Graduate Econometrics Recitations: 2021-2022

Alex Houtz[*]

March 25, 2023

[*] Graduate Teaching Assistant | University of Notre Dame

Contents

Preface

This book is a compilation of recitations given throughout the 2021-22 academic year at the University of Notre Dame for the first-year PhD econometrics sequence in the Department of Economics. They were originally composed by Alex Houtz, who was the graduate teaching assistant for Drew Creal in Fall 2021 and Marinho Bertanha in Spring 2022. The materials within may be distributed at any time to any audience for individual or instructional use, given proper citation to the author. These materials may not be used for commercial purposes.

Many problems were taken or adapted from Bruce Hansen's two econometrics books, which can be found here. The first book is referred to as simply "Hansen" while the second book is referred to as "Hansen II." Other textbooks that are drawn from include Wooldridge (found here) and Hayashi (found here). The remaining problems are either cited in-text or are taken from the homework given in the lecture class by either Drew Creal or Marinho Bertanha.

The following list documents the chain of teaching assistants at the University of Notre Dame that used and/or edited this book, from earliest to latest:

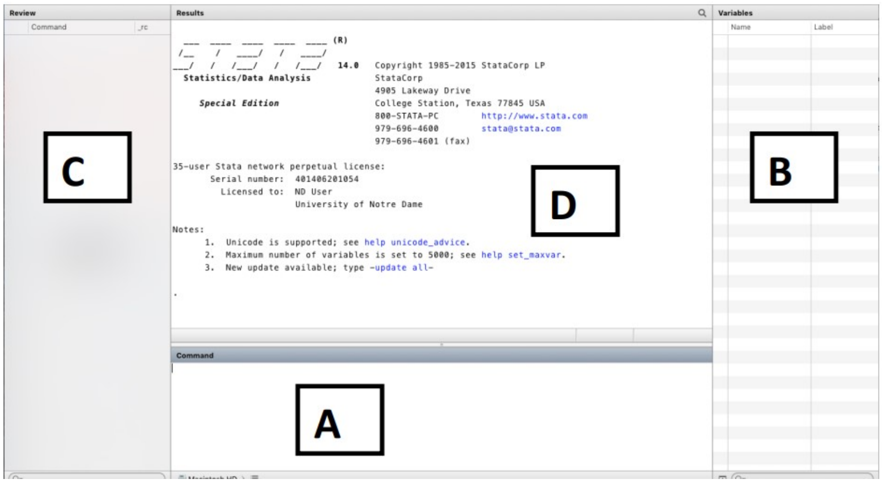
1. Alex Houtz (2021-2022)

Chapter 1

Stata and Linear Algebra Review

1.1 Stata Layout

After learning Matlab last semester, Stata should be fairly intuitive. Here is Stata's interface:



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   14.0   Copyright 1985-2015 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

35-user Stata network perpetual license:
       Serial number:  401406201054
         Licensed to:  ND User
                       University of Notre Dame

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

.
```

A denotes the command window - where you will input code. **B** is the list of variables saved in the program. **C** contains the history of executed code. **D** outputs the results of your code.

1.2 Establishing a Working Directory

Similarly to Matlab, we need to tell Stata where to find our data. Conveniently, we use the same idea as in Matlab (using my directory as an example):

cd "C:\Users\alexh\OneDrive\Documents\Notre Dame\Second Year
\Metrics TA\Spring\Recitations\Recitation 1"

Note that because my directory has spaces in the name I must put quotation marks around the directory. If there are no spaces, quotation marks are not necessary.

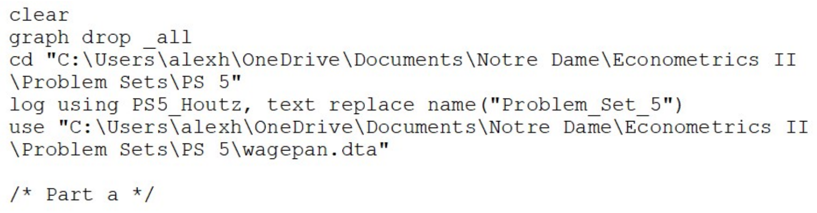
1.3 Constructing .log and .do Files

The .do file is a file that contains the commands you want Stata to run. By compiling a .do file, we can (1) save the work we have done and (2) ensure that our code runs without errors. To start a .do file, go to File→New→Do. Then copy and and paste your history into the .do file. I recommend having coding issues sorted out up to the point saved in the file. After saving the .do file, you can then run it in Stata by typing "do DO FILE NAME" if your saved .do file is in your directory.

Log files save the output and code you run while the log is opened. To open a log, type:

log using YOUR FILE, text replace name("RESEARCH")

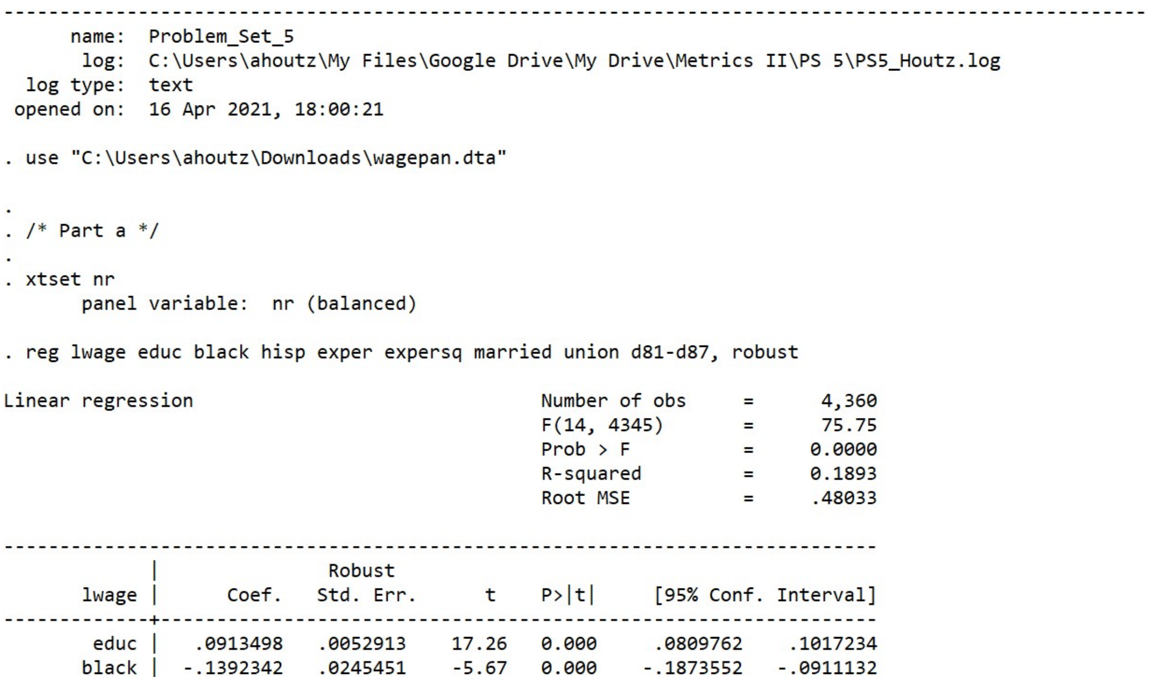
where YOUR FILE is the name you want your .log file to save as and "RESEARCH" is the label you want at the top of your .log file. Let's look at a .do file I wrote:



```
clear
graph drop _all
cd "C:\Users\alexh\OneDrive\Documents\Notre Dame\Econometrics II
\Problem Sets\PS 5"
log using PS5_Houtz, text replace name("Problem_Set_5")
use "C:\Users\alexh\OneDrive\Documents\Notre Dame\Econometrics II
\Problem Sets\PS 5\wagepan.dta"

/* Part a */
```

Note how I establish a directory then open my .log file. The .log file associated with this code is:



```
----------------------------------------------------------------------------------------------------
      name:  Problem_Set_5
       log:  C:\Users\ahoutz\My Files\Google Drive\My Drive\Metrics II\PS 5\PS5_Houtz.log
  log type:  text
 opened on:  16 Apr 2021, 18:00:21

. use "C:\Users\ahoutz\Downloads\wagepan.dta"

. 
. /* Part a */
. 
. xtset nr
       panel variable:  nr (balanced)

. reg lwage educ black hisp exper expersq married union d81-d87, robust

Linear regression                               Number of obs     =      4,360
                                                F(14, 4345)       =      75.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1893
                                                Root MSE          =     .48033

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0913498   .0052913    17.26   0.000     .0809762    .1017234
       black |  -.1392342   .0245451    -5.67   0.000    -.1873552   -.0911132
```

The name of the file matches the place denoted as "RESEARCH" above, while the file name in my computer matches the place denoted as YOUR FILE above. Also note that the .log file displays the output of the regression, whereas in the .do file, only the regress command would be shown. Remember to always close the .log file. Otherwise, Stata will output an error. To close a .log file, type:

log close RESEARCH

1.4 Reading in Data

If your data is in a Stata .dta form already, you can type:

use FILENAME

if your file is in the same place as the directory you set. If the file is in a different place, type:

use "FILE DIRECTORY"

What happens if your data is in a spreadsheet? First, save your spreadsheet as a comma separated file (.csv). Then type:

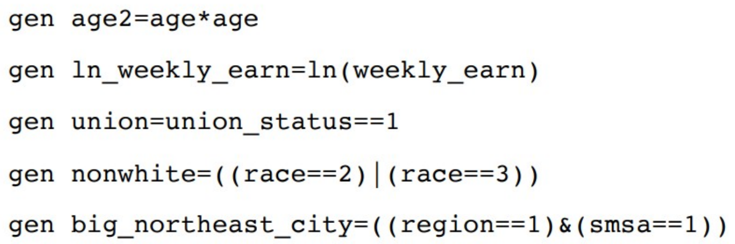
insheet using FILE.csv, comma

Variables are often not labelled after reading them into Stata. It is good practice to give each variable a short description. Let's consider a variable "age". If age was not labelled, I would write:

label var age "age in years"

which would attach that description to the variable age. If you want a table of your variables and the labels for each variable, type "describe".

In addition to labelling variables, sometimes we want to generate new variables. Here is a list of examples:



```
gen age2=age*age
gen ln_weekly_earn=ln(weekly_earn)
gen union=union_status==1
gen nonwhite=((race==2)|(race==3))
gen big_northeast_city=((region==1)&(smsa==1))
```

The first two examples demonstrate basic mathematical transformations: squaring a variable and logging a variable, respectively. The third example generates a dummy variable, "union", that is equal to one (or true) when "union_status" is one. The fourth example creates a dummy variable, "nonwhite", if race is either two OR three. The fifth example creates a dummy variable, "big_northeast_city", if the region is one AND smsa is one.

1.5 Descriptive Statistics Commands

We often want to learn about our raw data before performing more sophisticated statistical tests on the data. To generate a table that provides the mean, minimum, maximum, and standard deviation for each variable, type "sum". If you want these statistics for only a few specific variables, type those variables after the sum command. For example:

sum age educ

will provide a summary for only age and educ. For more information, such as the median, skew, and kurtosis, type:

```
sum, detail
```

We can also look at subsamples of our data. If we want to see summary statistics by race, for example, we first type "sort race" (assuming race is a variable we have). Then type "by race: sum". If we want to truncate our data summary to include only those that have more than 12 years of education, type "sum if educ >= 12".

Another summary command is the "tab" command. This command produces distributions for the data. If we want to see the distribution of race in our data, we type "tab race". We can include two variables as well: "tab race educ, row column". The row column option will produce totals for the rows and columns.

1.6 T-Test

If we want to compare means across groups, we can use a t-test. For example, suppose we wanted to test if there is a difference between salaries across race. We would then type "ttest salary, by(race)".

1.7 OLS

Running regressions in Stata is much simpler than in Matlab. Suppose we wanted to run a simple OLS regression of wages on age, education, race, region, and gender. We would then use the following command:

```
reg wages age educ i.race i.region i.gender
```

Note that our dependent variable is listed first with the independent variables following. By placing an i. in front of race, region, and sex, we create dummy variables for each race, region, and gender.

Recall from the first semester that sometimes we need standard errors robust to heteroskedasticity. To change how the standard errors are calculated, type ", robust" after listing all variables.

1.8 General Advice

We all get stuck while coding. If you need help, Stata has excellent documentation. Type "help COMMAND" to find information on how to use the code with which you are struggling. The internet is also helpful, especially the statalist website. If all else fails, come to me or Marinho and we will do our best to help you out.

There is a large variety of regression techniques that you can code in Stata. You will learn how to code IV, 2SLS, GMM, Difference-in-Differences, and more throughout this class. But the primary determinant for how well and how wide you will be able to code will be your own willingness to try new things and to struggle with the program. Keep at it and you'll be able to conduct your research with minimal technical barriers.

1.9 A Small Linear Algebra Review

1.9.1 Projection Matrix

Let $\boldsymbol{X}$ be an $n \times k$ matrix that is full rank. Then the projection matrix $\boldsymbol{P}$ is an $n \times n$ that results from:

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X'}\boldsymbol{X})^{-1}\boldsymbol{X}'$$

The projection matrix can be shown to have the following properties:

(i) $\boldsymbol{PX} = \boldsymbol{X}$

(ii) $\boldsymbol{P} = \boldsymbol{P'}$

(iii) $\boldsymbol{PP} = \boldsymbol{P}$

(iv) $tr(\boldsymbol{P}) = k$ and $rank(\boldsymbol{P}) = k$

We can use the projection matrix to find estimated $\hat{y}$ values in our regressions.

1.9.2 Annihilator Matrix

Let $\boldsymbol{X}$ be an $n \times k$ matrix that is full rank. Then the projection matrix $\boldsymbol{M}$ is an $n \times n$ that results from:

$$\begin{aligned}\boldsymbol{M} &= \boldsymbol{I} - \boldsymbol{P} \\ &= \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'}\boldsymbol{X})^{-1}\boldsymbol{X}'\end{aligned}$$

The annihilator matrix can be shown to have the following properties:

(i) $\boldsymbol{MX} = \boldsymbol{0}$

(ii) $\boldsymbol{MP} = \boldsymbol{0}$

(iii) $\boldsymbol{M} = \boldsymbol{M'}$

(iv) $\boldsymbol{MM} = \boldsymbol{M}$

(v) $tr(\boldsymbol{M}) = n - k$ and $rank(\boldsymbol{M}) = n - k$

We can use the annihilator matrix to remove parts of the regression that we are not interested in estimating. Why you would do this may not be obvious now but you will see its usefulness when you cover the Frisch-Waugh-Lovell Theorem. We can also use the annihilator matrix to find residuals.

1.9.3 Positive Semi-Definite Matrices

Throughout this semester, you will have to prove that certain results are positive semi-definite (see 2c and 4 on PS 1 for example). What is a positive semi-definite matrix? A matrix, M, is positive semi-definite if and only if it satisfies any of the following:

- M is congruent with a diagonal matrix, N, with non-negative real entries.
  - (i) Note that M is defined to be congruent with N if there exists an invertible matrix P such that $N = P'MP$.
  - (ii) For positive semi-definiteness, N must be diagonal with non-negative real entries.
- M is symmetric, and all its eigenvalues are real and non-negative.
- M is symmetric, and all its principal minors are non-negative.
- $\mathbf{z}'M\mathbf{z} \geq 0$ for *every* non-zero real column vector $\mathbf{z}$.
  - (i) This is the strict definition of positive semi-definiteness.

In practice, we usually use the fourth bullet point the most when proving that a matrix is PSD.

1.10 Closing Remarks

This semester will be very hard. The course is challenging, highly mathematical, and rigorous. A few tips:

(i) Make use of Marinho's office hours. He is very good at this and very helpful. Instead of struggling with a problem for many hours, talk to Marinho.

(ii) Focus on the main specification and proofs for each section. There is a lot to learn and all of it will be useful to you. But if you find yourself short on time, focus on learning the basic consistency, asymptotic normality, and identification proofs very well.

(iii) I am available for questions and assistance. Last semester I told you to email me with questions in advance because the material can be hard. That is doubly true this semester. I do not know this material nearly as well as Marinho, so I will need time to prepare good answers for your questions. That being said, feel free to stop by my office with questions you have in the moment and I will do my best.

Chapter 2

System OLS and Generalized Least Squares

2.1 System OLS Theory

Last semester, we saw how we could use OLS to estimate equations of interest. To begin this semester, we look at the case where we have multiple equations that appear to be unrelated at first glance. These Seemingly Unrelated Regressions can be written in a system of equations as follows:

$$\begin{cases} y_1 &= X_1'\beta_1 + \epsilon_1 \\ & \vdots \\ y_G &= X_G'\beta_G + \epsilon_G \end{cases}$$

where both X_g and β_g are $K \times 1$.

We can take advantage of possible correlation across the error terms, ϵ_g, by using SOLS. We want to do this to (1) eliminate possible bias in our estimates of our slope coefficients and to (2) increase the efficiency of our estimates. To do this, we first stack each part of our regression equations into one vector:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_G \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_G \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_G \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_G \end{bmatrix}$$

and write our normal regression equation: $\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

We have the same assumptions for SOLS to be identified, consistent, and asymptotically normal as we did for OLS. First, we assume that $\mathbb{E}[\boldsymbol{X}'\boldsymbol{\epsilon}] = \boldsymbol{0}$. This is called the orthogonality or exogeneity condition.

The second assumption we make is that $\mathbb{E}[\boldsymbol{X}'\boldsymbol{X}]$ has full rank. This assumption allows us to invert the matrix and identify β.

2.1.1 Identification Proof

Start with the regression equation:

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \\
\boldsymbol{X'Y} &= \boldsymbol{X'X\beta} + \boldsymbol{X'\epsilon} && \text{Take } E[\cdot]\text{:} \\
\mathbb{E}[\boldsymbol{X'Y}] &= \mathbb{E}[\boldsymbol{X'X}]\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{X'\epsilon}] && \text{Using assumption 1:} \\
\mathbb{E}[\boldsymbol{X'Y}] &= \mathbb{E}[\boldsymbol{X'X}]\boldsymbol{\beta} && \text{Using assumption 2:} \\
\mathbb{E}[\boldsymbol{X'X}]^{-1}\mathbb{E}[\boldsymbol{X'Y}] &= \boldsymbol{\beta}
\end{aligned}$$

2.1.2 Consistency Proof

We first start by applying the analogy principle to the identified $\boldsymbol{\beta}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'Y}\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'}(\boldsymbol{X\beta}+\boldsymbol{\epsilon})\right) \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'\epsilon}\right) && (*) \\
&\xrightarrow{P} \boldsymbol{\beta} + \mathbb{E}[\boldsymbol{X'X}]^{-1}\mathbb{E}[\boldsymbol{X'\epsilon}] && \text{Using assumption 1:} \\
&= \boldsymbol{\beta}
\end{aligned}$$

2.1.3 Asymptotic Normality Proof

Starting from equation (∗) from the consistency proof:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'\epsilon}\right) \\
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'\epsilon}\right) \\
\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X'X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X'\epsilon}\right) \\
&\xrightarrow{d} \mathbb{E}[\boldsymbol{X'X}]^{-1}\ \ N\left(0, \mathbb{E}[\boldsymbol{X'\epsilon\epsilon'X}]\right) \\
&= N\left(0, \mathbb{E}[\boldsymbol{X'X}]^{-1}\mathbb{E}[\boldsymbol{X'\epsilon\epsilon'X}]\mathbb{E}[\boldsymbol{X'X}]^{-1}\right)
\end{aligned}$$

Through these three proofs, we have shown that the SOLS estimator is identified, consistent, and asymptotically normal.

2.2 GLS Theory

Generalized Least Squares uses a transformation of the SOLS structure that we built above to produce a more efficient estimate than SOLS and a more accurate estimate than OLS equation by equation. Given that the assumptions listed further below are satisfied, GLS is BLUE (best linear unbiased estimator).

How do we begin setting up GLS? We start with our SOLS structure and pre-multiply by $\boldsymbol{\Omega}^{-1/2}$:

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \\
\boldsymbol{\Omega}^{-1/2}\boldsymbol{Y} &= \boldsymbol{\Omega}^{-1/2}\boldsymbol{X\beta} + \boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon} \\
\boldsymbol{X}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{Y} &= \boldsymbol{X}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{X\beta} + \boldsymbol{X}'\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon} \\
\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y} &= \boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X\beta} + \boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon} \\
\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}] &= \mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}]
\end{aligned}$$

Now, we need to find a matrix for $\boldsymbol{\Omega}$ such that the new model is homoskedastic $\Big(\text{i.e. } \mathbb{E}[\boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon\epsilon}'\boldsymbol{\Omega}^{-1/2}] = \boldsymbol{I}\Big)$. The solution is intuitively $\boldsymbol{\Omega}^{-1/2} = \mathbb{E}[\boldsymbol{\epsilon\epsilon}']^{-1/2}$.

To proceed with identification, and eventually to prove consistency and asymptotic normality, we use two key assumptions. First, we assume that $\mathbb{E}[\boldsymbol{X}_g \otimes \boldsymbol{\epsilon}_{g'}]$. This assumption states that there cannot be correlation between any right-hand side variable in any equation and any error term in any equation. This is a very strong assumption.

The second assumption is that $\boldsymbol{\Omega}$ is positive definite (to allow for the Cholesky decomposition during the derivation) and that $\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]$ is invertible. Assuming both of these, we continue on in identifying $\boldsymbol{\beta}$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}] &= \mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}] && \text{Using assumption 1:} \\
\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}] &= \mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]\boldsymbol{\beta} && \text{Using assumption 2:} \\
\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1}\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}] &= \boldsymbol{\beta}
\end{aligned}$$

To estimate, we simply apply the analogy principle:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}\right)$$

Let's next show that the GLS estimator is consistent. Starting with the analogy principle above:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}\right)\\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\right)\\
&= \boldsymbol{\beta}+\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\right) \qquad (**)\\
&\xrightarrow{P} \boldsymbol{\beta}+\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1}\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}]\\
&= \boldsymbol{\beta}+\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1}\mathbb{E}[\boldsymbol{\epsilon}\otimes\boldsymbol{X}]\cdot vec(\boldsymbol{\Omega}^{-1})\\
&= \boldsymbol{\beta}
\end{aligned}$$

So $\hat{\boldsymbol{\beta}}_{GLS}$ is consistent for $\boldsymbol{\beta}$. Let's look at asymptotic normality next. Starting from equation $(**)$ above:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= \boldsymbol{\beta}+\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\right)\\
\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\right)\\
\sqrt{n}(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}) &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\right)\\
&\xrightarrow{P} \mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1}\; N\left(0,\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]\right)\\
&= N\left(0,\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1}\;\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]\;\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}]^{-1}\right)
\end{aligned}$$

Through these proofs, we have shown that the GLS estimator is identified, consistent, and asymptotically normal after using our two assumptions. On your homework, you will show that the GLS estimator is more efficient than the SOLS estimator.

2.2.1 FGLS Theory

Going through these proofs, we should notice that GLS is actually not feasible. Let's look at the estimator $\hat{\boldsymbol{\beta}}_{GLS}$ again:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}\right)$$

Note that $\boldsymbol{\Omega}$ is not known! Recall that $\boldsymbol{\Omega}=\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']$. To conduct GLS, we need to estimate $\boldsymbol{\Omega}$. Let's use the analogy principle:

$$\tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}\boldsymbol{\epsilon}'$$

But we do not know $\boldsymbol{\epsilon}$ either, so we need to estimate that too:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'$$

and now we can do GLS. We call this formulation Feasible Generalized Least Squares, or FGLS:

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{Y}\right)$$

Marinho proved the consistency of FGLS in class and started the asymptotic normality proof. On your homework, you will show that the estimated asymptotic variance for $\hat{\boldsymbol{\beta}}_{FGLS}$ is consistent.

In practice, how do we go about estimating FGLS?

(1) Run SOLS on the model before transforming with $\boldsymbol{\Omega}$ and calculate the residuals: $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SOLS}$.

(2) Use the SOLS residuals, $\hat{\boldsymbol{\epsilon}}$, to find $\hat{\boldsymbol{\Omega}}$.

(3) Calculate $\hat{\boldsymbol{\beta}}_{FGLS}$ using $\hat{\boldsymbol{\Omega}}$.

(4) Calculate the estimated asymptotic variance using $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\epsilon}}$.

Before moving on to Matlab, let's look at a decision tree for when to use FGLS and when to use SOLS (where $\boldsymbol{u}$ in the figure is the same as the $\boldsymbol{\epsilon}$ we have been using):

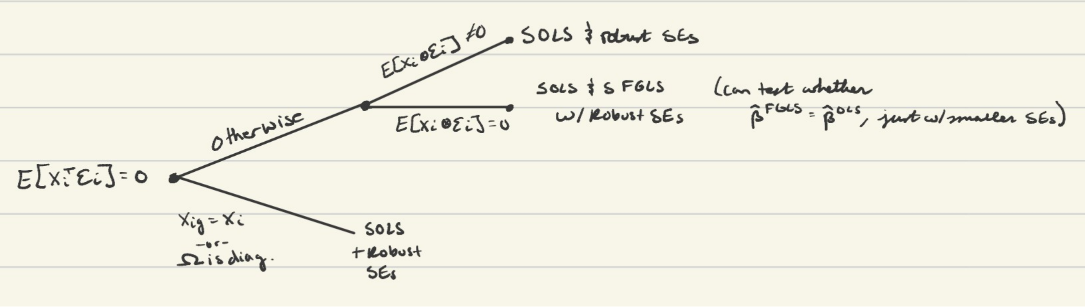
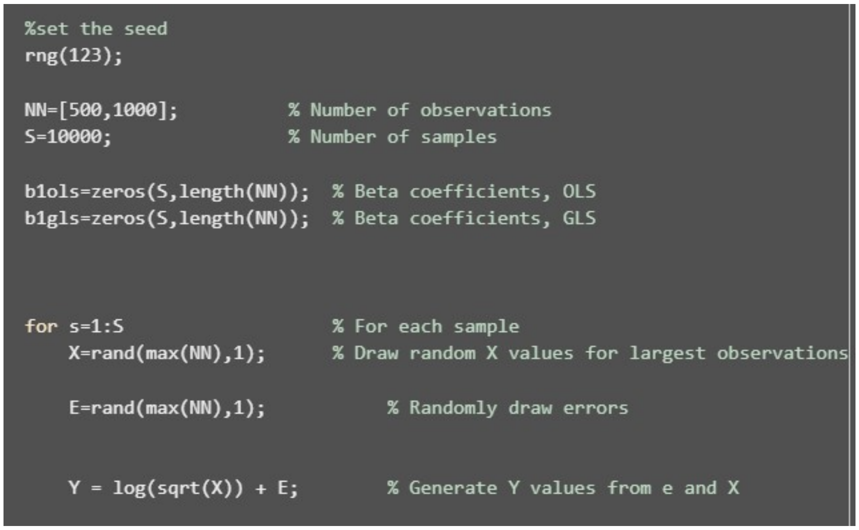


Figure 2.1: We can see that we always start with the exogeneity condition. If the right-hand side variables are the same in each equation, or if $\boldsymbol{\Omega}$ is the diagonal, then we use SOLS. Otherwise, we look to see if $\mathbb{E}[\boldsymbol{X} \otimes \boldsymbol{\epsilon}] = 0$. If this condition is satisfied, then we use FGLS. If not, we use SOLS.

2.3 Matlab Exercise

We went through a Matlab example that is similar to the exercise you are asked to do on the problem set. Come see me during office hours or send me an email if you need help. I have attached a snapshot of the data generating process below:



```
%set the seed
rng(123);

NN=[500,1000];          % Number of observations
S=10000;                % Number of samples

b1ols=zeros(S,length(NN));  % Beta coefficients, OLS
b1gls=zeros(S,length(NN));  % Beta coefficients, GLS



for s=1:S                   % For each sample
    X=rand(max(NN),1);      % Draw random X values for largest observations

    E=rand(max(NN),1);           % Randomly draw errors


    Y = log(sqrt(X)) + E;        % Generate Y values from e and X
```

Figure 2.2

Chapter 3

Instrumental Variables

3.1 Instrumental Variable Theory

In most equations that we want to estimate empirically, omitted variable bias or other biases end up violating one of our OLS assumptions: orthoganility between the regressors and the error term. This problem is known as the *endogeneity problem.*

How do we fix this issue? We can use what is known as an **instrumental variable**, often designated as z_k. A valid instrument satisfies three conditions:

1. z_k is not already part of the model.
2. z_k is uncorrelated with the error term ($\mathbb{E}[z_k u_i] = 0$).
3. z_k is correlated with the right-hand side regressors (so that $\mathbb{E}[z_k x_i] \neq 0$ and is full rank).

The following figure illustrates these three requirements:

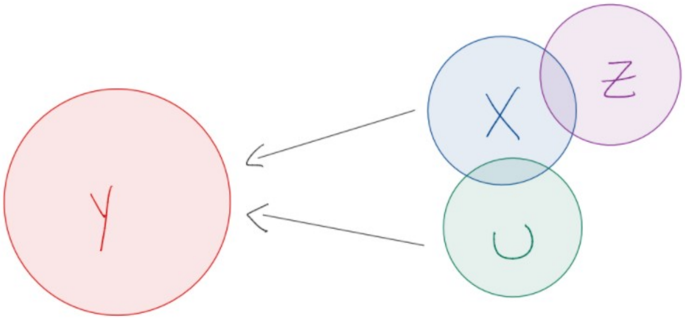


Figure 3.1: Note how both X_i and U_i impact Y_i. We can't disentangle the effect of X_i on Y_i from the effect of U_i on Y_i because of the intersection between X_i and U_i. The instrument only affects Y_i through X_i.

3.1.1 Identification

Let's go through the identification proof for IV. Define $x \equiv (1, x_1, x_2, ..., x_k)$, $\beta \equiv (\beta_0, \beta_1, ..., \beta_k)$, and $z \equiv (1, x_1, x_2, ..., x_{k-1}, z_k)$. Starting with our regression equation:

$$\begin{aligned}
y_i &= x_i\beta + u_i \\
z_i'y_i &= z_i'x_i\beta + z_i'u_i \\
\mathbb{E}[z_i'y_i] &= \mathbb{E}[z_i'x_i]\beta + \mathbb{E}[z_i'u_i] && \text{Using assumption 2:} \\
\mathbb{E}[z_i'y_i] &= \mathbb{E}[z_i'x_i]\beta && \text{Using assumption 3:} \\
\mathbb{E}[z_i'x_i]^{-1}\mathbb{E}[z_i'y_i] &= \beta
\end{aligned}$$

So β is identified.

3.1.2 Consistency

Using the analogy principle and proceeding from there:

$$\begin{aligned}
\hat{\beta}_{IV} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_i'x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i'y_i\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} z_i'x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i'(x_i\beta + u_i)\right) \\
&= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} z_i'x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i'u_i\right) \\
&\xrightarrow{P} \beta + \mathbb{E}[z_i'x_i]^{-1}\mathbb{E}[z_i'u_i] \\
&= \beta
\end{aligned}$$

The IV estimator is also asymptotically normal. I leave this to you to derive in your own time. If you need help, come see me during my office hours.

3.2 Quiz 2 Previous Problem

Consider the following system of equations:

$$\begin{aligned}
y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \\
x_{i1} &= \pi x_{i2} + v_i
\end{aligned}$$

where data are *i.i.d.* across $i = 1, ..., N$, all variables are scalars, $\mathbb{E}[u_i] = 0, \mathbb{E}[v_i] = 0$, $\mathbb{E}[u_i x_{i2}] = 0$, $\mathbb{E}[u_i x_{i1}] \neq 0$, $\mathbb{E}[u_i v_i] \neq 0$, $\mathbb{E}[x_{i2} v_i] = 0$, $\mathbb{E}[x_{i2}^2] \neq 0$, $\pi \neq 0$, and the researcher observes a sample of

(y_i, x_{i1}, x_{i2}).

3.2.1 Part a

Suppose $\beta_2 = 0$. Write down the IV estimator of β_1 that uses x_{i2} as an instrument for x_{i1}. Show this IV estimator is consistent and asymptotically normal. Write down a consistent estimator for the asymptotic variance of this IV estimator.

Solution

$$\begin{aligned}
y_i &= \beta_1 x_{i1} + u_i \\
x_{i2} y_i &= x_{i2} x_{i1} \beta_1 + x_{i2} u_i \\
\mathbb{E}[x_{i2} y_i] &= \mathbb{E}[x_{i2} x_{i1}] \beta_1 + \mathbb{E}[x_{i2} u_i] \\
\beta_1 &= \mathbb{E}[x_{i2} x_{i1}]^{-1} \mathbb{E}[x_{i2} y_i] \\
\hat{\beta}_1 &= \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} x_{i1} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} y_i \right)
\end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} x_{i1} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} (x_{i1} \beta_1 + u_i) \right) \\
&= \beta_1 + \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} x_{i1} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} u_i \right) && (*) \\
&\xrightarrow{P} \beta_1 + \mathbb{E}[x_{i2} x_{i1}]^{-1} \mathbb{E}[x_{i2} u_i] && = \beta_1
\end{aligned}$$

So we have shown that it is consistent. Let's show it is asymptotically normal starting from $(*)$:

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} x_{i1} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} u_i \right) \\
\sqrt{n}(\hat{\beta}_1 - \beta) &= \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} x_{i1} \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i2} u_i \right) \\
&\xrightarrow{d} \mathbb{E}[x_{i2} x_{i1}]^{-1} \, N\left(0, \mathbb{E}[x_{i2}^2 u_i^2]\right) \\
&= N\left(0, \mathbb{E}[x_{i2} x_{i1}]^{-2} \mathbb{E}[x_{i2}^2 u_i^2]\right) \\
\widehat{AVAR} &= \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2} x_{i1} \right)^{-2} \left(\frac{1}{n} \sum_{i=1}^{n} x_{i2}^2 \hat{u}_i^2 \right)
\end{aligned}$$

where $\hat{u}_i = y_i - x_{i1} \hat{\beta}_1$.

3.2.2 Part b

Suppose now $\beta_2 \neq 0$. x_{i2} was an exogenous excluded instrument in part(a) but now it is not excluded. The researcher creates a variable $z_{i2} = x_{i2} + \eta_i$, where η_i is randomly generated data that are independent of everything else and have mean zero. The researcher claims that he can consistently estimate (β_1, β_2) by using z_{i2} as an instrument for x_{i1}. Is the researcher correct?

Solution

Generating an instrument this way will fail the rank condition:

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2} \\ z_{i2} \end{bmatrix}\begin{bmatrix} x_{i1} & x_{i2} \end{bmatrix}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2} \\ z_{i2} \end{bmatrix} y_i\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2}x_{i1} & x_{i2}^2 \\ z_{i2}x_{i1} & z_{i2}x_{i2} \end{bmatrix}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2}y_i \\ z_{i2}y_i \end{bmatrix}\right)
\end{aligned}$$

If we just look at the first term:

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2}x_{i1} & x_{i2}^2 \\ z_{i2}x_{i1} & z_{i2}x_{i2} \end{bmatrix}\right)^{-1} &= \left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2}x_{i1} & x_{i2}^2 \\ (x_{i2}+\eta_i)x_{i1} & (x_{i2}+\eta_i)x_{i2} \end{bmatrix}\right)^{-1} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i2}x_{i1} & x_{i2}^2 \\ x_{i2}x_{i1}+\eta_i x_{i1} & x_{i2}x_{i1}+\eta_i x_{i2} \end{bmatrix}\right)^{-1} \\
&\xrightarrow{P} \begin{bmatrix} \mathbb{E}[x_{i2}x_{i1}] & \mathbb{E}[x_{i2}^2] \\ \mathbb{E}[x_{i2}x_{i1}+\eta_i x_{i1}] & \mathbb{E}[x_{i2}^2+x_{i2}\eta_i] \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \mathbb{E}[x_{i2}x_{i1}] & \mathbb{E}[x_{i2}^2] \\ \mathbb{E}[x_{i2}x_{i1}] & \mathbb{E}[x_{i2}^2] \end{bmatrix}^{-1}
\end{aligned}$$

Note that this is not invertible, so this instrument will not consistently estimate $\boldsymbol{\beta}$.

3.2.3 Part c

Consider the IV estimator in part(a). Is it possible to make that IV estimator more efficient if we use z_{i2} instead of x_{i2} as an instrument for x_{i1}? Assume $\mathbb{E}[u_i^2|z_{i2}]\mathbb{E}[u_i^2|x_{i2}] = \sigma^2$.

Solution

First, we find $AVAR_a$:

$$\begin{aligned}
AVAR_a &= \frac{\mathbb{E}[x_{i2}^2 u_i^2]}{\mathbb{E}[x_{i2}x_{i1}]^2} \\
&= \frac{\mathbb{E}_{x_{i2}}[x_{i2}^2 \mathbb{E}[u_i^2|x_{i2}]]}{\mathbb{E}[x_{i2}(x_{i2}\pi + v_i)]^2} \\
&= \frac{\sigma^2 \mathbb{E}[x_{i2}^2]}{\pi^2 \mathbb{E}[x_{i2}^2]^2} \\
&= \frac{\sigma^2}{\pi^2 \mathbb{E}[x_{i2}^2]}
\end{aligned}$$

Then we find $AVAR_b$ using the same asymptotic formula:

$$\begin{aligned}
AVAR_b &= \frac{\mathbb{E}[z_{i2}^2 u_i^2]}{\mathbb{E}[z_{i2}x_{i1}]^2} \\
&= \frac{\sigma^2 \mathbb{E}[z_{i2}^2]}{\mathbb{E}[x_{i1}(x_{i2} + \eta_i)]^2} \\
&= \frac{\sigma^2 \mathbb{E}[(x_{i2} + \eta_i)^2]}{(\mathbb{E}[x_{i2}x_{i1}] + \mathbb{E}[\eta_i x_{i1}])^2} \\
&= \frac{\sigma^2 \mathbb{E}[x_{i2}^2 + 2x_{i2}\eta_i + \eta_i^2]}{\pi^2 \mathbb{E}[x_{i2}^2]^2}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\sigma^2}{\pi^2 \mathbb{E}[x_{i2}^2]} + \frac{\sigma^2 \mathbb{E}[\eta_i^2]}{\pi^2 \mathbb{E}[x_{i2}^2]^2} \\
&= AVAR_a + \frac{\sigma^2 \mathbb{E}[\eta_i^2]}{\pi^2 \mathbb{E}[x_{i2}^2]^2}
\end{aligned}$$

Chapter 4

Two-Stage Least Squares

4.1 Two-Stage Least Squares

Technically, IV is a method for one endogenous regressor and one instrument. Two-stage least squares, 2SLS, allows for the use of multiple instruments. To do so, redefine $z \equiv (1, x_1, ..., x_{k-1}, z_1, ..., z_m)$.

2SLS, in effect, allows us to find the highest correlation between our instruments and our endogenous variable(s). A strong correlation helps ensure that the rank condition is satisfied and that our estimates are consistent.

4.1.1 Identification

Let's derive the 2SLS estimator. We start with what is called the 1st stage projection, where we project the instruments and exogenous regressors, z, on our endogenous regressors, x:

$$x^* = z\pi$$

where $\pi \equiv \mathbb{E}[z'z]^{-1}\mathbb{E}[z'x]$, the linear projection of z on x. Next we premultiply x^* with our standard regression equation:

$$\begin{aligned} y &= x\beta + u \\ x^{*\prime}y &= x^{*\prime}x\beta + x^{*\prime}u \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}[x^{*\prime}y] &= \mathbb{E}[x^{*\prime}x]\beta + \mathbb{E}[x^{*\prime}u] \\
\mathbb{E}[x^{*\prime}y] &= \mathbb{E}[x^{*\prime}x]\beta + \mathbb{E}[x'z(z'z)^{-1}z'u] \\
\mathbb{E}[x^{*\prime}y] &= \mathbb{E}[x^{*\prime}x]\beta \\
\beta &= \mathbb{E}[x^{*\prime}x]^{-1}\mathbb{E}[x^{*\prime}y] \\
\beta &= \left(\mathbb{E}[x'z]\mathbb{E}[z'z]^{-1}\mathbb{E}[z'x]\right)^{-1}\left(\mathbb{E}[x'z]\mathbb{E}[z'z]^{-1}\mathbb{E}[z'y]\right)
\end{aligned}
$$

And if we apply the analogy principle:

$$
\begin{aligned}
\hat{\beta}_{2SLS} = &\left(\left(\frac{1}{n}\sum_{i=1}^{n} x'z\right)\left(\frac{1}{n}\sum_{i=1}^{n} z'z\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z'x\right)\right)^{-1}\cdot \\
&\left(\left(\frac{1}{n}\sum_{i=1}^{n} x'z\right)\left(\frac{1}{n}\sum_{i=1}^{n} z'z\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z'y\right)\right)
\end{aligned}
$$

4.1.2 Consistency

To make notation easier, define $\hat{Q}_{xz} \equiv \left(\frac{1}{n}\sum_{i=1}^{n} x'z\right)$, $\hat{Q}_{zz} \equiv \left(\frac{1}{n}\sum_{i=1}^{n} z'z\right)$, $\hat{Q}_{zx} \equiv \left(\frac{1}{n}\sum_{i=1}^{n} z'x\right)$, and $\hat{Q}_{zu} \equiv \left(\frac{1}{n}\sum_{i=1}^{n} z'u\right)$. Starting from the analogy principle:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1}\left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zy}\right) \\
&= \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1}\left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\frac{1}{n}\sum_{i=1}^{n} z'(x\beta+u)\right) \\
&= \beta + \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1}\left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zu}\right) \qquad (*) \\
&\xrightarrow{P} \beta + \left(Q_{xz}Q_{zz}^{-1}Q_{zx}\right)^{-1}\left(Q_{xz}Q_{zz}^{-1}Q_{zu}\right) \\
&= \beta
\end{aligned}
$$

4.1.3 Asymptotic Normality

We start from equation $(*)$ above:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \beta + \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1}\left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zu}\right) \\
\hat{\beta}_{2SLS} - \beta &= \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1}\left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zu}\right) \\
\sqrt{n}(\hat{\beta}_{2SLS} - \beta) &= \left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx}\right)^{-1}\left(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\sqrt{n}\hat{Q}_{zu}\right) \\
&\xrightarrow{d} \left(Q_{xz}Q_{zz}^{-1}Q_{zx}\right)^{-1}\left(Q_{xz}Q_{zz}^{-1}\right)\ N\left(0, \mathbb{E}[z'uu'z]\right) \\
&= N\left(0, \left(Q_{xz}Q_{zz}^{-1}Q_{zx}\right)^{-1}\left(Q_{xz}Q_{zz}^{-1}\right)\mathbb{E}[z'uu'z]\left(Q_{zz}^{-1}Q_{xz}\right)\left(Q_{xz}Q_{zz}^{-1}Q_{zx}\right)^{-1}\right)
\end{aligned}
$$

4.1.4 Miscellaneous Information

First, a few brief warnings about 2SLS:

(i) OLS + OLS (running OLS on the first stage projection and then running OLS on the second stage with the projected x on the right hand side) is **NOT** the same as 2SLS. If you use OLS + OLS, your standard errors will not be accurate.

(ii) Always include all exogenous variables in the first stage projection. If you do not, 2SLS is not consistent.

Second, a small decision tree:

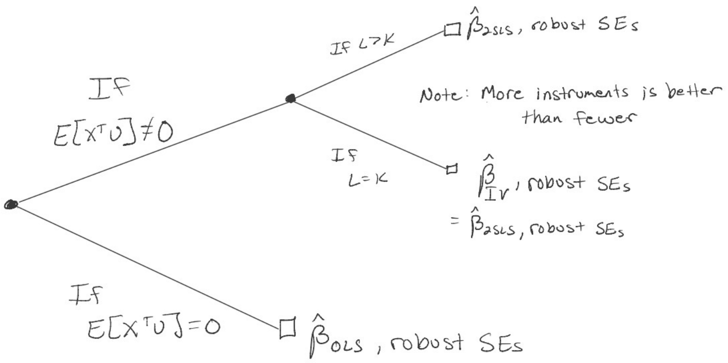

If $E[X^T U] \neq 0$
If $L > K$
$\hat{\beta}_{2SLS}$, robust SEs
Note: More instruments is better than fewer
If $L = K$
$\hat{\beta}_{IV}$, robust SEs
$= \hat{\beta}_{2SLS}$, robust SEs
If $E[X^T U] = 0$
$\hat{\beta}_{OLS}$, robust SEs

Figure 4.1: Where $L \equiv$ the number of instruments and $K \equiv$ the number of endogenous regressors.

4.2 Control Function

The goal of this approach is to determine if we have an endogenous regressor. Define $z \equiv [z_1, z_2]$, where $z_1 \equiv$ included exogenous regressors and the constant and $z_2 \equiv$ excluded instruments. We first set up a structural equation:

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + u_1$$

where $y_2 \equiv$ included potentially endogenous variable. Because of the way we structured z, we know that $\mathbb{E}[z'u_1] = 0$. Now we set-up a reduced-form equation for y_2:

$$y_2 = z\pi_2 + \nu_2$$

where $\pi_2 = \mathbb{E}[z'z]^{-1}\mathbb{E}[z'y_2]$. We assume that $\mathbb{E}[z'\nu_2] = 0$. Note that y_2 is endogenous if the structural error, u_1, is correlated with the reduced form error, ν_2. We can therefore set up an equation modelling this:

$$u_1 = \rho_1\nu_2 + \varepsilon_1$$

where ρ_1 is the linear projection of ν_2 on u_1. We assume that $\mathbb{E}[\nu_2\varepsilon_1] = 0$. Note that $\mathbb{E}[z'\varepsilon_1] = \mathbb{E}[z'(u_1 - \rho_1\nu_2)] = 0$. Subbing this equation into the structural equation yields:

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1\nu_2 + \varepsilon_1$$

The intuition: OLS is consistent if z_1, y_2, and ν_2 are uncorrelated with ε_1. Essentially, we are testing to see if $\rho_1 = 0$. But how do we do this?

We do not observe ν_2 directly. But we will use our usual trick for estimating errors: run OLS on the reduced form equation. After obtaining $\hat{\nu}_2$, we can use those estimated residuals in the transformed structural equation. Run OLS on this transformed equation using robust standard errors. Then use the Wald test to determine whether $\rho_1 = 0$. If $\rho_1 = 0$, then we conclude that y_2 is exogenous.

4.3 Practice Problem: Wooldridge 5.1

Consider the following model:

$$\begin{aligned} y_1 &= z_1\delta_1 + \alpha_1 y_2 + u_1 \\ y_2 &= z\pi_2 + v_2 \end{aligned}$$

where y_2 is the suspected endogenous variable and z is the vector of all exogenous variables. The second equation is the reduced form for y_2. Assume that z has at least one more element than z_1. We know that one estimator of $(\delta_1,\ \alpha_1)$ is the 2SLS estimator using instruments. Consider an alternative estimator: (a) estimate the reduced form by OLS and save the residuals $\hat{v}_2$ and (b) estimate the following equation by OLS:

$$y_1 = z_1\delta_1 + \alpha_1 y_2 + \rho_1\hat{v}_2 + \varepsilon_1 \tag{4.1}$$

Show that the OLS estimates of δ_1 and α_1 from this regression are identical to the 2SLS estimators. Hint: Use the partitioned regression algebra of OLS.

4.3.1 Solution

Define $x_1 \equiv [z_1\ y_2]$ and let $\hat{\beta} = [\hat{\beta}_1\ \hat{\rho}_1] = [\hat{\delta}_1\ \hat{\alpha}_1\ \hat{\rho}_1]$. To partition the regression we:

(i) Regress x_1 on $\hat{v}_2$ and get residuals $\ddot{x}$.

(ii) Regress y_1 on $\ddot{x}$.

We partition regression (1) in this way to specifically estimate the slope coefficients on z_1 and y_2.We start with (i) to find $\ddot{x}$:

$$\begin{aligned}
x_1 &= \hat{v}_2\mathcal{W} + \ddot{x} \\
\ddot{x} &= x_1 - \hat{v}_2\mathcal{W} \\
\begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} &= \begin{bmatrix} z_1 \\ y_2 \end{bmatrix} - \hat{v}_2(\hat{v}_2'\hat{v}_2)^{-1}\left(\hat{v}_2'\begin{bmatrix} z_1 \\ y_2 \end{bmatrix}\right)
\end{aligned}$$

$$= \begin{bmatrix} z_1 \\ y_2 \end{bmatrix} - \hat{v}_2(\hat{v}_2'\hat{v}_2)^{-1}\begin{bmatrix} 0 \\ \hat{v}_2'y_2 \end{bmatrix}$$

Remember that we can decompose a variable into its projection and the residual. Therefore, we can write y_2 as:

$$y_2 = \hat{y}_2 + \hat{v}_2$$

Subbing this in gives us:

$$\begin{aligned}
&= \begin{bmatrix} z_1 \\ y_2 \end{bmatrix} - \hat{v}_2(\hat{v}_2'\hat{v}_2)^{-1}\begin{bmatrix} 0 \\ \hat{v}_2'(\hat{y}_2 + \hat{v}_2) \end{bmatrix} \\
&= \begin{bmatrix} z_1 \\ y_2 \end{bmatrix} - \hat{v}_2(\hat{v}_2'\hat{v}_2)^{-1}\begin{bmatrix} 0 \\ \hat{v}_2'\hat{y}_2 + \hat{v}_2'\hat{v}_2 \end{bmatrix} \\
&= \begin{bmatrix} z_1 \\ y_2 \end{bmatrix} - \hat{v}_2(\hat{v}_2'\hat{v}_2)^{-1}\begin{bmatrix} 0 \\ \hat{v}_2'\hat{v}_2 \end{bmatrix} \\
&= \begin{bmatrix} z_1 \\ y_2 \end{bmatrix} - \hat{v}_2\begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} z_1 \\ \hat{y}_2 - \hat{v}_2 \end{bmatrix} - \begin{bmatrix} 0 \\ \hat{v}_2 \end{bmatrix} \\
\begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} &= \begin{bmatrix} z_1 \\ \hat{y}_2 \end{bmatrix}
\end{aligned}$$

Now that we have regressed x_1 on $\hat{v}_2$ and gotten the residuals $\ddot{x}$, we can regress y_1 on $\ddot{x}$:

$$\begin{aligned} y_1 &= \ddot{x}'\beta_1 + u_1 \\ &= \begin{bmatrix} z_1 & \hat{y}_2 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \alpha_2 \end{bmatrix} + u_1 \\ &= z_1\delta_1 + \hat{y}_2\alpha_1 + u_1 \end{aligned}$$

But notice that this is the second stage of 2SLS! By doing OLS on this regression, we obtain the same estimator that 2SLS provides us. So we can conclude that $\hat{\beta}_{OLS} = \hat{\beta}_{2SLS}$ when using the control function approach.

4.4 Durbin-Wu-Hausman Test

The Durbin-Wu-Hausman test is another way to test the endogeneity of a regressor. The hypotheses are as follows:

$$\begin{array}{ll} H_0 : \text{Exogeneity} & H_a : \text{Endogeneity} \\ \hat{\beta}_{OLS} \xrightarrow{P} \beta & \hat{\beta}_{OLS} \xrightarrow{P} \beta + \Delta \\ \hat{\beta}_{2SLS} \xrightarrow{P} \beta & \hat{\beta}_{2SLS} \xrightarrow{P} \beta \end{array}$$

The basic idea is to compare the slope coefficients under OLS and 2SLS. Ideally, we would like to run a Wald test using $\sqrt{n}\left(\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}\right)$. The asymptotic variance of this function is not invertible though. As such, we must develop a more clever approach. We first look at the structural equation:

$$y_1 = z_1\delta_1 + y_2\alpha_1 + u_1$$

Estimate this equation using OLS first to get $\hat{\alpha}_1$. After finding $\hat{\alpha}_1$, manipulate the equation to get an expression for $\hat{\delta}_1^{OLS}$:

$$\begin{aligned} y_1 &= z_1\delta_1 + y_2\hat{\alpha}_1 + u_1 \\ y_1 - \hat{\alpha}_1 y_2 &= z_1\delta_1 + u_1 \\ z_1'(y_1 - \hat{\alpha}_1 y_2) &= z_1'z_1\delta_1 + z_1'u_1 \\ \mathbb{E}[z_1'(y_1 - \hat{\alpha}_1 y_2)] &= \mathbb{E}[z_1'z_1]\delta_1 + \mathbb{E}[z_1'u_1] \\ \delta_1 &= \mathbb{E}[z_1'z_1]^{-1}\,\mathbb{E}[z_1'(y_1 - \hat{\alpha}_1 y_2)] \\ \hat{\delta}_1^{OLS} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}'z_{i1}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}'(y_{i1} - \hat{\alpha}_1 y_{i2})\right) \end{aligned}$$

We can go through similar steps to find $\hat{\delta}_1^{2SLS}$, using the 2SLS estimator for $\hat{\alpha}_1$ and substituting in the first stage projection dependent variable:

$$\begin{aligned}\hat{\delta}_1^{2SLS} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}' z_{i1}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}'(y_{i1} - \hat{\alpha}_1 \hat{y}_{i2})\right)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}' z_{i1}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}'(y_{i1} - \hat{\alpha}_1 P_z y_{i2})\right)\end{aligned}$$

where $P_z \equiv z(z'z)^{-1}z'$. Now we can compare the two estimators by subtracting one from the other:

$$\begin{aligned}\hat{\delta}_1^{2SLS} - \hat{\delta}_1^{OLS} &= (z_1'z_1)^{-1}\left(z_1'(y_1 - P_z y_2 \hat{\alpha}_1^{2SLS})\right) - (z_1'z_1)^{-1}\left(z_1'(y_1 - y_2\hat{\alpha}_1^{OLS})\right)\\ &= (z_1'z_1)^{-1}z_1'\left(P_z y_2 \hat{\alpha}_1^{2SLS} - y_2\hat{\alpha}_1^{OLS}\right)\\ &= (z_1'z_1)^{-1}z_1' y_2(\hat{\alpha}_1^{2SLS} - \hat{\alpha}_1^{OLS})\end{aligned}$$

Test to see if $\hat{\alpha}_1^{2SLS} - \hat{\alpha}_1^{OLS} = 0$, as the covariance matrix of this difference is invertible. If we cannot statistically distinguish the difference from zero, we fail to reject the null hypothesis (that the variable is exogenous).

4.5 Overidentification Test

The overidentification test provides a check on the exogeneity condition on the instrument set - that $\mathbb{E}[z'u] = 0$. Before continuing, note that to conduct this test, we need more instruments than endogenous variables. The extra instruments overidentify the model and provide us with the information needed to verify the exogeneity assumption. We start again with the structural control function equation:

$$y_1 = z_1\delta_1 + y_2\alpha_2 + u_1$$

Define $x \equiv [z_1, y_1]$ and again define $z \equiv [z_1, z_2]$. The theory behind this test is complicated, so we focus on how to implement the test here:

(1) Run 2SLS on the structural model to find $\hat{u}_1$.

(2) Calculate $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_1^2$.

(3) Calculate $\hat{Q}_{zx} = \frac{1}{n}\sum_{i=1}^{n} z_i' x_i$.

(4) Calculate the annihilator matrix $M_{zx} = I - \hat{Q}_{zx}[\hat{Q}_{zx}'\hat{Q}_{zx}]^{-1}\hat{Q}_{zx}'$.

(5) Calculate $\hat{\eta} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i' u_i\right)$, the OLS estimator of z on u_1.

(6) Calculate the test statistic $S = \left(\frac{n}{\hat{\sigma}^2}\right)(\hat{\eta}' M_{zx}\hat{\eta})$.

(7) Compare S to χ^2_{L-K} distribution, where $L \equiv$ number of exogenous regressors and instruments and $K \equiv$ number of included regressors.

(8) Reject H_0 at the 1% level.

4.6 Weak Instruments

Sometimes our instruments barely satisfy the rank condition (see figure 2). In these cases, 2SLS and IV may behave somewhat erratically. We want to test to see if the instruments we are using are "strong" instruments.

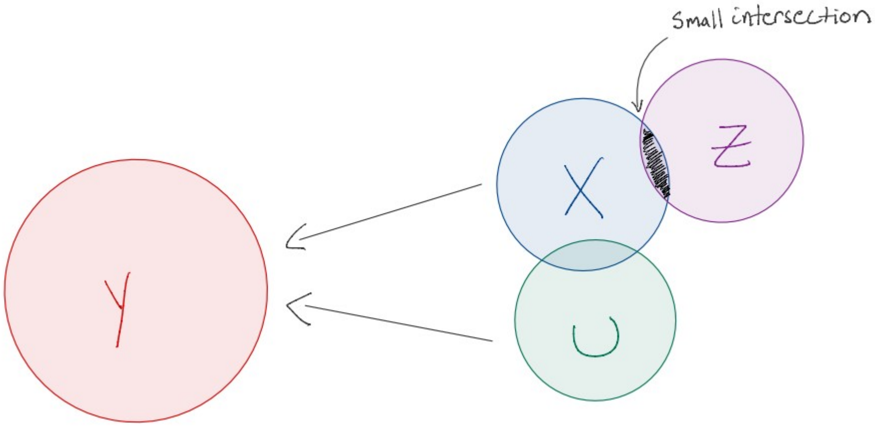


Figure 4.2

4.6.1 Stock-Yogo Test

The first test for weak instruments is the Stock-Yogo Test. Applied microeconomics papers commonly use this test. To begin with, we look at the first stage projection:

$$y_2 = z_1\pi_1 + z_2\pi_2 + \nu_2$$

Our null hypothesis, H_0, is that $\pi_2 = 0$. We do an F-test on the first stage after running 2SLS to determine whether z_2 should be included in the projection. Intuitively, we are seeing if the instruments are actually correlated with our endogenous variable (i.e. if $\mathbb{E}[z_2'y_2] = 0$). For a model with one endogenous variable and one instrument, we have a critical value of 16.4. At this level, we can be sure that the distortion level on our slope estimates is $\leq 5\%$. Figure 3 is a decision tree providing a guide through the Stock-Yogo procedure:

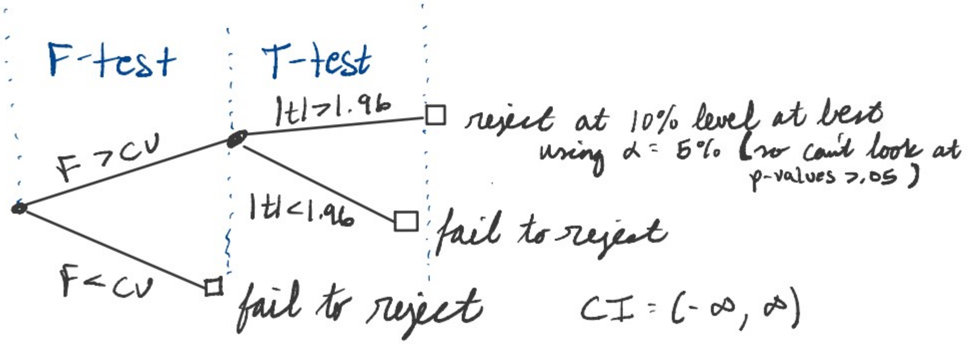


Figure 4.3

4.6.2 Lee-Moreira-McCrary-Porter (2020)

Lee-Moreira-McCrary-Porter (2020) updates the Stock-Yogo paper's guidelines for a model with one endogenous variable and one excluded instrument. For zero distortion of the second stage 5% t-test[4], we want a first stage projection F-test value of 104.7. The paper also provides a table of critical values for the t-test to be rejected at the 5% level if the F-statistic is not high enough. This paper is currently the paper you should use for testing weak instruments.

4.7 Problem Set 2, Question 6

Suppose you are given the following structural equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u$$

where $\mathbb{E}[u] = 0$. You are interested in β_1. The variables $(x_2, ..., x_K)$ are observable controls. All of these factors are correlated with x_1. Also suppose that x_1 is endogenous and that we have a valid instrument, z_1 for for x_1. We know that $\hat{\boldsymbol{\beta}}^{2SLS}$ is consistent for $\boldsymbol{\beta}$.

A common belief among applied researchers is that we do not need to worry about the exogeneity of $(x_2, ..., x_K)$ as long as we only care about x_1. Therefore, we only need a valid instrument for x_1 and the 2SLS estimator.

4.7.1 Part a

Suppose z_1 is independent of $(x_2, ..., x_K, u)$, z_1 is correlated with x_1, and $(x_2, ..., x_K)$ are exogenous. Does controlling for $(x_2, ..., x_K)$ matter for the consistency of $\hat{\beta}_1^{2SLS}$?

[4] Tests with size ≤ 0.455 will always be distorted

Solution

We know that $\hat{\beta}^{2SLS}$ is consistent with all of the control variables, as we showed last recitation. But what about without all the controls? Note that the error term, ε, is now:

$$\varepsilon = u + \beta_2(x_2 - \mathbb{E}[x_2]) + ... + \beta_K(x_K - \mathbb{E}[x_K])$$

and that the constant, $\tilde{\beta}_0$, is now:

$$\tilde{\beta}_0 = \beta_0 + \mathbb{E}[x_2] + ... + \mathbb{E}[x_K]$$

Let $\beta = [\tilde{\beta}_0 \;\; \beta_1]'$ and $x = [1 \;\; x_1]$. We rewrite our structural equation as:

$$y = x\beta + \varepsilon$$

Recall the two assumptions for 2SLS to be consistent. First, we need $\mathbb{E}[z_1\varepsilon] = 0$. Secondly, we need $\mathbb{E}[z_1x_1] \neq 0$. We start by verifying the first assumption:

$$\begin{aligned}
\mathbb{E}[z_1\varepsilon] &= \mathbb{E}[z_1\left(u + \beta_2(x_2 - \mathbb{E}[x_2]) + ... + \beta_K(x_K - \mathbb{E}[x_K])\right)] \\
&= \mathbb{E}[z_1u] + \mathbb{E}[z_1(x_2 - \mathbb{E}[x_2])\beta_2] + \mathbb{E}[z_1(x_K - \mathbb{E}[x_K])] \\
&= 0
\end{aligned}$$

So assumption one is satisfied. The second assumption, $\mathbb{E}[z_1x_1] \neq 0$, is satisfied through the given information in the problem. Therefore, we know that the 2SLS estimator is consistent for the true β_1.

4.7.2 Part b

Now assume that we only have two right-hand-side variables. The structural equation is now:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + u$$

Suppose that x_1 and x_2 are both endogenous, that z_1 is exogenous and affects x_1 after controlling for x_2, and that x_2 is independent of z_1. Is the 2SLS estimator, $\hat{\beta}_1^{2SLS}$, consistent for β_1 when x_2 is treated as exogenous?

Solution

We know that x_2 is endogenous. So start by looking at how x_2 is correlated with u:

$$u = \theta_0 + \theta_2 x_2 + v_2$$

Note that the error, v_2, will have a mean of zero. Plug this linear projection into the structural equation:

$$y = (\beta_0 + \theta_0) + \beta_1 x_1 + (\beta_2 + \theta_2) x_2 + v_2$$

By treating x_2 as exogenous, we are effectively estimating this equation. To determine if $\hat{\beta}_1^{2SLS}$ is consistent for β_1, we again look at the two sufficient assumptions for consistency. First, whether the instrument is exogenous:

$$\begin{aligned}
\mathbb{E}[z_1 v_2] &= \mathbb{E}[z_1(u - \theta_0 - \theta_2 x_2)] \\
&= \mathbb{E}[z_1 u] - \mathbb{E}[z_1]\theta_0 - \mathbb{E}[z_1 x_2]\theta_2 \\
&= -\mathbb{E}[z_1]\theta_0 - \mathbb{E}[z_1]\mathbb{E}[x_2]\theta_2 \\
&= -\mathbb{E}[z_1]\left(\theta_0 - \theta_2 \mathbb{E}[x_2]\right) \\
&= -\mathbb{E}[z_1]\mathbb{E}[u] \\
&= -\mathbb{E}[z_1 u] \\
&= 0
\end{aligned}$$

Once again assumption one is sastisfied. In addition, $\mathbb{E}[z_1 x_1] \neq 0$ is given, so $\hat{\beta}_1^{2SLS}$ is consistent for β_1.

4.7.3 Part c

Suppose now that z_1 is correlated with x_2 but still uncorrelated with u. Similar to before, x_1 and x_2 are still endogenous and z_1 still affects x_1 after controlling for x_2. Consider the same structural equation as in part (b). If the x_2 is treated as exogenous, is $\hat{\beta}_1^{2SLS}$ still consistent for β_1?

Solution

Since the structure of the problem is the same, we start by looking at assumption 1 again:

$$\begin{aligned}
\mathbb{E}[z_1 v_2] &= \mathbb{E}[z_1 v_2] - \mathbb{E}[z_1]\mathbb{E}[v_2] \\
&= Cov(z_1,\ v_2) \\
&= Cov\left(z_1,\ (u - \theta_0 - \theta_2 x_2)\right) \\
&= Cov(z_1,\ u) - Cov(z_1,\ \theta_0) - \theta_2 Cov(z_1,\ x_2) \\
&= -\theta_2 Cov(z_1,\ x_2)
\end{aligned}$$

This is not zero, so the first assumption fails. The covariance term cannot cancel out elsewhere in the proof for consistency, so the 2SLS estimator will now not be consistent.

4.8 Problem Set 3, Question 1 Parts (a)-(d)

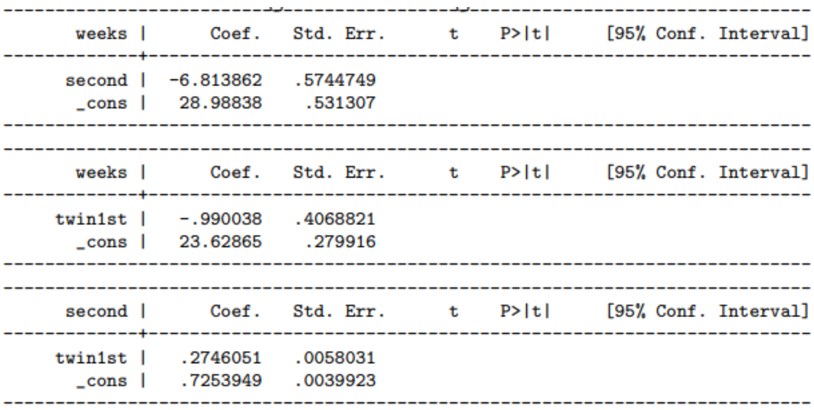
Suppose we are interested in the effect of having more than one child on the number of weeks worked by mothers:

$$weeks = \beta_0 + \beta_1 \; second + u$$

where *second* is a dummy that equals one if the mother has two or more children. We are concerned that *second* is correlated with the error term, u, so we use an instrument, *twin1st* - a dummy variable that equals one if the first pregnancy ended with twins.

4.8.1 Part a

Consider the following OLS regression outputs from Stata. Write down the OLS and IV estimate for β_1. Explain the difference between them intuitively.



```
------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      second |  -6.813862   .5744749
       _cons |   28.98838    .531307
------------------------------------------------------------------------------
------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twin1st |   -.990038   .4068821
       _cons |   23.62865    .279916
------------------------------------------------------------------------------
------------------------------------------------------------------------------
      second |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twin1st |   .2746051   .0058031
       _cons |   .7253949   .0039923
------------------------------------------------------------------------------
```

Solution

To find the OLS estimate, we look at the regression of *weeks* on *second*. The coefficient is -6.8.

To find the IV estimate, we look at an alternative IV estimator (which you will prove on PS 2):

$$\hat{\beta}^{IV} = \frac{\partial y / \partial z}{\partial x / \partial z}$$

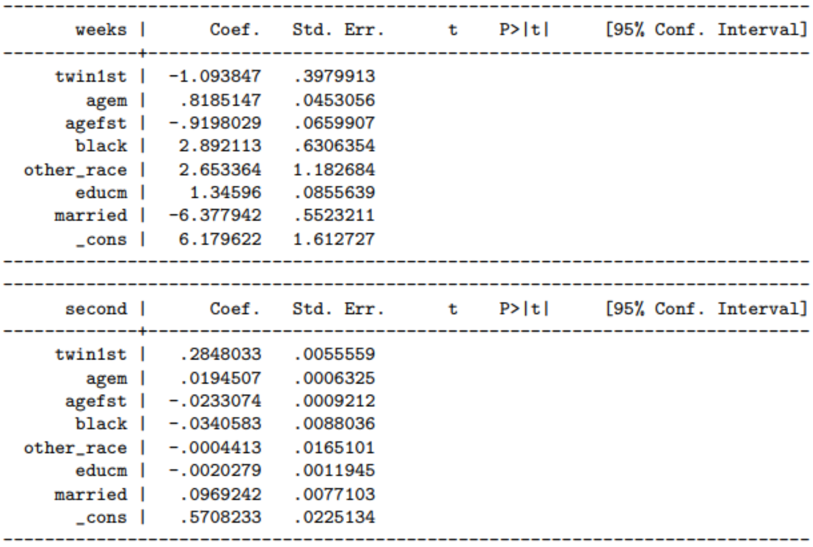
In this case, we want how *twin1st* affects *weeks* over how *twin1st* affects *second*. So:

$$
\begin{aligned}
\hat{\beta}^{IV} &= \frac{-0.99}{.2746} \\
&= -3.605
\end{aligned}
$$

The difference between the two estimates is 3.2 weeks per year. OLS overestimates the affect of more than one child on mother's weeks worked. Therefore, we can conclude there are omitted variables.

4.8.2 Part b

Consider the following two OLS regressions. What is the IV estimate for β_1 now? Why is this estimate not very different than the one found in part (a)?



```
------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twin1st |  -1.093847   .3979913
        agem |   .8185147   .0453056
      agefst |  -.9198029   .0659907
       black |   2.892113   .6306354
  other_race |   2.653364   1.182684
       educm |    1.34596   .0855639
     married |  -6.377942   .5523211
       _cons |   6.179622   1.612727
------------------------------------------------------------------------------
------------------------------------------------------------------------------
      second |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twin1st |   .2848033   .0055559
        agem |   .0194507   .0006325
      agefst |  -.0233074   .0009212
       black |  -.0340583   .0088036
  other_race |  -.0004413   .0165101
       educm |  -.0020279   .0011945
     married |   .0969242   .0077103
       _cons |   .5708233   .0225134
------------------------------------------------------------------------------
```

Solution

Using the same alternative estimator for IV:

$$
\begin{aligned}
\hat{\beta}^{IV} &= \frac{-1.09}{.285} \\
&= -3.841
\end{aligned}
$$

This estimate is not different than the estimate in part (a) because *twin1st* is probably uncorrelated with the rest of the mother's characteristics, and therefore still uncorrelated with u. So the IV estimator's consistency is not affected.

4.8.3 Part c

Based on the regressions above, do you think that the 5% t-test for the null hypothesis $H_0 : \beta_1 = 0$ will have any size distortion for weak instruments?

Solution

For no distortion in the hypothesis test on the slope coefficient, we need an F-stat of 104.7 according to Lee-Moreira-McCrary-Porter. We first need to find the t-statistic of *second* on *twin1st*:

$$\begin{aligned} t &= .2848033/.0055559 \\ &= 51.26 \end{aligned}$$

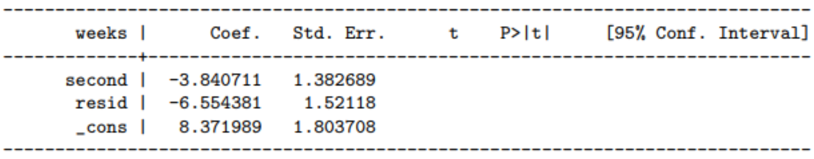
and then secondly square this t-statistic:

$$\begin{aligned} F &= (51.26)^2 \\ &= 2627.5876 \end{aligned}$$

This F-statistic is far larger than the required value of 104.7. Therefore, there should be no distortion in our test.

4.8.4 Part d

Now we create the variable *resid*, the OLS residuals from the reduced form equation in part (b). We then run OLS on an equations regressing *weeks* on *second*, *resid*, and all the controls. Test whether *second* is exogenous. Are the standard errors reported correct?



```
------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      second |  -3.840711   1.382689
       resid |  -6.554381    1.52118
       _cons |   8.371989   1.803708
------------------------------------------------------------------------------
```

Solution

Recall the control function from IV. We want to test whether the coefficient on the variable *resid* is zero. We construct the t-statistic:

$$\begin{aligned} t &= \frac{-6.554}{1.521} \\ &= -4.31 \end{aligned}$$

Therefore, we reject the null hypothesis and conclude that *second* is endogenous. We also know that if the coefficient on *resid* is not zero, then the standard errors are not correct because we are using a generated variable instead of the true value of the first stage projection residuals.

4.9 Analytical Derivations from Quiz 3

Quiz three required the Hausman test and Stock-Yogo procedure. Below are reminders of what these are.

4.9.1 Hausman Test

Recall the null hypothesis:

$$H_0 : \hat{\beta}_{2SLS} - \hat{\beta}_{OLS} = 0$$

Because we are only testing one value (i.e. β is not a vector), we can use a t-test. To conduct a t-test, we need to find the standard error of the asymptotic distribution of the difference between the $\hat{\beta}$'s. I think you saw this in class, but I will rederive the result from Hausman (1978):

Asymptotic Variance

Lemma: Consider two estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, which are both consistent and asymptotically normally distributed with $\hat{\beta}_0$ attaining the asymptotic Cramer-Rao bound so $\sqrt{n}(\hat{\beta}_0 - \beta) \sim N(0, V_0)$ and $\sqrt{n}(\hat{\beta}_1 - \beta) \sim N(0, V_1)$, where V_0 is the inverse of Fisher's information matrix. Consider $\hat{q} = \hat{\beta}_1 - \hat{\beta}_0$. Then the limiting distributions of $\sqrt{n}(\hat{\beta}_0 - \beta)$ and $\sqrt{n}\hat{q}$ have zero covariance, $C(\hat{\beta}_0, \hat{q}) = 0$.

Proof (by contradiction): Suppose $\hat{\beta}_0$ and $\hat{q}$ are not orthogonal. Define a new estimator $\hat{\beta}_2 = \hat{\beta}_0 + rA\hat{q}$, where r is a scalar and A is an arbitrary matrix to be chosen. The new estimator is consistent and asymptotically normal with asymptotic variance:

$$\begin{aligned} V(\hat{\beta}_2) &= V(\hat{\beta}_0) + rAC(\hat{\beta}_0, \hat{q}) + rC'(\hat{\beta}_0, \hat{q})A' + r^2AV(\hat{q})A' \\ V(\hat{\beta}_2) - V(\hat{\beta}_0) &= rAC(\hat{\beta}_0, \hat{q}) + rC'(\hat{\beta}_0, \hat{q})A' + r^2AV(\hat{q})A' \end{aligned}$$

Let $F(r) = V(\hat{\beta}_2) - V(\hat{\beta}_0)$. Now take the derivative with respect to r:

$$F'(r) = AC + C'A' + 2rAV(\hat{q})A'$$

Now choose $A = -C'$. Noting that C is symmetric, we get:

$$F'(r) = -2C'C + 2rC'V(\hat{q})C$$

Set $r = 0$. Then $F'(r) = -2C'C \leq 0$, meaning that we are at a maximum value. By setting $r = 0$, $F(0) = 0$. Therefore, for small r, $F(r) < 0$. But this is a contradiction, as $\hat{\beta}_0$ is asymptotically efficient. Therefore, $C = 0$.

Corollary: $V(\hat{q}) = V(\hat{\beta}_1) - V(\hat{\beta}_0)$

Proof: Since $\hat{q} + \hat{\beta}_0 = \hat{\beta}_1$:

$$\begin{aligned} V(\hat{q}) + V(\hat{\beta}_0) &= V(\hat{\beta}_1) \\ V(\hat{q}) &= V(\hat{\beta}_1) - V(\hat{\beta}_0) \end{aligned}$$

Back to the Problem

Now we can set-up the t-test:

$$t = \frac{\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}}{\frac{1}{\sqrt{n}}\sqrt{\sigma^2_{2SLS} - \sigma^2_{OLS}}}$$

4.9.2 Stock-Yogo Procedure

The Stock-Yogo Procedure tells us to first find the F-statistic for the instruments in the *first-stage projection*. If the F-statistic is higher than the critical value we are looking for, we conclude that the rank condition is satisfied. We then move on and find the t-statistic for the coefficient of interest in the *second-stage*.

Chapter 5

Generalized Method of Moments

5.1 Generalized Method of Moments Theory

The Generalized Method of Moments (GMM) comes from the method of moments that we covered in the first semester. Recall that in the method of moments, we want to set some function, $g_n(\theta)$, equal to zero and then solve for θ. As an example, let's take the IV estimator:

$$\begin{aligned}
\mathbb{E}[z_i' u_i] &= \mathbb{E}[z_i'(y_i - x_i\beta)] \\
0 &= \mathbb{E}[z_i' y_i - z_i' x_i \beta] \\
\mathbb{E}[z_i' x_i]\beta &= \mathbb{E}[z_i y_i] \qquad (*) \\
\beta &= \mathbb{E}[z_i' x_i]^{-1}\mathbb{E}[z_i y_i]
\end{aligned}$$

To estimate, we apply the analogy principle. But note that this only works if the number of unknowns equals the number of equations we have. What happens if we are over-identified (we have more equations than unknowns, so L $>$ K)? We cannot exactly identify β. To see this, label equation $(*)$ with the dimensions of the matrices:

$$\mathbb{E}[z_i' x_i]_{L\times K}\ \beta_{K\times 1} = \mathbb{E}[z_i y_i]_{L\times 1}$$

Now, we cannot invert $\mathbb{E}[z_i' x_i]$ as it is no longer a square matrix. How do we estimate β in this case? We can minimize the sum of squared residuals:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}[z'(y - x\beta)]'\ [z'(y - x\beta)]$$

This estimator will be consistent given our standard assumptions, but we are not guaranteed that this will be a particularly precise estimator. The key insight of GMM: weight observations by the inverse of their variance:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}[z'(y - x\beta)]' \, \Omega^{-1}[z'(y - x\beta)]$$

5.1.1 Criterion Function

Of course, we are not always estimating an IV model with overidentification. If we generalize the estimator, we get the GMM criterion function:

$$J_n(\beta) = n \cdot \bar{g}_n(\beta)' W \bar{g}_n(\beta)$$

where $W \equiv$ a positive definite, symmetric weighting matrix and $\bar{g}_n(\beta) \equiv$ estimator for the moment condition. Minimizing this function with respect to β gives us our GMM estimator, $\hat{\beta}_{GMM}$.

5.1.2 Example

Let's look at an IV model. First, we define our moment condition:

$$\begin{aligned} g_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} z_i' u_i \\ &= \frac{1}{n}\sum_{i=1}^{n} z_i'(y_i - x_i\beta) \end{aligned}$$

Now we put this in the criterion function:

$$\begin{aligned} J_n(\beta) &= n \cdot \frac{1}{n^2} \cdot \left(\sum_{i=1}^{n} z_i'(y_i - x_i\beta)\right)' \; W \; \left(\sum_{i=1}^{n} z_i'(y_i - x_i\beta)\right) \\ &= \frac{1}{n} \cdot \left(\sum_{i=1}^{n} z_i'(y_i - x_i\beta)\right)' \; W \; \left(\sum_{i=1}^{n} z_i'(y_i - x_i\beta)\right) \end{aligned}$$

and take the first-order conditions with respect to the β vector:

$$\begin{aligned} \frac{\partial J}{\partial \beta} &= 0 \\ \frac{2}{n}(-z'x)'W(z'(y - x\beta)) &= 0 \\ \frac{2}{n}(x'zW(z'(x\beta - y)) &= 0 \\ x'zWz'x\beta &= x'zWz'y \\ \hat{\beta} &= (x'zWz'x)^{-1}(x'zWz'y) \end{aligned}$$

If we choose $W = (z'z)^{-1}$, then:

$$\begin{aligned}
\hat{\beta} &= (x'z(z'z)^{-1}z'x)^{-1}(x'z(z'z)^{-1}z'y) \\
&= (x'P_z x)^{-1}(x'P_z y) \\
&= \hat{\beta}_{2SLS}
\end{aligned}$$

And if we are exactly identified ($K = L$):

$$\begin{aligned}
\hat{\beta} &= (x'zWz'x)^{-1}(x'zWz'y) \\
&= (z'x)^{-1}W^{-1}(x'z)^{-1}x'zWz'y \\
&= (z'x)^{-1}(z'y) \\
&= \hat{\beta}_{IV}
\end{aligned}$$

5.1.3 Consistency

To prove consistency of the GMM estimator (specifically for the IV case), we first define two structures:

$$\begin{aligned}
Q_{L\times R} &\equiv \mathbb{E}\left[\frac{\partial g_i}{\partial \beta}\right]' \\
&= \mathbb{E}[x'z]' \\
&= \mathbb{E}[z'x]
\end{aligned}$$

$$\begin{aligned}
\Omega_{L\times L} &\equiv \mathbb{E}[g_i(\beta)g_i(\beta)'] \\
&= \mathbb{E}[z'uu'z]
\end{aligned}$$

We start from the identified $\hat{\beta}$:

$$\begin{aligned}
\hat{\beta} &= (x'zWz'x)^{-1}(x'zWz'y) \\
&\xrightarrow{P} (Q'WQ)^{-1}(Q'W\mathbb{E}[z'y]) \\
&= (Q'WQ)^{-1}(Q'W\mathbb{E}[z'(x\beta + u)]) \\
&= \beta + (Q'WQ)^{-1}(Q'W\mathbb{E}[z'u]) \\
&= \beta
\end{aligned}$$

So the GMM estimator is consistent (similar proofs can be done for the other regression types).

5.1.4 Asymptotic Normality

$$\begin{aligned}
\hat{\beta}_{GMM} - \beta &= (x'zWz'x)^{-1}(x'zWz'u) \\
\sqrt{n}(\hat{\beta}_{GMM} - \beta) &= (x'zWz'x)^{-1}(x'zW\sqrt{n}z'u) \\
&\xrightarrow{d} (Q'WQ)^{-1}Q'W\, N\left(0, \mathbb{E}[z'uu'z]\right) \\
&= N\left(0, (Q'WQ)^{-1}Q'W\Omega WQ(Q'WQ)^{-1}\right)
\end{aligned}$$

Looking at this asymptotic variance, we can see that we want to choose $W = \Omega^{-1}$. Doing this will cause the asymptotic variance to shrink:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{GMM} - \beta) &\xrightarrow{d} N\left(0, (Q'\Omega^{-1}Q)^{-1}Q'\Omega^{-1}\Omega\Omega^{-1}Q(Q'\Omega^{-1}Q)^{-1}\right) \\
&= N\left(0, (Q'\Omega^{-1}Q)^{-1}Q'\Omega^{-1}Q(Q'\Omega^{-1}Q)^{-1}\right) \\
&= N\left(0, (Q'\Omega^{-1}Q)^{-1}\right)
\end{aligned}$$

We can prove that this asymptotic variance is better than any other choice for W. This amounts to proving that $AVAR\mid_W - AVAR\mid_{\Omega^{-1}}$ is PSD. To begin with, we use the property that $A - B$ is PSD iff $B^{-1} - A^{-1}$ is PSD:

$$\begin{aligned}
\left(AVAR\mid_{\Omega^{-1}}\right)^{-1} - \left(AVAR\mid_W\right)^{-1} &= (Q'\Omega^{-1}Q) - (Q'WQ)(Q'W\Omega WQ)^{-1}(Q'WQ) \\
&= Q'\left[\Omega^{-1} - WQ(Q'W\Omega WQ)^{-1}Q'W\right]Q \\
&= Q'\Psi\left[I - \Psi^{-1}WQ(Q'W\Psi^{-1}\Psi^{-1}WQ)^{-1}Q'W\Psi^{-1}\right]\Psi Q \\
&= Q'\Psi\left[I - D(D'D)^{-1}D'\right]\Psi Q \\
&= Q'\Psi[M_D]\Psi Q
\end{aligned}$$

This is PSD, so setting $W = \Omega^{-1}$ is the ideal weighting matrix. Fortunately, this is always the best weighting matrix. Also fortunately for us, $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \bar{g}_i(\beta)\bar{g}_i(\beta)' \xrightarrow{P} \Omega$.

In practice, when dealing with highly non-linear equations, we set $W = I$ and iterate GMM letting the algorithm eventually converge to the ideal W.

5.1.5 GMM for Time Series (from Hayashi Chapter 6)

Sometimes we do not have cross-sectional data. Instead, we could have longitudinal data. We need to prove asymptotic properties for GMM in this situation. To do so, we need a different central limit theorem.

Gordin's CLT for zero-mean ergodic stationary processes:

Suppose $\{y_t\}$ is stationary and ergodic[5] and suppose Gordin's condition is satisfied. Then $\mathbb{E}[y_t] = 0$, the autocovariances $\{\gamma_t\}$ are absolutely summable, and

$$\sqrt{n}\bar{y} \xrightarrow{d} N\left(0, \sum_{j=-\infty}^{\infty} \gamma_j\right)$$

A process meets Gordin's condition iff it satisfies three requirements:

(a) $\mathbb{E}[y_t^2] < \infty$. This condition simply says that the variance of the series exists.

(b) $\mathbb{E}[y_t \mid y_{t-j}, y_{t-j-1}, ...] \underset{j\to\infty}{\xrightarrow{m.s.}} 0$. This condition says that the conditional mean of the process is zero. It also implies that the unconditional mean is zero.

Before we go to the third requirement, we need to rewrite y_t:

$$\begin{aligned} y_t &= y_t - (\mathbb{E}[y_t \mid I_{t-1}] - \mathbb{E}[y_t \mid I_{t-1}]) - (\mathbb{E}[y_t \mid I_{t-2}] - \mathbb{E}[y_t \mid I_{t-2}]) \\ &\quad - ... - (\mathbb{E}[y_t \mid I_{t-j}] - \mathbb{E}[y_t \mid I_{t-j}]) \\ &= (y_t - \mathbb{E}[y_t \mid I_{t-1}]) + (\mathbb{E}[y_t \mid I_{t-1}] - \mathbb{E}[y_t \mid I_{t-2}]) \\ &\quad + ... + (\mathbb{E}[y_t \mid I_{t-j+1}] - \mathbb{E}[y_t \mid I_{t-j}]) + \mathbb{E}[y_t \mid I_{t-j}] \\ &= (r_{t,0} + r_{t,1} + ... + r_{t,j-1}) + \mathbb{E}[y_t \mid y_{t-j}, y_{t-j-1}...] \end{aligned}$$

where $r_{tk} = \mathbb{E}[y_t \mid I_{t-k}] - \mathbb{E}[y_t \mid I_{t-k-1}]$. Using assumption (b):

$$\begin{aligned} y_t &= (r_{t,0} + r_{t,1} + ... + r_{t,j-1}) \\ &= \sum_{j=0}^{\infty} r_{tj} \end{aligned}$$

So y_t is a telescoping sum.

(c) $\sum_{j=0}^{\infty}(\mathbb{E}[r_{tj}^2]^{1/2}) < \infty$. This conditions says that shocks that occurred long ago should not influence the current value of y too much.

We can easily extend this to the multivariate case:

$$\sqrt{T}(\bar{\boldsymbol{y}}) \xrightarrow{d} N\left(\mathbf{0}, \sum_{j=-\infty}^{\infty} \boldsymbol{\Gamma}_j\right)$$

where $\boldsymbol{\Gamma}_j = \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}'_{t-j}]$

[5] Ergodicity intuitively requires that two random variables sufficiently far apart in the sequence must be nearly independent.

Applied to GMM

Define $\boldsymbol{g_t}$ as the moment condition. Assuming our time series is ergodic and stationary, we can apply Gordon's CLT:

$$\begin{aligned}\sqrt{T}\bar{\boldsymbol{g}} &\xrightarrow{d} N\left(\mathbf{0},\ \sum_{j=-\infty}^{\infty}\boldsymbol{\Gamma}_j\right)\\ &= N\left(\mathbf{0}, \boldsymbol{\Gamma}_0 + \sum_{j=1}^{\infty}(\boldsymbol{\Gamma}_j + \boldsymbol{\Gamma}'_j)\right)\end{aligned}$$

To estimate the asymptotic variance, we replace $\boldsymbol{\Gamma}_j$ with $\hat{\boldsymbol{\Gamma}}_j = \frac{1}{T}\sum_{t=j+1}^{T}\hat{\boldsymbol{g}}_t\hat{\boldsymbol{g}}'_{t-j}$:

$$\widehat{AVAR}^{\boldsymbol{g}} = \hat{\boldsymbol{\Gamma}}_0 + \sum_{j=1}^{\infty}(\hat{\boldsymbol{\Gamma}}_j + \hat{\boldsymbol{\Gamma}}'_j)$$

But note that we cannot estimate for infinity periods and we do not want to weight all covariances the same (the further away the autocovariance is, the less we want it to impact our estimate). Therefore, we choose a number of periods to estimate (say 12 periods) and introduce a *kernel* to weight the autocovariances:

$$\mathcal{K}_j = 1 - \frac{j}{L+1}$$

where $j \equiv$ current lag and $L \equiv$ maximum lag we include. Putting this into $\widehat{AVAR}$ gives us the Newey-West estimator:

$$\widehat{AVAR}^{\boldsymbol{g}}_{NW} = \hat{\boldsymbol{\Gamma}}_0 + \frac{1}{T}\sum_{j=1}^{L}\sum_{t=j+1}^{T}\mathcal{K}_j\left(\hat{\boldsymbol{g}}_t\hat{\boldsymbol{g}}'_{t-j} + \hat{\boldsymbol{g}}_{t-j}\hat{\boldsymbol{g}}'_t\right)$$

Now recall the Criterion Function:

$$J = T\bar{\boldsymbol{g}}'\boldsymbol{W}\bar{\boldsymbol{g}}$$

We can use the Delta Method to find the asymptotic variance of the parameters estimated through GMM (equivalent to the normal way we find the asymptotic variance). The asymptotic distribution is thus:

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0},\ \left(\frac{\partial\boldsymbol{g}}{\partial\boldsymbol{\beta}}\right)'\boldsymbol{W}\left(\frac{\partial\boldsymbol{g}}{\partial\boldsymbol{\beta}}\right)\right)$$

We know that the optimal weighting matrix is the inverse of the variance from the CLT. In this case, that is our Newey-West estimator.

$$\widehat{AVAR} = \left(\frac{\partial\hat{\boldsymbol{g}}}{\partial\hat{\boldsymbol{\beta}}}\right)'\left(\widehat{AVAR}^{\hat{\boldsymbol{g}}}_{NW}\right)^{-1}\left(\frac{\partial\hat{\boldsymbol{g}}}{\partial\hat{\boldsymbol{\beta}}}\right)$$

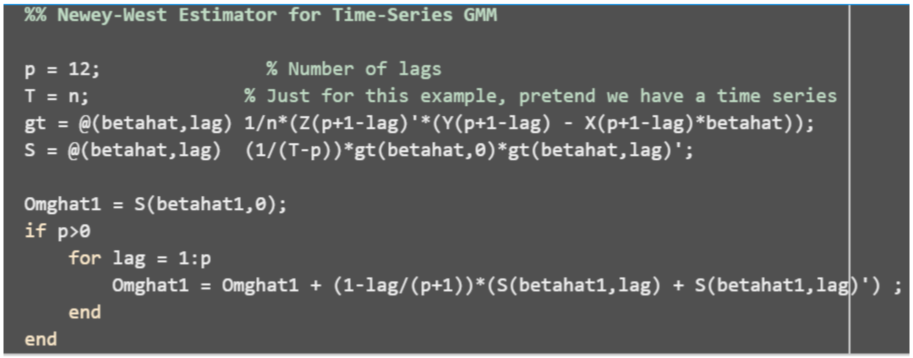
Note that the $\hat{\boldsymbol{\beta}}$ estimator remains the same as before (using IV as an example again):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x'z}\widehat{\boldsymbol{W}}\boldsymbol{z'x})^{-1}(\boldsymbol{x'z}\widehat{\boldsymbol{W}}\boldsymbol{z'y})$$

where we plug the inverse of the Newey-West estimator in for $\widehat{\boldsymbol{W}}$.

5.2 Matlab Help

In class, we went over an example using IV. In Matlab, we found $\hat{\beta}_{GMM}$ using a one-shot method, a two-step method, and an iterated method. We also talked about how to implement the Newey-West estimator using function handles. Below is a screenshot of the Newey-West code.



```
%% Newey-West Estimator for Time-Series GMM

p = 12;               % Number of lags
T = n;              % Just for this example, pretend we have a time series
gt = @(betahat,lag) 1/n*(Z(p+1-lag)'*(Y(p+1-lag) - X(p+1-lag)*betahat));
S = @(betahat,lag)  (1/(T-p))*gt(betahat,0)*gt(betahat,lag)';

Omghat1 = S(betahat1,0);
if p>0
    for lag = 1:p
        Omghat1 = Omghat1 + (1-lag/(p+1))*(S(betahat1,lag) + S(betahat1,lag)') ;
    end
end
```

Figure 5.1

Chapter 6

Binary Response Models

6.1 Binary Response Theory (Wooldridge Chapter 15)

We have primarily focused on cases where the left-hand side of our regressions, usually denoted as y_i, is a continuous random variable. Now, we will develop theoretical models for the case of a discrete y_i, particularly when y_i is binary (takes on the value of either 0 or 1). We will still have right-hand side explanatory variables, $\boldsymbol{x} = [x_1, x_2, ..., x_K]$.

With binary response models, we usually want to predict the probability of $y_i = 1$, also called the **response probability**:

$$p(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$$

for the various values of $\boldsymbol{x}$. In general, if x_j is continuous, the partial effect of x_j on the probability that $y_i = 1$ is:

$$\frac{\partial P(y = 1|\boldsymbol{x})}{\partial x_j} = \frac{\partial p(\boldsymbol{x})}{\partial x_j}$$

Note that if we have two variables on the right-hand side that take the form of $x_1 = z$ and $x_2 = z^2$, and we want the impact of z on $P(y = 1|\boldsymbol{x})$, then we need to take the derivative with respect to both x_1 and x_2. For a binary x_K, the partial effect is:

$$p(x_1, x_2, ..., x_{K-1}, 1) - p(x_1, x_2, ..., x_{K-1}, 0)$$

Before we go further in studying binary response models, four properties of Bernoulli random variables:

(1) $P(y = 1|\boldsymbol{x}) = p(\boldsymbol{x})$

(2) $P(y = 0|\boldsymbol{x}) = 1 - p(\boldsymbol{x})$

(3) $\mathbb{E}[y|\boldsymbol{x}] = p(\boldsymbol{x})$

(4) $Var(y|\boldsymbol{x}) = p(\boldsymbol{x})(1 - p(\boldsymbol{x}))$

6.1.1 Linear Probability Model

A linear probability model uses the same structure that we have been using up to this point. Namely, it takes the form of a standard regression equation:

$$P(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K$$

This form is convenient, as the partial effect of x_1, assuming it is not functionally related to any of the other explanatory variables, is β_1.

We must note that the linear probability model can return estimates outside of the unit interval. Since it is linear, the regression line will carry past $y = 0$ and past $y = 1$. This model should therefore be thought of as an approximation for the response probability.

To determine whether the linear probability model is the right estimation tool, we need to look at the conditional mean and variance of y. Because y is Bernoulli:

$$\begin{aligned} \mathbb{E}[y|\boldsymbol{x}] &= \beta_0 + \beta_1 x_1 + ... + \beta_K x_K \\ Var(y|\boldsymbol{x}) &= \boldsymbol{x\beta}(1 - \boldsymbol{x\beta}) \end{aligned}$$

Because the conditional expectation is linear, we know that the OLS regression of y on $[\mathbf{1}, \boldsymbol{x}]$ gives a consistent estimator for $\boldsymbol{\beta}$. But, because the variance term includes $\boldsymbol{x}$, we know that the variance is heteroskedastic. Therefore, we know that we need to use heteroskedasticity-robust standard errors. But because we know the analytic formula for the variance, we can use Weighted Least Squares:

(1) Run standard OLS for $y_i = \boldsymbol{x_i\beta} + \varepsilon_i$

(2) Find the fitted left-hand side variables $\hat{y}_i$ (ensure that all $0 < \hat{y}_i < 1$ for all $i \in I$)

(3) Derive $\hat{\sigma}_i = [\hat{y}_i(1 - \hat{y}_i)]^{1/2}$

(4) Find $\boldsymbol{\beta}^*$ from $\frac{y_i}{\hat{\sigma}_i} = \frac{\boldsymbol{x}_i}{\hat{\sigma}_i}\boldsymbol{\beta} + u_i$ using OLS

(5) Test $\boldsymbol{\beta^*}$ with the normal OLS standard errors in step (4)

So when is the linear probability model good? The basic answer is "when the conditional expectation is linear," which reliably occurs when:

(1) $\boldsymbol{x}$ contains dummy variables for mutually exclusive and exhaustive categories (the model is saturated) –or–

(2) We have a latent variable model where $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x\beta}$. To see why this is the case, start from the latent variable model:

$$y_i^* = \boldsymbol{x_i\beta} + u_i \text{ where } y_i = \mathbb{1}(y_i^* > 0)$$

Then:

$$\begin{aligned} P(y=1|\boldsymbol{x}) &= P(\boldsymbol{x\beta} + u > 0|\boldsymbol{x}) \\ &= P(u > -\boldsymbol{x\beta}|\boldsymbol{x}) \\ &= 1 - G(-\boldsymbol{x\beta}) && (*) \\ &= G(\boldsymbol{x\beta}) && (**) \end{aligned}$$

where $G(\boldsymbol{x\beta})$ is the underlying distribution we assume for the error term. So $\mathbb{E}[y|\boldsymbol{x}] = G(\boldsymbol{x\beta})$. For a linear probability model to be the correct specification, the conditional expectation must be linear, and so we need some distribution that will return $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{x\beta}$.

Note: we must have a symmetric (around zero) distribution for $G(\boldsymbol{x\beta})$ to go from equation $(*)$ to equation $(**)$. One such example is the uniform distribution centered at zero. If we don't have a symmetric distribution centered at zero, then the linear probability model will still work and the conditional expectation will still be linear. The proof is just more involved.

But what happens if neither of these two scenarios occur (or we do not want to make stringent assumptions or keep adding dummy variables)? Then we may want to consider probit and logit models.

6.1.2 Probit and Logit

For the probit model, we set $G(\boldsymbol{x\beta}) = \boldsymbol{\Phi}(\boldsymbol{x\beta})$, where $\boldsymbol{\Phi}$ stands for the cdf of a standard normal distribution. For the logit model, we set $G(\boldsymbol{x\beta}) = \boldsymbol{\Lambda}(\boldsymbol{x\beta})$, where $\boldsymbol{\Lambda}$ stands for the cdf of a standard logistic distribution.

New distributions means that partial effects will have a different calculation. For continuous right-hand side variables, the partial effect is:

$$\frac{\partial p(\boldsymbol{x})}{\partial x_j} = g(\boldsymbol{x\beta})\beta_j$$

where $g(\boldsymbol{x\beta})$ is the derivative of the cdf. If x_K is binary, then the partial effect is:

$$G(\beta_1 + \beta_2 x_2 + ... + \beta_{k-1}x_{K-1} + \beta_K) - G(\beta_1 + \beta_2 x_2 + ... + \beta_{K-1}x_{K-1})$$

And if x_K is discrete, then:

$$G(\beta_1 + \beta_2 x_2 + ... + \beta_{k-1}x_{K-1} + \beta_K(c_K + 1)) - G(\beta_1 + \beta_2 x_2 + ... + \beta_{K-1}x_{K-1} + \beta_K c_K)$$

where c_K is some value for variable x_K.

6.1.3 Estimating Probit and Logit

We know the underlying distribution for y, so the best method we can use is maximum likelihood estimation. We first define the conditional likelihood for i:

$$\begin{aligned} f(y_i|x_i,\beta) &= \begin{cases} G(x_i\beta) & y_i = 1 \\ 1 - G(x_i\beta) & y_i = 0 \end{cases} \\ &= G(x_i\beta)^{y_i}\left(1 - G(x_i\beta)\right)^{1-y_i} \end{aligned}$$

Now we define the conditional log-likelihood function for one i:

$$\begin{aligned} \ell_i(\beta) &= ln\left(f(y_i|x_i,\beta)\right) \\ &= y_i ln\left(G(x_i\beta)\right) + (1-y_i) ln\left(1 - G(x_i\beta)\right) \end{aligned}$$

Next, aggregate this across all observations and take the average:

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{i=1}^{n}\left[y_i ln\left(G(x_i\beta)\right) + (1-y_i) ln\left(1 - G(x_i\beta)\right)\right] \\ &= \sum_{i=1}^{n} \ell_i(\beta) \\ Q_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} \ell_i(\beta) \end{aligned}$$

Our estimator $\hat{\beta}^{MLE} = \underset{\beta}{\text{argmax}}\;\; Q_n(\beta)$ (it also maximizes $\mathcal{L}(\beta)$). We assume that there exists a unique β that maximizes $Q_n(\beta)$.

For consistency, we only need to assume that $Q_n(\beta) \xrightarrow{P} Q(\beta)$. This assumption implies that $\underset{\beta}{\text{argmax}}\; Q_n(\beta) \xrightarrow{P} \underset{\beta}{\text{argmax}}\; Q(\beta)$.

For asymptotic normality, we Taylor approximate around the score function:

$$\begin{aligned} \frac{\partial Q_n(\beta)}{\partial \beta} &= s_n(\hat{\beta}) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i(\beta)}{\partial \beta} &= 0 \\ s_n(\beta) + \frac{\partial s_n(\beta)}{\partial \beta}(\hat{\beta} - \beta) &= 0 && \text{(Taylor approx.)} \\ (\hat{\beta} - \beta) &= -s_n(\beta)\left(\frac{\partial s_n(\beta)}{\partial \beta}\right)^{-1} \\ \sqrt{n}(\hat{\beta} - \beta) &= -\sqrt{n} s_n(\beta)\left(\frac{\partial s_n(\beta)}{\partial \beta}\right)^{-1} \end{aligned}$$

Now note that because β maxes $Q(\beta)$ by definition:

$$\begin{aligned}\frac{\partial Q(\beta)}{\partial\beta} &= \frac{\partial}{\partial\beta}\mathbb{E}[\ell_i(\beta)]\\ &= \mathbb{E}\left[\frac{\partial}{\partial\beta}\ell_i(\beta)\right]\\ &= \mathbb{E}\left[s_i(\beta)\right]\\ &= 0\end{aligned}$$

Then we can apply the multivariate central limit theorem to $s_n(\beta)$:

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) &\xrightarrow{d} N\left(0,\mathbb{E}[s_i(\beta)s_i(\beta)']\right)\cdot\frac{-\partial}{\partial\beta}\mathbb{E}[s_i(\beta)]^{-1}\\ &= N\left(0,\mathbb{E}[s_i(\beta)s_i(\beta)']\right)\cdot-\mathbb{E}[H_i(\beta)]^{-1}\\ &= N\left(0,\mathbb{E}[H_i(\beta)]^{-1}\mathbb{E}[s_i(\beta)s_i(\beta)']\;\mathbb{E}[H_i(\beta)]^{-1}\right)\end{aligned}$$

where $H_i(\beta)$ denotes the Hessian. Remember back to first semester. If the model is correctly specified, then the expected Hessian is equal to the Fisher information matrix (score matrix “squared”). Imposing this, we note that the maximum likelihood estimator reaches the Cramér-Rao lower bound:

$$\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{d} N\left(0,\mathbb{E}[H_i(\beta)]^{-1}\right)$$

Note that in the real world, we do not know if we correctly specified the model, so we always use the full asymptotic variance to estimate (the robust sandwich estimator).

We now turn to the analytical expressions for the score function and Hessian, and to the expression for the sandwich estimator:

$$\begin{aligned}s_i(\beta) &= \frac{\partial}{\partial\beta}\left[y_i ln\left(G(x_i\beta)\right)+(1-y_i)ln\left(1-G(x_i\beta)\right)\right]\\ &= \frac{g_i(x_i\beta)[y_i-G(x_i\beta)]x_i'}{G(x_i\beta)(1-G(x_i\beta))}\end{aligned}$$

$$\begin{aligned}H_i(\beta) &= \frac{\partial s_i(\beta)}{\partial\beta}\\ &= \frac{-g_i(x_i\beta)^2x_i'x_i}{G(x_i\beta)(1-G(x_i\beta))}\end{aligned}$$

$$\widehat{AVAR}=\left(-\frac{1}{n}\sum_{i=1}^{n}H_i(\hat{\beta})\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}s_i(\hat{\beta})s_i(\hat{\beta})'\right)\left(-\frac{1}{n}\sum_{i=1}^{n}H_i(\hat{\beta})\right)^{-1}$$

6.2 Previous Quiz Question

This question tests your understanding of the probit model.

6.2.1 Part a

Write down the key components of the standard probit model seen in class for binary y_i, vector of explanatory variables x_i, and vector of parameters β. Your answer must cover: (i) model for latent y_i^*; (ii) model for observed y_i; (iii) likelihood function of a single observation; (iv) log-likelihood function of entire sample; (v) definition of the maximum likelihood estimator.

Solution

We go down the list:

(i) $y_i^* = x_i\beta + u_i$ where $u_i|x_i \sim N(0,1)$.

(ii) $y_i = \mathbb{1}\{y_i^* \geq 0\}$.

(iii) For a probit model, $G(x_i\beta) = \Phi(x_i\beta)$. So the likelihood of a single observation is:

$$f(y_i|x_i,\beta) = \Phi(x_i\beta)^{y_i}(1-\Phi(x_i\beta))^{1-y_i}.$$

(iv) We take the log of the likelihood function:

$$\begin{aligned}\ell_i(\beta) &= ln(f(y_i|x_i,\beta)) \\ &= y_i ln(\Phi(x_i\beta)) + (1-y_i)ln(1-\Phi(x_i\beta))\end{aligned}$$

Then aggregate it over the entire sample:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ell_i(\beta)$$

(v) The maximum likelihood estimator maximizes the aggregate log-likelihood:

$$\hat{\beta}^{MLE} = \underset{\beta}{\operatorname{argmax}}\, \mathcal{L}_n(\beta)$$

6.2.2 Part b

Derive the score function for the individual observation, that is, $s_i(\beta)$. Prove that $\mathbb{E}[s_i(\beta)] = 0$ at the true β.

Solution

We first take the derivative of the individual log-likelihood function:

$$\begin{aligned}s_i(\beta) = \frac{\partial \ell_i(\beta)}{\partial \beta} &= \frac{y_i\phi_i x_i'}{\Phi_i} + \frac{(1-y_i)(-\phi_i)x_i'}{1-\Phi_i} \\ &= \frac{\phi_i(y_i-\Phi_i)x_i'}{\Phi_i(1-\Phi_i)}\end{aligned}$$

Now we take the expectation of $s_i(\beta)$:

$$\begin{aligned}
\mathbb{E}[s_i(\beta)] &= \mathbb{E}_x\left[\mathbb{E}[s_i(\beta)|x_i]\right] \\
&= \mathbb{E}_x\left[\frac{\phi_i \mathbb{E}[y_i - \Phi_i|x_i]x_i'}{\Phi_i(1-\Phi_i)}\right] \\
&= \mathbb{E}_x\left[\frac{\phi_i\left(\mathbb{E}[y_i|x_i] - \mathbb{E}[\Phi_i|x_i]\right)x_i'}{\Phi_i(1-\Phi_i)}\right] \\
&= \mathbb{E}_x\left[\frac{\phi_i\left(\Phi_i - \Phi_i\right)x_i'}{\Phi_i(1-\Phi_i)}\right] \\
&= 0
\end{aligned}$$

6.2.3 Part c

Consider an OLS regression of y_i on x_i. Give an example of a distribution for the error term of the latent model that makes the OLS estimator here and the MLE both consistent for the same parameters.

Solution

We need a distribution that is symmetric around zero and that yields a linear conditional expectation function. Consider a uniform distribution:

$$\begin{aligned}
u_i|x_i &\sim U[-1/2, 1/2] \\
F_u(u) &= \frac{u + 1/2}{1/2 + 1/2} \\
&= u + \frac{1}{2}
\end{aligned}$$

First, we check if this is symmetric around zero. That is, does $F_u(-u) = 1 - F_u(u)$?

$$\begin{aligned}
1 - F_u(u) &= 1 - u - \frac{1}{2} \\
&= -u + \frac{1}{2} \\
&= F_u(-u)
\end{aligned}$$

Then we check to see if the conditional expectation is linear:

$$\begin{aligned}
\mathbb{E}[y_i|x_i] &= P(y_i = 1|x_i) \\
&= F_u(x_i\beta) \\
&= x_i\beta + \frac{1}{2}
\end{aligned}$$

This is linear, so OLS is consistent for the same parameter as MLE, excluding the constant term.

6.2.4 Part d

Start again with the probit model. Suppose that x_i is a vector of discrete random varaibles. Demonstrate how to consistently estimate β using an OLS regression. Explain how to obtain robust standard errors for your $\hat{\beta}$.

Solution

First, we create a dummy variable for all possible combinations of our right-hand side variables. Take a simple example where we have two right-hand side variables: [1, male]. Then we will have two dummies: one for the combination (1,0) and one for the combination (1,1). In general, we will have dummies $D_{i1}, ..., D_{iP}$ in the model. Run OLS on the equation $y_i = \sum_{j=1}^{P} \gamma_j D_{ij} + \varepsilon_i$. This is a fully saturated linear probability model. Therefore, $\mathbb{E}[y_i|x_i] = \Phi(x_i\beta) = \sum_{j=1}^{P} \gamma_j D_{ij}$. Then:

$$\begin{aligned} x_j\beta &= \Phi^{-1}(\gamma_j \\ x_j' x_j \beta &= x_j' \Phi^{-1}(\gamma_j) \\ \sum_{j=1}^{P} x_j' x_j \beta &= \sum_{j=1}^{P} x_j' \Phi^{-1}(\gamma_j) \\ \hat{\beta} &= \left(\sum_{j=1}^{P} x_j' x_j \right) \left(\sum_{j=1}^{P} x_j' \Phi^{-1}(\gamma_j) \right) \end{aligned}$$

This is consistent as $\hat{\gamma}$ is consistent for γ. Standard errors for $\hat{\beta}$ can be constructed using the delta method.

Chapter 7

Censoring and Selection

7.1 The First Problem: Data Censoring

Data censoring arises when observed data is partially continuous, but for some reason also consists of a mass of observations at a point. Two of the commons reasons are:

1. **Top-Coding:** Observations above an often arbitrary maximum value are recorded as that maximum value.
   - As an example, think of income. The Current Population Survey (CPS) sets a maximum limit (currently $200,000). Any observation of income above $200,000 will be recorded as $200,000.
2. **Corner-Solutions:** Observations bunch at the end-points of the support.
   - Here, think of donations to religious charities. A large number of observations will not donate to religious charities, so a mass of observations will be at $0.

So what exactly is the problem? Recall the issue with the linear probability model. To properly apply OLS, we need the underlying conditional expectation function to be linear. With data-censoring, the main issue is that the observed data is not generated from a linear process. How do we deal with this problem?

7.2 Tobit Models

We begin as we did with the probit and logit models - a latent variable set-up. For the mid-censored case (i.e. where the data process has a lower bound and an upper bound), our observed outcome variable is generated as follows:

$$y_i = \begin{cases} y_U & y_i^* \geq y_U \\ y_i^* & y_L < y_i^* < y_U \\ y_L & y_i^* \leq y_L \end{cases}$$

where we observe y_i but not y_i^*. Our goal, though, is to estimate the latent variable model $y_i^* = x\beta + u$, with $u \mid_{x,y_L,y_U} \sim N(0, \sigma^2)$ assumed. We can use maximum likelihood estimation (because we assume the underlying true distribution). But due to censoring, we need to think hard about how we construct the likelihood function. Consider the following figure:

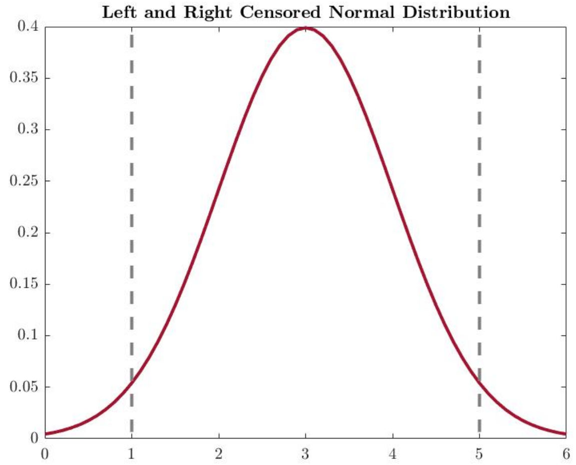


Figure 7.1: This standard normal distribution has been censored, with all values less than one being assigned the value one and all values greater than five being assigned the value five.

Let's first think about the case of left-censoring (when $y_i = y_L$). Then we want to look at:

$$\begin{aligned} P(y_i \le y_L \mid x_i, y_L, y_U) &= P(y_i^* \le y_L \mid x_i, y_L, y_U) \\ &= P(x_i\beta + u_i \le y_L \mid x_i, y_L, y_U) \\ &= P\left(\frac{u_i}{\sigma} \le \frac{y_L - x_i\beta}{\sigma} \middle| x_i, y_L, y_U\right) \\ &= \Phi\left(\frac{y_L - x_i\beta}{\sigma}\right) \end{aligned}$$

Then we look at the case of right-censoring (when $y_i = y_U$):

$$P(y_i \ge y_U \mid x_i, y_L, y_U) = 1 - \Phi\left(\frac{y_U - x_i\beta}{\sigma}\right)$$

Then we look at the middle portion where there is no censoring:

$$\begin{aligned} P(y_L < y_i < y_U \mid x_i, y_L, y_U) &= \Phi\left(\frac{y_i - x_i\beta}{\sigma}\right) \\ p(y_i \mid x_i, y_L, y_U) &= \phi\left(\frac{y_i - x_i\beta}{\sigma}\right) \cdot \frac{1}{\sigma} \end{aligned}$$

Constructing a likelihood function for an individual observation consists of combining all three cases using indicator functions:

$$
\begin{aligned}
f(y_i \mid x_i, y_L, y_U, \beta, \sigma^2) = \left[\Phi\left(\frac{y_L - x_i\beta}{\sigma}\right)\right]^{\mathbb{1}\{y_i = y_L\}} &\times \left[\phi\left(\frac{y_i - x_i\beta}{\sigma}\right) \cdot \frac{1}{\sigma}\right]^{\mathbb{1}\{y_i \in (y_L, y_U)\}} \\
&\times \left[1 - \Phi\left(\frac{y_U - x_i\beta}{\sigma}\right)\right]^{\mathbb{1}\{y_i = y_U\}}
\end{aligned}
$$

From here, we know what to do: take the log, sum up over all the observations, find the score vector, and lastly solve for $\hat{\beta}_{MLE}$. To find the partial effect, we need to do a little bit more work. We need to find the conditional expectation function:

$$
\begin{aligned}
\mathbb{E}[y \mid x] = y_L \cdot P(y_i = y_L \mid x) + \mathbb{E}[y \mid x, y_i \in \{y_L, y_U\}] \cdot P(y_L < y_i < y_U \mid x) \\
+ y_U \cdot P(y_i = y_U \mid x)
\end{aligned}
$$

then take the partial of the conditional expectation with respect to x_j. To see how this works, let's simplify the problem by looking at a special case called the "standard Tobit." The selection equation looks as follows:

$$
\begin{aligned}
y_i &= \begin{cases} y^* & y^* \geq 0 \\ 0 & y^* \leq 0 \end{cases} \\
&= \max\{0, y^*\}
\end{aligned}
$$

This selection equation returns the following graph:

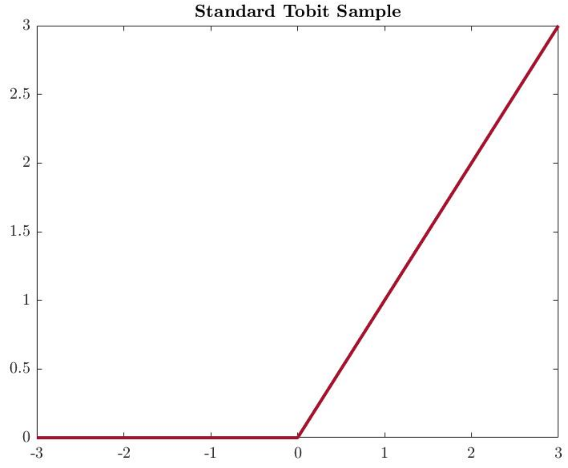


Figure 7.2: The standard Tobit sample.

Note that there is no top-censoring here. In addition, because the bottom-censored value is zero, the conditional expectation function boils down to one term:

$$\mathbb{E}[y \mid x] = \mathbb{E}[y \mid x, y_i > 0] \cdot P(y_i > 0 \mid x)$$

Let's first look at the probability that y_i is greater than zero:

$$\begin{aligned} P(y_i > 0 \mid x) &= P(x_i\beta + u_i > 0 \mid x) \\ &= P\left(\frac{u_i}{\sigma} > -\frac{x_i\beta}{\sigma}\middle| x\right) \\ &= 1 - \Phi\left(-\frac{x_i\beta}{\sigma}\right) \\ &= \Phi\left(\frac{x_i\beta}{\sigma}\right) \end{aligned}$$

where the last step comes from the symmetry of the normal distribution assumed on the error term. Next we find $\mathbb{E}[y \mid x, y > 0]$. To do so, first let z be distributed standard normal. Then note that $\mathbb{E}[z \mid z > c] = \frac{\phi(c)}{1-\Phi(c)}$. Keeping this result in mind, we proceed:

$$\begin{aligned} \mathbb{E}[y \mid x, y > 0] &= \mathbb{E}[x_i\beta + u_i \mid x, x_i\beta + u_i > 0] \\ &= x_i\beta + \sigma\mathbb{E}\left[\frac{u_i}{\sigma}\middle| x, \frac{u_i}{\sigma} > -\frac{x_i\beta}{\sigma}\right] \\ &= x_i\beta + \sigma\frac{\phi\left(-\frac{x_i\beta}{\sigma}\right)}{1-\Phi\left(-\frac{x_i\beta}{\sigma}\right)} \\ &= x_i\beta + \sigma\frac{\phi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)} \\ &= x_i\beta + \sigma\lambda\left(\frac{x_i\beta}{\sigma}\right) \end{aligned}$$

where $\lambda\left(\frac{x_i\beta}{\sigma}\right)$ denotes the inverse Mills ratio. Now we know that the conditional expectation function is:

$$\mathbb{E}[y \mid x] = \left[x_i\beta + \sigma\lambda\left(\frac{x_i\beta}{\sigma}\right)\right]\Phi\left(\frac{x_i\beta}{\sigma}\right)$$

Let's take the partial derivative with respect to x_j, the j^{th} right-hand side variable:

$$\begin{aligned} \frac{\partial\mathbb{E}[y \mid x]}{\partial x_j} &= \frac{\partial}{\partial x_j}\left(P(y > 0 \mid x)\mathbb{E}[y \mid x, y_i > 0]\right) \\ &= \frac{\partial P(y_i > 0 \mid x)}{\partial x_j} \cdot \mathbb{E}[y_i \mid x, y_i > 0] + P(y_i > 0 \mid x) \cdot \frac{\partial\mathbb{E}[y_i \mid x, y_i > 0]}{\partial x_j} \qquad (*) \end{aligned}$$

Before continuing, we will accept *ex ante* that the derivative of the inverse Mills ratio with respect to x_j is:

$$\frac{\partial \lambda \left(\frac{x_i\beta}{\sigma}\right)}{\partial x_j} = -\lambda\left(\frac{x_i\beta}{\sigma}\right)\left(\frac{x_i\beta}{\sigma} + \lambda\left(\frac{x_i\beta}{\sigma}\right)\right)$$

Therefore, the partial effect of x_j on $\mathbb{E}[y_i \mid x, y_i > 0]$ is:

$$\frac{\partial \mathbb{E}[y_i \mid x, y_i > 0]}{\partial x_j} = \beta_j\left[1 - \lambda\left(\frac{x_i\beta}{\sigma}\right)\left(\frac{x_i\beta}{\sigma} + \lambda\left(\frac{x_i\beta}{\sigma}\right)\right)\right]$$

Plugging this into equation $(*)$ yields:

$$\frac{\partial \mathbb{E}[y \mid x]}{\partial x_j} = \phi\left(\frac{x_i\beta}{\sigma}\right)\frac{\beta_j}{\sigma}\left(x_i\beta + \sigma\lambda\left(\frac{x_i\beta}{\sigma}\right)\right) + \Phi\left(\frac{x_i\beta}{\sigma}\right)\beta_j\left[1 - \lambda\left(\frac{x_i\beta}{\sigma}\right)\left(\frac{x_i\beta}{\sigma} + \lambda\left(\frac{x_i\beta}{\sigma}\right)\right)\right]$$

which simplifies to:

$$\frac{\partial \mathbb{E}[y \mid x]}{\partial x_j} = \Phi\left(\frac{x_i\beta}{\sigma}\right)\beta_j$$

This is the partial effect of x_j on the conditional expectation in a standard Tobit set-up.

7.3 The Second Problem: Sample Selection

Sometimes sample selection is not random or we are missing observations. Two common reasons for problematic sample selection are:

1. **Truncation:** We throw out observations outside of a certain range. Then the likelihood of observing an individual observation is:

$$f(y_i \mid x_i) = \begin{cases} \frac{f(y_i|x_i)}{F(b|x_i) - F(a|x_i)} & \text{if } y_i \in [a, b] \\ 0 & \text{if } y_i > b \text{ or } y_i < a \end{cases}$$

   - This case is essentially a truncated Tobit. We set the probabilities of observing a value less than a or a value greater than b to zero and continue with the Tobit.

2. **Incidental Truncation:** We are missing observations due to the nature of data gathering.

   - Think of trying to estimate the labor supply for women, where the outcome variable is the potential wage rate. The problem here is that we don't observe the wages of women that do not work.

How do we deal with incidental truncation? We use a Heckit model.

7.4 Heckit Models

We first define an equation of interest:

$$y_{i1} = x_{i1}\beta + u_{i1}$$

and then a selection equation:

$$y_{i2} = \mathbb{1}\{x_{i2}\delta + u_{i2} > 0\}$$

When this selection equation is true, meaning that $y_{i2} = 1$, we observe outcome variable y_{i1}. To proceed, we make three assumptions:

1. $\boldsymbol{u}$ is independent of $\boldsymbol{x}_i$ and has a mean of zero.

2. $u_2 \sim N(0, 1)$.

3. $\mathbb{E}[u_1 \mid u_2] = \gamma_1 u_2$.

The end goal here is to learn about β using what we observe. We cannot directly find $\mathbb{E}[y_1 \mid x, y_2 = 0]$, so we start off by finding $\mathbb{E}[y_1 \mid x, u_2]$:

$$\begin{aligned}
\mathbb{E}[y_1 \mid x, u_2] &= \mathbb{E}[x_1\beta + u_1 \mid x, u_2] \\
&= x_1\beta + \mathbb{E}[u_1 \mid u_2] \\
&= x_1\beta + \gamma_1 u_2
\end{aligned}$$

Now we return to $\mathbb{E}[y_1 \mid x, y_2 = 0]$:

$$\begin{aligned}
\mathbb{E}[y_1 \mid x, y_2 = 0] &= \mathbb{E}[y_1 \mid x, x_2\delta + u_2 > 0] \\
&= \mathbb{E}[y_1 \mid x, u_2 > -x_2\delta] \\
&= \mathbb{E}\left[\mathbb{E}[y_1 \mid x, u_2] \middle| x, u_2 > -x_2\delta\right] \\
&= x_1\beta + \gamma_1\mathbb{E}[u_2 \mid x, u_2 > -x_2\delta] \\
&= x_1\beta + \gamma_1\lambda(x_2\delta)
\end{aligned}$$

Note that because γ_1 is probably not zero, OLS will be inconsistent for the partial effect.[6] How do we estimate this conditional expectation then? We use Heckman's two-step procedure (the Heckit). The procedure is as follows:

1. Notice we have assumed $u_2 \sim N(0, 1)$ and that y_{i2} is a binary response variable. Then we can use a probit model of y_{i2} on x_{i2} to find $\hat{\delta}_2$. From $\hat{\delta}_2$ we can compute the estimated inverse Mills ratio, $\hat{\lambda}_{i2}$.

2. Now we can run OLS of y_{i1} on x_{i1} and $\hat{\lambda}_{i2}$. This regression will give us $\hat{\beta}$ and $\hat{\gamma}_1$, both of which are consistent estimates.

[6] If γ_1 is zero, then OLS will be consistent and the robust standard errors from that OLS regression will be correct.

- Note that the standard errors are incorrect because we are using a generated regressor ($\hat{\lambda}_{i2}$).
- To properly estimate the standard errors, we need to use the Delta Method.

Chapter 8

Panel Data

8.1 Panel Data Model

Panel data combines two different types of data: longitudinal (or time series) and cross-section. Longitudinal data follows one individual over multiple periods of time, while cross-sectional data follows many individuals across one period of time. So far, we have only dealt with cross-sectional data.

To construct a good set of panel data, we want to randomly sample n individuals, with each individual denoted with i, and track those individuals over time for T periods, with each time period denoted with t. We assume that x_{it} is independent across i, but dependent across t for each i.

Clearly, if there is a component of each observation that is constant across time that we do not account for, the estimates on our coefficients for x_{it} will be inconsistent (as our standard OLS assumption that $\mathbb{E}[x_{it}\varepsilon_{it}] = 0$ will likely not hold). In cross-sectional data, we solved this problem by using an instrument and estimating via 2SLS. What do we do here? Consider the following model, starting from our standard regression model:

$$\begin{aligned} y_{it} &= x_{it}\beta + \varepsilon_{it} \\ &= x_{it}\beta + c_i + u_{it} \end{aligned}$$

where c_i is a variable that is specific to each individual and constant over time. Some examples of this c_i could include innate ability, sex, race, etc. In panel data, we want to control for these variables. Before continuing, do note that we are assuming that the error term can be linearly broken up into a time-dependent term and a time-independent term. We also assume that $\mathbb{E}[x_{is}u_{it}] = 0$ for $s, t \in \{1, ..., T\}$. In words, this says that our right-hand side regressors are orthogonal to the time-dependent error term across all time periods. Otherwise, we would need an instrument.

8.2 Estimation Method I: Pooled OLS

Pooled OLS is the simplest method conceptually. Basically, we just run OLS after stacking the observations across time (chronologically) into one matrix. Effectively, we are estimating the following model:

$$y_i = x_i\beta + \varepsilon_i$$

Therefore, The estimator is as follows:

$$\hat{\beta}^{POLS} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} x'_{it}x_{it}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T} x'_{it}y_{it}\right)$$

For this estimator to be consistent we must assume that $\mathbb{E}[x'_{it}\varepsilon_{it}] = 0$, as we are not controlling for the time-independent portion of the error.

Use robust standard errors when finding the asymptotic variance, as observations will be correlated across time (due to c_i in the error term).

8.3 Estimation Method II: Random Effects

To apply random effects, we need a number of other assumptions:

(1) Strict Exogeneity: $\mathbb{E}[u_{it} \mid x_i, c_i] = 0$

(2) Random Effects: $\mathbb{E}[c_i \mid x_i] = \mathbb{E}[c_i] = 0$

(3) Rank Condition: $Rank\left(\mathbb{E}[x'_i\Omega^{-1}x_i]\right) = k$, where $\Omega = \mathbb{E}[\varepsilon_i\varepsilon'_i]$ and k is the number of right-hand side variables

(4) Conditional Homoskedasticity: $\mathbb{E}[u_iu'_i \mid x_i, c_i] = \sigma_u^2 I_T$

(5) No Serial Correlation: $\mathbb{E}[c_i^2 \mid x_i] = \sigma_c^2$

Ω is used to improve the efficiency of the estimator, and assumptions (4) and (5) make estimation of Ω relatively simple. How do we estimate Ω? First start by finding Ω:

$$\Omega = \begin{bmatrix} \mathbb{E}[\varepsilon_{i1}^2] & \mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}] & ... & \mathbb{E}[\varepsilon_{i1}\varepsilon_{iT}] \\ \mathbb{E}[\varepsilon_{i2}\varepsilon_{i1}] & \mathbb{E}[\varepsilon_{i2}^2] & ... & \mathbb{E}[\varepsilon_{i2}\varepsilon_{iT}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\varepsilon_{iT}\varepsilon_{i1}] & ... & ... & \mathbb{E}[\varepsilon_{iT}^2] \end{bmatrix}$$

Note that by using assumptions (4) and (5):

$$\begin{aligned} \mathbb{E}[\varepsilon_{it}^2] &= \mathbb{E}\left[(c_i + u_{it})^2\right] \\ &= \mathbb{E}[c_i^2] + 2\mathbb{E}[c_iu_{it}] + \mathbb{E}[u_{it}^2] \\ &= \sigma_c^2 + \sigma_u^2 \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\varepsilon_{it}\varepsilon_{is}] &= \mathbb{E}[(c_i + u_{it})(c_i + u_{is})] \\ &= \mathbb{E}[c_i^2] + \mathbb{E}[]c_i u_{is}] + \mathbb{E}[c_i u_{it}] + \mathbb{E}[u_{it}u_{is}] \\ &= \sigma_c^2\end{aligned}$$

So Ω becomes:

$$\Omega = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & ... & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & ... & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & ... & ... & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

To estimate Ω, we need to estimate σ_c^2 and σ_u^2. To do so, we run the following algorithm:

(1) Run POLS to find $\hat{\varepsilon}_{it}$

(2) Estimate $\hat{\sigma}_\varepsilon^2 = \frac{1}{nT-k}\sum_{i=1}^n \sum_{t=1}^T \hat{\varepsilon}_{it}^2$

(3) Estimate $\hat{\sigma}_c^2 = \frac{1}{nT(T+1)/2-k}\sum_{t=1}^{T-1}\sum_{s=t+1}^T \sum_{i=1}^n \hat{\varepsilon}_{it}\hat{\varepsilon}_{is}$

(4) Back out $\hat{\sigma}_u^2 = \hat{\sigma}_\varepsilon^2 - \hat{\sigma}_c^2$

We can now construct $\hat{\Omega}$. Finally, we estimate the random effects estimator:

$$\hat{\beta}^{RE} = \left(\sum_{i=1}^n x_i'\hat{\Omega}^{-1}x_i\right)^{-1}\left(\sum_{i=1}^n x_i'\hat{\Omega}^{-1}y_i\right)$$

which, using what we learned from GLS, is consistent. Flip back to the GLS notes to find the asymptotic variance.

8.4 Estimation Method III: Fixed Effects

For fixed effects models, we do not assume that c_i is uncorrelated with x_{it}. Rather, we control for c_i by demeaning the data across time:

$$\begin{aligned} y_{it} - \bar{y}_i &= (x_{it} - \bar{x}_i)\beta + c_i - c_i + u_{it} - \bar{u}_i \\ \ddot{y}_{it} &= \ddot{x}_{it}\beta + \ddot{u}_{it}\end{aligned}$$

We can then stack these across time:

$$\ddot{y}_i = \ddot{x}_i\beta + \ddot{u}_i$$

Notice that this looks very similar to the POLS set-up. Can we apply the POLS estimator? We need two assumptions for consistency and another for unbiased-ness:

(1) Orthogonality (for consistency): $\mathbb{E}[\ddot{x}_i'\ddot{u}_i] = 0$

(2) Rank (for consistency): $Rank\left(\mathbb{E}[\ddot{x}_i'\ddot{x}_i]\right) = k$

(3) Strict Exogeneity (for unbiased): $\mathbb{E}[\ddot{u}_i \mid \ddot{x}_i] = 0$

Then the fixed effects estimator is:

$$\hat{\beta}^{FE} = \left(\sum_{i=1}^{n} \ddot{x}_i'\ddot{x}_i\right)^{-1} \left(\sum_{i=1}^{n} \ddot{x}_i'\ddot{y}_i\right)$$

To find asymptotic variance, we need to slightly rewrite the model. Define $Q_T \equiv I_T - P_T$, where $P_T \equiv \xi_T(\xi_T'\xi_T)^{-1}\xi_T'$ and $\xi_T \equiv [1, ..., 1]_{1\times T}$. In effect, when multiplied by a matrix of data, P_T finds the mean. Thus, Q_T demeans the data. Also note that Q_T is an annihilator matrix, meaning that it is idempotent. Turning now to the proof for asymptotic normality:

$$\begin{aligned}
\hat{\beta}^{FE} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i'Q_T'Q_T x_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i'Q_T'Q_T y_i\right) \\
&= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} x_i'Q_T x_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i'Q_T u_i\right) \\
\sqrt{n}(\hat{\beta}^{FE} - \beta) &\xrightarrow{d} \mathbb{E}[x_i'Q_T x_i]^{-1}\, N\left(0, \mathbb{E}[x_i'Q_T u_i u_i' Q_T' x_i]\right) \\
&= N\left(0, \mathbb{E}[\ddot{x}_i'\ddot{x}_i]^{-1}\mathbb{E}[\ddot{x}_i' u_i u_i' \ddot{x}_i']\ \mathbb{E}[\ddot{x}_i'\ddot{x}_i]^{-1}\right)
\end{aligned}$$

If we assume conditional homoskedasticity (that $\mathbb{E}[u_i u_i' \mid x_i, c_i] = \sigma_u^2 I_T$), then the asymptotic variance becomes $\mathbb{E}[x_i'Q_T x_i]^{-1}$.

8.4.1 Dummy Variables in Place of Fixed Effects

If you're experience in undergrad was like mine, you were told that fixed effects were simply dummy variables in an OLS regression. Fortunately for us, we can run OLS on the following specification:

$$y_{it} = x_{it}\beta + \sum_{q=0}^{1} c_q \mathbb{1}\{q = 1\} + u_{it}$$

and $\hat{\beta}^{OLS}$ will be equal to $\hat{\beta}^{FE}$. In a linear model, when estimating β and c together, $\hat{\beta}$ is consistent (c is only unbiased, not consistent).

8.5 Estimation Method IV: First-Differences

Here, the idea is to lag the model once and then subtract it from the original model:

$$\begin{array}{c}
y_{it} = x_{it}\beta + c_i + u_{it} \\
\underline{-(y_{it-1} = x_{it-1}\beta + c_i + u_{it-1})} \\
\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}
\end{array}$$

We then run POLS on this differenced model, stacked across time. So the first-differences estimator is:

$$\hat{\beta}^{FD} = \left(\frac{1}{n} \sum_{i=1}^{n} \Delta x_i' \Delta x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \Delta x_i' \Delta y_i \right)$$

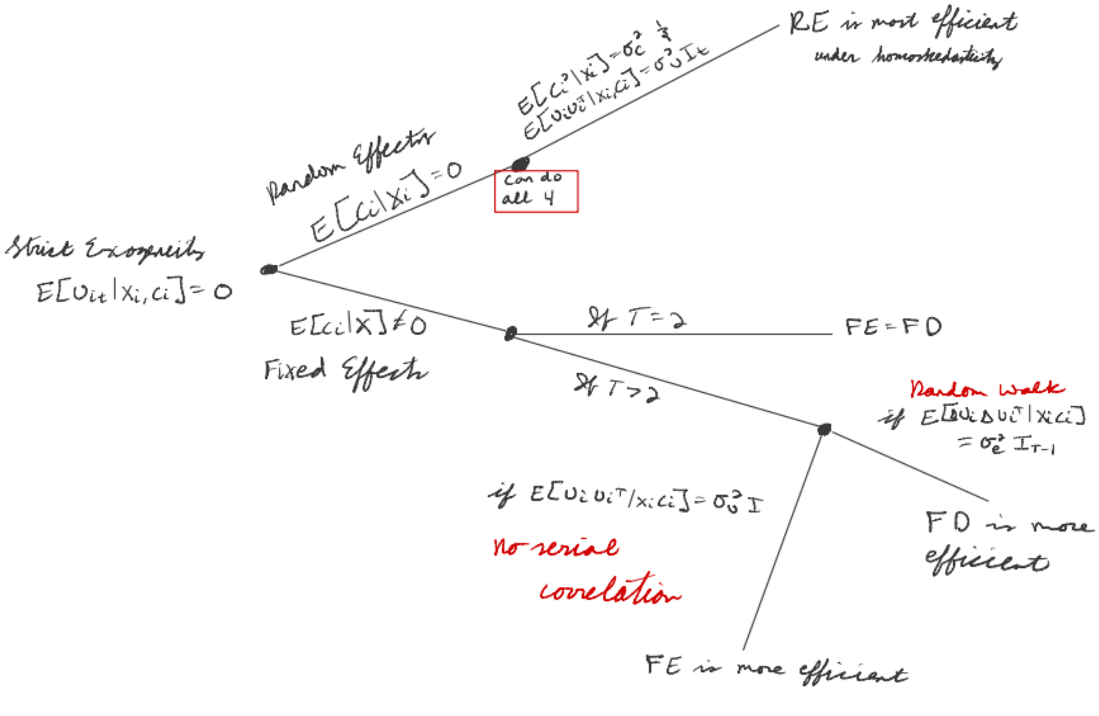
The asymptotics are derived in the same way as from OLS. For the asymptotic variance, we get:

$$AVAR_{FD} = \mathbb{E}[\Delta x_i' \Delta x_i]^{-1} \mathbb{E}[\Delta x_i' \Delta u_i \Delta u_i' \Delta x_i] \; \mathbb{E}[\Delta x_i' \Delta x_i]^{-1}$$

Under a random walk, first-differences is BLUE.

8.6 Comparing Methods

This is a lot of material very quickly. So to boil it down, how do we decide which method we should use? Follow the tree diagram below:



8.7 Previous Quiz Question

This question centers on the Heckit procedure.

8.7.1 Part a

Write down the key components of the sample selection model seen in class. there is an outcome variable y_{1i}, a selection dummy variable y_{2i}, and a vector of explanatory variables x_i that enters both equations. Your answer must cover: (i) outcome and selection equations, (ii) explain what is and what is not observed, (iii) assumption on the error terms, (iv) conditional mean of outcome equation in selected sample, (v) Heckit two-step procedure.

Solution

(i) Outcome: $y_{1i} = x_i\beta + u_i$

Selection: $y_{2i} = \mathbb{1}\{x_i\delta + \nu_i \geq 0\}$

(ii) Observed: $y_{2i},\ x_i,\ y_{1i}$ if $y_{2i} = 1$

Unobserved: $u_i,\ \nu_i,\ y_{1i}$ if $y_{2i} = 0$

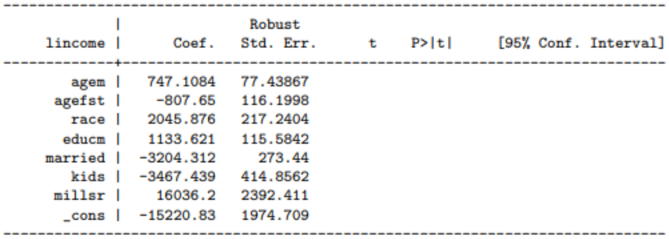
(iii) $(u_i,\ v_i)$ is independent of x_i and has zero mean, $\mathbb{E}[u_i \mid \nu_i] = \gamma\nu_i$, and $v_i \sim N(0,1)$

(iv) $\mathbb{E}[y_{1i} \mid x_i,\ y_{2i} = 1] = x_i\beta + \gamma\lambda(x_i\delta)$, where $\lambda(\cdot)$ denotes the inverse Mills ratio

(v) First, we estimated δ in the selection equation using a probit model. Then we calculate $\hat{\lambda}_i = \lambda(x_i\hat{\delta})$ for every i. Second, we estimate β via OLS by regressing y_{1i} on x_i and $\hat{\lambda}$.

8.7.2 Part b

Consider a selection model for labor supply of mothers. Below you have the OLS Stata output of the second step of the Heckit procedure with robust standard errors. The outcome variable is log income in 1979. We have 6 mother characteristics in the right-hand side plus "millsr" - the estimated inverse Mills ratio. Test the null hypothesis that selection is exogenous. Are the standard errors correct in this regression output?



```
------------------------------------------------------------------------------
             |               Robust
     lincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agem |   747.1084   77.43867
      agefst |    -807.65   116.1998
        race |   2045.876   217.2404
       educm |   1133.621   115.5842
     married |  -3204.312     273.44
        kids |  -3467.439   414.8562
      millsr |    16036.2   2392.411
       _cons |  -15220.83   1974.709
------------------------------------------------------------------------------
```

Solution

We are testing whether γ, the coefficient on the inverse Mills ratio, is zero. Under the null hypothesis, the standard errors in this regression are accurate, so we use them to test the hypothesis in a t-test:

$$\begin{aligned} t &= \frac{16036.2 - 0}{2392.411} \\ &= 6.70 \\ &> 1.96 \end{aligned}$$

Therefore, we reject the null hypothesis that $\gamma = 0$. But if $\gamma \neq 0$, then the standard errors of this regression output are not correct. A significant generated regressor has been included in the model.

8.7.3 Part c

Go back to part (a). Suppose the vector of covariates is:

$$x_i = \begin{bmatrix} 1 & \text{married}_i & \text{grad}_i & \text{married}_i \times \text{grad}_i \end{bmatrix}$$

where married and grad are dummy variables indicating whether the individual is married or has a graduate degree, respectively. A researcher uses the vector x_i for both the outcome and selection equations but is unable to compute the estimator in the second step. Explain why and propose a solution.

Solution

There is perfect colinearity in the second-step regression equation. The vector x_i take four different values:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Therefore, the inverse Mills ratio estimate, $\hat{\lambda}_i$ can take four values: $\lambda_1, \ldots, \lambda_4$. But then we can write $\hat{\lambda}_i$ as a linear combination of x_i vectors:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \lambda_4 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + (\lambda_2 - \lambda_4) \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + (\lambda_3 - \lambda_4) \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + (\lambda_1 - \lambda_2 - \lambda_3 + \lambda_4) \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Therefore, $\hat{\lambda}$ is perfectly colinear with x_i and the second-stage regression cannot be run.

A solution to the problem is to exclude one dummy from either equation. Then, x_i and $\hat{\lambda}_i$ would not be perfectly colinear.

8.8 Violating Assumptions

We know what to do when strict exogeneity holds. But what happens when strict exogeneity fails and only contemporaneous exogeneity holds? That is, suppose:

$$\mathbb{E}[v_{it} \mid x_i, c_i] \neq 0 \text{ but } \mathbb{E}[x_{it}' u_{it}] = 0$$

Which specification is more robust? Let's compare fixed effects and first differences. To begin with, look at the asymptotic normality proof for fixed effects:

$$\begin{aligned}\hat{\beta}^{FE} &= \beta + \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\ddot{x}_{it}'u_{it}\right)\\ &\xrightarrow[n\to\infty]{P} \beta + \left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ddot{x}_{it}'\ddot{x}_{it}]\right)^{-1}\underbrace{\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ddot{x}_{it}'u_{it}]\right)}_{(1)}\end{aligned}$$

Now, if we look at just term (1):

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ddot{x}_{it}'u_{it}] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(x_{it}-\bar{x}_i)'u_{it}]\\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[x_{it}'u_{it}] - \mathbb{E}[\bar{x}_i'u_{it}]\\ &= -\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\bar{x}_i'u_{it}]\\ &= -\mathbb{E}[\bar{x}_i'\bar{u}_i]\end{aligned}$$

Invoke the Cauchy-Schwartz Inequality (that $|\mathbb{E}[xy]| \leq \left(\mathbb{E}[x]^2\mathbb{E}[y]^2\right)^{1/2}$):

$$\mathbb{E}[\bar{x}_i\bar{u}_i] \leq \left[\mathbb{E}[\bar{x}_i^2]\underbrace{\mathbb{E}[\bar{u}_i^2]}_{(2)}\right]^{1/2}$$

Looking just at term (2):

$$\begin{aligned}\mathbb{E}[\bar{u}_i^2] &= \frac{1}{T}\sigma_u^2\\ &\xrightarrow[T\to\infty]{} 0\end{aligned}$$

So all together:

$$\begin{aligned}|\mathbb{E}[\bar{x}_i'\bar{u}_i]| &\leq \left(\frac{1}{T}A\sigma_u^2\right)^{1/2}\\ &\xrightarrow[T\to\infty]{} 0\end{aligned}$$

where A denotes a constant term. Note that as T gets larger, the inconsistency gets smaller. Now, let's look at first differences' asymptotic proof:

$$\begin{aligned}\hat{\beta}^{FD} &= \beta + \left(\frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{n}\sum_{i=1}^{n}\Delta x'_{it}\Delta x_{it}\right)^{-1}\left(\frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{n}\sum_{i=1}^{n}\Delta x'_{it}\Delta u_{it}\right)\\ &\xrightarrow[n\to\infty]{P} \beta + \left(\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}[\Delta x'_{it}\Delta x_{it}]\right)^{-1}\left(\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}[\Delta x'_{it}\Delta u_{it}]\right)\end{aligned}$$

Assuming the data is stationary (but still weakly dependent) gives us:

$$\beta + \underbrace{\left(\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}[\Delta x'_{it}\Delta x_{it}]\right)^{-1}}_{\text{bounded}}\underbrace{\left(\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}[\Delta x'_{it}\Delta u_{it}]\right)}_{(3)}$$

The first braced term is bounded (meaning it has a minimum value) due to the autocorrelation over time. What about the (3)?

$$\begin{aligned}\left(\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}[\Delta x'_{it}\Delta u_{it}]\right) &= \mathbb{E}[x'_{it}u_{it} - x'_{it}u_{it-1} - x'_{it-1}u_{it} + x'_{it-1}u_{it-1}]\\ &= \mathbb{E}[-x'_{it}u_{it-1} - x'_{it-1}u_{it}] \neq 0\end{aligned}$$

So overall, as $T \to \infty$, the first differences error does not disappear. The fixed effects method is more robust to violations of the strict exogeneity assumption. This result gives us an opportunity to test the strict exogeneity using the Hausman test:

$$\mathcal{H} = \sqrt{n}(\hat{\beta}^{FE} - \hat{\beta}^{FD})'(AVAR)^{-1}\sqrt{n}(\hat{\beta}^{FE} - \hat{\beta}^{FD}) \xrightarrow{d} \chi^2_k$$

where k is the number of coefficients on variables that vary across individuals, and $AVAR$ is the asymptotic variance of the difference between $\hat{\beta}^{FE}$ and $\hat{\beta}^{FD}$.

8.9 Random Effects vs Fixed Effects

8.9.1 Relationship between RE and FE

To begin with, let's calculate Ω for random effects:

$$\begin{aligned}\Omega &= \sigma_u^2 I_T + \sigma_c^2 \xi_T \xi'_T\\ &= \sigma_u^2 I_T + \sigma_c^2 T \xi_T \underbrace{(\xi'_T\xi_T)^{-1}}_{1/T}\xi_T\\ &= \sigma_u^2 + \sigma_c^2 T P_T\\ &= \frac{\sigma_u^2 I_T + T\sigma_c^2 P_T}{\sigma_u^2 + T\sigma_c^2}\cdot(\sigma_u^2 + T\sigma_c^2)\end{aligned}$$

$$
\begin{aligned}
&= \left[\underbrace{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2}}_{\equiv \eta} I_T + \left(1 - \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2}\right) P_T \right] (\sigma_u^2 + T\sigma_c^2) \\
&= [\eta I_T + (1-\eta) P_T](\sigma_u^2 + T\sigma_c^2) \\
&= \underbrace{(\eta Q_T + P_T)}_{\equiv S_T}(\sigma_u^2 + T\sigma_c^2)
\end{aligned}
$$

Manipulating S_T:

$$
\begin{aligned}
S_T^{-1} &= P_T + \frac{1}{\eta} Q_T \\
S_T^{-1/2} &= P_T + \frac{1}{\sqrt{\eta}} Q_T \\
&= P_T + \frac{1}{\sqrt{\eta}}(I_T - P_T) \\
&= P_T \left(1 - \frac{1}{\sqrt{\eta}}\right) + \frac{1}{\sqrt{\eta}} I_T \\
&= P_T \left(\frac{\sqrt{\eta} - 1}{\sqrt{\eta}}\right) + \frac{1}{\sqrt{\eta}} I_T \\
&= P_T \left(\frac{-\lambda}{1-\lambda}\right) + I_T \left(\frac{1}{1-\lambda}\right)
\end{aligned}
$$

where $\lambda = 1 - \sqrt{\eta}$. Raise both sides of the Ω expression to the -1/2:

$$
\begin{aligned}
\Omega^{-1/2} &= S_T^{-1/2}(\sigma_u^2 + T\sigma_c^2)^{-1/2} \\
&= \frac{(\sigma_u^2 + T\sigma_c^2)^{-1/2}}{1-\lambda}[I_T - \lambda P_T] \\
&= \frac{(\sigma_u^2 + T\sigma_c^2)^{1/2}}{(\sigma_u^2 + T\sigma_c^2)^{1/2}(\sigma_u^2)^{1/2}}[I_T - \lambda P_T] \\
&= \frac{1}{\sigma_u}[I_T - \lambda P_T]
\end{aligned}
$$

Okay, now we are ready to transform the regression model as in GLS:

$$
\begin{aligned}
y_i &= x_i \beta + u_i \\
\Omega^{-1/2} y_i &= \Omega^{-1/2} x_i \beta + \Omega^{-1/2} u_i \\
(I_T - \lambda P_T) y_i &= (I_T - \lambda P_T) x_i \beta + (I_T - \lambda P_T) u_i \\
(y_i - \lambda \bar{y}_i) &= (x_i - \lambda \bar{x}_i)\beta + (u_i - \lambda \bar{u}_i)
\end{aligned}
$$

Note that λ consists of population moments. We therefore estimate λ with $\hat{\lambda} = \left(1 - \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + T\hat{\sigma}_c^2}\right)^{-1/2}$. But this means that random effects quasi-demeans the data, weighting by relative variance, whereas fixed effects just demeans the data.

8.9.2 Testing random effects

We can use the Hausman test to check whether random effects hold (that $\mathbb{E}[c_i \mid x_i] = 0$). Because under the null hypothesis random effects is BLUE, the Hausman principle applies (see recitation 6). Therefore:

$$\mathcal{H} = \sqrt{n}(\hat{\beta}^{FE} - \hat{\beta}^{RE})' \left[\widehat{AVAR}^{FE} - \widehat{AVAR}^{RE}\right]^{-1} \sqrt{n}(\hat{\beta}^{FE} - \hat{\beta}^{RE}) \xrightarrow{d} \chi^2_k$$

Sometimes, though, we do not want to assume that we have conditional homoskedasticity (a key assumption for GLS to be BLUE). Then we need a control function approach. The structural equation is:

$$y_{it} = x_{it}\beta + \varepsilon_{it}$$

Note that POLS is inconsistent:

$$\begin{aligned}
\hat{\beta}^{POLS} - \beta &= \left(\frac{1}{nT}\sum_{it} x'_{it}x_{it}\right)^{-1}\left(\frac{1}{nT}\sum_{it} x'_{it}\varepsilon_{it}\right) \\
&\xrightarrow{P} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}x_{it}\right]^{-1} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}\varepsilon_{it}\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}x_{it}\right]^{-1} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}(c_i + u_{it})\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}x_{it}\right]^{-1} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}c_i\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} x'_{it}x_{it}\right]^{-1} \mathbb{E}\left[\bar{x}'_i c_i\right] \\
&\neq 0
\end{aligned}$$

So we take the best linear projection of c_i on $\bar{x}_i$:

$$c_i = \bar{x}_i\varphi + a_i$$

Now the control function becomes:

$$y_{it} = x_{it}\beta + \bar{x}_i\varphi + a_i + u_{it}$$

Because $\mathbb{E}[\bar{x}'_i a_i] = 0$ by construction, we can test whether $\varphi = 0$ by running POLS and using robust standard errors. If c_i is a vector, we need to run a joint Wald test on φ. If $\varphi = 0$, then c_i is not correlated with x_{it}.

8.10 Problem Set 5, Question 5 (Wooldridge 10.17)

Consider a standard unobserved effects model, but where we separate out aggregate time effects, say $\boldsymbol{d}_t$, a $1 \times R$ vector, where $R \leq T - 1$. Therefore, the model is:

$$y_{it} = \alpha + d_t \eta + w_{it}\delta + c_i + u_{it}$$

where w_{it} is the $1 \times M$ vector of explanatory variables that vary across i and t. Because d_t does not change across i, we take them to be nonrandom. We can also assume that $\mathbb{E}[c_i] = 0$, as the model has an intercept term. Let $\lambda = 1 - \left(\frac{1}{1+T(\sigma_c^2/\sigma_u^2)}\right)^{1/2}$ be the quasi time-demeaning parameter for the random effects estimation. We assume that λ is known.

8.10.1 Part a

Show that we can write the equation for random effects as:

$$y_{it} - \lambda \bar{y}_i = \mu + (d_t - \bar{d})eta + (w_{it} - \lambda \bar{w}_i)\delta + (v_{it} - \lambda \bar{v}_i)$$

where $\mu = (1-\lambda)\alpha + (1-\lambda)\bar{d}\eta$ and $v_{it} = c_i + u_{it}$.

Solution:

We quasi-demean across all variables first:

$$\begin{aligned} y_{it} - \lambda \bar{y}_i &= (1-\lambda)\alpha + (d_t - \lambda \bar{d})\eta + (w_{it} - \lambda \bar{w}_i)\delta + (v_{it} - \lambda \bar{v}_i) \\ &= [(1-\lambda)\alpha + (1-\lambda)\bar{d}\eta] + (d_t - \bar{d})\eta + (w_{it} - \lambda \bar{w}_i)\delta + (v_{it} - \lambda \bar{v}_i) \\ &= \mu + (d_t - \bar{d})\eta + (w_{it} - \lambda \bar{w}_i)\delta + (v_{it} - \lambda \bar{v}_i) \end{aligned}$$

which is what we wanted to show.

8.10.2 Part b

Now assume that $\mu = 0$. Define $g_{it} = [d_t - \bar{d},\ w_{it} - \lambda \bar{w}_i]$ and $\beta = [\eta',\ \delta']'$. Show that by assuming $\mathbb{E}[u_{it} \mid x_i, c_i] = 0$, $\mathbb{E}[c_i \mid x_i] = \mathbb{E}[c_i] = 0$, and $rank\,(\mathbb{E}[x_i'\Omega x_i]) = k$:

$$\sqrt{n}(\hat{\beta}^{RE} - \beta) = \frac{A_1^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} g_{it}'(v_{it} - \lambda \bar{v}_i) + o_p(1)$$

where $A_1 = \sum_{t=1}^{T} \mathbb{E}[g_{it}' g_{it}]$. Also verify that for any i:

$$\sum_{t=1}^{T} (d_t - \bar{d})(v_{it} - \lambda \bar{v}_i) = \sum_{t=1}^{T} (d_t - \bar{d}) u_{it}$$

Solution:

The random effects estimator is a POLS estimator with quasi-demeaning:

$$\begin{aligned}
\hat{\beta}^{RE} &= \left(\sum_{i=1}^{n} g_i' g_i\right)^{-1} \left(\sum_{i=1}^{n} g_i'(y_i - \lambda \bar{y}_i)\right) \\
\sqrt{n}(\hat{\beta}^{RE} - \beta) &= \left(\frac{1}{n}\sum_{i=1}^{n} g_i' g_i\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i'(v_i - \lambda \bar{v}_i)\right) \\
&= \frac{\mathbb{E}[g_i' g_i]^{-1}}{\sqrt{n}} \left(\sum_{i=1}^{n} g_i'(v_i - \lambda \bar{v}_i)\right) + o_p(1)
\end{aligned}$$

Now to show the second part:

$$\begin{aligned}
\sum_{t=1}^{T}(d_t - \bar{d})(v_{it} - \lambda \bar{v}_i) &= \sum_{t=1}^{T}(d_t - \bar{d})(1-\lambda)c_i + \sum_{t=1}^{T}(d_t - \bar{d})u_{it} - \sum_{t=1}^{T}(d_t - \bar{d})\lambda \bar{u}_i \\
&= (1-\lambda)c_i \sum_{t=1}^{T}(d_t - \bar{d}) + \sum_{t=1}^{T}(d_t - \bar{d})u_{it} - (\lambda \bar{u}_i)\sum_{t=1}^{T}(d_t - \bar{d}) \\
&= \sum_{t=1}^{T}(d_t - \bar{d})u_{it}
\end{aligned}$$

8.10.3 Part c

Show that by assuming $\mathbb{E}[u_{it} \mid h_i, c_i] = 0$ and $rank\left(\sum_{t=1}^{T} \mathbb{E}[\ddot{h}_{it}' \ddot{h}_{it}]\right) = k$:

$$\sqrt{n}(\hat{\beta}^{FE} - \beta) = \frac{A_2^{-1}}{\sqrt{n}} \sum_{i=1}^{n}\sum_{t=1}^{T} h_{it}' u_{it} + o_p(1)$$

where $A_2 = \sum_{t=1}^{T} \mathbb{E}[h_{it}' h_{it}]$ and $h_{it} = [d_t - \bar{d},\ w_{it} - \bar{w}_i]$.

Solution:

From both class and recitation, we know that:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}^{FE} - \beta) &= \left(\frac{1}{n}\sum_{i=1}^{n} h_i' h_i\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_i' u_i\right) \\
&= \frac{\mathbb{E}[h_i' h_i]^{-1}}{\sqrt{n}} \left(\sum_{i=1}^{n} h_i' u_i\right) + o_p(1)
\end{aligned}$$

8.10.4 Part d

Show that $A_1\sqrt{n}(\hat{\beta}^{RE} - \beta) - A_2\sqrt{n}(\hat{\beta}^{FE} - \beta)$ has an asymptotic variance matrix of rank M.

Solution:

Using part b, we know that:

$$A_1\sqrt{n}(\hat{\beta}^{RE}-\beta)=\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}g_i'(v_i-\lambda\bar{v}_i)\right)+o_p(1)$$

and from part c, we know that:

$$A_2\sqrt{n}(\hat{\beta}^{FE}-\beta)=\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}h_i'u_i\right)+o_p(1)$$

Note that the first R elements of both equations are the same: $(d_t-\bar{d})u_i$. Therefore:

$$\begin{aligned}&A_1\sqrt{n}(\hat{\beta}^{RE}-\beta)-A_2\sqrt{n}(\hat{\beta}^{FE}-\beta)\\&\qquad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix}0\\(w_{it}-\lambda\bar{w}_i)'(v_i-\lambda\bar{v}_i)-(w_i-\bar{w}_i)'u_{it}\end{bmatrix}+o_p(1)\end{aligned}$$

Therefore, the asymptotic variance of the difference can only have rank M, the size of w_{it}.

8.10.5 Part e

What implications does part d have for a Hausman test that compares fixed effects and random effects when the model contains aggregate time variables? Does it matter whether $\mathbb{E}[u_iu_i' \mid x_i, c_i]=\sigma_u^2 I_T$ and $\mathbb{E}[c_i^2 \mid x_i]=\sigma_c^2$?

Solution:

Part d implies that the Hausman test will have only M degrees of freedom, not $M+R$. Getting the degrees of freedom correct is important, as it affects the limiting distribution and thus the critical values. Assuming the two additional assumptions in the question will not change this fact.

Additionally, part d implies that if the model only includes time aggregates, the limiting distributions for random effects and fixed effects are the same. We would not have the rank M part of the distribution. Therefore, the asymptotic difference is zero.

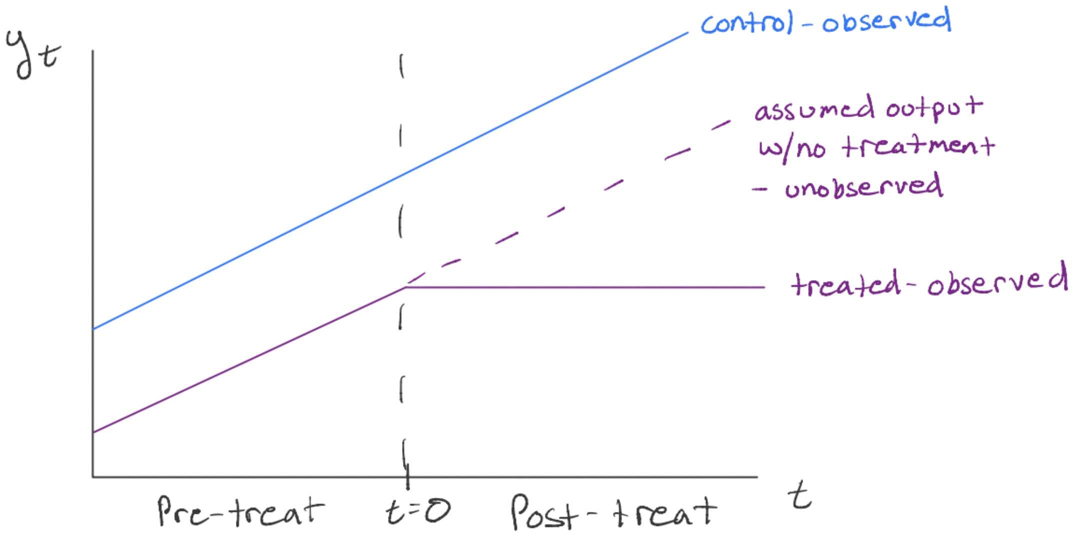
Chapter 9

Difference-in-Differences

9.1 Intuition

The idea behind difference-in-differences is to take subtract the average of the difference between the treatment and control groups before the treatment time from the average of the difference after the treatment time. This estimator, given a set of assumptions, will causally identify the impact of the treatment.

We will look at a simple two time period, two group model for intuition. Graphically, an ideal set-up to apply difference-in-differences would like like:



Here, we can see that our control group goes forward with the same trend as before the treatment. Our treatment group, hereto after denoted by $T = 1$, is impacted only after the treatment at event time $t = 0$. We can also see that we assume that without treatment, the treated group's trend would have remained the same. This assumption is called the **strong parallel trends assumption**. Mathematically:

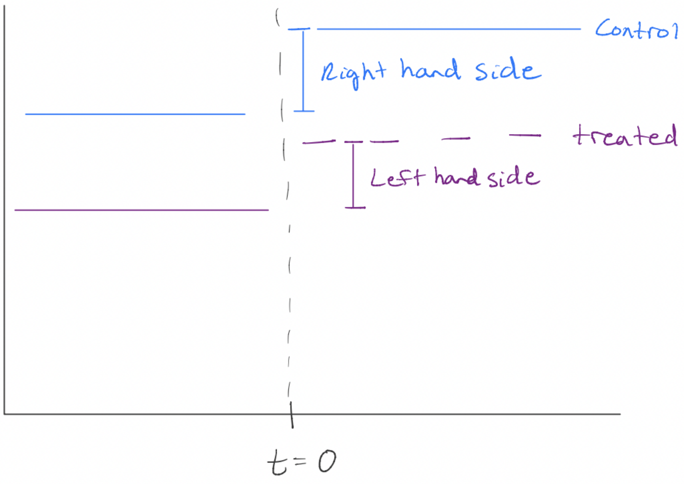
$$\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 1] = \mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 0]$$

where $(\cdot)$ denotes whether an individual actually received treatment, denoted by D.

This equation seems confusing, so let's go through it little-by-little. The first term on the left-hand side denotes our outcome, y, at time $t = 1$. So we are looking at the post-treatment. time-period. It also has the subscript that $D = 0$. Therefore, this outcome is *not* treated. Note though that the expectation is conditional on $T = 1$. Therefore, we are looking at the hypothetical outcome in the post-treatment time period of the subjects in the treatment group had they not received treatment. Graphically, the first term is the mean of the dashed line.

After having gone through the first term slowly, the remaining terms follow naturally. The second term on the left looks at those in the treated group before the treatment time and without treatment. The first term on the right looks at those in the control group after the treatment time, while the second term on the right looks at those in the control group before the treatment time.

We can draw a picture visualizing the parallel trend assumption as written above:



The strong parallel trends assumption says that the two gaps highlighted in the picture are equal.

9.2 Structural Equation

Now that we have an intuitive understanding of the model, let's write down a structural equation:

$$\begin{aligned}y_t(D) = \beta_0 + \beta_1 * T + \beta_2 * Post_t + \beta_3 * D + \beta_4 * T * Post_t \\ + \beta_5 * T * D + \beta_6 * Post_t * D + \beta_7 * T * Post_t * D + u_t\end{aligned}$$

Let's think carefully about this equation. β_0 denotes the mean for those in the untreated group, before treatment time. β_1 denotes the mean for those in the treated group, before the treatment time. β_3 denotes the mean for those actually treated before the treatment time. We then interact all of these terms. Note that because we have a fully saturated model, $\mathbb{E}[u_t \mid T] = 0$.

We can simplify this structural equation though by invoking the parallel trends assumption. Using the left hand side of the assumption:

$$\begin{aligned}\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 1] &= (\beta_0 + \beta_1 + \beta_2 + \beta_4) - (\beta_0 + \beta_1) \\ &= \beta_2 + \beta_4 \qquad (1)\end{aligned}$$

Now the right hand side:

$$\begin{aligned}\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 0] &= (\beta_0 + \beta_2) - (\beta_0) \\ &= \beta_2 \qquad (2)\end{aligned}$$

Setting (1) and (2) equal to each other:

$$\begin{aligned}\beta_2 + \beta_4 &= \beta_2 \\ \beta_4 &= 0\end{aligned}$$

So by invoking the parallel trends assumption, we assume that $\beta_4 = 0$. Our structural equation becomes:

$$\begin{aligned}y_t(D) = \beta_0 + \beta_1 * T + \beta_2 * Post_t + \beta_3 * D \\ + \beta_5 * T * D + \beta_6 * Post_t * D + \beta_7 * T * Post_t * D + u_t \qquad (3)\end{aligned}$$

9.3 Estimation

We can't estimate the altered structural model though. We can't observe D. So, we impose that $D_{it} = T_{it} \times Post_t$. Substituting this into equation (3):

$$\begin{aligned}y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Post_t + \beta_3 * Post_t * T_{it} + \beta_5 * T_{it} * T_{it} * Post_t \\ + \beta_6 * Post_t * T_{it} * Post_t + \beta_7 * T_{it} * Post_t * T_{it} * Post_t + u_{it}\end{aligned}$$

Note that a dummy variable squared maps to itself (e.g. $T_{it} * T_{it} = T_{it}$). Therefore:

$$y_{it} = \beta_0 + \beta_1 * T_{it} + \beta_2 * Post_t + (\beta_3 + \beta_5 + \beta_6 + \beta_7) * Post_t * T_{it} + u_{it} \tag{4}$$

We now have a viable equation to estimate. What parameter do we want to find from this equation though? Ideally, we would want to find the average treatment effect (ATE):

$$ATE_t = \mathbb{E}[y_t(1) - y_t(0)]$$

But we can't find this, as we do not know the unconditional expectations.[7] Instead, we can find what is known as the average treatment on the treated (ATT):

$$ATT_t = \mathbb{E}[y_t(1) - y_t(0) \mid T = 1]$$

Evaluating the $ATT_{t=1}$ using equation (3):

$$\begin{aligned} ATT_{t=1} &= \mathbb{E}[y_{t=1}(1) - y_{t=1}(0) \mid T = 1] \\ &= [\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_5 + \beta_6 + \beta_7] - [\beta_0 + \beta_1 + \beta_2] \\ &= \beta_3 + \beta_5 + \beta_6 + \beta_7 \end{aligned}$$

Fortunately, this is the coefficient on $Post_t * T_{it}$ in equation (4)! Rewriting (4) in a recognizable regression format yields:

$$y_i = x_i\delta + u_i$$

$$\begin{bmatrix} y_{i,t=-1} \\ y_{i,t=1} \end{bmatrix} = \begin{bmatrix} 1 & T_{i,t=-1} & Post_{t=-1} & Post_{t=-1} * T_{i,t=-1} \\ 1 & T_{i,t=1} & Post_{t=1} & Post_{t=1} * T_{i,t=1} \end{bmatrix} \begin{bmatrix} \delta_0 \\ \delta_1 \\ \delta_2 \\ \delta_3^* \end{bmatrix} + \begin{bmatrix} u_{i,t=-1} \\ u_{i,t=1} \end{bmatrix}$$

This is a regression stacked across t. Therefore, we can use POLS with robust standard errors to estimate the coefficient of interest. Note that GLS does not deliver efficiency gains because we have no correlation of error terms across time.

9.4 Covariates

What happens if we add covariates to the equation (such that $x_{i,t=1} \overset{d}{=} x_{i,t=-1}$)? The model becomes:

$$y_t(D) = x\beta_t(D, T) + u_t \qquad \text{where } \mathbb{E}[u_t \mid T, x] = 0$$

with x being a $1 \times k$ vector of covariates and $\beta_t(D, T)$ is a $7k \times 1$ vector of slope coefficients. Now that we have covariates, we can use a weaker, less restrictive version of the parallel trends assumption.

[7] The idea here is that we cannot observe a random person receiving treatment. Those in the control group will never receive treatment.

Only after conditioning on x must the trends be parallel. This is called the **conditional parallel trends assumption**. Mathematically:

$$\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 1, x] = \mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 0, x] \qquad \forall x$$

This assumption allows us to set β_4 to zero again:

$$\begin{aligned}
\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 1, X] &= x[\beta_{t=1}(0,1) - \beta_{t=-1}(0,1)] && \text{Using PT|x:} \\
x[\beta_{t=1}(0,1) - \beta_{t=-1}(0,1)] &= x[\beta_{t=1}(0,0) - \beta_{t=-1}(0,0)] \\
x'x[\beta_{t=1}(0,1) - \beta_{t=-1}(0,1)] &= x'x[\beta_{t=1}(0,0) - \beta_{t=-1}(0,0)]
\end{aligned}$$

$$\begin{aligned}
\beta_{t=1}(0,1) - \beta_{t=-1}(0,1) &= \beta_{t=1}(0,0) - \beta_{t=-1}(0,0) \\
\beta_0 + \beta_1 + \beta_2 + \beta_4 - \beta_0 - \beta_1 &= \beta_0 + \beta_2 - \beta_0 \\
\beta_2 + \beta_4 &= \beta_2 \\
\beta_4 &= 0
\end{aligned}$$

9.4.1 Identification

Once again, we cannot do average treatment effects or conditional average treatment effects. We turn to conditional average treated on the treated:

$$CATT_t(x) = \mathbb{E}[y_{t=1}(1) - y_{t=1}(0) \mid T = 1, x]$$

Before continuing, note that the ATT is simply the expected value of the $CATT$:

$$\begin{aligned}
ATT_t &= \mathbb{E}[y_{t=1}(1) - y_{t=1}(0) \mid T = 1] && \text{Using LIE:} \\
&= \mathbb{E}_x[CATT_{t=1}(x)]
\end{aligned}$$

With that in mind, let's derive which β's the $CATT$ corresponds to:

$$\begin{aligned}
CATT_{t=1}(x) &= \mathbb{E}[y_{t=1}(1) \mid T = 1, x] - \underbrace{\mathbb{E}[y_{t=1}(0) \mid T = 1, x]}_{\text{Can't observe}} \\
&= \mathbb{E}[y_{t=1} \mid T = 1, x] - \{\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 1, x] \\
&\qquad + \mathbb{E}[y_{t=-1}(0) \mid T = 1, x]\} && \text{Using PT|x:} \\
&= \mathbb{E}[y_{t=1}(1) \mid T = 1, x] - \mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T = 0, x] \\
&\qquad - \mathbb{E}[y_{t=-1}(0) \mid T = 1, x] \\
&= \{\mathbb{E}[y_{t=1}(1) \mid T = 1, x] - \mathbb{E}[y_{t=1}(0) \mid T = 0, x]\} \\
&\qquad - \{\mathbb{E}[y_{t=-1}(0) \mid T = 1, x] - \mathbb{E}[y_{t=-1}(0) \mid T = 0, x]\}
\end{aligned}$$

$$
\begin{aligned}
&= x\{\beta_0+\beta_1+\beta_2+\beta_3+\beta_5+\beta_6+\beta_7-\beta_0-\beta_2\} - x\{\beta_0+\beta_1-\beta_0\} \\
&= x\,\underbrace{[\beta_3+\beta_5+\beta_6+\beta_7]}_{ATT_{t=1}}
\end{aligned}
$$

Since we know x, we can find the ATT.

9.4.2 Estimation

Like in the previous case without covariates, we put our model into regression form:

$$
\begin{aligned}
y_{it} &= x_{it}\beta_t(D_{it}, T_{it}) + u_{it} \\
&= x_{it}[\beta_0 + \beta_1 * T_{it} + \beta_2 * Post_t + (\beta_3+\beta_5+\beta_6+\beta_7) * Post_t * T_{it}] + u_{it}
\end{aligned}
$$

We assume that $\mathbb{E}[u_{it} \mid T_{it}, x_{it}] = 0$ so that POLS is unbiased and consistent. In the case of a discrete x, this assumption can actually be proven. Once again, use robust standard errors when doing POLS. This time though, we can run pooled FGLS and get more precise estimates.

To estimate $CATT$ and ATT, we calculate:

$$
\begin{aligned}
\widehat{CATT}_{t=1}(x_0) &= x_0\left(\hat\beta_3+\hat\beta_5+\hat\beta_6+\hat\beta_7\right) \\
\widehat{ATT}_{t=1} &= \bar{\bar{\bar{x}}}\left(\hat\beta_3+\hat\beta_5+\hat\beta_6+\hat\beta_7\right) \\
\text{where } \bar{\bar{\bar{x}}} &= \frac{\sum_{i,t} x_{it}\mathbb{1}\{T_{it}=1\}}{\sum_{it}\mathbb{1}\{T_{it}=1\}}
\end{aligned}
$$

where x_0 denotes that the value of the covariates before treatment.

Before moving ahead, let's first be more explicit in our regression specification. Define $x_{it} = [1_{1\times 2}, \tilde{x}_{it_{1\times k}}]$ and $\beta = [\alpha_{1\times 2}, \gamma_{1\times k}]'$. Then the full regression would be:

$$
\begin{aligned}
y_{it} = \alpha_0 + \alpha_1 * T_{it} + \alpha_2 * Post_t + \alpha_3^* * Post_t * T_{it} \\
+ \tilde{x}_{it}[\gamma_0 + \gamma_1 * T_{it} + \gamma_2 * Post_t + \gamma_3^* * Post_t * T_{it}] + u_{it}
\end{aligned}
$$

Writing the model this way gives us:

$$
\begin{aligned}
CATT_{t=1}(x_0) &= \alpha_3^* + x_0 * \gamma_3^* \\
ATT_{t=1} &= \alpha_3^* + \bar{\bar{\bar{x}}}\gamma_3^*
\end{aligned}
$$

How do researchers estimate these effects practically? They make four assumptions:

(1) Replace $\tilde{x}_{it}$ with $\tilde{x}_{it} - \bar{\bar{\bar{x}}}$ to demean the treatment group. Then:

$$
\begin{aligned}
ATT_t &= \alpha_3^* \\
CATT_t(x_0) &= \alpha_3^* + (x_0 - \bar{\bar{\bar{x}}})\gamma_3^*
\end{aligned}
$$

(2) In addition, assume homogenous time effects across x. This then assumes that

$$\gamma_3^* = 0$$

and so:

$$CATT_t(x_0) = ATT_t = \alpha_3^*$$

(3) Then assume that the effect of x does not vary over time (that $\gamma_2 = 0$ and $x_{it} = x_{i0}$). This effectively means that we are strengthening the conditional parallel trends assumption to the strong parallel trends assumption:

$$\begin{aligned}\mathbb{E}[y_{t=1}(0) - y_{t=-1}(0) \mid T, \tilde{x}] &= [\alpha_{t=1}(0, T) - \alpha_{t=-1}(0, T)] + \tilde{x}[\gamma_{t=1}(0, T) - \gamma_{t=-1}(0, T)] \\ &= [\alpha_{t=1}(0, T) - \alpha_{t=-1}(0, T)]\end{aligned}$$

(4) Lastly, also assume that $\gamma_1 = \gamma_3 = 0$. Under this assumption, x does not help with identification, but it does help with precision.

After making all these assumptions, we are left with the following specification:

$$y_{it} = \alpha_0 + \alpha_1 * T_{it} + \alpha_2 * Post_t + \alpha_3^* * Post_t * T_{it} + \gamma_0(\tilde{x}_{it} - \bar{\bar{\tilde{x}}}) + u_{it}$$

where α_3^* is the ATT_t and $CATT_t(x)\ \forall\ x$.

9.5 Previous Quiz Question

This question tests your knowledge of the difference-in-differences model.

9.5.1 Part a

Consider a difference-in-differences model for the case of two periods, two groups, and pooled cross-sectional data, but no covariates. Write down the key components of the model seen in class. Your answer must cover: (i) the correctly specified model for the conditional mean of potential outcomes $\mathbb{E}[y_t(D) \mid T]$; (ii) the equation for observed treatment D_t and the equation for observed outcome y_t; (iii) the parallel trend assumption stated in terms of (i); and (iv) the regression equation for y_{it} as a function of T_i, $Post_t$, and u_{it}.

Solution

Starting with (i):

(i) The conditional expectation of potential outcomes is:

$$\begin{aligned}\mathbb{E}[y_t(D) \mid T] = \beta_0 + \beta_1 T + \beta_2 Post_t + \beta_3 D + \beta_4 T * Post_t \\ + \beta_5 T * D + \beta_6 Post_t * D + \beta_7 T * Post_t * D\end{aligned}$$

(ii) The observed treatment and outcomes are:

$$D_t = T * Post_t \text{ and } y_t = y_t(D_t).$$

(iii) The parallel trends assumption is:

$$\mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 1] = \mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 0]$$

(iv) The regression equation is:

$$y_{it} = \delta_0 + \delta_1 T_i + \delta_2 Post_t + \delta_3 T_i * Post_t + u_{it}$$

9.5.2 Part b

Suppose now that the researcher has a $(1 \times k)$ vector of pre-treatment covariates x. (i) How should we use these covariates to “weaken” the parallel trend assumption? (ii) Write down the parallel trend assumption with covariates and without covariates. (iii) Explain why PT|x does not imply PT.

Solution

Starting with (i):

(i) We can now assume that parallel trends holds after conditioning on the covariates instead of having to hold by itself. After controlling for more variation, we are more likely to get parallel trends in the data.

(ii) The two assumptions are:

$$\begin{aligned}\mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 1, x] &= \mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 0, x] && \text{(PT|x)} \\ \mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 1] &= \mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 0] && \text{(PT)}\end{aligned}$$

(iii) Looking at the left-hand side of PT|x and using LIE:

$$\mathbb{E}_x \left[\underbrace{\mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 1, x]}_{\text{Mean over } x \text{ with treatment}} \middle| T = 1 \right]$$

Then looking at the right-hand side of PT|x and using LIE:

$$\mathbb{E}_x\left[\underbrace{\mathbb{E}[y_t(0)-y_{t-1}(0)\mid T=0,x]}_{\text{Mean over x w/out treatment}}\middle| T=0\right]$$

These distributions are not usually the same, so the conditional parallel trends assumption does not imply the strong parallel trends assumption.

9.5.3 Part c

Suppose a researcher estimates the following regression specification:

$$y_{it} = \alpha_0 + \alpha_1 T_i + \alpha_2 * Post_t + \alpha_3 Post_t * T_i + x_{it}\gamma + u_{it}$$

List all the simplifying assumptions that this researcher is implicitly making relatively to the flexible model in class.

Solutions:

The assumptions the researcher makes are:

(a) The effect of covariates are the same across groups, implying that $\gamma_1 = 0$.

(b) The effect of covariates are the same across time, implying that $\gamma_2 = 0$.

(c) That there is no treatment effect heterogeneity with respect to x ($CATT_t(x) = ATT_t$), implying that $\gamma_3 = 0$.

9.6 Adding Panel Data

When we add panel data to the model, the structural equation remains:

$$y_t(D) = x\beta_t(D,T) + u_t \qquad \text{where } \mathbb{E}[u_t \mid T, x] = 0$$

The theory behind this case is relatively simple: the parallel trends assumptions are the same as in the previous case. Identification is derived in the same manner as case 2 as well. How do we estimate this new model? Let's look at the estimation equation:

$$y_{it} = x_i[\beta_0 + \beta_1 T_i + \beta_2 Post_t + \beta_3 T_i * Post_t] + u_{it}$$

This model is the same as before, with one exception: u_{it} is now correlated with $u_{it'}$. To make sure we account for this, we need to cluster standard errors at the individual level.

In applied micro, researchers will commonly use fixed effects to weaken the parallel trends assumptions. Does this actually help though? Let's add an unobserved individual fixed effect, w, to the structural model:

$$y_t(D) = x\beta_t(D,T) + w\delta_t(D,T) + u_t$$

Then the conditional parallel trends assumption becomes:

$$\mathbb{E}[y_t(0) - y_{t-1}(0) \mid x, w, T = 1] = \mathbb{E}[y_t(0) - y_{t-1}(0) \mid x, w, T = 0] \tag{PT|w,x}$$

If we estimated this structural equation, our ideal regression equation would be:

$$\begin{aligned} y_{it} = x_i[\beta_0 + \beta_1 T_i + \beta_2 Post_t + \beta_3 T_i * Post_t] \\ + w_i[\delta_0 + \delta_1 T_i + \delta_2 Post_t + \delta_3 T_i * Post_t] + u_{it} \end{aligned}$$

But we cannot actually observe $w\delta_t(D,T)$. So we approximate it with a time-invariant individual fixed effect c_i. Therefore, we can estimate:

$$y_{it} = x_i[\beta_2 Post_t + \beta_3 T_i * Post_t] + c_i + u_{it}$$

where c_i absorbs all things that do not vary across time. But then PT|w,x is just equal to PT|x, as w will be differenced out. Therefore, fixed effects do not actually help in identification or efficiency.

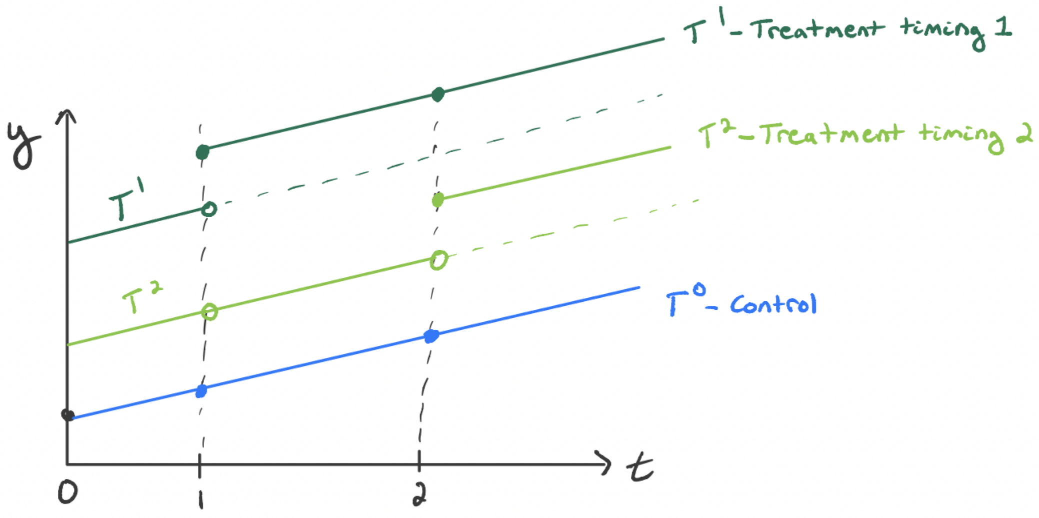
We can use the fixed effects idea, though, to greatly simplify the structural model. Including c_i gives us:

$$\begin{aligned} y_{it} &= \alpha_2 + \alpha_3^* T_i * Post_t + \tilde{x}_i[\gamma_2 Post_t + \gamma_3^* T_i * Post_t] + c_i + u_{it} \\ &\text{where } \ c_i = \alpha_0 + \alpha_1 T_i + \gamma_0 \tilde{x}_i + \gamma_1 T_i \tilde{x}_i \end{aligned}$$

As we just showed, estimation using fixed effects does not improve the regression, so estimating this equation using either POLS or fixed effects is identical.

9.7 Adding More Groups and Time Periods

Suppose now that we have more than two groups and more than two time periods. Also suppose that different groups could be treated at different times. How do we analyze this situation? Let's first start with a picture that shows a three group, three time period model so we can understand this scenario:



Going forward, we will write the potential outcomes framework as:

$$Y_t(\tau) \qquad \text{where } \ \tau \in \{0, 1, 2\}$$

where 0 denotes no treatment, 1 denotes treatment at time $t = 1$, and 2 denotes treatment at time $t = 2$. We can write the population structural model as:

$$y_t(\tau) = x\beta_t(\tau, T) + u_t \qquad \text{with } \ \mathbb{E}[u_t \mid T, x] = 0$$

Remember that T, $t \in \{0, 1, 2\}$. If we listed all the parameters, we would have $3 \times 3 \times 3 = 27$. We would have 9 dummies:

$$T^0,\ T^1,\ T^2 \tag{Groups}$$

$$\tau^0,\ \tau^1,\ \tau^2 \tag{Pot. Outcomes}$$

$$P_t^0,\ P_t^1,\ P_t^2 \tag{Years}$$

Of course, because we cannot observe τ, we set $\tau = T \times \mathbb{1}\{T \geq t\}$.

9.7.1 Parallel Trend Assumption

With more than two groups and two time periods, how do we construct the parallel trend assumption? First, denote the group label as $g \in \{0, 1, 2\}$. Second, note that we will need two parallel trend assumptions: one with respect to the group that is never treated, and one with respect to a group that is not *yet* treated.

$$\mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = g, x] = \mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = 0, x] \qquad \text{(PT, never treated)}$$
$$\mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = g_1, x] = \mathbb{E}[y_t(0) - y_{t-1}(0) \mid T = g_0, x] \qquad \text{(PT, not yet treated)}$$

where g_0 denotes the group that is not yet treated and g_1 denotes the group that has already been treated.

9.7.2 Identification

Our causal effects do not really change:

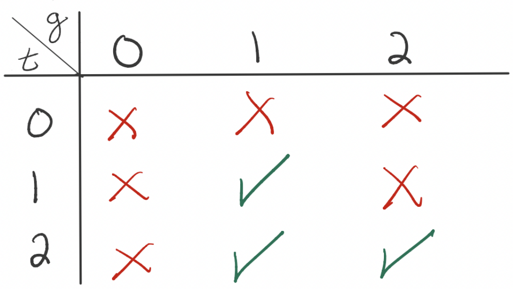
$$ATT_t(g) = \mathbb{E}[y_t(g) - y_t(0) \mid T = g]$$
$$CATT_t(g, x) = \mathbb{E}[y_t(g) - y_t(0) \mid T = g, x]$$

We can see by these definitions that the causal effects will differ by the groups analyzed. Let's look at $ATT_2(1)$ as an example.

$$\begin{aligned} CATT_2(1, x) &= \mathbb{E}\left[y_2(1) - \underbrace{y_2(0)}_{\text{Unobserved}} \middle| T = 1, x \right] \qquad \text{using PT, never treat:} \\ &= \mathbb{E}[y_2(1) \mid T = 1, x] - \{\mathbb{E}[y_2(0) \mid T = 0, x] \\ &\qquad + \mathbb{E}[y_0(0) \mid T = 1, x] - \mathbb{E}[y_0(0) \mid T = 0, x]\} \\ &= \mathbb{E}[y_2(1) \mid T = 1, x] - \mathbb{E}[y_2(0) \mid T = 0, x] \\ &\qquad - \{\mathbb{E}[y_0(0) \mid T = 1, x] - \mathbb{E}[y_0 \mid T = 0, x]\} \\ &= ATT_2(1) \end{aligned}$$

where the last equality comes from the proof we did for the initial conditional parallel trend assumption last recitation.

It turns out that we can do this proof if $g = 1, 2$ and $t \geq g$. The following chart gives the table for the $ATT_t(g)$ we can identify.



$t \backslash g$	0	1	2
0	✗	✗	✗
1	✗	✓	✗
2	✗	✓	✓

9.7.3 Regression Specification

The following equation details the full regression set-up:

$$\begin{aligned} y_{it} = {} & \underbrace{\alpha_0 P_t^0 + \alpha_1 P_t^1 + \alpha_2 P_t^2}_{\text{Time trend for T=0}} + \overbrace{\underbrace{\beta_1 T_i^1 + \beta_2 T_i^2}_{\text{with no treatment}}}^{\text{Time trend for T=1,2}} + \underbrace{ATT_1(1) T_i^1 P_t^1 + ATT_2(1) T_i^1 P_t^2}_{ATTs \text{ for } g\,=\,1} \\ & + \underbrace{ATT_2(2) T_i^2 P_t^2}_{ATT \text{ for } g\,=\,2} + (x_i - \bar{x}_i)[\alpha_0^x P_t^0 + \alpha_1^x P_t^1 + \alpha_2^x P_t^2 + \beta_1^x T_i^1 \\ & + \beta_2^x T_i^2 + \gamma_{1,1} T_i^1 P_t^1 + \gamma_{2,1} T_i^1 P_t^2 + \gamma_{2,2} T_i^2 P_t^2] + u_{it} \end{aligned} \tag{9.1}$$

The aquamarine-colored terms denote those terms that would be absorbed by individual fixed effects. We can estimate equation 1 using either POLS with clustered robust standard errors or fixed effects with clustered robust standard errors.

Two-Way Fixed Effects

What if we wanted to simplify the model by using a fixed effect term for individuals and a fixed effect term for time? Then we could set up the regression as:

$$y_{it} = \mu_t + \beta D_{it} + c_i + (x_i - \bar{x}_i)\gamma + u_{it} \qquad \text{where } \; D_{it} = \begin{cases} 1 \text{ if } \tau_{it} \geq 1 \\ 0 \text{ if } \tau_{it} = 0 \end{cases} \tag{9.2}$$

where μ_t is the year fixed effect term and c_i is the individual fixed effect term. Note that by setting up the model as in equation 2, we are implicitly assuming that all the ATTs in equation 1 (colored red) are the same. We are also assuming that the effects of covariates are homogenous (meaning that $\alpha_i^x = \beta_i^x = 0$ and $\gamma_{1,1} = \gamma_{2,1} = \gamma_{2,2}$ in equation 1).

While nice, this equation is a gross simplification of the underlying model. Because of that, there are some potential issues. Goodman-Bacon (2021) demonstrates that asymptotically:

$$\hat{\beta}^{FE} \xrightarrow{P} VWATT + VWCT - \Delta ATT$$

where $VWATT$ stands for the variance-weighted ATT and $VWCT$ stands for the variance weighted change in trends. In his paper, Goodman-Bacon derives these three objects:

$$\begin{aligned} VWATT &= \frac{\sigma_{10}}{2}\,[ATT_1(1) + ATT_2(1)] + \sigma_{20} ATT_2(2) + \sigma_{12}^{(1)} ATT_1(1) + \sigma_{12}^{(2)} ATT_2(2) \\ \Delta ATT &= \sigma_{12}^{(2)} [ATT_2(1) - ATT_1(1)] \\ VWCT &= 0 \end{aligned}$$

where $\sigma_{..} > 0$, $\sum \sigma = 1$. Note that if the ATT is constant across time, $\Delta ATT = 0$. $VWCT = 0$ only under the parallel trends assumption with respect to a group that is not yet treated. If we put all of this together, only assuming parallel trends with respect to a group not yet treated:

$$\hat{\beta}^{FE} \xrightarrow{P} ATT_1(1)\left[\frac{\sigma_{10}}{2}+\sigma_{12}^{(1)}+\sigma_{12}^{(2)}\right]+ATT_2(1)\left[\frac{\sigma_{10}}{2}-\sigma_{12}^{(2)}\right]+ATT_2(2)[\sigma_{20}+\sigma_{12}^{(2)}]$$

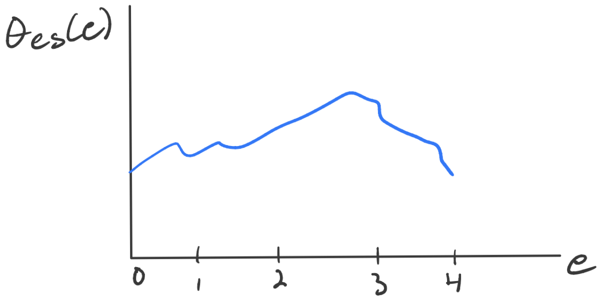
This is just a weighted average of the ATTs. But note that the coefficient on $ATT_2(1)$ has a $-\sigma_{12}^{(2)}$. Therefore, even if each ATT is positive, $\hat{\beta}^{FE}$ could end up negative! This result is known as the Goodman-Bacon critique of Two-Way Fixed Effects models.

Chapter 10

Event Studies, Clustering, and LATE

10.1 Event Studies

What if we could track the outcome, y, for each time-period before and after a treatment? Event study designs allow this. Let $\theta_{es}(e)$ denote the average effect of treatment "e" periods after treatment. A graph depicting $\theta_{es}(e)$ could look like:



where $e = 0$ is the treatment time. How do we graph this? We use the formula:

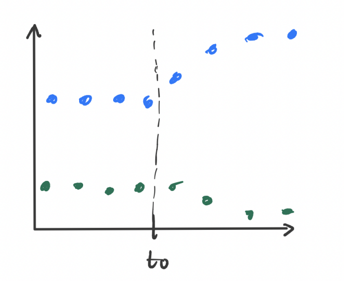
$$\theta_{es}(e) = \sum_{g=0}^{G} \overbrace{\mathbb{1}\{g + e \leq H\} P(T = g \mid T + e \leq H)}^{\text{weights, } w(g,t)} ATT_{g+e}(g)$$

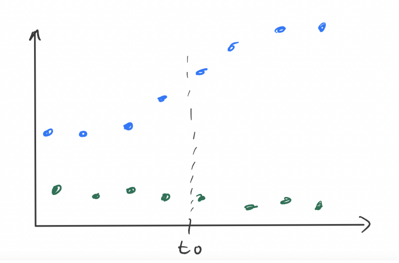
Using our example from last time (g=3, H=3), we get:

$$\theta_{es}(0) = P(T = 1 \mid T \in \{1, 2\}) ATT_1(1) + P(T = 2 \mid T \in \{1, 2\}) ATT_1(2)$$
$$\theta_{es}(1) = ATT_2(1)$$

because group 1 is observed during the period it is treated and one period after, while group 2 is observed only during the period it is treated.

Now that we have this tool, we can gather data before treatment to plot y prior to $t = 0$. In this way, we can test the parallel trends assumption. Two examples include:





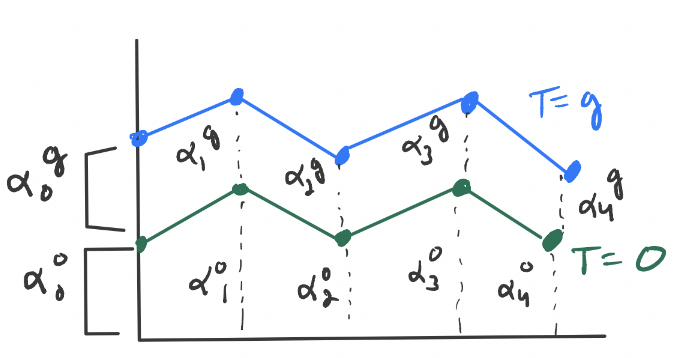
The figure on the left has parallel trends prior to treatment, while the figure on the left clearly does not. To formally test this, we can run a regression with the following specification:

$$y_{it} = \sum_{u=0}^{H} \left\{ \alpha_u^0 P_t^u + \sum_{g=1}^{G} \alpha_u^g P_t^u * T_i^g \right\} + \varepsilon_{it}$$

and test the following hypothesis:

$$H_0 : \alpha_0^g = \alpha_1^g = ... = \alpha_{g-1}^g$$
$$H_A : \text{Any } \alpha_u^g \text{ is different from another}$$

where groups are ordered by treatment time. The following picture demonstrates the idea:



If all the α_u^g's are the same, then we conclude that the parallel trends assumption holds. If they are not, then we must use a different regression design or attempt a triple difference-in-differences.

10.2 Clustering

Clustering is a method for addressing correlations within a hierarchical data set (this simply means that observations are linked to each other). As an example, think of observing ACT test scores within New York City. There is a high probability that students in the same school are going to have correlated test scores. Therefore, each observation inside a school is linked to other observations inside that school. To obtain correct standard errors, we would want to cluster by school.

Clustering eliminates the cross-correlation at the level of clustering. By clustering at the school level, we eliminate cross-correlation at the school level. How do we show this mathematically? First consider the following regression:

$$y_{ig} = \beta_0 + \beta_1 x_g + u_{ig}$$

Naturally, we would estimate this using POLS:

$$\hat{\beta} = \left(\sum_{g=1}^{G} x_g' x_g\right)^{-1} \left(\sum_{g=1}^{G} x_g' y_g\right)$$

Here, we make the assumption that the clustering structure is known (that we know g). We also assume that $(y_g,\ x_g)$ are independent *across* g. The proof for the asymptotic distribution is the same as the standard POLS proof. Hence, we get the following asymptotic variance:

$$AVAR = G\left(\sum_{g=1}^{G} x_g' x_g\right)^{-1} \left(\sum_{g=1}^{G} x_g' u_g u_g' x_g\right) \left(\sum_{g=1}^{G} x_g' x_g\right)^{-1}$$

This is the sandwich estimator, robust to both conditional heteroskedasticity and in-cluster correlation. Note though, that this requires us to assume that $G \longrightarrow \infty$ while n_g (the number of observations in each group) grows at a fixed rate. We might worry about precision if G is small.

This caveat raises an important question: What if there are multiple levels to at which we might cluster? In our example, now suppose we have multiple cities in our data set. We can now cluster at either the school level or the city level. There are two considerations:

(1) If we cluster at too fine a level, the variance estimate will be both biased and inconsistent.

(2) If we cluster at too coarse a level, we reduce G and the variance becomes much less precise.

Note that just because the standard error increases, this does not mean we are accounting for the cross-correlation, especially if G is very small. The variance estimator could just be picking up random noise. In this sense, choosing the clustering level is more of an art than a science.

Panel Data

Suppose instead we have panel data, so the model looks like:

$$\begin{aligned} y_{it} &= x_{it}\beta + c_i + u_{it} \\ y_i &= x_i\beta + \underbrace{c_i\xi_T + u_i}_{\equiv v_i \text{ using POLS}} \end{aligned}$$

We can do POLS again, if we want to assume that $\mathbb{E}[v_i \mid x_i] = 0$. After clustering, the robust variance would be:

$$AVAR_{POLS} = G\left(\sum_{g=1}^{G} x_g' x_g\right)^{-1}\left(\sum_{g=1}^{G} x_g' v_g v_g' x_g\right)\left(\sum_{g=1}^{G} x_g' x_g\right)^{-1}$$

We can also estimate using fixed effects. Then the robust variance would be:

$$AVAR_{FE} = G\left(\sum_{g=1}^{G} \ddot{x}_g' \ddot{x}_g\right)^{-1}\left(\sum_{g=1}^{G} \ddot{x}_g' \ddot{u}_g \ddot{u}_g' \ddot{x}_g\right)\left(\sum_{g=1}^{G} \ddot{x}_g' \ddot{x}_g\right)^{-1}$$

Let's consider two scenarios. First, suppose we cluster at the individual level. Then:

- Fixed Effects eliminate c_{ig}.
- There is correlation across time (between u_{igt} and u_{igs}).
- There is no need to cluster if c_{ig} eliminates time correlation (essentially means that time correlation only exists in individuals and/or groups).

Second, suppose that we cluster at the group level. Then:

- Fixed Effects still eliminates c_{ig}.
- Within group $corr(u_{igt},\ u_{jgt}) \neq 0$ or $corr(u_{igt},\ u_{igs}) \neq 0$.
- Across groups, there will be no correlation across individuals nor across time.
- We do not cluster at the group level if $corr(c_{ig},\ c_{jg})$ captures the spatial correlation across individuals in the same group. Then there is no need to cluster at g.

10.3 LATE

LATE, or local average treatment effects, is a model that provides identified causal effects while requiring few assumptions on the underlying data structures. Suppose we have two variables, an outcome y_i and a treatment dummy x_i. Also suppose that these two variables are generated using some DGP such that:

$$\begin{aligned} y_i &= h(x_i, u_i) \\ x_i &= g(z_i, v_i) \end{aligned}$$

where z_i is a binary instrument. Previously, when looking for a treatment effect, we essentially used the following strategy:

$$\begin{array}{c} y_i = \beta + u_i \\ -(y_i = u_i) \\ \hline \beta \end{array}$$

Now, we will use another potential outcomes framework (from Rubin 1974) to get different treatment effects for each i:

$$\begin{array}{c} y_i(1) = h(1, u_i) \\ -(y_i(0) = h(0, u_i)) \\ \hline h(1, u_i) - h(0, u_i) \end{array}$$

We run into the same problem as in difference-in-differences though: we cannot actually observe $y_i(0)$. So we cannot observe the average treatment effect $\mathbb{E}[y_i(1) - y_i(0)]$. We also cannot take CATE:

$$CATE = \mathbb{E}[y_i \mid x_i = 1] - \mathbb{E}[y_i \mid x_i = 0]$$

as there may be bias inherent in group selection (remember, we imposed no restrictions on the DGP for x_i). We can, however, find the average treatment effect over a subpopulation. Using the notation of $x(T) = D$, where D denotes actual treatment and T denotes treatment group. Using this, we can develop a table:

Table 1

LATE Subpopulations

	$x(0) = 0$	$x(0) = 1$
$x(1) = 0$	Never-Taker	Defier
$x(1) = 1$	Complier	Always-Taker

If we assume that the "Defiers" do not exist, then we can identify the average treatment effect on "Compliers." LATE is defined as:

$$\text{LATE} = \mathbb{E}[y_i(1) - y_i(0) \mid x_i(0) = 0,\ x_i(1) = 1]$$

Let's first decompose x_i, noting that we can write x_i in the potential outcome framework by defining

$x_i(1) = g(1, v_i)$ and $x_i(0) = g(0, v_i)$:

$$\begin{aligned} x_i &= (1 - z_i)x_i(0) + z_i x_i(1) \\ &= x_i(0) + z_i(x_i(1) - x_i(0)) \qquad (1) \end{aligned}$$

Expand y_i:

$$\begin{aligned} y_i &= (1 - x_i)y_i(0) + x_i y_i(1) \\ &= y_i(0) + x_i(y_i(1) - y_i(0)) \qquad (2) \end{aligned}$$

Plug (1) into (2):

$$y_i = y_i(0) + x_i(0)(y_i(1) - y_i(0)) + z_i(x_i(1) - x_i(0))(y_1(1) - y_i(0)) \qquad (3)$$

Assuming that the instrument z is valid, we can subtract the two conditional expectations of (3).[8]

$$\begin{array}{r} \mathbb{E}[y \mid z = 1] = \mathbb{E}[y(0)] + \mathbb{E}[x(0)(y(1) - y(0))] + \mathbb{E}[(x(1) - x(0))(y(1) - y(0))] \\ -\,(\mathbb{E}[y \mid z = 0] \;=\; \mathbb{E}[y(0)] + \mathbb{E}[x(0)(y(1) - y(0))]) \qquad\qquad\qquad\qquad\qquad \\ \hline \mathbb{E}[y \mid z = 1] - \mathbb{E}[y \mid z = 0] = \mathbb{E}[(x(1) - x(0))(y(1) - y(0))] \qquad\quad \end{array}$$

We can then take this difference and write the right-hand side in terms of the possible values of $x(1) - x(0)$:

$$\begin{aligned} \mathbb{E}[y \mid z = 1] - \mathbb{E}[y \mid z = 0] &= (-1)\overbrace{P(x(1) - x(0) = -1)}^{\text{Defiers, so } = 0}\mathbb{E}[y(1) - y(0) \mid x(1) - x(0) = -1] \\ &\quad + (1)P(x(1) - x(0) = 1)\mathbb{E}[y(1) - y(0) \mid x(1) - x(0) = 1] \\ &= P(x(1) - x(0) = 1)\mathbb{E}[y(1) - y(0) \mid x(1) - x(0) = 1] \end{aligned}$$

But notice that the expectation is simply the definition of LATE. So:

$$\begin{aligned} \text{LATE} &= \frac{\mathbb{E}[y \mid z = 1] - \mathbb{E}[y \mid z = 0]}{P(x(1) - x(0) = 1)} \\ &= \frac{\mathbb{E}[y \mid z = 1] - \mathbb{E}[y \mid z = 0]}{\mathbb{E}[x \mid z = 1] - \mathbb{E}[x \mid z = 0]} \end{aligned}$$

Interestingly, this equation for LATE is basically the ILS estimator for IV. Thus, using the model

$$\begin{aligned} y_i &= x_i\beta + u_i \\ x_i &= z_i\pi + v_i \end{aligned}$$

we can estimate LATE. Although surprising at first, notice that the right-hand side of each equation satisfies the model assumptions of LATE: y_i is a function of x_i and u_i while x_i is a function of z_i and v_i.

[8] I drop the i subscripts from here on to make the proof cleaner.

An important note on finding the average treatment effect on compliers: depending on how the instrument is determined, the local average treatment effect will change.

Chapter 11

Summary of Econometrics II

11.1 Semester Review

This semester you have covered a lot of material. Below is a list of topics to help you prepare for the comprehensive exams that take place in approximately one month.

I. System OLS and Generalized Least Squares

  A. Identification

  B. Consistency

  C. Asymptotic normality

  D. Feasible GLS

  E. Assumptions and decision tree

II. Instrumental Variables

  A. OLS assumption violated

  B. Additional assumptions for IV

  C. Identification

  D. Consistency

  E. Asymptotic normality

III. Two-Stage Least Squares

  A. Difference between IV and 2SLS

  B. Identification

  C. Consistency

  D. Asymptotic normality

  E. Control function approach/endogeneity test

F. Durbin-Wu-Hausman test

G. Overidentification test

H. Weak instruments and Stock-Yogo test

IV. Generalized Method of Moments

A. Criterion function

B. Identification given a weighting matrix

C. Consistency

D. Asymptotic normality and optimal weighting matrix

E. Iterated GMM

V. Binary Response Models

A. Linear probability model

B. Probit

C. Logit

D. Partial effects

VI. Censoring and Sample Selection

A. Tobit

B. Partial effects

C. Heckit two-step procedure

D. Heckman selection test

VII. Panel Data

A. Pooled OLS

B. Random effects

C. Fixed effects

D. First differences

E. Assumptions and decision tree

F. Comparing methods, testing assumptions

VIII. Difference-in-Differences

(a) Structural equation

(b) Parallel trends assumptions

(c) ATT and CATT estimation

(d) Two-way fixed effects and the Goodman-Bacon critique

(e) Event studies and parallel trends test

IX. Clustering

A. Level of clustering

B. Clustering with fixed effects

X. LATE

A. Compliers, defiers, never-takers, always-takers

B. LATE identification

C. IV and LATE

D. Assumptions