519.4	100.08	18	664	82.13
9	543.3	127.75	19	704.2	182.28
10	548.7	104.94	20	704.8	139.13

Scatterplot of household food expenditure dataset



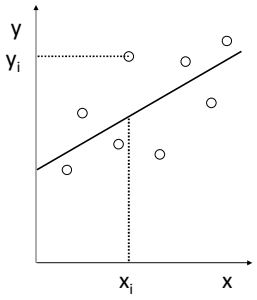
Fitting a line to data

- Economic data rarely lie exactly on a straight line.
- So we must find a line that fits the data "best."
- How to choose the "best"-fitting line?



Deviations from the line

- The "best" line would come as close as possible to the actual data points.
- Suppose we measure deviations from the line vertically.



The least-squares principle

- Choose the line that minimizes the sum of the squared vertical deviations.
- In other words, find values of β_1 and β_2 that minimize the following:

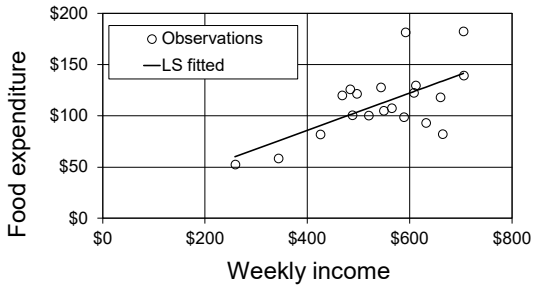
$$\sum_{i=1}^{n}\left(y_i - \left[\beta_1 + \beta_2 x_i\right]\right)^2$$

- MS Excel has a "regression tool" for this.

OLS estimates for the household food expenditure dataset

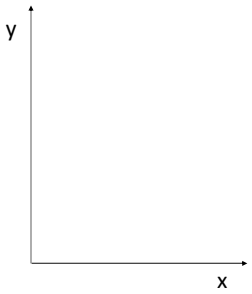
- Y-intercept (β_1): 12.94
- Slope (β_2) : 0.18
- Fitted line: y = _____ + _____ x
- Interpretation: if income increases by one dollar, spending on food increases by $_____.

OLS fitted line for household food expenditure dataset



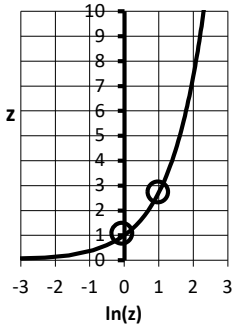
Nonlinear relationships

- If the relationship between x and y is not likely to be linear, we can still use OLS regression if we first apply a trick.
- Convert one or both variables to natural _____ (ln).



Natural logs

- Definition of natural logarithm: ln(z) = logarithm of z to base e = 2.711828…
- In other words $z = e^{\ln(z)}$.
- Thus if ln(z) increases by 1 unit, then z itself is _____ by a factor of e = 2.711828…



Small changes in natural logs

- From calculus we know that $\frac{d\ln(z)}{dz} = \frac{1}{z}$ or $d\ln(z) = \frac{dz}{z}$.
- This implies that if ln(z) increases by a small amount, then z itself changes by approximately that _____.

Change in ln(z)	Approx. change in z
0.01	%
0.03	%
0.05	%
0.10	%

Percent change ≈ change in natural log: example

- How much more are engineers paid than managers, on average? _____%.
- How much more are engineers paid than mathematicians and computer scientists? _____%.

Occupation	Hourly wage	Log hourly wage
Managers	$23.57	3.16
Engineers	$29.08	3.37
Mathematical and computer scientists	$28.79	3.36

Meaning of slope when y is in natural logarithms

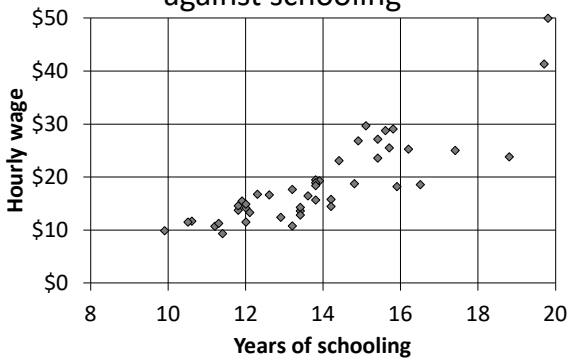
- Suppose $\ln y = \beta_1 + \beta_2 x$.
- If x increases by a small amount, then ln(y) changes by β_2 times that amount.
- But y itself changes by a _____ about equal to β_2 times that amount.
- Example: Suppose $\beta_2 = 0.03$ and x increases by 2. Then y increases by (approximately) _____%.

Example 2: average hourly wage and average schooling by occupation

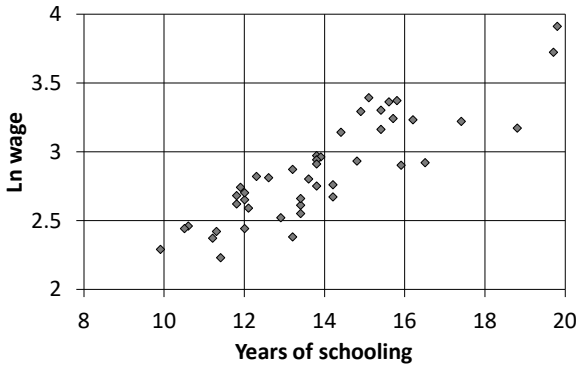
Occupation	Mean log hourly wage of male workers	Mean years of schooling for male workers
Administrators and officials, public admin.	3.24	15.7
Other executives, administrators, and managers	3.29	14.9
Management-related occupations	3.16	15.4
Engineers	3.37	15.8
Mathematical and computer scientists	3.36	15.6
Etc.	Etc.	Etc.
Forestry and fishing occupations	2.70	12.0

Borjas 5th edition, page 15, table 1-1. 45 observations total.
SOURCE: Annual demographic files of the CPS, 2002.

Scatterplot of average hourly wage against schooling



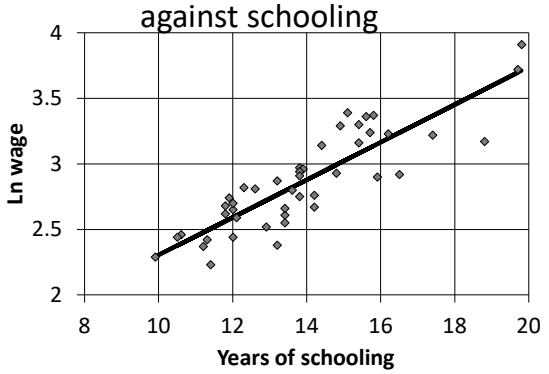
Scatterplot of **ln**(wage) against schooling



OLS estimates for the wage and schooling dataset

- Using Excel, I computed OLS estimates of $\ln w = \beta_1 + \beta_2 s$.
- Intercept (β_1): 0.869
- Slope (β_2) : 0.143
- Fitted line: ln w = _____ + _____ s
- Interpretation: if schooling increases by one year, wage increases by about _____ %.

OLS fitted line of ln(wage) against schooling



Standard errors and confidence intervals

- OLS estimates are not precise because datasets are, at best, _____ samples from a much larger population.
- The "standard error" of a OLS estimate is a measure of its precision.
- Roughly speaking, an estimate is within _____ standard errors of its true value 95% of the time.

Standard errors for the wages and schooling dataset

- Excel also reported the standard errors shown below in parentheses:

  ln w = 0.869 + 0.143 s
  (0.172) (0.012)

- So a 95% confidence interval (or "margin of error") for the slope would be
  $0.143 \pm 2 \times 0.012 = 0.143 \pm$ _____.

Statistical significance

- We say that an estimate is "statistically significant at the 5% level" if
  - the 95% confidence interval does not include zero,
  - or equivalently, $t = \frac{estimate}{standard\ error} > 2$.
- For the slope estimate, t = 0.143/0.012 = _____.
- Clearly the slope estimate _____ statistically significant.

Multiple regression

- In the real world, any variable y is affected by multiple variables.
- If an important variable (other than x) is changing in our dataset, then our simple OLS estimate of the slope may be _____.
- To get an unbiased estimate, we must include that other variable in a _____–regression equation:

  $y = \beta_1 + \beta_2 x + \beta_3 z$.

Example 2: percent female employment by occupation

Occupation	Mean log hourly wage of male workers	Mean years of schooling for male workers	Female share (%)
Administrators and officials, public admin.	3.24	15.7	52.4
Other executives, administrators, and managers	3.29	14.9	42.0
Management-related occupations	3.16	15.4	59.4
Engineers	3.37	15.8	10.7
Mathematical and computer scientists	3.36	15.6	32.2
Etc.	Etc.	Etc.	Etc.
Forestry and fishing occupations	2.70	12.0	3.7

OLS estimates for the multiple-regression equation

- ln w = 0.924 + 0.150 s - 0.003 f
  (0.154) (0.011) (0.001)
- Interpretation: holding female share (f) constant, if schooling increases by one year, wage increases by about _____ %.
- Holding schooling constant, if female share increases by one percentage point, wage _____ by about _____%.

Conclusions

- OLS regression allows us to test and measure real-world relationships between variables.
- If the y variable is in natural logs, then the estimate of β_2 shows the _____ increase in y from a one-unit increase in x.
- If an important variable (other than x) is changing in our dataset, we should use _____ regression.

TYPES OF DATA SETS

•What four forms do economic data sets usually take?

Structure of economic datasets

- *Datum* = a single number, like 47.5. Plural of datum is _____.
- *Dataset* = array of data to be analyzed.
- Datasets are often arranged so that the rows are _____ and the columns are _____.
- Datasets differ in how observations are related to each other.

Types of datasets

1. Cross-sections.
2. Time series.
3. Pooled cross-sections.
4. Panels.

1. Cross-sectional datasets

- All observations collected at roughly the same point in time.
- Observations can be people, firms, industries, cities, countries, etc.

Obs. #	Name	Age	Education	Income
1	B. Smith	34	12 years	$38,845
2	C. Valdez	47	16 years	$65,150
3	J. Huang	24	18 years	$45,275

Cross-sectional datasets are easiest to analyze

- Often we can plausibly assume that observations are a _____ *sample* from some larger population.
- Each new observation is a fresh draw from the population, unrelated to other observations.
- Observations are thus _____.

2. Time-series datasets

- Same individual (person, firm, country) is observed repeatedly over time.
- Frequency might be weekly, monthly, quarterly, or annual.

Obs. #	Year	Unempl. rate	Inflation (CPI)	GR RGDP per capita
1	2000	4.0	3.4	2.5
2	2001	4.7	2.8	-0.6
3	2002	5.8	1.6	1.1

Patterns in time-series

- Time-series data often show _____ patterns (unless the data are annual).
  - Electricity use peaks in July or August every year (most places).
  - Unemployment peaks in June most years.
- Time series data often show long-run _____ (usually upward).
  - GDP, employment, and the price level all trend upward.

Time series datasets are harder to analyze

- Time-series observations are *not* usually independent over time.
- Example: If GDP is above trend in one quarter, there is a good chance GDP will be _____ trend in the next quarter, too.
- Each new observation is *not* a fresh draw from the population. Time-series data sets cannot be considered _____ samples.

3. Pooled cross-section datasets

- Several cross-section datasets are combined (or pooled).
- Example: Surveys from several different years, covering different individuals, might be combined into one dataset.
- Observations in the same year might be related to each other, but not to observations in another year.

What a pooled dataset looks like

Obs. #	Year	Name	Age	Income
1	2000	B. Smith	34	$38,845
2	2000	C. Valdez	47	$65,150
3	2000	J. Huang	24	$45,275
4	2002	P. Abdul	65	$55,250
5	2002	A. O'Toole	19	$22,750
6	2002	H. Schmidt	29	$44,500

Why pooled datasets can be useful

- Can include _____ observations than a single cross-section. The more observations, the more precise the statistical estimates.
- Can estimate relationships _____ each cross section, and compare to see whether relationship has changed over time.

4. Panel (or longitudinal) datasets

- Same cross-section is followed over time.
- Same individuals appear, period after period.
- Example: Statistical Abstract contains data on the same 50 states, year after year.
- Example: Some government surveys collect information from the same people month after month.

What a panel dataset looks like

Obs. #	Year	Name	Age	Income
1	2000	B. Smith	34	$38,845
2	2001	B. Smith	35	$40,150
3	2002	B. Smith	36	$42,750
4	2000	C. Valdez	47	$65,150
5	2001	C. Valdez	48	$68,275
6	2002	C. Valdez	49	$70,155

What a panel dataset looks like (another example)

Obs. #	Year	Country	Consumer price index	Employment
1	2017	France	106.9	26.8
2	2018	France	108.8	27.0
3	2019	France	110.1	27.1
4	2000	UK	114.9	32.0
5	2001	UK	117.6	32.4
6	2002	UK	119.6	32.7

Why panel datasets can be useful

- Sometimes can get better *ceteris paribus* measures.
- Extraneous differences between individuals (if constant over time) can be removed by focusing on *changes* over time in the same individual.

Conclusions

- A _____ dataset observes many individuals (persons, firms, states, countries, etc.) at one point in time.
- A _____ dataset observes one individual repeatedly at many points in time.
- A _____ dataset combines several cross-sections.
- A _____dataset observes the same set of individuals at different points in time.

MEASURING INEQUALITY

- How can we compare inequality between countries and over time?

Differences in inequality

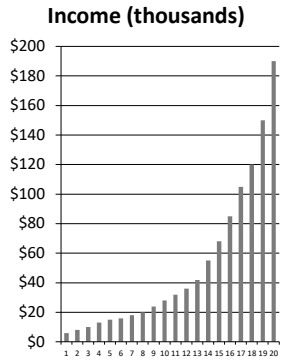
- Is income inequality greater in the U.S. than in other countries?
- Has inequality increased over time?
- To answer these questions we need a way to measure inequality.

Possible measures

- Range
- Variance or standard deviation
- Variance or standard deviation of logarithm
- Coefficient of variation
- etc.

Another approach

- How much income goes to each segment of the distribution?
- Begin by ordering all households from lowest income to highest income.
- Here is an artificial example with 20 households.

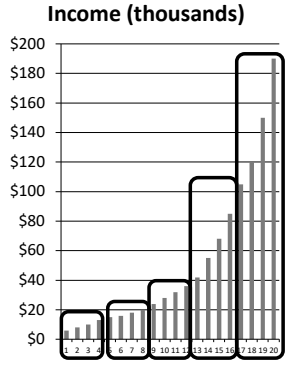


Quantiles

- Then divide population into groups of equal size.
- Same number of households in each group.
  - 4 groups: "_____."
  - 10 groups: "_____."
  - 100 groups: "_____."

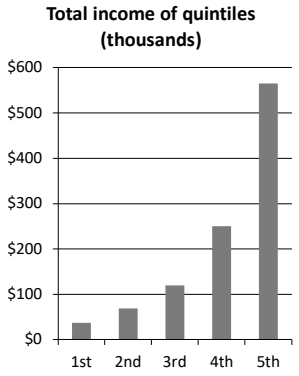
Quintiles

- If we divide observations into _____ groups of equal size, groups are called "quintiles."



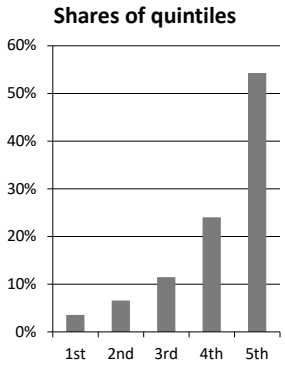
Income of quintiles

- Then sum the income of each group.



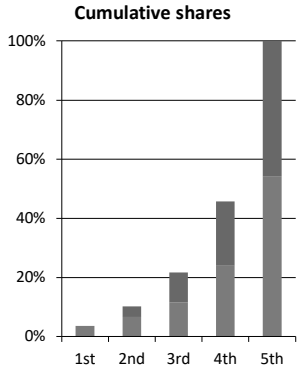
Income shares of quintiles

- Then divide the income of each group by the total of all groups.
- In this example, the 1st quintile's share is less than _____ percent.
- 5th quintile's share is over _____ percent.



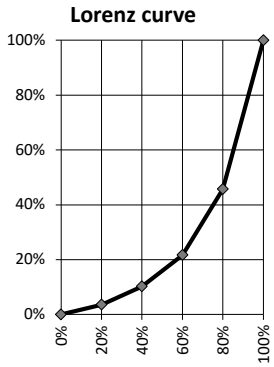
Cumulative shares of quintiles

- Now *cumulate* the income shares.
- 1st is unchanged.
- New 2nd = 1st + 2nd.
- New 3rd = 1st + 2nd + 3rd.
- Etc.
- Last share must equal _____%.



Lorenz curve

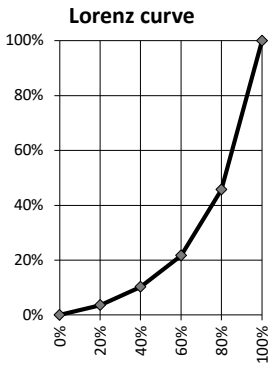
- Graph cumulative income share against cumulative population share.
- Result is a curve with increasing slope.



Lorenz, M. O. (1905). *Methods of measuring the concentration of wealth*. Publications of the American Statistical Association. Vol. 9 (New Series, No. 70) 209-219.

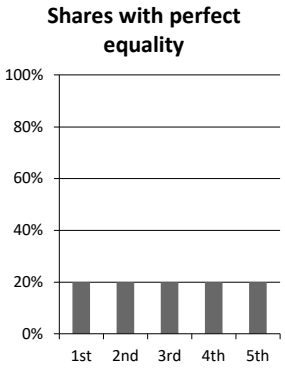
Lorenz curve with other quantiles

- What happens if we divide population into smaller groups (e.g., percentiles)?
- Lorenz curve gets smoother, more accurate.
- Still goes through _____.



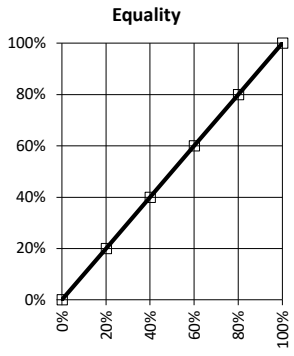
Extreme cases: perfect equality

- Suppose every household had exactly the same income.
- Then each quintile would have exactly a _____ share.



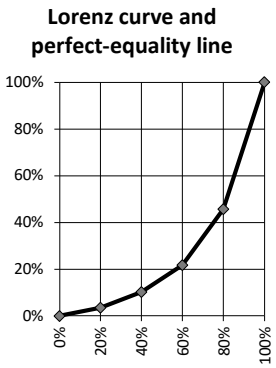
Extreme cases: perfect equality (cont'd)

- Graph cumulative income share against cumulative population share.
- With perfect equality, Lorenz curve would be a _____ at 45 degrees.



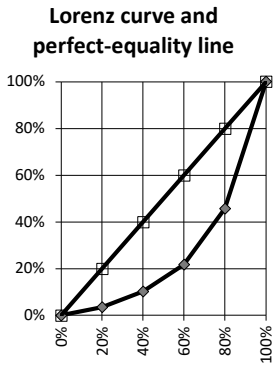
Actual curve versus perfect equality

- The farther the actual Lorenz curve is from the 45-degree line, the more _____ the distribution.



Extreme cases: perfect inequality

- Perfect inequality would occur if _____ household had all the income.
- Lorenz curve would look like a backwards "L".

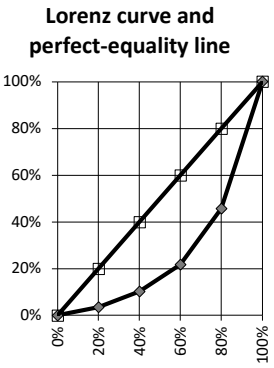


Gini coefficient

= area between Lorenz curve and equality line, divided by total area under equality line.

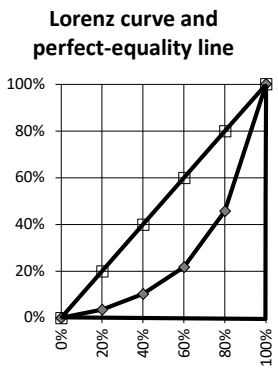
= _____ in this example.

Gini, C. (1912) Italian: Variabilità e mutabilità (Variability and Mutability), C. Cuppini, Bologna, 156 pages.



Extreme values of Gini coefficient

- With perfect equality, Gini coefficient = 0/0.5 = _____.
- With perfect inequality, Gini coefficient = 0.5/0.5 = _____.



Quintiles for US household income, 2022

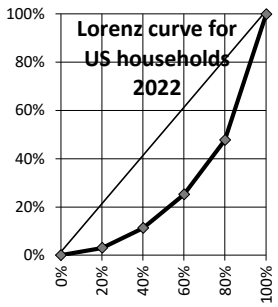
Quintile	Share of income	Cumulative share of income
First	0.030	
Second	0.082	
Third	0.140	
Fourth	0.226	
Fifth	0.522	

SOURCE: U.S. Census Bureau, "Income in the United States: 2022." Table A-4b, p. 33. Issued September 2023.

Lorenz curve and Gini coefficient for US household income, 2022

Gini coefficient can be computed from areas of triangles and rectangles between Lorenz curve and equality line.

Gini coefficient = _____.

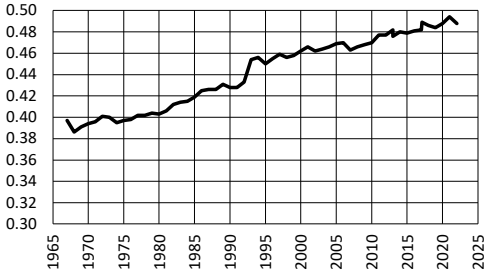


Gini coefficients of family income

Country	Gini	Country	Gini
Canada	33.3	China	38.2
Germany	31.7	India	35.7
Sweden	29.3	Malaysia	41.1
United Kingdom	35.1	Mexico	45.4
United States	41.5	South Africa	63.0

SOURCE: CIA World Factbook, https://www.cia.gov/the-world-factbook/field/gini-index-coefficient-distribution-of-family-income/country-comparison/, accessed December 2023. Note: household ≠ family.

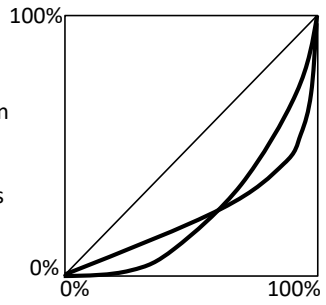
Gini index of US household income inequality



SOURCE: U.S. Census Bureau, "Income in the United States: 2022." Table A-4b, p. 34. Issued September 2023.

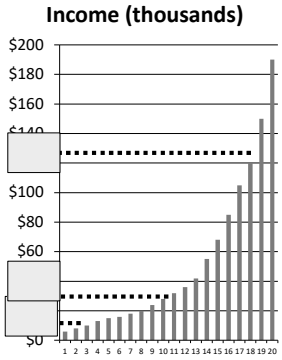
Limitations of Gini coefficient

- Gini coefficient is a _____ measure, a single number.
- Cannot distinguish between inequality in different parts of the distribution.
- Alternative measures of inequality are needed.



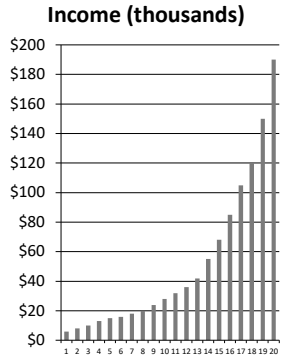
Alternative measures

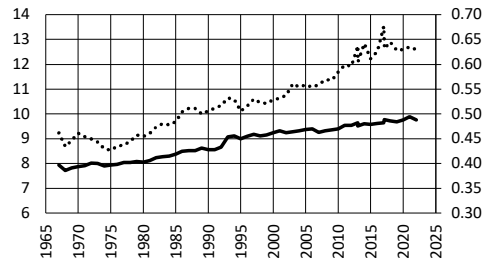
- Let w_{10} = income of 10th percentile.
- Let w_{50} = income of 50th percentile (or _____).
- Let w_{90} = income of 90th percentile.



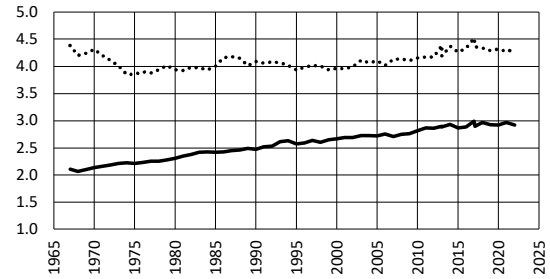
Alternative measures: income gaps

- 90-50 income gap = $(w_{90}-w_{50})/w_{50}$.
- 50-10 income gap = $(w_{50}-w_{10})/w_{10}$.
- 90-10 income gap = $(w_{90}-w_{10})/w_{10}$.





SOURCE: U.S. Census Bureau, "Income in the United States: 2022." Tables A-4a and A-4b, pp. 33-34. Issued September 2023.



SOURCE: U.S. Census Bureau, "Income in the United States: 2022." Tables A-4a and A-4b, pp. 33-34. Issued September 2023.

Conclusions

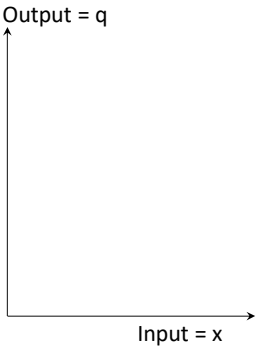
- To compare inequality across countries or over time, we need a way to measure it.
- A _____ curve graphs cumulative shares against cumulative population.
- A _____ coefficient = area between Lorenz curve and perfect equality line / 0.5.
- Other measures of inequality are the 90-50 gap and the 50-10 gap.

SKILLED VERSUS UNSKILLED WAGES

- What determines the relative pay of skilled and unskilled workers?

Production with 1 input

- Production function: $q = q(x)$.
- x = input or factor of production.
- Marginal product of x: $MP = dq/dx$.
- Example: $q(x) = 200\ x^{1/2}$.
  $MP =$ _____.



Profit maximization with 1 input and all prices taken as given

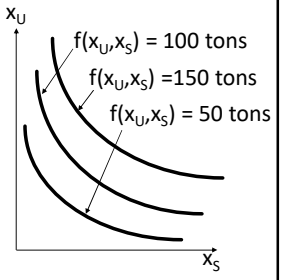
- Let p = price of output, w = price of input.
- Profit = rev – cost = $p\ q - w\ x = p\ q(x) - w\ x$.
- To maximize profit, set derivative = 0: $p\ (dq/dx) - w = 0$, or **$p\ MP = w$**.
- p MP called "value of marginal product," VMP.
- So profit maximization implies _____.
- (w/p) called "real wage," so alternatively _____.

Production with 2 inputs

- Production function: $q = q(x_1, x_2)$.
- Can use this framework for many issues related to income inequality. Examples:
  - x_1 = unskilled workers, x_2 = skilled workers.
  - x_1 = labor, x_2 = capital.

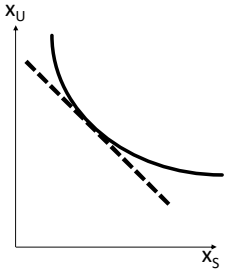
Graphing production with 2 inputs

- Graph with *isoquants*, connecting input combinations yielding same output.
- Isoquants usually slope down and are curved.
- Position depends on technology.



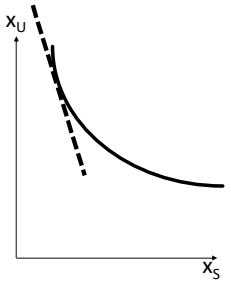
Marginal rate of substitution

- Marginal products:
  $MP_U = \partial q/\partial x_U$
  $MP_S = \partial q/\partial x_S$.
- MP_S/MP_U called "marginal rate of substitution" (MRS) = |slope| of isoquant.



Diminishing marginal rate of substitution

- MRS usually diminishes x_U decreases and x_S increases.
- Isoquants usually are curved.

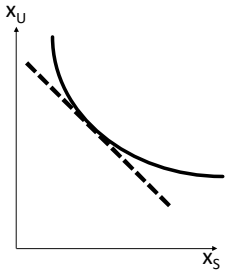


Profit maximization with 2 kinds of workers

- Let w_U , w_S = wages of workers.
- Profit = rev – cost = $p\, q(x_U,x_S) - w_U x_U - w_S x_S$.
- To maximize profit, set derivatives = 0:
  $p\,(\partial q/\partial x_U) - w_U = 0$ or $\mathbf{p\, MP_U = w_U}$
  $p\,(\partial q/\partial x_S) - w_S = 0$ or $\mathbf{p\, MP_S = w_S}$.
- So again profit maximization implies _____.

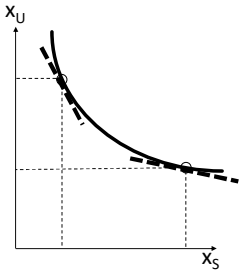
Marginal rate of substitution equals ratio of input prices

- Divide second equation by first:
  $$\frac{p\ MP_S}{p\ MP_U} = \frac{w_S}{w_U}$$
- Profit maximization implies MRS = |slope| of isoquant = ratio of input prices.



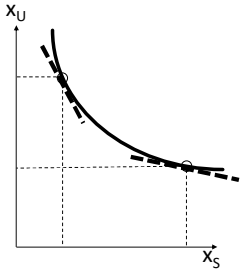
Relative input prices negatively related to relative input quantities

- If isoquants are curved, MRS is **negatively** related to (x_S/x_U).
- So (w_S/w_U) **negatively** related to (x_S/x_U), along an isoquant.



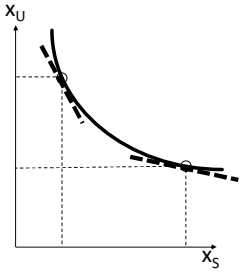
What if unskilled workers become relatively more expensive?

- Then |slope| of isoquants (w_S/w_U) decrease.
- Cost-minimizing firm chooses less x_U and more x_S.
- Firm **substitutes away** from unskilled workers.



What if skilled workers become relatively more expensive?

- Then |slope| of isoquants (w_S/w_U) increase.
- Cost-minimizing firm chooses less x_S and more x_U.
- Firm **substitutes away** from skilled workers.

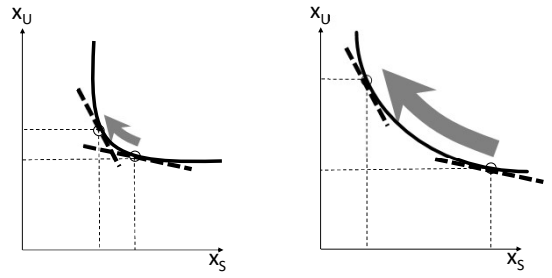


Elasticity of substitution (σ)

- Strength of firm's response is called "elasticity of substitution":
$$\sigma = -\frac{d\ln(x_S/x_U)}{d\ln(MRS)} = -\frac{d\ln(x_S/x_U)}{d\ln(w_S/w_U)} > 0$$
- Effectively, σ ("sigma") measures straightness of isoquants, or ease of substitution.

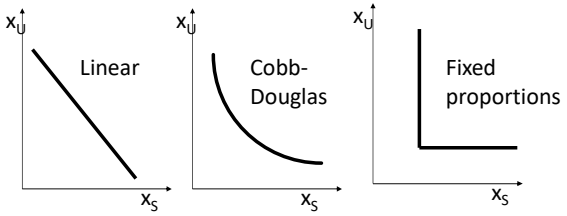
Strength of response to changes in (w_S/w_U) depends on straightness of isoquant

Weak response, _____ σ. **Strong response, _____ σ.**



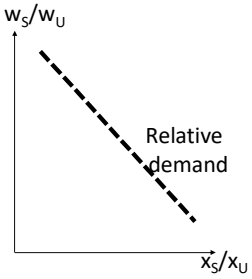
Elasticity of substitution (σ) for various production functions

σ = _____ σ = _____ σ = _____



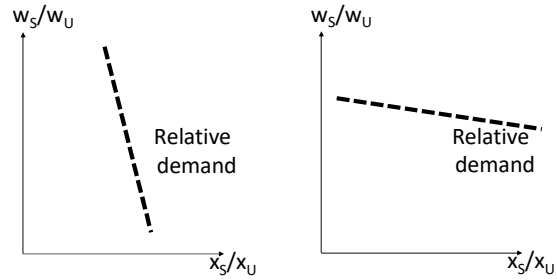
Same idea graphed as "relative demand curve"

- Wage ratio (w_S/w_U) must be **negatively** related to relative demand for workers (x_S/x_U).
- In this diagram, σ measures flatness of relative demand curve.



Elasticity of substitution measures flatness of "relative demand curve"

Steep curve, _____ σ. **Flat curve, _____ σ.**

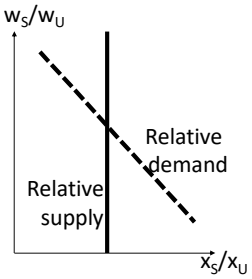


Economy as a whole

- Individual firms take wages as given and choose their (x_S/x_U) to minimize cost of production.
- But for economy as a whole, total relative supply **(x_S/x_U) is fixed** in short run.
- So at least in short run, total (x_S/x_U) determines (w_S/w_U).

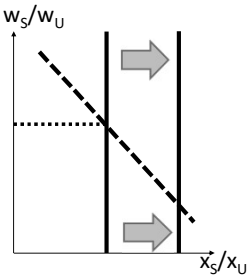
Economy as a whole: relative demand and supply

- Relative supply is vertical, at least in short run.
- Relative wages (w_S/w_U) determined by intersection of relative demand curve with vertical relative supply curve (x_S/x_U).



Change in relative supply in the long run

- In long run, if relative supply (x_S/x_U) increases, then relative price (w_S/w_U) decreases.
- Amount of decrease depends on flatness of relative demand, measured by elasticity of substitution (σ).



Suppose elasticity of substitution (σ) is constant

- Recall definition: $\sigma = -\frac{d\ln(x_S/x_U)}{d\ln(w_S/w_U)}$
- Taking reciprocals: $\left(\frac{1}{-\sigma}\right) = \frac{d\ln(w_S/w_U)}{d\ln(x_S/x_U)}$
- If σ is constant, then: $\left(\frac{1}{-\sigma}\right) = \frac{\Delta\ln(w_S/w_U)}{\Delta\ln(x_S/x_U)}$

Using σ to compute effect of relative supplies on relative wages

$\Delta(x_S/x_U)$	Elasticity of substitution (σ)	$\left(\frac{1}{-\sigma}\right)$	$\Delta(w_S/w_U)$
+10%	0.8		%
+10%	1.0		%
+10%	1.5		%
+10%	2		%

Measuring elasticity of substitution (σ)

- If σ is constant, then: $\left(\frac{1}{-\sigma}\right) = \frac{\Delta\ln(w_S/w_U)}{\Delta\ln(x_S/x_U)}$
- Suppose we have data on relative wages and relative supplies of skilled and unskilled workers over time.
- Could fit this equation to data by LS:

$$\ln\left(\frac{w_S}{w_U}\right) = a + \left(\frac{1}{-\sigma}\right)\ln\left(\frac{x_S}{x_U}\right)$$

Conclusions

- The _____ of workers equals the increase in output from one more worker.
- Competition implies that, for any skill level, workers' wage equals the value of marginal product**.**
- The _____ is the |slope| of isosquants, and equals the ratio of marginal products.
- The _____ **(σ)** measures the straightness of isoquants, or the ease of substitution.
- **Relative worker wages** (w_S/w_U) are negatively related to **relative supplies** (x_S/x_U) of workers, *ceteris paribus*. This negative relationship is stronger if σ is smaller.

OMITTED VARIABLE BIAS FORMULA

- What happens when an important regressor is omitted from a regression equation?

What happens if a variable is omitted from a regression equation?

Suppose wages depend on education and ability:

$$W = \beta_1 + \beta_2 E + \beta_3 A + \varepsilon_W ,$$

where $\beta_2 > 0$ and $\beta_3 > 0$. We want to measure the effect of E, the treatment variable, on W. But suppose we don't have data on A.

Is there a problem if we just leave A out of the regression equation?

A little algebra

Again:

$$W = \beta_1 + \beta_2 E + \beta_3 A + \varepsilon_W ,$$

Now suppose that education and ability are correlated. We can express this correlation as another regression equation:

$$A = \pi_1 + \pi_2 E + \varepsilon_A .$$

Substitute:

$$W = \beta_1 + \beta_2 E + \beta_3 \mathbf{(\pi_1 + \pi_2 E + \varepsilon_A)} + \varepsilon_W$$

$$W =$$

Formula for omitted variable bias

So instead of estimating :

$$W = \beta_1 + \beta_2 E + \beta_3 A + \varepsilon_W ,$$

ordinary least squares will estimate:

$$W = (\beta_1+\beta_3\pi_1) + (\beta_2+\beta_3\pi_2) E + (\varepsilon_W+\beta_3 \varepsilon_A).$$

Our estimate of the coefficient of E is *biased* due to an omitted variable (A). Bias = _____.

In words, omitted variable bias equals

{effect of omitted [A] in long equation} × {regression of omitted [A] on included [E]}.

Formula for omitted variable bias: numerical example

Example in Angrist and Pischke, pp. 69-72:

$$Y = \beta_1 + \beta_2 P + \beta_3 A + \varepsilon_Y ,$$

$$A = \pi_1 + \pi_2 P + \varepsilon_A .$$

Here, Y = earnings, P = attendance at private college, and A = ability (or "applicant group").

If A is omitted in the top equation, formula for omitted variable bias = _____.

Suppose $\pi_2 = (1/6)$ and $\beta_3 = 60{,}000$.

Then bias = _____.

Formula for omitted variable bias: qualitative example

Often we can only guess the sign (+/-) of omitted variable bias.

$$\text{productivity} = \beta_1 + \beta_2 \text{workday} + \beta_3 \text{machines} + \varepsilon_Y ,$$

$$\text{machines} = \pi_1 + \pi_2 \text{workday} + \varepsilon_A .$$

If machines are omitted from first equation, formula for omitted variable bias = _____.

Suppose we guess $\pi_2 < 0$ and $\beta_3 > 0$.

Is the estimate of β_2 biased up or biased down? _____

Put differently, is omitted variable bias positive or negative? _____

When not to worry

But it seems like we always omit *some* variables from regression equations. Is it ever OK to omit them?

The formula for the bias is $\boldsymbol{\beta_3 \pi_2}$, where

$$W = \beta_1 + \beta_2 E + \beta_3 A + \varepsilon_W ,$$
$$A = \pi_1 + \pi_2 E + \varepsilon_A .$$

So no bias if either $\beta_3 = 0$ or $\pi_2 = 0$. In words, there is no bias if either

When not to worry: example

Again, suppose we have

$$W = \beta_1 + \beta_2 E + \beta_3 A + \varepsilon_W ,$$
$$A = \pi_1 + \pi_2 E + \varepsilon_A .$$

Suppose E is *randomly assigned*.

Then A and E are uncorrelated and $\pi_2 = 0$.

So omitted variable bias $\beta_3 \pi_2 =$ _____.

Conclusions

"Omitted variable bias" occurs in regression equation if a regressor is omitted which is both

- _____ —that is, it has a nonzero effect on the dependent variable—and
- _____ __ with the included regressor.

The bias occurs on the coefficient of the included regressor, which picks up some of the effect of the omitted regressor.

ERROR TERM CORRELATED WITH REGRESSOR

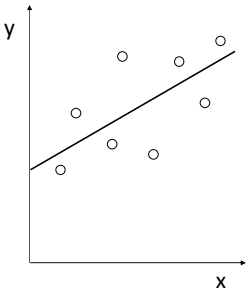
- What key assumption, if violated, causes LS estimates to be biased and inconsistent?

Key assumption

- Suppose we are trying to fit a model like
$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$
- A key assumption when using ordinary least squares (OLS) is:
$$E(\varepsilon_i | x_i) = 0 .$$
- Value of error term ε_i is not affected by value of regressor x_i.
- Or, x_i is uncorrelated with all other unobserved factors (ε_i) that affect y_i .

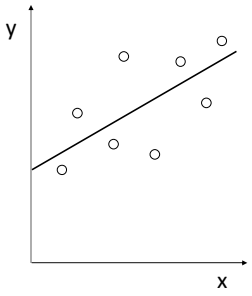
Technical meaning of assumption: $E(\varepsilon_i | x_i) = 0$

- This assumption implies $E(\varepsilon_i x_i) = 0$.
- If x_i is viewed as a random variable, then this further implies $Cov(x_i,\varepsilon_i) = Corr(x_i,\varepsilon_i) = 0$.



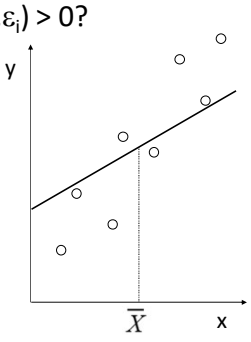
Graphical meaning of assumption: $E(\varepsilon_i | x_i) = 0$

- Data are scattered evenly on either side of regression line.
- Scattering is unrelated to value of x_i .



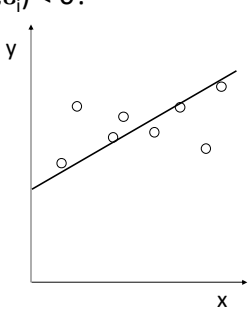
What if assumption were violated, with $Cov(x_i,\varepsilon_i) > 0$?

- Suppose x_i and ε_i are positively correlated.
- Then observations tend to be above line for large x_i, but below line for small x_i.
- So LS slope estimate is biased _____.



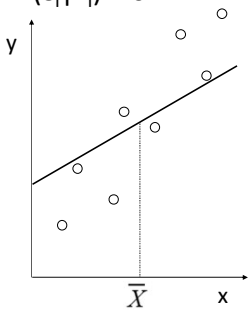
What if assumption were violated, with $Cov(x_i,\varepsilon_i) < 0$?

- Suppose x_i and ε_i are negatively correlated.
- Then observations tend to be below line for large x_i, but above line for small x_i.
- So LS slope estimate is biased _____.



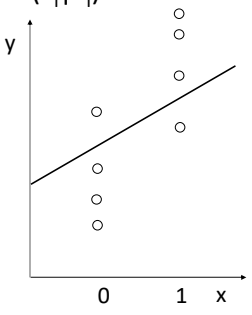
Omitted variable bias (OVB) is an example of $E(\varepsilon_i | x_i) \neq 0$

- Suppose y = earnings and x = schooling.
- Suppose ability also affects earnings, but is omitted from equation.
- Anything omitted is in ε_i .
- So if x and ability are positively correlated, then $Cov(x_i,\varepsilon_i) > 0$.
- OLS slope estimate is biased _____.



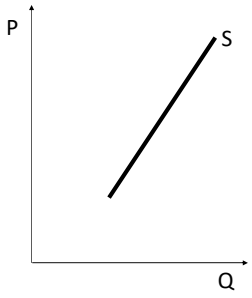
Selection bias is an example of $E(\varepsilon_i | x_i) \neq 0$

- Suppose y = test score and x = attends charter school.
- Suppose students in charter schools are more motivated or have more involved parents.
- "Motivation" is in ε_i .
- So then $Cov(x_i,\varepsilon_i) > 0$.
- OLS slope estimate is biased _____.



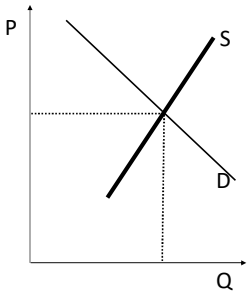
Endogeneity is an example of $E(\varepsilon_i | x_i) \neq 0$

- Suppose we wish to estimate supply equation for corn: $Q = \beta_1 + \beta_2 P_i + \varepsilon_i$
- Supply depends on other things beside price: weather, pests, etc. These are in ε_i .



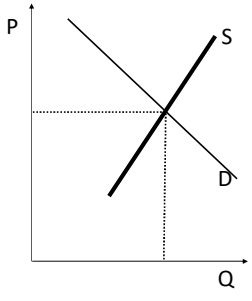
Endogeneity is an example of $E(\varepsilon_i | x_i) \neq 0$ (cont'd)

- Economic theory says Q and P are jointly determined by supply and demand, so P is *endogenous.*
- Suppose ε_i is negative (bad weather).
- Then Q decreases and P _____.



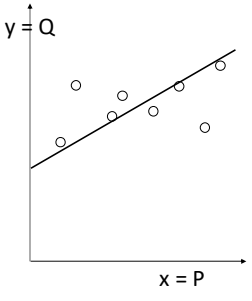
Endogeneity is an example of $E(\varepsilon_i | x_i) \neq 0$ (cont'd)

- Alternatively, suppose ε_i is positive (good weather).
- Then Q increases and P _____.
- So $Cov(P_i,\varepsilon_i) < 0$.



Endogeneity is an example of $E(\varepsilon_i | x_i) \neq 0$ (cont'd)

- When we estimate supply equation for corn: $Q = \beta_1 + \beta_2 P_i + \varepsilon_i$, we have $Cov(P_i,\varepsilon_i) < 0$.
- LS slope estimate is biased _____.



Conclusions

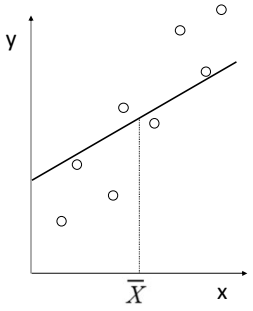
- A key assumption for ordinary least squares is $E(\varepsilon_i | x_i) = 0$.
- That is, x_i is _____ with all other unobserved factors that affect y (in ε_i).
- If violated, OLS slope estimator is _____.
- Examples of violations include
  - Omitted variable bias
  - Selection bias
  - Endogeneity

INSTRUMENTAL VARIABLES

- What is the method of "instrumental variables"?
- When is it better than the method of ordinary least squares (OLS)?

Error term correlated with regressor

- If unobserved error term is correlated with regressor, OLS slope estimator is *biased*.
- OLS is also *inconsistent*, meaning problem does not go away as sample size increases.



What to do?

1. If possible, make x *randomly assigned*, so that x is not correlated with error term (or anything else).
2. If possible, include relevant *control variables* as extra regressors.
3. If possible, use method of *instrumental variables*.

What is an instrument?

- Suppose we are trying to estimate the slope coefficient of x in
  $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$.
- An "instrument" is a variable that is correlated with x_i but _____ with ε_i .

Instrumental variables (IV) estimation

- Suppose w_i is an instrument.
- First, estimate ϕ_1 and ϕ_2 in this "first-stage" equation:
  $x_i = \phi_1 + \phi_2 w_i + \varepsilon_{1i}$.
- Save the fitted values $\hat{x}_i$.
- Second, estimate β_1 and β_2 in this "second-stage" equation:
  $y_i = \beta_1 + \beta_2 \hat{x}_i + \varepsilon_{2i}$.

IV chain from cause to effect

$w \rightarrow x \rightarrow y$

Equation	Relationship
First stage	
Second stage (LATE)*	
Reduced form (ITT)**	

*The second-stage equation is the relationship of interest, sometimes called the "structural equation." If the instrument is randomly assigned, the second-stage coefficient of x is called the "local average treatment effect" (LATE).
**If the instrument is randomly assigned, the reduced-form coefficient of w is called the "intention to treat effect" (ITT).

Properties of IV estimators

- It can be shown that second-stage estimators are _____, meaning they converge in probability to true values of β_1 and β_2 as sample size increases, unlike OLS.
- However, IV estimators may have some bias in small samples.
- Also, IV estimators have larger _____ than OLS estimators.

Extension: "two-stage least squares"

- Strictly speaking, term "instrumental variables" is usually applied only to case of _____ regressor and one instrument.
- But method can be extended to situations with _____ instruments and controls.

2SLS: example

- Suppose we are trying to estimate β_2 in
  $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$,
  where x_{3i} is a control variable.
- We fear that x_{2i} is correlated with ε_i .
- But we have two instruments w_{1i} and w_{2i} that are correlated with x_{2i} but not with ε_i .

2SLS: example (cont'd)

- First, estimate ϕ_1, ϕ_2, ϕ_3 in this "first-stage" equation:
  $x_{2i} = \phi_1 + \phi_2 w_{1i} + \phi_3 w_{2i} + \phi_4 x_{3i} + \varepsilon_{1i}$.
- Save the fitted values $\hat{x}_{2i}$.
- Second, estimate β_1 and β_2 in this "second-stage" equation:
  $y_i = \beta_1 + \beta_2 \hat{x}_{2i} + \beta_3 x_{3i} + \varepsilon_{2i}$.

Computing IV estimates

- If you compute first and second stage using `regress` (or Excel), coefficient estimates will be correct but standard errors will be wrong.
- Instead, use special Stata command:
  ```
  ivregress 2SLS y x3 (x2=w1 w2),
    first vce(robust)
  ```
  which automatically computes standard errors correctly.

Finding good instruments (the hard part)

An "instrument" is a variable that is correlated with x_i but not with ε_i . This means:

1. Has *causal effect* on x.
2. Is *unrelated to any omitted variables* (in ε_i).
3. Can be *excluded* from equation explaining y. That is, instrument affects y only through x.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: the path from cause to effect*. Princeton, New Jersey: Princeton University Press, p. 106.

Example: omitted variable bias (OVB)

- Suppose y = earnings and x = schooling.
- Suppose schooling is correlated with "ability" (in ε_i), so $Cov(x_i,\varepsilon_i) > 0$.
- Need instrument that affects earnings only through schooling.
- Possible instrument: _____.

Example: selection bias

- Suppose y = test score and x = attended charter school (dummy variable).
- Suppose charter-school attendance correlated with "motivation" (in ε_i), so $Cov(x_i,\varepsilon_i) > 0$.
- Need instrument that affects test score only through charter-school attendance.
- Possible instrument: _____ _____.

Example: endogeneity

- Suppose y = quantity of corn supplied and x = price of corn.
- Suppose price negatively correlated with "good weather" (in ε_i), so $Cov(x_i,\varepsilon_i) < 0$.
- Need instrument that affects quantity supplied only through price.
- Possible instrument: _____.

Beware weak instruments

Does instrument really have *causal effect* on x?

- If effect is weak, IV estimates are too biased in small sample to be useful.
- Check first-stage regression:
  $$x_i = \phi_1 + \phi_2 w_{1i} + \phi_3 w_{2i} + \phi_4 x_{3i} + \varepsilon_{1i}.$$
- Rule of thumb: the F-statistic (testing H_0: $\phi_2=\phi_3=\phi_4=0$) should be at least _____.

Beware invalid instruments

Is instrument really *unrelated to any omitted variables* (in ε_i)?

- If instrument not valid, then IV inconsistent.
- If instrument is randomly assigned, no worries, guaranteed valid!
- Otherwise, perhaps check correlation of instrument with included controls.

Conclusions

- The method of instrumental variables (IV) is a technique for estimating equations where error term ε_i is correlated with regressor x_i.
- An "instrument" is a variable that is correlated with x_i but _____ with ε_i.
- IV estimators can be computed in ___ stages.
- If instrument is valid and not weak, IV estimator is consistent even when OLS is not.

SECTION D
STATA LABS

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Frequently-Used Stata Commands

The general form for a Stata command is

[prefix:] command [in …] [if …] [, options]

The following commands are frequently used in Stata *.do* files. Most commands can be abbreviated. For detailed information and examples, in Stata, choose *Help>Stata command…*

Command	**What it does**
`* blah`	Comment at beginning of line, not executed by Stata.
command `// blah`	Comment at end of line, not executed by Stata.
command `///`	Continue command onto next line.
`log`	Create a file of results for all commands that follow.
`clear`	Remove existing data from memory.
`infile`	Read data in free format (separated by spaces in a text file).
`import`	Read data from a spreadsheet.
`use`	Read data from a file already in Stata format (.dta).
`describe`	List variables in the data set and their labels.
`summarize`	Compute descriptive statistics.
`tabulate`	Tabulate a variable or cross-tabulate several variables that each take a limited number of different values.
`histogram`	Draw a histogram for a variable.
`list`	Peek at the data by listing a specified number of observations of specified variables.
`label`	Attach an explanatory label to a variable, or to particular values of a variable.
`generate`	Compute a new variable from existing ones.
`replace`	Modify values of an existing variable.
`drop`	Delete specified unneeded variables or observations (helpful when working with very large data sets).
`keep`	Keep only specified variables or observations.
`sort`	Sort the observations on one or more variables.
`egen`	Compute a new variable using multiple observations of existing variables.
`destring`	Convert alphabetical values to numerical values.
`ttest`	Test whether two subsamples have identical means.
`regress`	Estimate ordinary least-squares regression.
`ivregress`	Estimate instrumental-variables regression.
`test`	Immediately following a regression command (“postestimation”), test the hypothesis that one or more coefficients are zero.
`log close`	Close file of results.
`exit`	End of Stata do file.

Also see *Help>Stata command…>operators* to see how to write expressions like “variable 1 divided by variable 2” or “variable 1 greater than variable 2”.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Tips for Using Stata Efficiently

(1) While using Stata, **close** any other applications and windows that you don't need.

(2) **File extensions** (.xyz) indicate what kind of file it is. Working with Stata is easier if you can see file extensions. For example, when downloading a data extract from IPUMS, you will get a group of files: a compressed data file, a Stata do-file, a codebook file, and so forth. All files in the same IPUMS extract will have the same name, but different file extensions (.dat.gz, .do, .cbk, etc.).

- On a Windows computer, in File Manager, choose *View* tab and check the box for *file name extensions*.
- On a Mac computer, in Finder, choose *Preferences>Advanced>Show all filename extensions*.

(3) Keep all your files for a project in **one folder**. After you start Stata, choose *File>Change working directory…* to point Stata at that folder.

(4) For any complicated work in Stata, create a **do-file** of your Stata commands so you can easily replicate or revise your work later if needed.

(5) Make sure your do-file creates a **log file of results**, preferably in text format that can be read without Stata.

(6) Sprinkle **comments** throughout your do-file, so that (a) you can come back to it later and remember what you were doing and (b) someone else, such as your professor or your classmate, can understand your code. In Stata, * at the beginning of a line indicates that the entire line is a comment, and // in the middle of a line indicates that the remainder of the line is a comment.

(7) When you exit Stata, it will ask you whether you want to save your data. In most cases you should **just say no!** If you save the data, then any variables you "generated" will be added to the Stata dataset, and the next time you run your Stata do-file, you will get an error message. This is because Stata will not allow you to "generate" a variable that already exists.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Topics: Downloading IPUMS data

National Health Interview Survey Lab
Part 1 of 4: Extract NHIS data from IPUMS

IPUMS, a service located at the University of Minnesota, houses and organizes large samples of person-level and household-level data, including the *National Health Interview Survey* (NHIS). Anyone can access ("extract") the data free from IPUMS, though you must register. Let's extract some data.

1. Go to www.ipums.org , choose "IPUMS Health Surveys" and then "NHIS-Get Data" to access data from the National Health Interview Survey. Choose "Create an extract—get data."

2. The NHIS collects both household-level data (H) and person-level data (P). Both kinds of variables are organized in categories. We will extract "person" variables needed to replicate table 1.1 in the book by Angrist and Pischke. Under "Select Variables/Person" add the following variables to your cart.
    - Category: Person>Demographic>Core Demographic . Add the variables AGE, SEX, and MARSTAT to your cart by clicking on the plus (+) signs.
    - Category: Person>Demographic>Ethnicity/Nativity . Add the variable RACENEW to your cart.
    - Category: Person>Demographic>Family interrelationships . Add the variable FAMKIDNO to your cart.
    - Category: Person>Socio-Economic Status>Education . Add the variable MAXEDUC to your cart.
    - Category: Person>Socio-Economic Status>Work . Add the variable EMPSTAT to your cart.
    - Category: Person>General Health>General Health . Add the variable HEALTH to your cart.
    - Category: Person>Health Insurance>General Coverage . Add the variable HINOTCOVE to your cart.

3. Choose "Select Samples." Let's use only the latest available year.
    - Check only the box for **2022**.
    - Click on "Submit Sample Selections."

4. View your data cart. You should have a number of "preselected" variables (YEAR through CSTATFLG) and the above variables selected by you (AGE through HINOTCOVE).

5. Choose "Create Data Extract." In the box labeled "Describe your extract," type something that will help you remember why you created this extract—for example "ECON 190 NHIS LAB EXERCISE."

6. Choose "Submit Extract." You will be asked to sign in with your email address and password. If you do not already have an account at IPUMS, you will be asked to create one and to agree to the terms of use.

7. You have now requested data for only one year. So is your extract a *cross section*, a *time-series*, a *pooled* dataset, or a *panel* (or longitudinal dataset)? _____
   If you had checked boxes for several years, what kind of dataset would you get? _____
   _____

8. Log out of IPUMS.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `summarize`

National Health Interview Survey Lab
Part 2 of 4: Open IPUMS-NHIS data in Stata

Let's save and view the data we extracted from IPUMS-CPS.

1. Go to www.ipums.org , choose "IPUMS Health Surveys" and then "NHIS-Get Data" to access data from the National Health Interview Survey. Choose "Create an extract—get data." Log into your account. Then choose "MY DATA."

2. You will see that IPUMS has created several files for you for download. Create a folder on your computer for this lab. Then save the following files to your computer in that folder. (In Windows, right-click and choose "Save link as...")
   - Download DAT: This is a large fixed-width text file of the data you requested, in compressed (.dat.gz) format to speed downloading.
   - Stata: This do-file (.do) will read the data file.
   - Basic Codebook. This codebook-file (.cbk) contains short descriptions of each variable you requested. You can read it with a text editor (such as Windows Notepad or MacOS TextEdit) or with MS-Word.
   - DDI Codebook. This file contains more detailed descriptions of each variable, but do *not* download it as is. Instead, click on the link to view it and then print or save it as a PDF file.

   After downloading the four files, log out of IPUMS.

3. All your IPUMS files have the same file name but different file *extensions* (.something). Set your computer so that you can see file extensions.
   - Mac users: In the Finder>Preferences menu, select the Advanced tab and check the box "Show all filename extensions."
   - Windows users: In the File Manager window, select the View tab and check the box "File name extensions."

4. You must decompress your data file (.dat.gz) before Stata can read it.
   - Mac users: You should be able to decompress the file by double-clicking on it.
   - Windows users: You will need special software to decompress the file: 7-Zip (free, open source) is recommended. It runs under Windows 10, 8, or 7 and there are versions for 32-bit and 64-bit machines. Get it at 7-zip.org . After you run the installation file, decompress the dat.gz file by right-clicking on it and choosing *7-Zip>Extract here*.

5. Launch Stata. Then choose *File>Change working directory…* to point Stata at the folder where your downloaded files are located.

6. The do-file (.do) provided by IPUMS is a series of Stata commands to read the data. Edit the do-file by choosing *Window>Do-file Editor>New Do-file Editor* . When the Do-file Editor window opens, choose *File>Open* to access the do-file provided by IPUMS. You will see that this do-file contains many commands that tell Stata how to interpret the data in the raw data file (.dat) that you just decompressed.

7. At the end of the do-file, add the following command:

   ```
   summarize
   ```

   Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

8. Look at the summary statistics generated by the "summarize" command in your results window. You should have over 72,000 observations.
   - Why is the standard deviation of YEAR equal to zero? _____
   - What is the average AGE in your sample? _____
   - What is the average FAMKIDNO in your sample? _____

   The other variables are difficult to interpret without consulting the codebooks.

9. Save your do-file. In the Do-file Editor window, choose *File>Save as…* and give it a sensible name.

10. The data are now in memory but need to be saved to a file. Save the data in Stata's own proprietary .dta format, which includes the auxiliary information provided by IPUMS in the do-file. In the big window, choose *File>Save as…* and give it a sensible name. In what follows, I will assume you gave it the name `nhislabdata`.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `summarize`

National Health Interview Survey Lab
Part 3 of 4: Interpret data values using codebook

At first glance, the values of the some of the variables in our extract do not make sense. That is because people give a variety of verbal answers to the NHIS, all of which are coded as numbers even if their answer was not a number. The way those answers are coded can be confusing, but the codebooks can help us.

1. Launch Stata. Then choose *File>Change working directory…* to point Stata at the folder where your do-file (.do) and Stata data file (.dta) are located.

2. Choose *Window>Do-file Editor>New Do-file Editor* . When the new window opens, choose *File>Open* to access the do-file you saved in the last lab. Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

3. Look at the summary statistics generated by the `summarize` command in your results window. Simultaneously, open one of the codebook files.

4. In the codebook file, scroll down to the documentation for SEX. What does it mean if the variable SEX=1? If SEX = 2? _____
   According to the summary statistics, what fraction of the sample is female? What fraction is male?
   _____

5. What does it mean if the variable RACENEW=100? If RACENEW = 400?
   _____

6. What does it mean if the variable MAXEDUC=0? If MAXEDUC=08? If MAXEDUC=11?
   _____

7. What values of EMPSTAT indicate the person is currently working?
   _____

8. What value of HEALTH indicates excellent health? _____
   What value indicates poor health? _____
   What values indicate that the health status question was not asked, or that the respondent did not answer? _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `log, use, describe, summarize, generate, replace, ttest, exit`

National Health Interview Survey Lab
Part 4 of 4: Replicate table 1.1 in Angrist and Pischke

Table 1.1 in Angrist and Pischke (p. 5) compares health status and other variables for people with and without health insurance. Let's write a Stata do-file to compute similar numbers with our data.

Note: our numbers may differ from Angrist and Pischke's for at least four reasons. First, Angrist and Pischke's table 1.1 uses data from 2008, whereas we use more recent data. Second, table 1.1 uses data on husbands and wives, so that the number of females is exactly equal to the number of males in their data—as we saw in the previous part, that is not quite true for our data. Third, we did not collect data on family income. Respondents in the NHIS report their income in broad categories (zero to $35 thousand, $35 thousand to $50 thousand, etc.). Angrist and Pischke use these categories to impute numerical family income using an elaborate procedure (see pp. 249-250) which we will skip. Finally, the NHIS is not a representative sample, so sampling weights must be applied to estimate values for the U.S. population as a whole. Angrist and Pischke apply these weights, but we will skip them.

1. Now let's write a Stata do-file to compute our numbers. In Stata, choose *Window>Do-file Editor>New Do-file Editor* . Every do-file should begin with the following: one or more "*" comments to explain the purpose of the do-file, a `log` command to direct all Stata calculations to a file, a command to open a dataset and clean out any prior data from memory, and commands to check the dataset. So type the following lines into the blank window:

```
* Program to analyze NHIS
* log using log file name, text replace
use nhislabdata.dta, clear
describe
summarize
tab1 hinotcove
```

   The *log file name* should be the same name you plan to give this do-file.

2. In the previous part of this lab, we found that some values of the variables actually indicate that the data are missing. We need to exclude any observations with missing data. There are many ways to do this. I like to generate a special zero-one variable, which I call `ok2use`, to flag useable observations as follows:


```
generate ok2use = age>=26 & age<=59 ///
  & sex<7 ///
  & famkidno<98 ///
  & racenew<900 ///
  & maxeduc>0 & maxeduc<96 ///
  & empstat>0 & empstat<900 ///
  & health<7 ///
  & hinotcove>0 & hinotcove<7
```

3. Next, referring to the codebook, we need to create variables that match (at least approximately) the variables in Angrist and Pischke's table 1.1. We use the following important Stata commands.
   - The `generate` command creates a new variable for the first time.
   - The `replace` command modifies the values of an existing variable.
   - The `if` clause at the end of each command restricts the command to certain observations.

   Type the following lines.

```
generate insured = 1 if hinotcove==1     // has health insurance
  replace insured = 0 if hinotcove==2    // no health insurance
generate healthindex = 6 - health        // 5=excellent, 1=poor
generate nonwhite = racenew>100
generate hsdiploma = maxeduc>2           // has HS diploma
generate bachelors = maxeduc>7           // has bachelors degree
generate employed = 1 if empstat>=100 & empstat<200
  replace employed = 0 if empstat>=200 & empstat<900
```

   Note that everything following `//` is a comment. It is good practice to sprinkle comments everywhere to make your Stata code easier to read.

4. Now let's compute the numbers corresponding to table 1.1. Starting at the top left of the table, compute the mean health status of men with and without health insurance, the difference in the means, and the standard errors. This can be done with a single `ttest` command, which tests for a difference in the means of two samples:

```
ttest healthindex if ok2use & sex==1, by(insured)  // men
```

5. Finally, end your do-file with the following commands. The `clear` command deletes data from memory, avoiding problems the next time the do-file is run. The `log close` command stops writing output to the log file and releases the log file to other applications.

```
clear
* log close
```

6. Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

7. Look at the output from `ttest` in the Stata output window.
   What is the average value of health status for men with insurance? _____
   What is the average value for men without insurance? _____
   Is the difference statistically significant? _____ Why or why not? _____

   _____
   Are the estimates computed by `ttest` similar to the corresponding numbers in Angrist and Pischke's table 1.1? Comment.
   _____
   _____

8. Now insert additional `ttest` commands to compare the health status of women with and without health insurance. Then insert additional `ttest` commands to compare other variables (`nonwhite, hsdiploma, bachelors, employed`) for men and women with and without health insurance. Run your do-file again and compare the results to the corresponding numbers in Angrist and Pischke's table 1.1.
   _____
   _____
   _____

9. If everything looks good, erase the stars in front of the commands `log using`... and `log close` and run the do-file one more time to create a log file of your output. Save your do-file, giving it the same name as *log file name* above (it will be assigned a different extension). Upload your log-file to Blackboard.

10. When exiting Stata, do NOT save your data. Saving your data in memory sounds like a good idea, but it is not and here is why. Your data in memory now includes a slew of new variables, including `insured, healthindex, nonwhite,` etc. If you save your data in memory, all these variables will be added to your .dta file. But then you will get errors the next time you run your do-file, because one cannot "generate" a variable that already exists in the .dta file.

[end of lab]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Topics: Downloading CPS data

Current Population Survey Lab
Part 1 of 6: Extract CPS data from IPUMS

IPUMS, a service located at the University of Minnesota, houses and organizes large samples of person-level and household-level data, including the *Current Population Survey* (CPS). Anyone can access ("extract") the data free from IPUMS, though you must register. Let's extract some data.

1. Go to www.ipums.org . Choose "IPUMS CPS" to access data from the Current Population Survey.

2. Choose "Create an extract—get data."

3. The CPS collects both household-level data (H) and person-level data (P). Both kinds of variables are organized in categories. We will extract "person" variables needed to estimate a Mincer wage equation. Under "Select Variables/Person" add the following variables to your cart.

    - Category: Person>Core>Demographics . Add the variables AGE, SEX, and RACE to your cart by clicking on the plus (+) signs.
    - Category: Person>Core>Work . Add the variable EMPSTAT to your cart.
    - Category: Person>Core>Education . Add the variable EDUC to your cart.
    - Category: Person>Annual Social and Economic Supplement (ASEC)>Work . Add the variables WORKLY, WKSWORK1 and UHRSWORKLY. (Note that variables in this group are only available in the ASEC.)
    - Category: Person> Annual Social and Economic Supplement (ASEC)>Income . Add the variable INCWAGE. (Note that variables in this group are only available in the ASEC.)

4. Choose "Select Samples." We will collect just one sample to keep the exercise simple and minimize the file size.

    - Uncheck "All Default Samples at the top left.
    - Under the ASEC tab, check only the box for **2023.**
    - Make sure everything else is unchecked under both the ASEC tab and the BASIC MONTHLY tab.
    - Click on "Submit Sample Selections."

5. View your data cart. You should have a number of "preselected" variables (YEAR through ASECWT), the above variables selected by you (AGE through INCWAGE), and one sample (ASEC **2023**).

6. Choose "Create Data Extract." In the box labeled "Describe your extract," type something that will help you remember why you created this extract—for example "ECON 190 CPS lab exercise."

7. Choose “Submit Extract.” You will be asked to sign in with your email address and password. If you do not already have an account at IPUMS, you will be asked to create one and to agree to the terms of use.

8. You have now requested data for only one year. So is your extract a cross section, a time-series, a pooled dataset, or a panel (or longitudinal dataset)? _____
   If you had checked boxes for several years, what kind of dataset would you get? _____
   _____

9. Log out of IPUMS.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `summarize`

Current Population Survey Lab
Part 2 of 6: Open IPUMS-CPS data in Stata

Let's save and view the data we extracted from IPUMS-CPS.

1. If you are still viewing your extract page, refresh it. If you are logged out of IPUMS, go to www.ipums.org , choose "IPUMS CPS" and then "CPS-Get Data" to access data from the Current Population Survey. Choose "Create an extract—get data." Log into your account. Then choose "MY DATA."

2. You will see that IPUMS has created several files for you for download. Create a folder on your computer for this lab. Then save the following files to your computer in that folder. (In Windows, right-click and choose "Save link as...")
   - Download DAT: This is a large fixed-width text file of the data you requested, in compressed (.dat.gz) format to speed downloading.
   - STATA: This do-file (.do) will read the data file.
   - Basic Codebook. This codebook-file (.cbk) contains short descriptions of each variable you requested. You can read it with a text editor (such as Windows Notepad or MacOS TextEdit) or with MS-Word.
   - DDI Codebook. This file contains more detailed descriptions of each variable, but do *not* download it as is. Instead, click on the link to view it and then print or save it as a PDF file.

   After downloading the four files, log out of IPUMS.

3. All your IPUMS files have the same file name but different file *extensions* (.something). Set your computer so that you can see file extensions.
   - Mac users: In the Finder>Preferences menu, select the Advanced tab and check the box "Show all filename extensions."
   - Windows users: In the File Manager window, select the View tab and check the box "File name extensions."

4. You must decompress your data file (.dat.gz) before Stata can read it.
   - Mac users: You should be able to decompress the file by double-clicking on it.
   - Windows users: You will need special software to decompress the file: 7-Zip (free, open source) is recommended. It runs under Windows 10, 8, or 7 and there are versions for 32-bit and 64-bit machines. Get it at 7-zip.org . After you run the installation file, decompress the dat.gz file by right-clicking on it and choosing *7-Zip>Extract here*.

5. Launch Stata. Then choose *File>Change working directory…* to point Stata at the folder where your downloaded files are located.

6. The do-file (.do) is a series of Stata commands to read the data. Edit the do-file by choosing *Window>Do-file Editor>New Do-file Editor* . When the Do-file Editor window opens, choose *File>Open* to access the do-file provided by IPUMS. You will see that this do-file contains many commands that tell Stata how to interpret the data in the raw data file (.dat) that you just decompressed.

7. At the end of the do-file, add the following command:

   ```
   summarize
   ```

   Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

8. Look at the summary statistics generated by the "summarize" command in your results window. You should have over 150,000 observations.
   - Why is the standard deviation of YEAR equal to zero? _____
   - What is the average AGE in your sample? _____
   - What is the smallest AGE ? _____
   - What is the largest? _____

   The other variables are difficult to interpret without consulting the codebooks.

9. Save your do-file. In the Do-file Editor window, choose *File>Save as…* and give it a sensible name.

10. The data are now in memory but need to be saved to a file. Save the data in Stata's own .dta format, which includes the auxiliary information provided by IPUMS in the do-file. In the big window, choose *File>Save as…* and give it a sensible name. In what follows, I will assume you gave it the name `cpslabdata` .

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `summarize`

Current Population Survey Lab
Part 3 of 6: Interpret data values using CPS codebook

At first glance, the values of the some of the variables in our extract do not make sense. That is because people give a variety of verbal answers to the CPS, all of which are coded as numbers even if their answer was not a number. The way those answers are coded can be confusing, but the codebooks can help us.

1. Launch Stata. Then choose *File>Change working directory…* to point Stata at the folder where your do-file (.do) and Stata data file (.dta) are located.

2. Choose *Window>Do-file Editor>New Do-file Editor* . When the new window opens, choose *File>Open* to access the do-file you saved in the last lab. Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

3. Look at the summary statistics generated by the `summarize` command in your results window. Simultaneously, open one of the codebook files.

4. In the codebook file, scroll down to the documentation for SEX. What does it mean if the variable SEX=1? If SEX = 2? _____

   According to the summary statistics, what fraction of the sample is female? What fraction is male?
   _____

5. What does it mean if the variable RACE=100? If RACE = 830?
   _____

6. What value of EMPSTAT indicates the person is currently working? What value indicates the person has a job but was not at work last week?
   _____

7. What does it mean if the variable EDUC=125? What value of EDUC indicates a high school diploma but no more schooling?
   _____

8. What does it mean if the variable UHRSWORKLY=999?
   _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `log, use, tabulate, describe, generate, replace, recode, exit`

Current Population Survey Lab
Part 4 of 6: Calculate years of schooling from CPS data

There is no variable in the CPS that represents years of schooling. How can we compute one?

1. Open one of the codebook files and look for the documentation on "EDUC." There are about 36 possible values of EDUC! Fortunately, not all of these values appear in the sample that we have downloaded.

2. Let's see how many different values we actually have in our sample. Launch Stata. Then choose *File>Change working directory…* to point Stata at the folder where your files are located. Choose *Window>Do-file Editor>New Do-file Editor* . When the new window opens, choose *File>Open* to access the do-file you saved in the last lab. At the end of the do-file, below the "summarize" command, add the following commands:

   ```
   tabulate educ
   tabulate educ, nolabel
   ```

   Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file. Now check the results window. It appears there are about 17 possible values of EDUC in our sample. We need to translate only those values into years of schooling. Use the output in Stata's results window to complete the following table.

Value of Schooling	Value(s) of EDUC
Missing	1
0	2
4	10
6	20
8	30
9	40
10	
11	
12	71 or 73
13	
14	or
16	
18	
19	
20	

3. Now let's write Stata code to implement this table. In Stata, choose *Window>Do-file Editor>New Do-file Editor* . When the new window opens, type the following lines into the blank window:

```
* Program to analyze CPS-ASEC
* log using log file name, text replace
use cpslabdata.dta, clear
describe
summarize
tabulate  educ
```

"Log file name" can be any name you want, but to minimize confusion, use the same name that you plan to give this do-file. In what follows, I will assume you chose the name `myprog` .

We will now create a new variable `schooling` from the CPS variable `educ`. There are at least two ways to do this. The **first way** uses the `replace` command repeatedly:

```
generate  schooling = .
replace  schooling = 0 if educ==2
replace  schooling = 4 if educ==10
etc.
replace  schooling = 12 if  educ== 71 | educ== 73
etc.
```

The vertical bar means "or". The **second way** uses the `recode` command just once, but the command is so long that it spills over multiple lines. Be sure to add three slashes whenever you continue to the next line:

```
recode educ  2=0  10=4  etc.  71 73=12  etc.   ///
  , generate(schooling)
```

Choose whichever method you like and finish your program with the following commands.

```
tabulate  schooling
clear
* log close
```

4. How are your Stata skills? Do you know what each command above does? You can look up any of them by choosing *Help>Stata command…*

5. Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file. Now check the results window. Does the tabulation for the variable "schooling" make sense?

6. If all is well, in the Do-file Editor window, choose *File>Save as…* and give it the same name as the *log file name* above (it will be assigned a different extension).

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `histogram`, `regress`

Current Population Survey Lab
Part 5 of 6: Estimate a Mincer wage equation

A Mincer wage equation relates the logarithm of a person's wage (*lwage*) to schooling (*schooling*) *and* labor-market experience (*exper*) as follows:

$$lwage = \beta_1 + \beta_2\, schooling + \beta_3\, exper + \beta_4\, exper^2 ,$$

Let's estimate a Mincer equation on your data extract, using the measure of schooling that you calculated in the last lab.

1. Launch Stata. Then choose *File>Change working directory…* to point Stata at the folder where your files are located. Choose *Window>Do-file Editor>New Do-file Editor* . When the new window opens, choose *File>Open* to access the `myprog.do` file you saved in the last lab. We will now add some statements after `tabulate schooling` and before `log close` .

2. Labor-market experience is not available in the CPS-ASEC. Let's estimate it the same way that many researchers do, as age minus years of schooling minus 6. Write a `generate` command to compute the new variable `exper` .

3. Write another `generate` command to compute `expersq` , the square of `exper` .

4. The hourly wage is not available in the ASEC. Let's estimate it as wage income divided by weeks worked and by hours per week, that is, `incwage/uhrsworkly/wkswork1`. Stata's natural log function is either `log( )` or `ln( )` . Write a `generate` command to compute `lwage` , the natural log of the estimated wage.

5. We need to exclude some of the observations with missing or irrelevant data. There are many ways to do this. I like to generate a special zero-one variable, which I call `ok2use`, to flag useable observations as follows:

```
generate  ok2use =  educ>1 & age>15 & wkswork1>0  ///
  & uhrsworkly<99 & incwage<9999998
```

6. Before running the regression, add these commands to check the data:

```
summarize schooling exper lwage  if ok2use
histogram lwage  if ok2use
```


7. Finally, add the regression command for the Mincer equation:

```
regress lwage schooling exper expersq if ok2use, vce(robust)
```

Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

8. Do the signs of the estimated coefficients make sense? Comment on the sign of each estimated coefficient:

schooling _____

exper _____

expersq _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `test`, `lincom`, `nlcom`

Current Population Survey Lab
Part 6 of 6: Stata "postestimation" commands

Stata has numerous commands to analyze the results of least-squares regression. Let's use a few of these commands to analyze the results of our estimates of the Mincer equation:

$$lwage = \beta_1 + \beta_2\, schooling + \beta_3\, exper + \beta_4\, exper^2 ,$$

1. First, just use the output from `regress` to test the null hypothesis that schooling has no effect on the wage—in other words, the null hypothesis that β_2 equals zero—by completing the following table:

Test H_0: $\beta_2 = 0$	
t-statistic	
p-value	
Can you reject the null hypothesis at 5%?	

2. The coefficient of *schooling* is the increase in *lwage* for a one-unit increase in schooling, commonly referred to as the "return to schooling." Since the dependent variable is in logarithms, that increase in lwage , converted to a percent, is the percent increase in the wage. What is your estimated return to schooling, to the nearest tenth of a percentage point? _____%

3. Next, let's test the null hypothesis that experience has no effect on the wage. This will take more work because experience appears twice in the regression equation. We need to test the joint null hypothesis that both β_3 and β_4 equal zero, an F test.[1] The `regress` command does not automatically compute this test. We need the Stata "postestimation" command `test` . A bit confusingly, `test` refers to coefficients by their corresponding *variable names* (not β_1, β_2, etc.). So to test the joint null hypothesis β_2=0 and β_3=0, insert the following immediately after `regress` :

```
test  (exper=0) (expersq = 0)
```

[1] Alternatively, a chi-square test can be used. The F test and the chi-square test give the same answers if the sample size is large.

Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file. Complete the following table:

Test H_0: $\beta_3 = 0$ and $\beta_4 = 0$	
F-statistic	
p-value	
Can you reject the null hypothesis at 5%?	

4. What is the return to another year of experience? Using the equation at the top of the previous page, find the derivative:

$\frac{d\ lnwage}{d\ exper} =$ _____

Evidently, the return to another year of experience depends on the amount of experience the worker already has. What is the return to experience when a worker has 10 years of experience? Rewrite your previous answer, substituting **10** for *exper* :

$\left.\frac{d\ lnwage}{d\ exper}\right|_{exper=10} =$ _____

Now, you could use a pocket calculator to compute the answer using this formula, but let's make Stata compute the answer and also a standard error. The Stata command `lincom` computes linear combinations of the coefficients from the previous `regress` command. Like `test`, `lincom` refers to coefficients by their corresponding *variable names* (not β_1, β_2, etc.). So insert the following immediately after `test` :

```
lincom  exper + 2*expersq*10
```

Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

5. At what level of experience does the return to experience fall to zero? Set $\frac{d\ lnwage}{d\ exper} = 0$, and solve for *exper* as a function of the β coefficients:

exper = _____

Again, you could use a pocket calculator to compute the answer, but let's make Stata compute the answer and also a standard error. The Stata command `lincom` will not work because this formula is not a linear combination of the coefficients (one coefficient appears in the denominator). We need the more powerful Stata command `nlcom`, which computes nonlinear combinations of the coefficients. Less confusingly but more cumbersomely, `nlcom` refers to the coefficient of *variable* as `_b[`*variable*`]` . So insert the following immediately after `lincom` :


```
nlcom  - _b[exper] / (2* _b[expersq])
```

Choose *Tools>Execute (do)* or click on the corresponding button to run your do-file.

6. If the output looks OK, erase the stars in front of the commands `log using…` and `log close` and run the do-file one more time to create a log file of your output. Save your do-file, giving it the same name as *log file name* above (it will be assigned a different extension). **Upload your log-file to Blackboard.**

7. When exiting Stata, do NOT save your data. Otherwise, you will get errors the next time you run your do-file.

[end of lab]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `regress`
Topics: requirements for an instrument
(see Angrist & Pischke, 2015, p.106)

Instrumental Variables Lab
Part 1 of 2: Finding an instrument

Many people have hypothesized that estimates of the return to schooling, from ordinary least squares (OLS) applied to an equation like

$$lwage = \beta_1 + \beta_2\, schooling + controls,$$

suffer from "ability bias," a type of selection bias or omitted-variable bias. The most able people get more schooling, they hypothesize, so the OLS estimate of schooling coefficient confounds both the true return to more schooling and the return to greater (but unobserved) ability. The implication of this hypothesis is that OLS estimates of the return to schooling, which typically range from 7% to 12%, are biased up.

One way to tackle ability bias is to use the method of instrumental variables (IV). In this lab, we replicate IV estimates originally made by labor economist (and Nobel Prize winner) David Card.

1. From Blackboard, download and save the raw data file `card.raw` and the Stata do-file `ivexample.do`.

2. From Stata, open the do-file, which already contains an `infile` command to read the raw data. Evidently, the data file contains lots of variables, most of which we will ignore. For this exercise, the outcome is a variable called `lwage` and the treatment is called `educ`. What do these variables measure? [Hint: check the `label` statements in the do-file.]
   `lwage` = _____
   `educ` = _____

3. The do-file already contains a `regress` command to estimate the relationship between `lwage` and `educ` using ordinary least squares (OLS), with some control variables. In words, what do these variables control for?
   _____
   _____
   _____

4. Run the do-file. What is the OLS estimated rate of return to schooling? _____

5. Card's idea was to use the variable `nearc4` as an instrument. What does `nearc4` measure?
   `nearc4` = _____
   Why did Card measure `nearc4` in 1966 but measure `wage` and `lwage` in 1976?
   _____
   _____
6. Do you think `nearc4` meets the three requirements of an instrument for `educ` ?
   (i) **Relevance**: Does `nearc4` likely have a causal effect on `educ`? Explain. _____
   _____
   _____
   _____
   (ii) **Independence**: Is `nearc4` likely uncorrelated with unobserved ability? Explain. _____
   _____
   _____
   _____
   (iii) **Exclusion**: Could `nearc4` realistically affect `lwage` through any channel except `educ`?
   Explain. _____
   _____
   _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `regress`, `ivregress`

Instrumental Variables Lab
Part 2 of 2: Estimation

We will now estimate the equation of interest,

$$lwage = \beta_1 + \beta_2\, schooling + controls,$$

using the method of two-stage least squares with `nearc4` as an instrument.

1. Write out the reduced-form equation for `lwage`.

   `lwage` = _____

2. Insert a Stata command to estimate the reduced form equation using `regress` (be sure to include the option for robust standard errors) and rerun the do-file. What is the estimated coefficient of `nearc4`? _____ What is its standard error? _____
   Angrist and Pischke (2015, p. 146) advise that "if you can't see it in the reduced form, it ain't there." Is something there? _____
   _____

3. We will now estimate the return to schooling using the method of two-stage least squares (2SLS), a variant of instrumental variables. Write out the first-stage equation.

   `educ` = _____
   Write out the second-stage equation.

   `lwage` = _____

4. Insert a Stata command to estimate both stages at once using `ivregress`. The syntax is as follows:

   ```
   ivregress 2sls lwage control variables (educ = nearc4),
   first vce(robust)
   ```

   Remember to use `///` if you must continue on the next line. Rerun the do-file.

5. The first stage equation can be used to check the first requirement of an instrument, often called "relevance." In this case, we want to check whether `nearc4` really has a causal effect on `educ`. A popular rule of thumb is that the overall F-statistic should be at least 10. What is the value of the F-statistic for the first stage? F = _____

7. The second stage estimates the parameter of interest—in this case, the return to schooling. According to the second-stage results, what is the estimated rate of return to schooling? _____ What is its standard error? _____

8. Compare the 2SLS estimate of the return to schooling with the OLS estimate from part 1. Which is larger? _____

9. Comment on the hypothesis of ability bias. Does it appear that our OLS estimate was biased *up*, as predicted by this hypothesis? Explain. _____
_____
_____
_____
_____
_____
_____

7. Remove the comment stars (*) from the `log using` and `log close` statements. Rerun your do-file to create a log file of results. Then save your do-file. Upload your log-file to Blackboard.

[end of lab]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `summarize, twoway`

Regression Discontinuity Lab
Part 1 of 3: Look at the data

If treatment changes abruptly according to some rule, we have a *regression discontinuity design* (RDD) which offers the possibility of causal inference. In this lab, we reproduce the results for a sharp RDD model described in Angrist and Pischke (2015, chapter 4, section 4.1). The data are taken from an article by Carpenter and Dobkin (2009) studying the effects of the minimum legal drinking age (MLDA) on death rates in the U.S. In this lab, we will reproduce some of the results of that study as summarized by Angrist and Pischke (table 4.1, figure 4.2, and figure 4.4).

1. From Blackboard, download and save the data file `AEJfigs.dta` and the Stata do-file `rddexample.do`.

2. Using Stata, open the do-file, which already contains a `use` command to read the data and several other commands. Run the do-file and look at the output from the `summarize` command. Evidently, the data set includes many causes of death, any one of which could be an interesting outcome variable to study. In this lab, we will estimate the effect of the legal drinking age on the outcome variable `all`, the death rate from all causes. What is its mean value? _____ deaths per 100,000.

3. The variable `agecell` shows the age at which each death rate is calculated.
   What is the oldest age in the data set? _____ years and _____ months.
   What is the youngest age in the data set? _____ years and _____ months.
   What is the average age in the data set? _____ years and _____ months.
   How many months are evidently included in the data set? _____ months.

4. The `twoway` command generates a `scatter` plot of all deaths against age (the variable `agecell`). Do you see any relationship between age and death rates?
   _____
   _____
   _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `regress, predict, twoway, graph export`

Regression Discontinuity Lab
Part 2 of 3: "Simple" RDD estimation

We will now edit the do-file to estimate a "simple" model of the death rate, using a linear parametric specification. Then we will replicate figure 4.2 (page 150) in Angrist and Pischke (2015). The equation to estimate is

$$outcome = \beta_1 + \beta_2\, treatment\ variable + \beta_3\, running\ variable + error\ term\ .$$[1]

1. What should be our running variable for this RDD? _____
   The coefficient estimates are easier to interpret if the running variable is centered to equal zero at the point of treatment—that is, age 21. Open the do-file. After the `twoway` command, but before the second `clear` command, insert this Stata command:

   ```
   generate  age = agecell – 21
   ```

2. In this data set, the treatment variable does not yet exist. To create it, insert the following Stata command:

   ```
   generate  over21 = (agecell >= 21)
   ```

   Note that the expression in parentheses is a logical expression.[2] If the expression is true, then `over21` equals one. If false, then `over21` equals zero. Thus `over21` is a zero-one dummy variable representing treatment.

3. Now insert a Stata command to regress `all` on the treatment variable and the running variable. Run the do-file. What is your estimated effect of the MLDA? _____ deaths per 100,000. Does your estimate match the one reported in Angrist and Pischke (2015) in table 4.1 on page 160, top of column (1)? _____

4. Let's create a scatter plot with both the original data and the fitted value from the regression. First, immediately after your regression command insert a command to save the fitted values of the regression:

   ```
   predict allfitsimple
   ```

[1] This is similar to equation (4.2) in Angrist and Pischke (2015, p. 152).
[2] The parentheses are not necessary. They are used here simply to make the Stata code easier to read.

5. Then insert a command to create a graph similar to figure 4.2 in the Angrist and Pischke book:

```
twoway (scatter all agecell) ///
  (line allfitsimple agecell if age<0) ///
  (line allfitsimple agecell if age>=0)
```

The `twoway` command can overlay many plots, each enclosed in parentheses. Here, we overlay a scatter plot of the raw data for all ages, a line plot of the predicted values before age 21 years, and a line plot after age 21 years. Run the do-file. Does your graph resemble figure 4.2 (page 150) in Angrist and Pischke (2015)?

6. Your graph should resemble figure 4.2 in substance but not in formatting. Fortunately, Stata has a huge range of graphing options to adjust the formatting of graphs. Remember that options to Stata commands follow a comma. Edit your `twoway` command to add formatting options as follows.[3]

```
twoway (scatter all agecell) ///
  (line allfitsimple agecell if age<0) ///
  (line allfitsimple agecell if age>=0), ///
  legend(off) ytitle("Deaths per 100,000") ylabel(80(5)115) ///
  title("Figure 4.2" ///
  "A sharp RD estimate of MLDA mortality effects")
```

Then run the do-file again. Does the formatting match figure 4.2 now? Can you figure out what all these options do?

7. Add one more command to save your graph. Stata can save graphs in a variety of file formats.[4] Let's use JPEG format, indicated automatically by the .jpg file extension. Add the following command:

```
graph export fig42.jpg, replace
```

Then run the do-file again.

[3] Add your own formatting options! Choose *Help > Stata command ...* and type "twoway options" to see formatting possibilities.

[4] Choose *Help > Stata command ...* and type "graph export" to see possible file formats.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `regress`, `predict`, `twoway`, `graph export`

Regression Discontinuity Lab
Part 3 of 3: "Fancy" RDD estimation

In Section 2 of this lab, we estimated a "simple" RDD model using a simple linear control for age. Now we estimate a "fancy" RDD model that allows the control to be quadratic and to differ before and after treatment. Then we will replicate figure 4.4 (page 158) in Angrist and Pischke (2015).

1. Open the do-file. After the `graph export` command, but before the second `clear` command, create the square of `age`:

```
generate  age2 = age^2
```

2. Next, create interactions between the `over21` dummy variable and the age variables. When these interactions are included in the regression, the coefficients of age and age-squared are allowed to differ before and after the treatment.

```
generate over_age = over21*age
generate over_age2 = over21*age2
```

3. Then regress the death rate from all causes on the treatment variable, age and age-squared, and the interactions:

```
regress all over21 age age2 over_age over_age2
```

Run the do-file. What is your estimated effect of the MLDA? _____ deaths per 100,000. Does your estimate match the one reported in Angrist and Pischke (2015) in table 4.1 on page 160, top of column (2)? _____

4. Immediately after your regression command insert a command to save the fitted values of the regression:

```
predict allfitfancy
```

5. Then insert a command to create a second graph similar to figure 4.4 in the Angrist and Pischke book. This graph needs to overlay a scatter plot of the raw data for all ages, line plots of fitted values of the "simple" model estimated in the last section of this lab, and line plots of fitted values of the "fancy" model just estimated, for a total of 5 plots.


```
twoway (scatter all agecell) ///
  (line allfitsimple allfitfancy agecell if age<0) ///
  (line allfitsimple allfitfancy agecell if age>=0)
```

Run the do-file. Does your graph resemble figure 4.4 (page 158) in Angrist and Pischke (2015)?

6. Your graph should resemble figure 4.4 in substance but not in formatting. Edit your `twoway` command to add formatting options as follows.

```
twoway (scatter all agecell) ///
  (line allfitsimple allfitfancy agecell if age<0, lpattern(dash)) ///
  (line allfitsimple allfitfancy agecell if age>=0, lpattern(dash)), ///
  legend(off) ytitle("Deaths per 100,000") ylabel(80(5)115) ///
  title("Figure 4.4" "Quadratic control in an RD design")
```

Then run the do-file again. Does the formatting match figure 4.4 now? Can you figure out what all these options do?

6. Add one more command to save your second graph in JPEG format:

```
graph export fig44.jpg, replace
```

7. Remove the comment stars (*) from the `log using` and `log close` statements. Rerun your do-file to create a log file of results. Then save your do-file. Upload your log-file to Blackboard.

[end of lab]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Topics: Downloading NLS data

National Longitudinal Surveys Lab
Part 1 of 4: Extract data using NLS "Investigator"

The *National Longitudinal Surveys* are a collection of survey studies sponsored by the U.S. Bureau of Labor Statistics. Each survey follows a group of people roughly the same age, called a "cohort," over time. The survey data are housed at the Center for Human Resources Research at Ohio State University. Anyone can access the data for free through a website called the "NLS Investigator," though you must register. For this lab we will use the **NLSY79**, which follows a cohort of about 12,000 people born between 1957 and 1964 and first interviewed in 1979. Although you can download an entire survey, the data set is enormous so this is inadvisable. It is easier to download only the variables, individuals, and years that you need.

1. Go to www.nlsinfo.org/investigator . Click "Register" in the top right corner of the screen and complete the information. Remember your password. You will be sent an email to confirm your registration. Follow the instructions.

2. Log in using your new username and password. You will be asked to select the study you want to work with. For this lab, choose the NLSY79 (National Longitudinal Survey of Youth 1979).

   To access data from the NLS Investigator, you must check (or "tag") the variables you need, thereby creating a "**tagset**" or list of wanted variables (similar to a "cart" at IPUMS). Like other data websites, the NLS Investigator automatically "tags" certain standard variables.

   The unit of observation in NLS data is the individual person. Data for different years on the same individual are stored as different variables in the same observation—so-called "wide" format. As you might expect, this arrangement yields an enormous number of variables for each individual. To get the years you want, you must search through many similar variables for different years. For this exercise, we will search the NLSY79 for variables for the years **1985 through 1993.**

   Let's find some data to estimate the effect of education, experience, and union membership on pay. The NLSY79 has remarkably detailed data regarding work experience. Information about all jobs held during the previous year is recorded. For this exercise, we will use data only on "**Job #1**," which is the current or most recent job. Note that in variable titles, the letter "R" stands for "Respondent," the individual person.

3. Choose the tab "Variable Search." The second set of tabs shows several tools for finding variables in the NLS—we will start with the "Browse Index" tool.

- Let's find education level. In the "Index of Selected Variables," choose *Education, Training & Achievement Scores>Education>Summary measures>All schools>By year>Highest grade completed.* Check the variable R1890901 "Highest grade completed as of May 1 Survey year (revised)" in 1985. (We will assume for this lab that "highest grade completed" does not change over time—a longer research project should verify this.)
- Notice, by the way, that when you hover your mouse over a variable name, a handy box appears with information from the codebook, including a tabulation of values.
- Let's find hourly wage. In the "Index of Selected Variables," choose *Employment>Summary measures>By job>Hourly wages.* Then click the header "VARIABLE TITLE" to sort the results. Check all nine variables with title "Hourly rate of pay job #01" for the years 1985 through 1993.

4. Let's also get the age of the respondent, this time using the "Browse Index with Search" tool.
   - Under "Create search criteria below," remove the gray line boxes referring to "Hourly wages."
   - In the remaining white line of boxes, choose "Word in Title (enter search term)," "contains," and in the empty box type at right, type "age of R at interview date". Click "Display Variables." Then click the header "VARIABLE TITLE" to sort the results.
   - Check all nine variables with title "Age of R at interview date" for the years 1985 through 1993.

5. Finally, let's also get union status, this time using the third tool, "Search."
   - Pick "Word in Title (pick from list)," "is," "COLLECTIVE." Click "Display Variables." Then click the header "VARIABLE TITLE" to sort the results.
   - Check all nine variables with the title "Wages set by collective bargaining? Job #01" for the years 1985 through 1993.

6. Let's see what variables are now in our "tagset" (cart). Choose "Review Selected Variables" in the top row of tabs. You should see four default variables ("Identification code" through "Sex of R") and the 28 variables we have tagged (checked). If anything is missing, redo the previous steps.

7. Now choose the tab "Save/Download" in the top row of tabs.
   - Choose "Save Tagset" in the second row of tabs to store your list of desired variables (similar to a "cart" at IPUMS). Click the radio button for "Save on our server" and the radio button for "By Rnum". Then give it a sensible "Filename" and click "Save."
   - Now choose "Basic Download" in the second row of tabs and click "Download." The NLS Investigator will create a compressed zip folder of your data and related files. It might take a few seconds.
   - Choose "Manage Downloads" in the second row of tabs and click "Download" to download and save the zipped folder. (If you forgot to give it a name, it will be called "default.zip.") Move the zipped folder to a convenient folder on your computer.
   - Click "Logout" in the upper-right corner of the NLS Investigator web page.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `log`, `infile`, `describe`, `rename`,

National Longitudinal Surveys Lab
Part 2 of 4: Open NLS data in Stata

The folder you downloaded from NLS Investigator is in a standard compression format, so you will not need any special software to unzip it. If you click on the downloaded zip folder, you will see that it contains many files. Which ones do you need? Many of the files have the same file name but different file *extensions* (.something). Set your computer so that you can see file extensions.

- Mac users: In the Finder>Preferences menu, select the Advanced tab and check the box "Show all filename extensions."
- Windows users: In the File Manager window, select the View tab and check the box "File name extensions."

The data are in three files which contain the same information in different formats as indicated by their extensions: free-form (.dat) or spreadsheet (.csv) or "Stata dictionary" (.dct). You could read the data into Stata from any one of these files, but the "Stata dictionary" is probably the easiest.

1. Create a folder on your computer for this lab. Copy the .dct file to that folder.

2. Open Stata. Choose *File>Change working directory...* and make sure Stata is pointing to the folder with the .dct file.

3. Choose *Window>Do-file editor>New do-file editor.* Write the following simple program to check your download:

```
* Program to analyze NLSY79
* log using  your program file name , text replace
infile using  your data file name.dct, clear
describe
clear
* log close
```

To minimize confusion, the log file should be given the same name that you plan to give this do-file. Run the do-file. The command `describe` should report in the results window that you have 32 variables and over 12,000 observations. Recall that each observation is an individual person—that is, the data are in "wide" format.

4. The output from `describe` also shows the names of the variables, which all consist of R followed by 7 digits, and the variable labels (which are more informative). Let's give better names to the variables we will use. Be sure to include the year at the end of the new name of any variable observed in multiple years—Stata will use this information later. So insert the following statements in your do-file after `describe` and before `clear`:

```
rename R0000100 personid
rename R0214800 sex
rename R1810710 wage85
rename R1811000 union85
rename R1890901 schooling
rename R1891010 age85
(etc.)
rename R4418700 age93
rename R4200000 union93
rename R4416900 wage93
```

   To speed this tedious coding task and reduce errors, I recommend you copy the output from "describe" directly into your do-file, and then edit it.

5. Add a `summarize` command after the last `rename` and rerun the do-file.

6. Do the descriptive statistics make sense? Consult the codebook (.cdb) that was included in the zip file. Why are there so many negative values, as revealed by the "Min" column?

   _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `foreach`, `sort`, `generate`, `regress`

National Longitudinal Surveys Lab
Part 3 of 4: Estimate the effect of unionism on wages using cross-section data

Let's clean up the data a bit and estimate a Mincer wage equation with unionism.

1. Last time, we discovered that many of our variables have negative values. According to the codebook, these indicate data that are missing for some reason. To prevent Stata from using these negative values, we need to change them to Stata's own code for missing values, a decimal point. This could be done with a long series of "replace" commands, one for each variable, but there is an easier way using a `foreach` loop. Insert the following command in your do-file after the `summarize` command:

```
foreach x of varlist sex schooling wage* union* age*{
  replace `x' = .  if `x'<0
  }
```

   In a Stata `foreach` loop, any statements between the curly braces (here there is only one) are executed for each of the variables following `varlist`. The expression `wage*` stands for all variables beginning with `wage`, and similarly for `union*` and `age*`. Note that in the second line, `x` is bounded on the right by a single quote and on the left by a *left single quote*, which is usually found in the upper-left corner of a computer keyboard.

   After the `foreach` loop, insert another `summarize` command to check the results. Are the negative values gone from the "Min" column? _____

2. Suppose we are interested in the effect of unions on wages. Let's compare the average wages of union and nonunion workers in, say, the year 1985. At the same time, let's check for **balance** between the two groups of workers. Add the following commands and rerun your do-file.

```
sort union85
by union85: summarize wage85 sex schooling age85
```

   Are the average wages similar? _____
   But are union and nonunion workers similar in other respects? _____

3. Multivariate regression provides one way of controlling for differences between groups. Recall that a Mincer wage equation relates a person's log wage to schooling and labor-market experience. Let's estimate a variation on a Mincer equation that includes the effect of union coverage, a zero-one variable:

$$\ln(\text{wage}) = \beta_1 + \beta_1 \text{schooling} + \beta_3 \text{experience} + \beta_4 \text{experience}^2 + \beta_5 \text{union} .$$

Let's estimate this equation using *data for 1985 only*, a cross-section regression. In your do-file, `generate` the variables `lwage85`, `exper85`, and `expersq85`. Then with insert a `regress` command and run the do-file.

Do the estimated values of the coefficients make sense? _____

Which ones are statistically significant? _____

4. Stata's default standard errors are not robust to heteroskedasticity. For robust standard errors, add a comma and the option `vce(robust)` to the `regress` command and rerun the do-file. Do the standard errors change? _____

5. Rerun your do-file. Examine the regression results and think about whether the estimated coefficient of `union` has a **causal interpretation**.
What is the treatment group? _____
What is the control group? _____
Do you believe the groups are alike, on average (aside from differences in education and experience, which we controlled for)? Or do you suspect people who join unions might command higher wages even if they were not union members? _____
_____
_____
If the latter, we face a problem of **selection bias** in estimating the effects of unionism on wages.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `drop`, `reshape`, `list`, `summarize`, `generate`, `xtset`, `xttab`, `xtreg`

National Longitudinal Surveys Lab
Part 4 of 4: Estimate the effect of unionism on wages using longitudinal data

One possible way to eliminate selection bias is to track the *same individuals* over time, before and after treatment. National Longitudinal Survey data are well-suited for this approach because the same individual is tracked for many years, as indicated by the adjective "**longitudinal**." To measure the effect of unionism on wages, we can compare the wages of the *same worker* before and after a change of union status. This is easiest in Stata if we first rearrange the data set from wide format to so-called **long format** where each observation is a person-year rather than simply a person.

1. First, drop the variables created earlier for cross-section regression. In your do-file, before the `clear` command, insert the following:

   ```
   drop lwage85 exper85 expersq85
   ```

   Insert the following powerful Stata command to simultaneously rearrange the data into long format, strip the years from the variable names `union`, `age`, and `wage`, and create a new time variable to be called `year`:

   ```
   reshape long union age wage, i(personid) j(year)
   ```

   Let's check whether the `reshape` command actually worked as intended. Insert a command to list the first 30 or so lines of data and another command to compute summary statistics:

   ```
   list in 1/30
   summarize
   ```

   Rerun your do-file. The `list` command should display all the observations for person #1, followed by the observations for person #2, etc. The `summarize` command should show that you have nearly nine times as many observations as before, because there are potentially nine years of data (1985-1993) for each person.

2. Insert more commands to generate the new variables we need for panel regression:

   ```
   generate lwage = ln(wage)
   generate exper = age - schooling - 6
   generate expersq = exper * exper
   ```


3. Now our longitudinal approach will only work if some individuals in our data actually did switch from union jobs to nonunion jobs or vice versa. We can check this by inserting the following commands.

```
xtset personid year
xttab union
```

In Stata, `xt` at the beginning of a command stands for "cross-section time-series"—that is, longitudinal or panel data. The `xtset` command tells Stata which variable denotes the individual and which variable denotes the time period. It must precede all other Stata `xt` commands.

The `xttab` command asks for a tabulation of the variable `union` showing its variation both **between** individuals and **within** individuals (over time). Rerun your do-file. Examine the output from the `xttab` command, focusing on the column headed "Between/Freq." It should show that roughly 10,000 individuals had a least one year with a nonunion job (union=0), and roughly 4000 individuals had at least one year with a union job (union=1). But since n equals only about 11,000 individuals, as shown at the bottom of the table, there must have been several thousand individuals in both categories—that is, individuals who actually switched union status. Using a calculator, compute the exact number of such individuals: switchers = _____.

4. If we have multiple observations on each person, we can control for all differences across people (that do not change over time) by introducing dummy variables for each person. However, it would be impractical to write a `regress` command with 10,000 or so dummy variables (and the computer memory requirement would be prohibitive). Fortunately, there is a practical alternative. Panel regression with so-called "**fixed effects**" is equivalent to including dummy variables for each individual person. Insert the Stata command for panel regression with fixed effects:

```
xtreg lwage  exper expersq union, fe vce(robust)
```

The option `fe` specifies fixed effects. The option `vce(robust)` requests that standard errors be computed using a formula robust to serial correlation of errors (within each individual), and to heteroskedasticity (of any kind). These are sometimes called "cluster standard errors" or more precisely "standard errors clustered on individuals."

Note that the variable `schooling` is omitted from this regression equation—coefficients of variables that vary across individuals but not over time cannot be estimated because those variables are perfectly collinear with individual dummies (implicit in fixed effects). If you insist on including `schooling` as a regressor, Stata will refuse to compute its coefficient.

5. Rerun your do-file. Compare the estimated coefficient of `union` with the coefficient of `union85` in the cross-section regression for 1985 as estimated in part 3 of this lab.

Which is larger? _____

Do the results suggest **positive or negative selection bias** in the cross-section regression? Put differently, would people working at union jobs command higher or lower wages than average, even if they were not working at union jobs? _____
_____

6. Remove the comment stars (*) from the `log using` and `log close` statements. Rerun your do-file to create a log file of results. Then save your do-file. Upload your log-file to Blackboard.

[end of lab]

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `summarize, tab1, describe`

Consumer Expenditure Surveys Lab
Part 1 of 4: Download data from BLS-CE

The *Consumer Expenditure Surveys* (CE), conducted by the U.S. Bureau of Labor Statistics, survey American "consumer units" (CUs, or households) on their spending. The CE actually consists of *two* quarterly surveys—an *Interview Survey*, in which respondents are interviewed about spending on major items over the last quarter, and a *Diary Survey*, in which respondents are asked to record all spending over a week for two weeks in succession.[1] Both surveys also collect a large amount of data on household characteristics, including various forms of income, family size and configuration, and education.

The data collected by the CE have a variety of uses. In particular, they are used to weight the prices in the monthly Consumer Price Index. They are also published, cross-tabulated by household characteristics like age, income, education, and family size, in detailed tables available at https://www.bls.gov/cex/tables.htm .

Finally, the underlying data are used to create "Public Use Micro Data" (PUMD) files, available for free download. The PUMD files are slightly redacted through top-coding and similar edits so that individual respondents cannot be identified, but they are still useful for research. Let's download a recent year of PUMD files from the Diary Survey.

1. Go to https://www.bls.gov/cex/pumd_data.htm . (Note the underscore between "pumd" and "data".) Choose the "STATA" tab. In the row for "**2022**," click on "Diary (zip)" and download the zipped (compressed) folder to your computer.

2. Just above, download "Dictionary for Interview and Diary Surveys (XLSX)". This file lists the variables available in the CE Survey (though not every variable is available in every year). Save it for future reference. Then close your browser.

3. Unzip the Diary data folder. In Windows, right-click on it and choose "Extract all". On a Mac, double click on it. You should see 20 files ending with the extension " .dta" which indicates a Stata data file (set your computer so that you can see file extensions). The first three letters of each file name indicate its contents (see the Dictionary XLSX file for more information). The fourth letter is "D" for diary. The last three numbers of each file name indicate the year and quarter.

[1] The Interview Surveys have been shown to be more accurate, but the Diary Surveys are easier to use, so we will use a Diary Survey for this lab.

4. Let's inspect the FMLD file for the first quarter of 2022: fmld221.dta . Double-click on the unzipped version of fmld221.dta . Stata should start and open the file. Stata's "Properties" window should show that there are over 300 variables and nearly 3000 observations. A list of variable names should appear in Stata's "Variables" window, but unfortunately there are no labels describing the variables. In Stata's "Command" window, type

   `summarize` FINCBEFX FOODHOME ALCBEV FRSHVEG

   Note that CE variables are all caps in the .dta file we downloaded. Since Stata is case-sensitive, the CE variables must be written all caps in Stata commands. Check the Dictionary XLSX file, "Variables" tab, to find out what these CE variables stand for. Note that income data in the CE are annual but spending data are weekly (all in dollars). Do the mean values look reasonable?

   _____

5. For continuous variables such as those above, summary statistics are helpful, but for discrete variables, a tabulation can be more informative. In Stata's "Command" window type

   `tab1 FAM_SIZE VEHQ CUTENURE HIGH_EDU`

   Check the Dictionary XLSX file, "variables" tab, to find out what these variables stand for. Do the tabulated values look reasonable?

   _____

6. Some discrete variables in the CE are stored as numbers ("double" or "float"), while others are stored as text ("str" or string) even if they appear to be numbers. Find out which are which by typing the following in Stata's "Command" window:

   `describe FAM_SIZE VEHQ CUTENURE HIGH_EDU`

   Stata can tabulate any variable, but it can only do calculations (like `summarize`) with numbers. To see this, type the following in Stata's "Command" window:

   `summarize FAM_SIZE VEHQ CUTENURE HIGH_EDU`

   The output should report zero observations for the two string variables. But the data are obviously not missing—they are just in string format so `summarize` cannot process them.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `log, clear, append, summarize, log, exit, generate, sort, egen`

Consumer Expenditure Surveys Lab
Part 2 of 4: Estimate income elasticities

We will now write a Stata do-file to compute the average income elasticity of demand for food, using the FMLD files.

1. To get a bigger sample, we will combine four quarters of data into a single file using Stata's `append` command, which is used to tack on more observations (not more variables). In Stata, choose *Window>Do-file Editor>New Do-file Editor* . When the new window opens, type the following lines into the blank window:

```
* Analyze Consumer Expenditure Diary Survey PUMD
* log using log file name, text replace
clear
append using fmld221.dta fmld212.dta fmld213.dta fmld214.dta
summarize FINCBEFX FOODHOME ALCBEV FRSHVEG
clear
* log close
```

   As usual, "log file name" can be any name you want, but to minimize confusion, use the same name that you plan to give this do-file when you save it.

2. Point Stata at the folder with the unzipped data files (File>Change working directory…). Then run the do-file. `summarize` should now report about four times as many observations as before, but the mean values of the variables should be roughly similar.

3. An interesting feature of the Diary data files is that they contain two (weekly) observations for every consumer unit (CU). Repeated observations on the same CU are surely serially correlated, so regression standard errors computed without taking this into account would be biased. One solution to this problem is to combine serially-correlated observations by computing average weekly spending over the two weeks for each CU.

   As you know, Stata's `generate` command can create new variables from old ones, but it operates only within observations. To make calculations across multiple observations, we need Stata's `egen` (for "extension to generate") command. To compute average weekly food expenditures for each consumer unit, insert the following commands after `summarize` and before `clear`:


```
sort CUID
by CUID:  egen avgfoodhome = mean(FOODHOME)
summarize FOODHOME avgfoodhome
```

The `sort` command is necessary because the `by` prefix requires that the data set be sorted first. Re-run the do-file. The mean of `avgfoodhome` should be identical to the mean of `foodhome`, but with smaller standard deviation and smaller maximum.

4. Now let's regress spending on income and family size. First let's put all variables except education in logarithms, so that coefficients can be interpreted as elasticities. This will create missing values for any observations with zero or negative values, but hopefully there will be few of these. Insert this command after `summarize` and before `clear`:

```
generate  lincome = ln(FINCBEFX)
generate  lfamsize = ln(FAM_SIZE)
generate  lfood = ln(avgfoodhome)
```

Second, to avoid duplication, we will only use the first week (not both weeks) from each consumer unit. The variable `WEEKI` takes the values 1 or 2 to indicate the week (according to the Dictionary XLSX file, "variables" tab). Let's create a flag variable based on `WEEKI` to ensure that we will only use week 1. Now the Stata command `describe WEEKI` reveals that its type is actually "str2", not numeric. So we must use Stata's `real` function, which converts string variables to numbers, as follows:

```
generate ok2use =  real(WEEKI)==1
```

Note that a double equal sign is used for testing equality, whereas a single equal sign assigns values to a variable in `generate` or `replace` statements.

Third, insert the regression command after the `generate` commands and before `clear`:

```
regress lfood lincome lfamsize if ok2use, vce(robust)
```

Re-run the do-file. Now if all households face the same prices, then spending on food is proportional to the quantity they buy, and so the coefficient of log-income is the income elasticity of demand for food. What is the estimated income elasticity? _____
What is the estimated elasticity of demand for food with respect to family size? _____
Do these values make sense? _____
Are the estimates significantly different from zero? _____

5. Suppose we suspect that education affects household preferences and therefore spending on food. We want to add one more regressor, `HIGH_EDU` . Although this variable appears to take numerical values, the Stata command `describe HIGH_EDU` reveals that its type is actually "str2". Stata can tabulate string variables, but it cannot use them in a regression. So insert these commands after the first `regress` command and before `clear`:


```
generate schooling = real(HIGH_EDU)
regress lfood lincome lfamsize schooling if ok2use, vce(robust)
```

Re-run the do-file. Did the coefficient estimates change much? _____
Does education have a positive or negative effect on spending on food? _____
Is it statistically significant? _____

6. If you have time, use `generate` and `regress` to estimate the income elasticity of demand for
   - alcoholic beverages (`ALCBEV`) _____
   - fresh vegetables (`FRSHVEG`) _____

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `describe`, `tab1`, `generate`, `summarize`, `regress`

Consumer Expenditure Surveys Lab
Part 3 of 4: Linear probability model

Suppose we have a binary, yes-no dependent variable, such as whether the person is a homeowner. How can we estimate and interpret an econometric model explaining such a variable?

1. The CE Diary Survey has a variable `CUTENURE` indicating homeownership. In your Stata do-file, insert the following commands just before the last `clear`:

```
describe CUTENURE
tab1 CUTENURE
```

   Then re-run your do-file. You should see that `cutenure` is a string ("str2") variable that takes six different code values, with 1, 2, and 4 being the most frequent. Check the Dictionary XLSX file, "codes" tab, to find out what these six values mean.

Value of `CUTENURE`	Meaning
1	
2	
3	
4	
5	
6	

2. Now let's generate a new binary variable equal to 1 if the consumer unit (CU) owns their home, and equal to 0 if not. Insert the following commands just before `clear`:

```
generate ownhome = real(CUTENURE)<4
summarize ownhome
```

   Re-run the do-file. Note that the mean of `ownhome` equals the fraction of CUs that own their homes. What fraction of the CUs in our sample own their homes? Is this value plausible?

A zero-one random variable is called a **Bernoulli random variable**. Let Y be a Bernoulli random variable that equals 1 with probability p and equals zero with probability 1-p. Then using the definitions of mean and variance, we have

$$E(Y) = p\ 1 + (1-p)\ 0 = p \qquad Var(Y) = p\ (1-p)^2 + (1-p)\ (0-p)^2 = p\ (1-p)\ .$$

Now the `summarize` command reports both the sample mean and the standard deviation, which is $\sqrt{Var}$. Verify with a calculator that the standard deviation of `ownhome` is in fact equal to , $\sqrt{mean\ (1 - mean)}$.

3. Let's estimate a linear regression equation to explain `ownhome` . Insert the following command just before `clear`:

```
regress ownhome lincome lfamsize if ok2use, vce(robust)
```

Re-run the do-file.

What does this linear regression model mean, given that `ownhome` is a zero-one variable? The mean of the dependent variable is the fraction of CUs that own their own home, or alternatively, the probability that a particular CU owns their own home. So a predicted value of the model is a predicted *probability* that a CU owns their own home, for any given values of `lincome` and `lfamsize` . Because the regression function is linear, this is called a **linear probability model,** and each coefficient gives the change in the probability of a CU owning their own home, for a one-unit change in the regressor.

Now in our model, the regressors are in logarithms, and changes in logarithms are approximately equal to percent changes in the underlying variable. So let's reinterpret the coefficients in terms of percent changes in the underlying variables. For example, suppose income increases by 10%. Then `lincome` (its logarithm) increases by 0.10. So the probability of homeownership increases by 0.10 times the coefficient of `lincome` . How much would that be, according to your estimates?

Increase in probability of homeownership for a 10% increase in income = _____.

ECON 190 – Seminar: Inequality
Drake University, Spring 2024
William M. Boal

Stata commands: `predict`, `count`, `logit`, `probit`

Consumer Expenditure Surveys Lab
Part 4 of 4: Logit and probit models

When modeling a zero-one variable y_i, linear probability models like

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

are attractive because they are simple and easy to interpret. However, they have one awkward feature. The predicted (or fitted) values are often greater than one or less than zero, values which are nonsensical for a probability.

1. Let's see if our linear probability model produced any out-of-bounds predictions. Stata's `predict` command computes and names the predicted values from the most recent `regress` command.[2] Let's call our predicted value variable `predlpm`. Insert the following commands immediately after the last `regress` command:

```
predict predlpm
count if predlpm > 1 & predlpm < .
count if predlpm < 0
```

(A dot represents a missing value, which Stata represents internally as an extremely large number. Since we don't want to include missing values, we ask Stata to count only values less than the dot.) Re-run the do-file. How many out-of-bounds predictions were there?
Number of predicted values > 1: _____ Number of predicted values < 0: _____.

Usually, interest lies in the coefficients, not the predicted values, but if sensible predicted values are important to our project, then we need an alternative model where the predicted values are constrained to lie between zero and one. The usual approach is to transform the right-hand side of the regression equation using an upward-sloping function F(z) that is bounded between zero and one, and then to estimate

$$y_i = F(z) + \varepsilon_i \ , \quad \text{where } z_i = \beta_1 + \beta_2 x_i$$

The F(z) function should be some kind of S-curve that bends z to lie between zero and one.

There are two popular choices for a F(z) function: the logistic function $F(z) = \frac{exp(z)}{1+exp(z)}$ and the standard normal cumulative distribution function $F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-w^2}{2\sigma^2}\right) dw$. When the

[2] The `predict` command is thus what Stata calls a "postestimation" command.

logistic function is used for F(z), the model with the binary y-variable is called a **logit model**. When the standard normal cumulative distribution function is used for F(z), the model is called a **probit model**.

The advantage of logit and probit models is that the predicted values are *always* greater than zero and less than one, as probabilities should be, but there are some complications. First, because logit and probit models are nonlinear, there are no simple formulas for the coefficient estimates. Instead, estimation proceeds essentially by computerized trial-and-error. If there are many β coefficients to estimate, it may take noticeably longer to compute the estimates than with a linear probability model.

Second, the marginal effect of a one-unit change in x on y is more complicated than in the linear probability model, and it depends on z. Using the chain rule for differentiation, $\frac{dy}{dx} = F'(z)\,\beta_2$. But all these complications are automated in Stata and other econometric software, so for the user, estimating a logit or probit model is almost as easy as estimating a linear probability model.

2. Let's estimate a **logit model** for home ownership. In your Stata do-file, insert these commands just before the last `clear`:

```
logit ownhome lincome lfamsize if ok2use
margins, dydx(lincome lfamsize) atmeans
predict predlogit
count if predlogit > 1 & predlogit < .
count if predlogit < 0
```

   The `logit` command estimates the logit model. The `margins` command computes $F'(z)\,\beta_2$ for each regressor, using the value of z that corresponds to the sample means of `lincome` and `lfamsize`. The `predict` and `count` commands will (hopefully) demonstrate that the logit model never produces out-of-bounds predictions. Run the do-file again to check for errors.

3. Let's also estimate a **probit model** for home ownership. In your Stata do-file, insert these commands just before the last `clear`:

```
probit ownhome lincome lfamsize if ok2use
margins, dydx(lincome lfamsize) atmeans
predict predprobit
count if predprobit > 1 & predprobit < .
count if predprobit < 0
```

   The `probit` command estimates the probit model. The `margins` command computes $F'(z)\,\beta_2$ for each regressor, using the value of z that corresponds to the sample means of `lincome` and `lfamsize`. Run the do-file again.

4. Now let's compare the estimates for the linear probability model of the previous part of this lab with these new logit and probit estimates. As you can see, the three sets of coefficient estimates are quite different from each other. But this is no surprise—they are not comparable because of the

F(z) functions. The linear probability model coefficient estimates represent the estimated marginal effects of one-unit changes in the regressors, but the logit and probit coefficient estimates do not.

To compare models fairly, we should instead compare the linear probability model's coefficients with the estimated **conditional marginal effects** for the logit and probit models as computed by the `margins` command. These are usually quite close.

Complete the following table to summarize your results. Put the coefficients or marginal effects at the top of each box and put the standard errors of the coefficients or marginal effects underneath in parentheses (as you would in a research paper). Also compute the implied increase in the probability of homeownership for a 10% increase in income.

	Linear probability model	Logit model	Probit model
	Estimated coefficients	Estimated marginal effects	Estimated marginal effects
`lincome`	( )	( )	( )
`lfamsize`	( )	( )	( )
Increase in probability of homeownership for a 10% increase in income	%	%	%

5. Remove the comment stars (*) from the `log using` and `log close` commands. Rerun your do-file to create a log file of results. Then save your do-file. Upload your log-file to Blackboard.

[end of lab]

