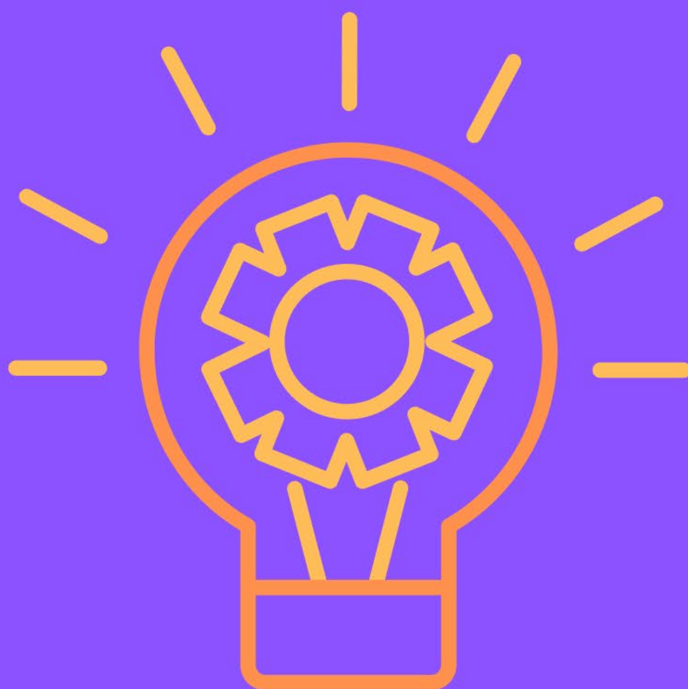
POLIMETRICS

A COMPANION TO INTRODUCTION TO POLITICAL SCIENCE REESARCH METHODS



Josh Franco, Ph.D.

AN OPEN EDUCATION RESOURCE

Polimetrics: A Companion to Introduction to Political Science Research Methods

1st Edition

Page Intentionally Left Blank

Polimetrics

A Companion to Introduction to Political Science Research Methods

1st Edition

Josh Franco, Ph.D., Cuyamaca College

Website
Please visit https://www.ipsrm.com/polimetrics for the latest PDF version.

PDF Version ISBN-13: Pending.

This open education resource is dedicated to Ethan, my son, and future generations

Brief Table of Contents

Table of Contents

Preface

When I was a first-generation student at community college, I never heard of Stata or any other statistical and data analysis software. I had a challenging enough time in my computer programming, statistics, and calculus courses during my first two years, that having a professor try to teach me about such data analysis software would likely have gone in one ear and out the other ear.

Once I transferred to UC Merced, I was introduced to Stata in my econometrics (think economics and statistics) course. I earned a D+ in that course because I was confused by the underlying mechanics of the best, linear, unbiased estimate (good old "BLUE") and having to use Stata. Also, I spent time organizing California Students for Barack Obama during spring 2007, so my attendance in econometrics was not consistent.

I ended up successfully retaking econometrics at Sacramento State after my employer, Lieutenant Governor John Garamendi, told me that I needed to finish my bachelor's degree. I did not mention this in the prior paragraph, but that D+ needed to be a C- so the course would count towards my graduation requirements. While it cost me about $500 to retake the course, I was able to leave work early on Tuesdays and Thursdays so I could drive from downtown to the university campus.

Once I started working in the U.S. House of Representatives, I encountered more reports with data analysis. A practical link was established after I met with representatives of the RAND Corporation to learn more about reports they published. That is when I realized the utility of my econometrics training and it's real-world application to public policy.

I left my real-world career as a congressional policy advisor and returned to UC Merced to start the political science Ph.D. program. I was confronted with data analysis again, and I struggled so much that I had to retake the whole core methods sequence during my second year in the program; this was not a small setback, as it consisted of 3 courses. I powered through and successfully completed the quantitative analysis-rich program. I resolved that when I became a professor, I would work to introduce my mostly first-generation community college students to statistical and data analysis software. And so, Polimetrics is a result of that resolve.

Josh Franco, Ph.D.
April 2021

List of Figures

Acknowledgements

Thank you to my wife Mayra and son Ethan for letting me have time to work on this workbook over the last year.

Thank you to Gabriela Ortiz, Applied Econometrician at StataCorp, for reviewing a draft of this workbook and providing insightful comments and feedback.

Thank you to Alexandra Alcantara, Daniel Amodeo-Chavez, Lava Khurshid, Octavio Vicencio, and James Zillo from my spring 2020 Introduction to Political Science Research Methods. They were, and will forever be, the inaugural political science research methods student at Cuyamaca College. They got the kernel of this workbook rolling since we worked with Stata until the COVID-19 pandemic resulted in stay-at-home orders in March 2020.

Chapter 1 - Overview

About this Workbook

Polimetrics is designed as a companion to *Introduction to Political Science Research Methods, 1st Edition*, or IPSRM for short.

I co-authored IPSRM, which is an Open Education Resource textbook as well, with my colleagues Dr. Charlotte Lee at Berkeley City College, Kau Vue at Fresno City College, Dr. Dino Bozonelos at Victor Valley College, Dr. Masahiro Omae at San Diego City College, and Dr. Steven Cauchon at Imperial Valley College.

Visit https://ipsrm.com/ to download your PDF copy of IPSRM 1st edition.

This workbook provides a tour of the Stata software, an introduction to cross-sectional, time series, and panel data, and an introduction to a variety of models. We review models where the outcome is linear, binary, ordinal, categorical, and count. Additionally, we have an interpretation chapter on survival models.

Each "Models" chapter has a similar organizational structure: about, estimated time, what is the model, how are models run in Stata, how do we interpret the model results, and a real-world example of model results in a Creative Commons licensed, peer-reviewed journal article. Additionally, mini-assignment instructions and a rubric will be included so students can practice their interpretation skills.

Why an OER workbook on non-free Stata?

Open Education Resources, by definition, are free in electronic form. This means this Workbook is free. However, Stata costs money. For students, you can buy a 6-month license of Stata/IC (entry-level) for $48, an annual license for $94, and a perpetual (aka lifetime) license for $225. For teachers, a Stata/IC annual license is $125, and a 3-year license is $365.

I prefer Stata to RStudio (which is free) because of syntax. There are simply fewer characters to type when using Stata versus RStudio. And fewer characters to type means fewer syntax errors. And fewer syntax errors mean less frustration for the beginning user, which is my target audience.

Relatedly, I took two computer programming classes when I was in community college: BASIC and C++. Both were difficult, because I did not understand at the time, that I was learning a language, logic, and mathematics simultaneously. Therefore, I prefer not to conflate the learning of how to interpret the results of statistical and data analysis models with programming syntax-heavy software.

How to Use this Workbook

This workbook has two audiences: faculty and students. Below I describe how to use this workbook from each perspective.

Faculty

From the faculty perspective, this workbook serves as an unpretentious introduction to Stata targeted for first-year or second-year college-level students. As I mentioned in the Preface, when I was a community college student, I never heard of Stata or any other statistical and data analysis software. But, in retrospect, I would have benefitted from knowing about, being explained its utility, and letting that knowledge inform my journey.

This Workbook does not cover the underlying mathematics or statistics of these models: no expressions, no equations, and no proofs. If I feel the need, the need for mathematical expressions, then I will relegate it to a footnote and citation for self-exploration. I use plain language to communicate seemingly complex concepts because statistical and data analysis should be accessible to anyone, not just those attending or working in the proverbial ivory tower.

Faculty who are teaching an introductory level course in research methods or statistics are encouraged to use this Workbook, since it dovetails nicely with increasing students' awareness and knowledge of statistical and data analysis software, like Stata.

I think chapter 2 (Software Tour and Getting Started) and chapter 3 (Datasets) are necessary for clearly orientating a student. From there, you can pick and choose which Models chapters you want to explore with your students. Each Model is based on the dependent variable, or outcome variable, of interest. For example, if you and your students are exploring why someone votes or not, this is a binary outcome, meaning you should use chapter 8 (Binary Outcome Models). Of, if you and your students are exploring the length of time it takes a student to earn their Associates degree, then chapter 17 (Survival Models) can be utilized.

Student

I will forever be a first-generation, community college student. I decided to attend college because I believed it offered me a path for a better life. And it has.

This Workbook is designed for you: the new college student who is starting their adult life or the returning college student who wants to live a better life for themselves and their family. I bridge theory with practice, having worked in government and politics for five years before earning my Ph.D. I hope that comes through in each chapter, so that your honest "Why does this matter?" question has an answer for it.

Also recall that each "Models" chapter has a similar organizational structure: about, estimated time, what is the model, how are models run in Stata, how do we interpret the model results, and a real-world example of model results in a Creative Commons licensed, peer-reviewed journal article. Additionally, mini-assignment instructions and a rubric will be included so you can practice your interpretation skills.

This similar organizational structure should help you develop a rhythm as you work your way through chapters assigned by your professor, or that you are discovering on your own.

Your Feedback

I would appreciate feedback that you have for me regarding this Workbook, so feel free to send me an email at josue.franco@gcccd.edu.

If you find a spelling error, send me an email. If you think something is poorly written, send me an email. If you believe I do not explain something well enough, send me an email. If you think everything is great, send me an email.

Chapter 2 – Stata Software Tour and Getting Started

About

Stata is statistical and data analysis software. According to Stata, "In January 1985, Stata 1.0 was released. In June 2019, Stata 16 was released. For over thirty years, StataCorp has been a leader in statistical software, dedicated to providing the tools professional researchers need to analyze their data." Stata even has a Political science | Stata webpage.

Software is a computer program that you install on your computer. Software is commonly called "apps" these days. Much like you would download apps, like Facebook app or Discord app, you can download and install Stata software on your PC or Mac computer.

The purpose of this chapter is to take a brief software tour of Stata, which is the latest version, and explore how to Get Started. I offer a written, truncated version of the Tour and Getting Started below. The Mini-Assignment ask you to watch videos to reinforce what your read.

Estimated Time

An estimated 90-120 minutes is needed to complete this activity.

Tour of Stata Interface

We interact with any software using its interface. An interface includes menu bars, panels, and buttons.

Below is a screenshot of the standard Stata user interface. The interface includes seven elements:

1. Top left menu that lists: File, Edit, Data, Graphics, Statistics, User, Window, and Help
2. Top left buttons that include icons for Open, Save, Print, and so on
3. Left Panel labeled "History".
4. Center top panel, called the "Results" window, that contains output of commands. For example, you can see **`update query`** in bold, and the following text was produced from that command.
5. Center bottom panel, labeled "Command" window, is where you type Stata commands that produce output. For example, I typed **`update query`** in this field.

6. Right top panel, labeled "Variables" window, that list the variables if you had variables defined or a dataset loaded.
7. Right bottom panel, labeled "Properties" window, that provide specific information about Variables and Data.

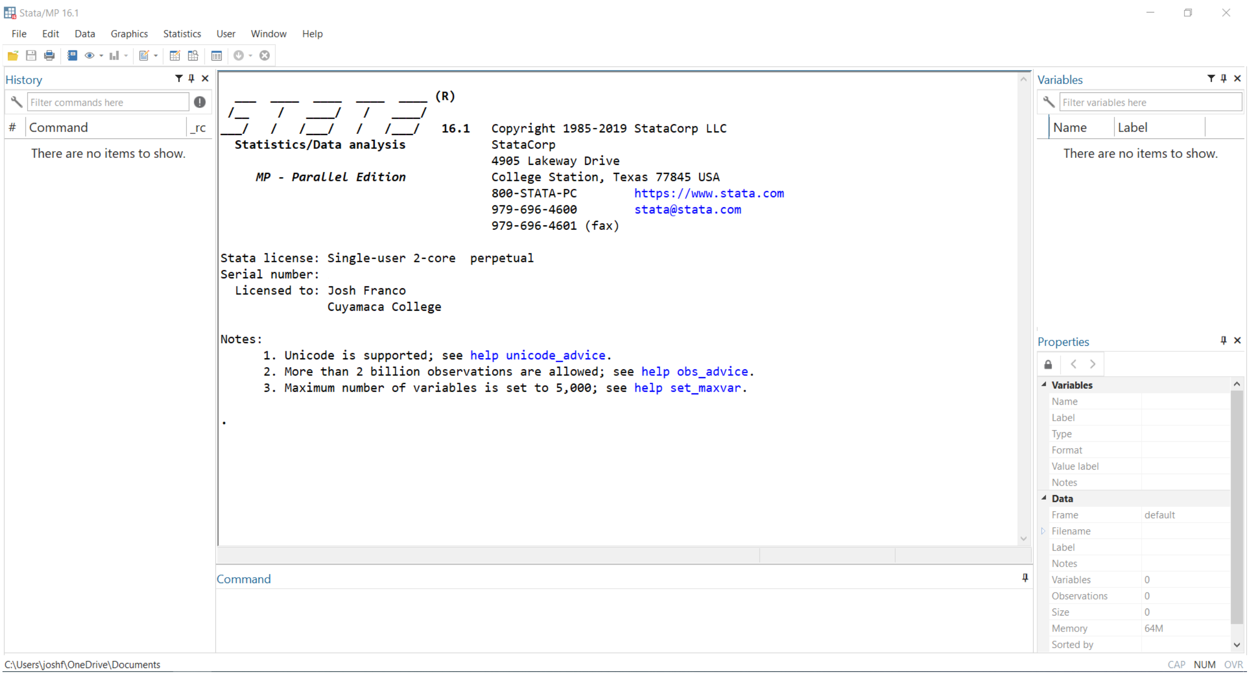


Figure 2-1: User interface of Stata

Now, you might be telling yourself: this is too complicated. Really?

Let us compare Stata to another, more common, user interface: Facebook.com. This interface has six elements:

1. Top Left Search field
2. Top Center Buttons that include Home, Watch, Marketplace, and so on
3. Left navigation bar that includes my name, Friends, and so on
4. Center "What's on your mind, Josh?" field.
5. Center Wall or Stream or whatever it is called now.
6. Right side Birthdays and Contacts

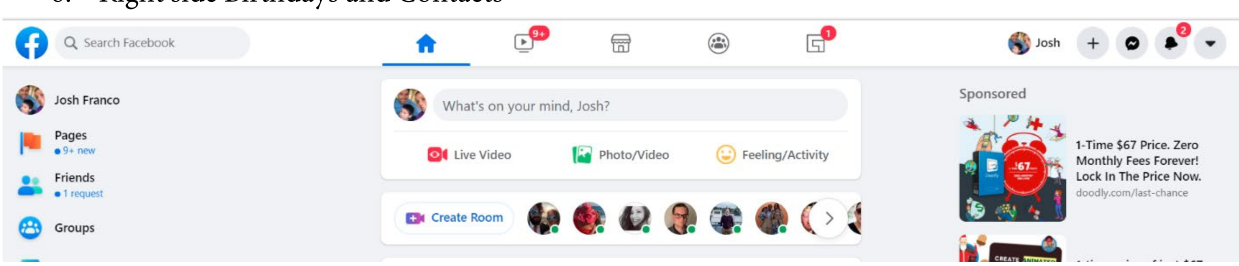


Figure 2-2: User Interface of Facebook.com

Both are readily complicated if you are encountering them for the first time. But, if you take a moment to see the forest (interface) for the trees (elements of the interface), then it does not seem as complicated, or daunting.

Mini-Assignment #1: Instructions

Step 1: Watch Tour of the Stata interface - YouTube

Step 2: Share two parts you found interesting.

- In 4 or more sentences, share what two parts about Stata Tour of the Stata 16 interface - YouTube you found most interesting and state why.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Video: 1st Part: Interesting: What	Yes Missing	25 0
Video: 1st Part: Interesting: Why	Yes Missing	25 0
Video: 2nd Part: Interesting: What	Yes Missing	25 0
Video: 2nd Part: Interesting: Why	Yes Missing	25 0
Explanation: Quantity: # Sentences	4 or more Less than 4	50 0
Quality: Subjective evaluation by Professor	01 – Superb 02 – Excellent 03 – Great 04 – Good 05 – Insufficient	0 0 0 0 0

Getting Started in Stata

Stata, like any software, offers a host of options. Continuing with our comparison of Stata and Facebook.com, there are parts of Facebook.com I use all the time, like the scrolling through the Wall or whatever it is called now. However, there are other parts of Facebook.com that are completely unknown to me. For example, what is this Gaming button all about?

Stata has a mountain of features for cutting-edge statistical and data analysis, and importantly for us, has introductory tools for budding data analysts.

Find Data to Import

Stata itself has Example datasets that you can import. You can access these example datasets by clicking on “File” in the Top Left Menu bar and selecting “Example datasets...”

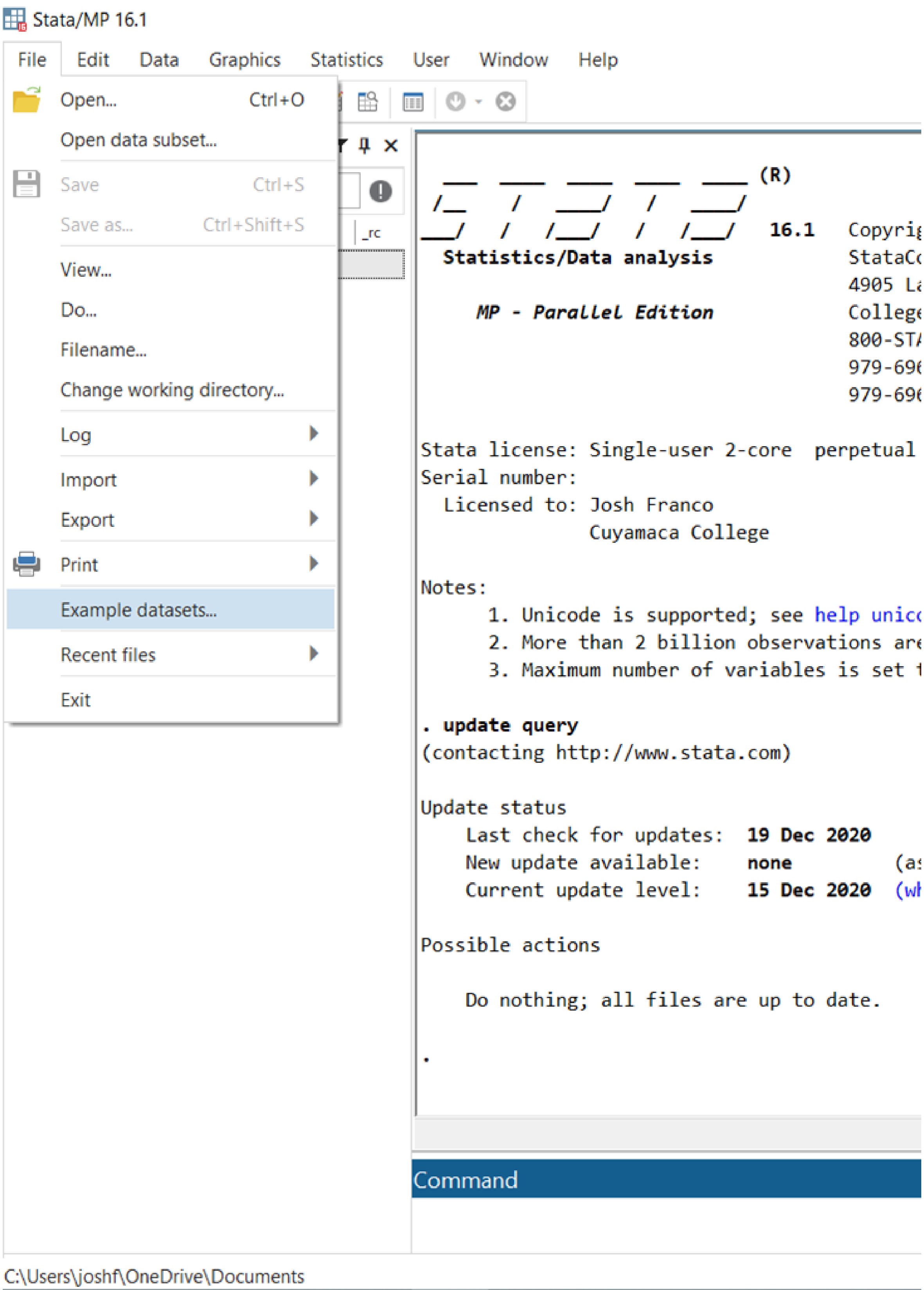


Figure 2-3: Accessing Example Datasets installed with Stata.

Finding non-Stata provided data to import can be chore. Where does someone even begin to find such datasets? Here are at least two repositories of social science related datasets:

- Inter-university Consortium for Political and Social Research (ICPSR)
- The Dataverse Project - Dataverse.org

I will be utilizing the Public Policy Institute of California (PPIC)'s Statewide Survey Data from 2020 throughout the Workbook.

- PPIC Statewide Survey Data - 2020 - Public Policy Institute of California

Import Data into Stata

I went to PPIC Statewide Survey Data - 2020 - Public Policy Institute of California and downloaded the January 2020 Survey Data and the May 2020 Survey Data.

Both Survey Data files download as ZIP (file format) - Wikipedia. ZIP files allow creators to package and compress multiple files into a single ZIP files. In other words, ZIP files are like your multiple item Amazon package that got doubly bubbled wrapped.

After I download and unzip both files, I find two files: a codebook and .sav file. The codebook I can readily open using Microsoft Word or some other word processor. However, the .sav file would be tricky since older versions of Stata could not import this file type. Luckily for us, Stata 16 allows you to import .sav files and convert them into Stata-native .dta files.

You can Import these .sav files by clicking on "File" in the Top Left Menu bar and selecting "Import" then selecting "SPSS data (*.sav). For this example, we will just import the January 2020 Survey Data.

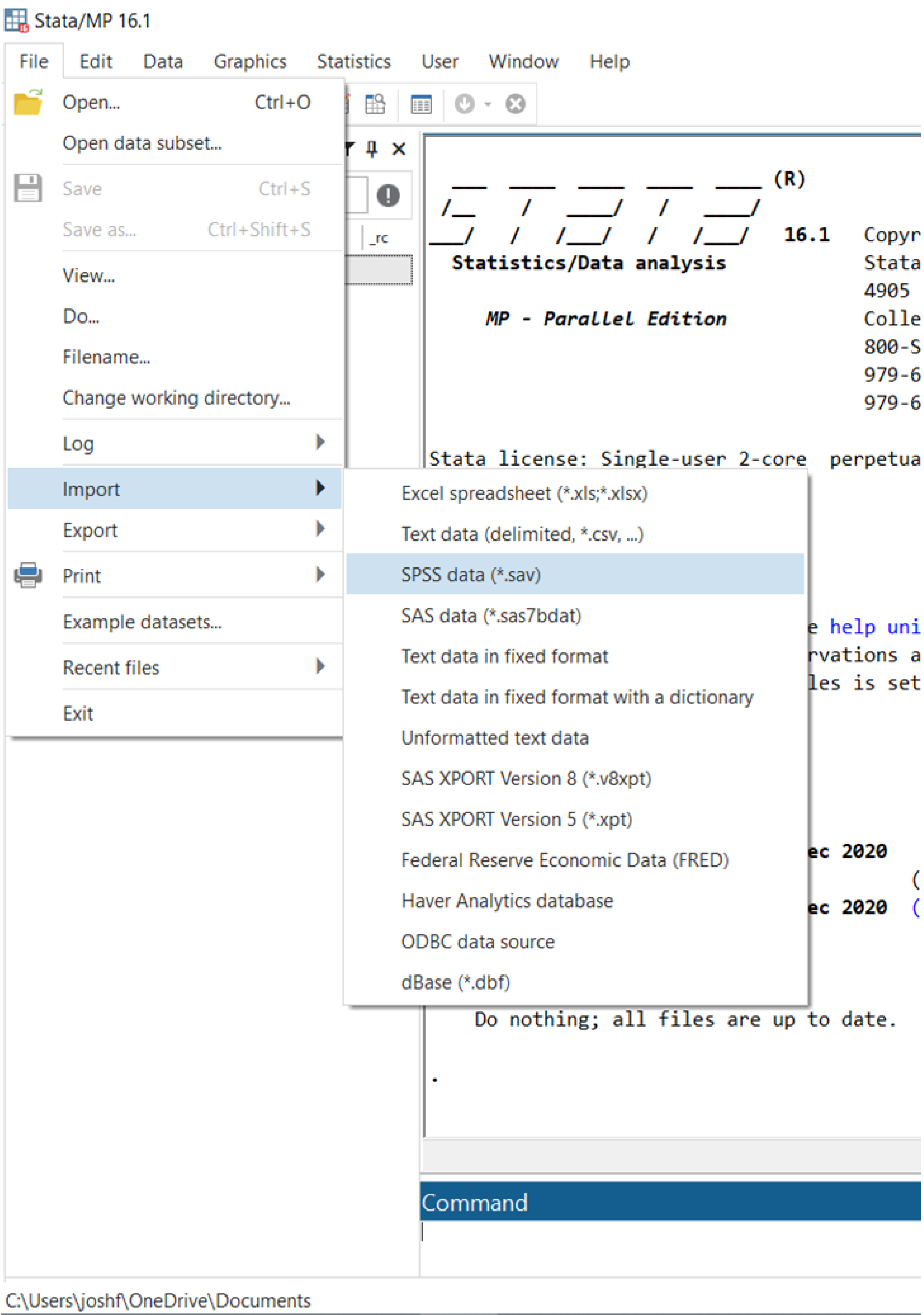


Figure 2-4: Importing a .sav file into Stata.

A dialog box, or a small popup window, titled "import spss – Import SPSS files" will appear. As a side note: dialog boxes are common when you are interactively using Stata, so get use to them appearing as you point-and-click on different menus and cons. Within the dialog box, you can click "Browse…" button near the top right to find the unzipped .sav file on your computer.

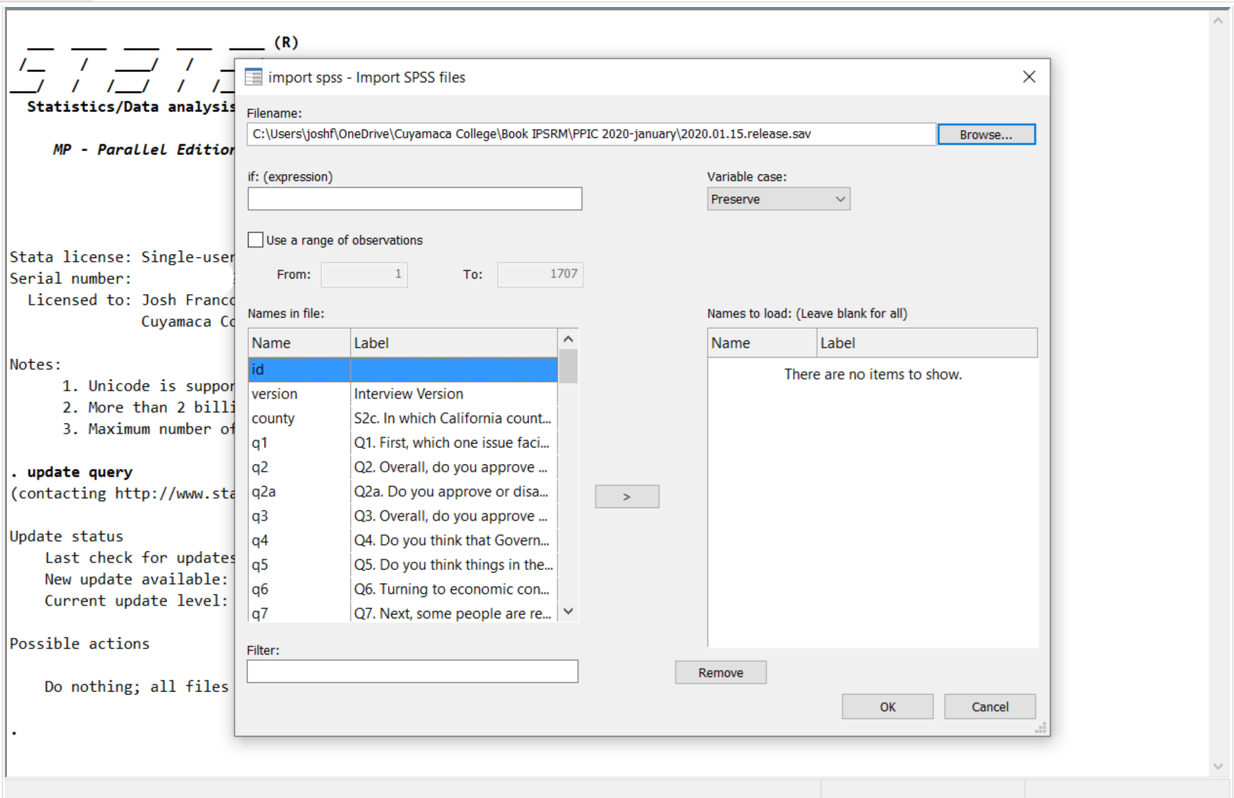


Figure 2-5: Screenshot of the "import spss" Dialogue Box

After you find the file, the location of the file will appear in the "Filename:" field at the top, along with the "Names in file:" field now populated with the variables in the dataset. You are welcome to scroll through the "Names in file:" field to get a sense of the variables and their labels. After that, you can click the "OK" button in the bottom right of the dialog box, and now you have successfully imported your data into Stata !

Review the Data in Stata

For a seasoned data analyst, what we went through above is unremarkable. But do not pay any attention to these haters who have forgotten where they started. Remember your beginnings, and never forget them, because it keeps us humble as to where we have been, where we are, and where we want to go.

By clicking the "OK" button in the prior dialog box, you essentially typed the following text in the Command field at the Command window, located in the center bottom panel:

```
import spss using "C:\Users\joshf\OneDrive\Cuyamaca College\Book IPSRM\PPIC 2020-january\2020.01.15.release.sav"
```

Immediately after that command is executed, the following output appears below:
`(73 vars, 1,707 obs)`

Let us re-examine the Stata user interface. The interface includes seven elements:

1. Top left menu: no change
2. Top left buttons: no change
3. Left panel labeled "History now includes two lines: #1 `update query` and #2 `import spss…`
    a. It is clear to us now that the "History" panel list the history of commands we typed or executed by pointing and clicking.
4. The Results window, located in the center top panel, which contains output of commands. In addition to **`update query`**, we now see **`import spss using "C:\Users\joshf\OneDrive\Cuyamaca College\Book IPSRM\PPIC 2020-january\2020.01.15.release.sav"`**
5. Center bottom panel labeled "Command": no change.
6. Right top panel labeled "Variables" now includes a list of variable Names and their Labels.
    a. Whenever you import a dataset, you'll see the variables listed in the Variables window
7. Right bottom panel labeled "Properties" now has information in the "Data" portion.
    a. We see the Variables field change from blank to 73; Observations change from blank to 1,707; Size change from blank to 138.36k; and Memory change from blank to 64M.

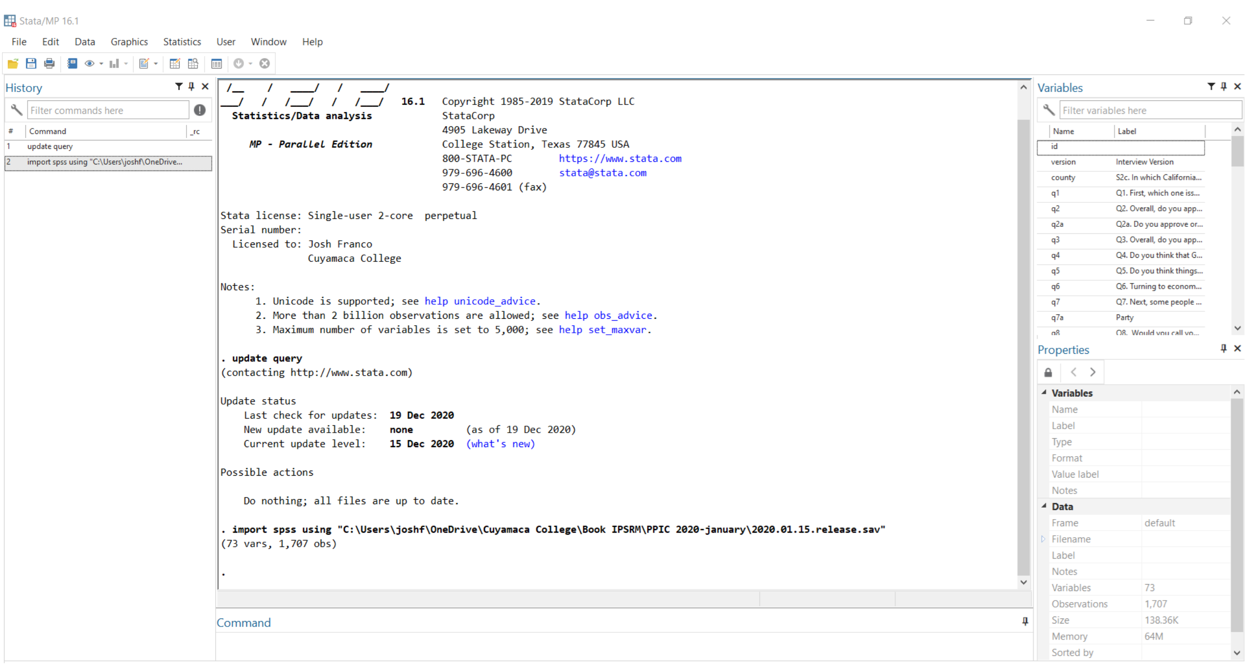


Figure 2-6: Revisiting of the User interface of Stata after importing data.

Describe the Data

In the Command field, we type **`describe`** and the following appears:

```
Contains data
Obs: 1,707
Vars: 73
```

`Obs` stands for observations. In this case, there are 1,707 observations. And `Vars` stands for variables. There are 73 variables in this data.

After this output, we see a table with five columns: variable name, storage type, display format, value label, and variable label. The two columns we are most interested is variable name and variable label. The variable name identifies the column of observations; for example, the variable county records the county in which the respondent lives. Whenever you want to refer to that column of data, you'll use the variable name, regardless if you're typing in the Command window, using a dialog box, or making a selection in the Variables window. The variable label is attached to the variable, and it describes the contents of the variable.

For example, the variable **`county`** has the variable label **`S2c. In which California county do you live?`** Or the variable **`q8`** has the variable label **`Q8.  Would you call yourself a strong Democrat or not a very strong Democrat?`**

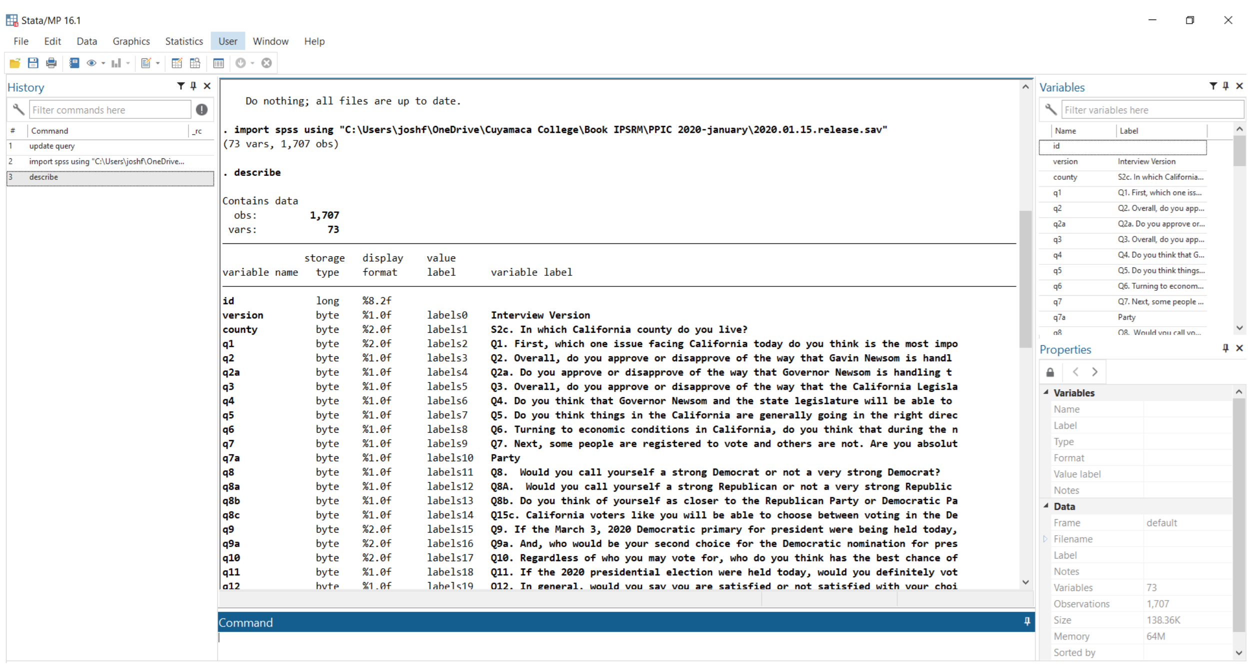


Figure 2-7: The output after typing the command "describe"

Mini-Assignment #2: Instructions

Step 1: Watch What's it like–Getting started in Stata - YouTube.

Step 2: Share two parts you found interesting.

- In 4 or more sentences, share what two parts you found interesting about What's it like–Getting started in Stata - YouTube and explain why.

Mini-Assignment #2: Rubric

Criteria	Ratings	Points
Video: 1st Part: Interesting: What	Yes Missing	25 0
Video: 1st Part: Interesting: Why	Yes Missing	25 0
Video: 2nd Part: Interesting: What	Yes Missing	25 0
Video: 2nd Part: Interesting: Why	Yes Missing	25 0
Explanation: Quantity: # Sentences	4 or more Less than 4	50 0
Quality: Subjective evaluation by Professor	01 – Superb 02 – Excellent 03 – Great 04 – Good 05 – Insufficient	0 0 0 0 0

Chapter 3 - Datasets: Cross-section, Time Series, and Panel

About

Data are typically stored in a tabular manner, with rows and columns. One of the most common datasets are spreadsheets that contain rows and columns. The intersection of rows and columns creates cells. Numeric, alpha, and alphanumeric data can reside in these cells.

The image below is a screenshot of a Microsoft Excel spreadsheet, a very common software. There are four rows marked: 1, 2, 3, and 4; and there are four columns marked: A, B, C, and D. These 4 rows and 4 columns create 16 cells. Cells A1, B1, and C1 are populated with the following data: "123" (numeric), "abc" (alpha), and "123abc" (alphanumeric), respectively. Note that the remaining 13 cells are empty.

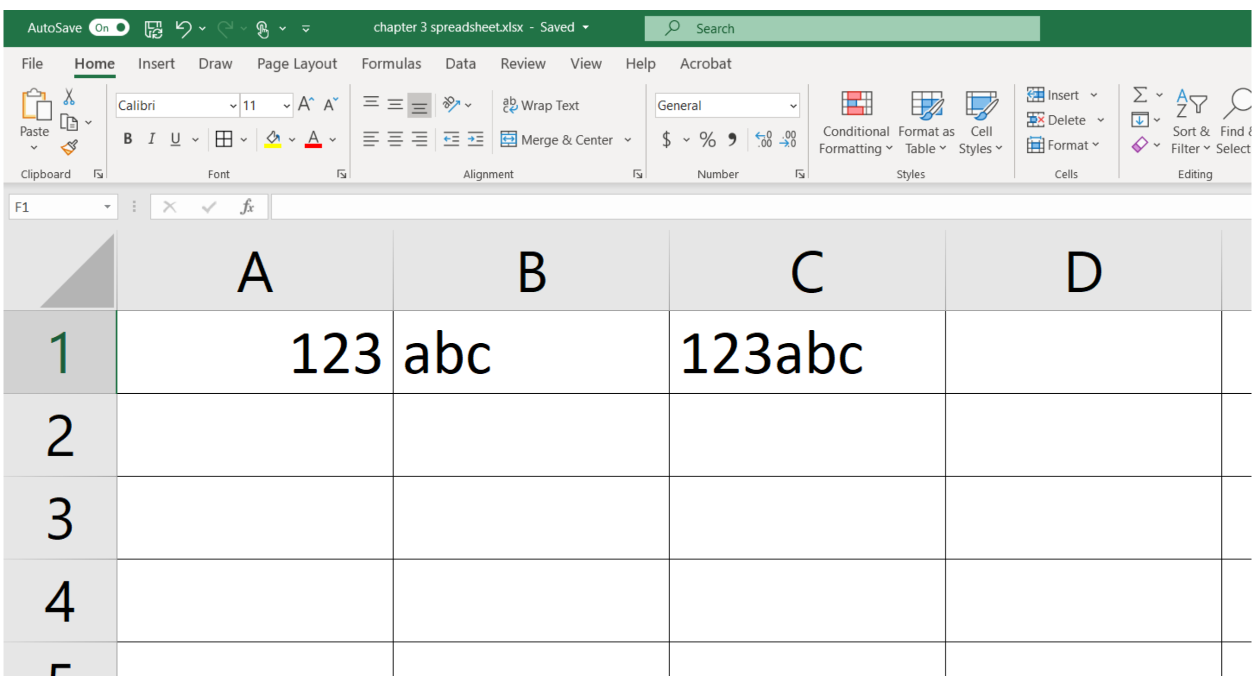


Figure 3-1: Screenshot of Excel spreadsheet with 4 rows and 4 columns

Estimated Time

An estimated 90-120 minutes is needed to complete this activity.

Cross-Sectional dataset

Cross-sectional data contain information on many subjects, or units, for a single time period.

Observations can be persons, cities, states, countries, legislation, committees, schools, and so on. Variables are concepts that are being measure, or observed, and they have at least two values. For example, the variable `age` can have values from 0 to 100+. Or the variable `race` can have the values African American, White, Hispanic, Asian American, and so on.

For an example of a cross-sectional dataset, I updated the Microsoft Excel spreadsheet from above. In cells A1 through E1 included the variable name. It is common to use the 1st row of cells to state the variable name of each column. In rows 2 through 5, I have four notable people listed: Cardi B, Joe Biden, Dolores Huerta, and Andrew Yang. For each person, I have information about their `gender`, `age`, `race`, and `year` the data was collected.

The data is cross-sectional because we are looking at many objects (notable persons) in a single time period (year 2020).

	A	B	C	D	E
1	**name**	**gender**	**age**	**race**	**year**
2	Cardi B	Female	28	African American	2020
3	Joe Biden	Male	78	White	2020
4	Dolores Huerta	Female	90	Hispanic	2020
5	Andrew Yang	Male	45	Asian American	2020

Figure 3-2: Example of a cross-sectional dataset

Time Series dataset

With time-series data, we are looking at a single subject, or object, over multiple time periods.

Below we have some time-series data on Cardi B, one of the notable individuals from the cross-sectional datasets. In cells A1 through F1, we see six variables: `name`, `gender`, `age`, `race`, `year`, and `singlerecords`. The variable `singlerecords` refers to the number of single songs with Cardi B as lead artist ([Cardi B discography - Wikipedia](#)).

The data is time series because we are looking at one object (Cardi B) over multiple time periods (years 2017 to 2020). And in this case, our variables `age`, `year`, and `singlerecords` change for each row of data.

	A	B	C	D	E	F
1	**name**	**gender**	**age**	**race**	**year**	**singlerecords**
2	Cardi B	Female	25	African American	2017	3
3	Cardi B	Female	26	African American	2018	4
4	Cardi B	Female	27	African American	2019	3
5	Cardi B	Female	28	African American	2020	1

Figure 3-3: Example of a time series dataset

Panel dataset

Panel data contains information on multiple objects for multiple time periods.

For an example of a panel dataset, I updated the time series dataset to include a second musical artist: Harry Styles. Again, in cells A1 through F1, we see six variables: `name`, `gender`, `age`, `race`, `year`, and `singlerecords`.

The data is panel because we are looking at multiple objects (Cardi B and Harry Styles) over multiple time periods (years 2017 to 2020). And again, our variables `age`, `year`, and `singlerecords` change for each row of data for each artist. For example, for year 2017, both Cardi B and Harry Styles ([Harry Styles discography - Wikipedia]) had 3 single records. But in year 2019, Cardi B had 3 compared to Harry's 2 singles.

	A	B	C	D	E	F
1	**name**	**gender**	**age**	**race**	**year**	**singlerecords**
2	Cardi B	Female	25	African American	2017	3
3	Cardi B	Female	26	African American	2018	4
4	Cardi B	Female	27	African American	2019	3
5	Cardi B	Female	28	African American	2020	1
6	Harry Styles	Male	23	White	2017	3
7	Harry Styles	Male	24	White	2018	0
8	Harry Styles	Male	25	White	2019	2
9	Harry Styles	Male	26	White	2020	3

Figure 3-4: Example of a panel dataset

Mini-Assignment #1: Instructions

Step 1: Select 1 dataset type that interests you.

Your dataset choices are:

- Cross-sectional
- Time series
- Panel

Step 2: In 4 or more sentences, explain why you selected this dataset type.

- To help write your explanation, consider the following questions:
  - What is one strength of the dataset you selected?
  - What is one weakness of the dataset your selected?
  - How does your dataset compare to one of the other datasets?

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Dataset Type: Selected	Yes Missing	50 0
Why Dataset Type: # sentences	4 3 2	100 75 50

	1	25
	Missing	0

361

Chapter 4 - Data Management

About

Data management is the structure and process by which you organize and manage your data and datasets. Often overlooked, data management is a key process to be aware of and implement for projects, small and large. There are at least seven features to be aware of related to data management: storing, sourcing, folders, files, version control, base dataset, and variable data.

Estimated Time

An estimated 90-120 minutes is needed to complete this activity.

Big Picture

Data management sounds complicated because it consists of several part. However, it is useful to see the forest for the trees. In other words, the forest seems vast and mighty, but when you look at individual trees, you can begin to appreciate the simplicity and complexity of both the forest and the trees.

The basic radial diagram helps us visualize the big picture. At the center is Data Management and radiating out from this are the seven features mentioned above: storing, sourcing, folders, files, version control, base dataset, and variable data. Keep this visualization in mind as we proceed in learning about each feature.

Figure 4-1: Seven features of data management

Storing Data

Storing data is about where you are going to save the data you will be working with. You can store data in three places: 1) on your personal computer, 2) on an external thumb drive or hard drive, or 3) in the "cloud" via the Internet.

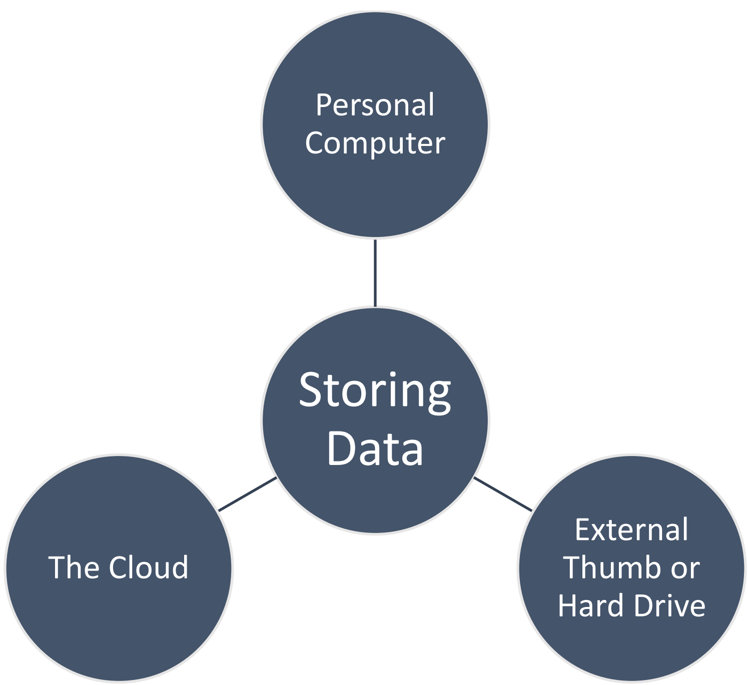


Figure 4-2: Three options for storing data.

Your personal computer is the logical place to store your data because it is your computer, and you use it regularly. One drawback to storing the data only on your computer is that if the computer fails, then all your data are likely lost, or very costly to retrieve.

The second place to store your data is on an external thumb drive or hard drive. Thumb drives are common these days, and not expensive, but you can lose since they are small devices. External hard drives are also available, but a bit costly.

The third place, which I am going to recommend storing your data, is in the cloud. The growth of cloud computing, such as Google Drive or DropBox, over the last 10 years is changing the nature of how we interact with our computers and data.

A drawback of the cloud is that you need an internet connection to retrieve the data. Thus, if your internet is spotty or you have lost power, then you will not be able to access your data. However, the advantage the cloud has to the other storage forms is that there is a backup of your data.

Sourcing Data

Sourcing data is about the source of where you find the data you want to analyze. There is a growing mountain of data sources. Recall from Chapter 2, I mentioned three specific data sources:

- Inter-university Consortium for Political and Social Research (ICPSR)
- The Dataverse Project
- PPIC Statewide Survey Data - 2020 - Public Policy Institute of California

And there are other data sources as well. For example, consider the following:

- Data.gov
- U.S. Census Data
- California Open Data
- GSS General Social Survey | NORC
- ANES | American National Election Studies
- Cooperative Congressional Election Study
- San Diego County Data Portal

Up until this point, I have used data and datasets interchangeably. But, after introducing you to sourcing data, it is important to make a distinction between these terms. Data is a general term used to describe text (alpha, numeric, alphanumeric), images, audio, and video. All these can be considered data.

However, a dataset is a meaningful collection of data organized by an individual or team. Datasets can be created by you, not created by you, or a combination of the two.

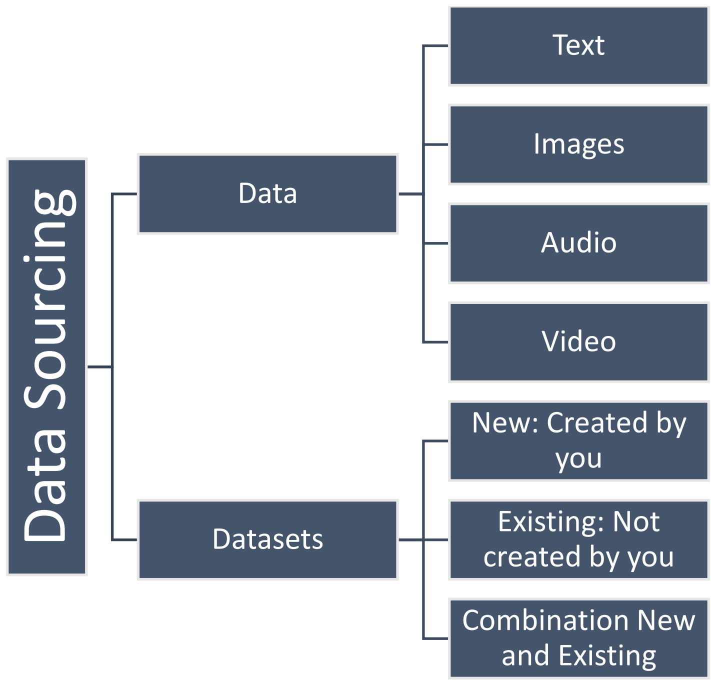


Figure 4-3: Making a distinction between data and datasets.

Non-academic example of using an existing dataset.

For example, I attended UC Merced during 2005-2007 and again from 2012-2018. During this time, I met a fellow Bobcat named Michael Urner. Michael co-founded Tergis Technologies, "a company developing new medical devices to reduce the number of hospital-acquired infections."[1] During a UC Merced Venture Lab presentation, Michael shared how he used Centers for Disease Control and Prevention's National Vital Statistics System datasets to quantify the demand for his medical device. I thought this was a novel way of how a business entrepreneur can use an existing dataset.

Academic example of creating a new dataset and using an existing dataset.

Another example, this time from my Ph.D. dissertation titled *Judicial Pork: The Congressional Allocation of Districts, Seats, Meeting Places, and Courthouses to the U.S. District Courts*. And I collected data from the Federal Judicial Center | (fjc.gov) to create a new dataset of federal court districts, seats, meeting places, and courthouses. I combined this dataset with existing datasets, such as Charles Stewart's congressional committee data, to form a "super dataset" that I then analyzed for my research.

[1] Alum Wins Opportunity to Pitch at Venture Summit | Newsroom (ucmerced.edu)

Folders

Folders are the folders on your cloud drive that you place other folders and files into. You can think of folders like containers where you store files, like Word docs, pictures, and datasets.

Figure 4-4: Extra-large icon of my Google Drive "00 Stata Workbook" Folder

Files

Files are the files that you store in folders located on your cloud drive. Below is a list of files that are likely to work with when conducting data analysis:

1. Text files (.txt)
2. Comma separated files (.csv)
3. Word Documents (.doc or .docx)
4. Pictures (.gif, .jpg, .png)
5. Audio (.mp3)
6. Videos (.mp4, .mov)
7. Excel spreadsheets (.xls or .xlsx)
8. Stata project files (.stpr)
9. Stata Do-files (.do)
10. Stata datasets (.dta)
11. PowerPoints (.pptx)

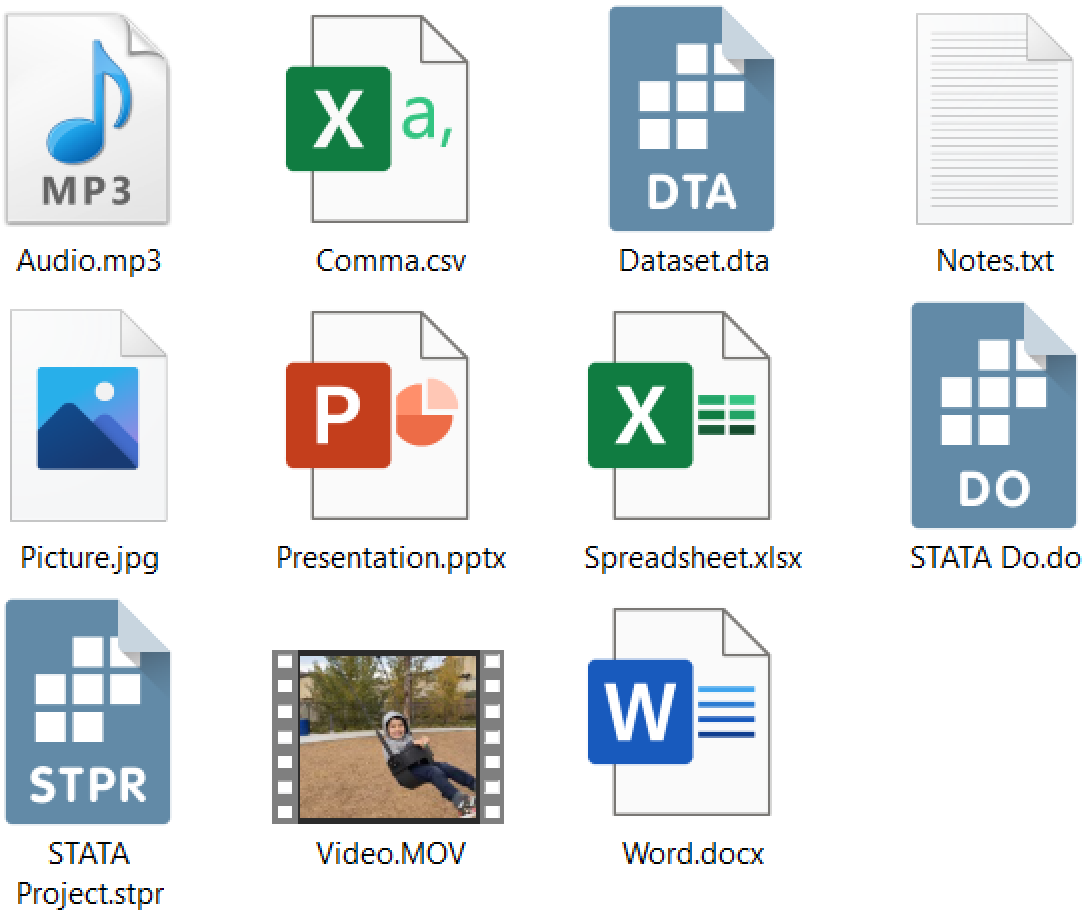


Figure 4-5: Large icons of different file types you are likely to use for data analysis.

Version Control

According to [Wikipedia](), "In software engineering, version control (also known as revision control, source control, or source code management) is a class of systems responsible for managing changes to computer programs, documents, large web sites, or other collections of information. Version control is a component of software configuration management.[1]"

Version control is important for data management because we need a systematic way of managing the folders and files that we are working with. Any data analysis project, simple or complex, requires organization and processes that are thought out and through in advance.

Students can use the -version- command in Stata to make sure that their programs will continue to work with future releases of Stata.

Folder Structure

Luckily for us, Stata has a Project Manager that helps us organize our folders and files. Below is sample outline of folders and files that you can use for a data analysis project:

- Project Folder
    - .stpr (Stata Project Manager file)
    - 00 Log folder
    - 01 Do folder
        - 00 Master.do
        - Other .do files
    - 02 Data Source folder
        - Websites links
        - .csv
        - .xlsx
        - .dta
    - 03 Variables folder
        - .do
    - 04 Datasets folder
        - .dta
    - 05 Models folder
        - .do
    - 06 Graphs folder
        - .gph
        - .png
    - 07 Tables folder
        - .rtf
        - .doc
    - 08 Papers folder
        - .doc
    - 09 Presentations folder
        - .pptx

Google Drive > 00 STATA Workbook > Chapter 4 > Example of Project

Name	Date modified	Type	Size
00 Log	12/20/2020 4:32 PM	File folder	
01 Do	12/20/2020 4:32 PM	File folder	
02 Data source	12/20/2020 4:32 PM	File folder	
03 Variables	12/20/2020 4:32 PM	File folder	
04 Datasets	12/20/2020 4:32 PM	File folder	
05 Models	12/20/2020 4:32 PM	File folder	
06 Graphs	12/20/2020 4:32 PM	File folder	
07 Tables	12/20/2020 4:33 PM	File folder	
08 Papers	12/20/2020 4:33 PM	File folder	
09 Presentations	12/20/2020 4:33 PM	File folder	
Project Name.stpr	12/20/2020 4:31 PM	Stata Project	0 KB

Figure 4-6: Visual representation of Project folder structure

As you read through the list and view the visual representation of a project's folder structure, you may be asking "Why number folders starting with 00, 01, 02, and so on?" The reason is that computers sort folder and files alphabetically. However, I need folders sorted according to data management principles and data analysis processes. The nine folders included in this example follow a linear process for completing data analysis projects:

- keep a log of what you are doing.
- do what you need to do.
- catalog your data sources.
- define and organize your variables.
- organize and use your datasets.
- apply models to analyze datasets.
- store graphs and charts created from your models.
- store tables resulting from your models.
- write papers based on your findings.
- present your findings to others.

File Naming

File naming is another way of maintaining version control. I use the two-number prefix (aka 00 File name, 01 File name, etc.) for do files, variable files, and dataset files. For example, here is a list of eight different file naming conventions:

- 00 Master.do: runs all the other .do files.
- 01 Base dataset.do: prepares the theoretically possible base dataset (more on this in the next section)
- 02 Data source 01.do: converts a 1st data source from .xlsx to .dta
- 03 Variables.do: brings in variables and variable labels into the base dataset.
- 04 Dataset 01.do: combines or merges converted datasets into the Base dataset.
- 05 Dataset Analysis Ready.do: ensures the combined dataset is ready for analysis.
- 06 Descriptive.do: produces descriptive statistics and cross-tabulations on the Analysis Ready dataset
- 07 Models.do: runs theoretically informed statistical and data analysis models on the Analysis Ready dataset

Files with a two-number suffix could have additional similarly named files. For example, "02 Data source 01.do" is referring to a single data source. However, what if you are relying on two or more data sources? In this case, you can create a second file named "02 Data source 02.do" and so forth.

While no data analysis project is completed in a linear fashion, it can be organized in a linear fashion so that you can replicate the process for yourself, and others can examine your work, if needed or required.

Base Dataset

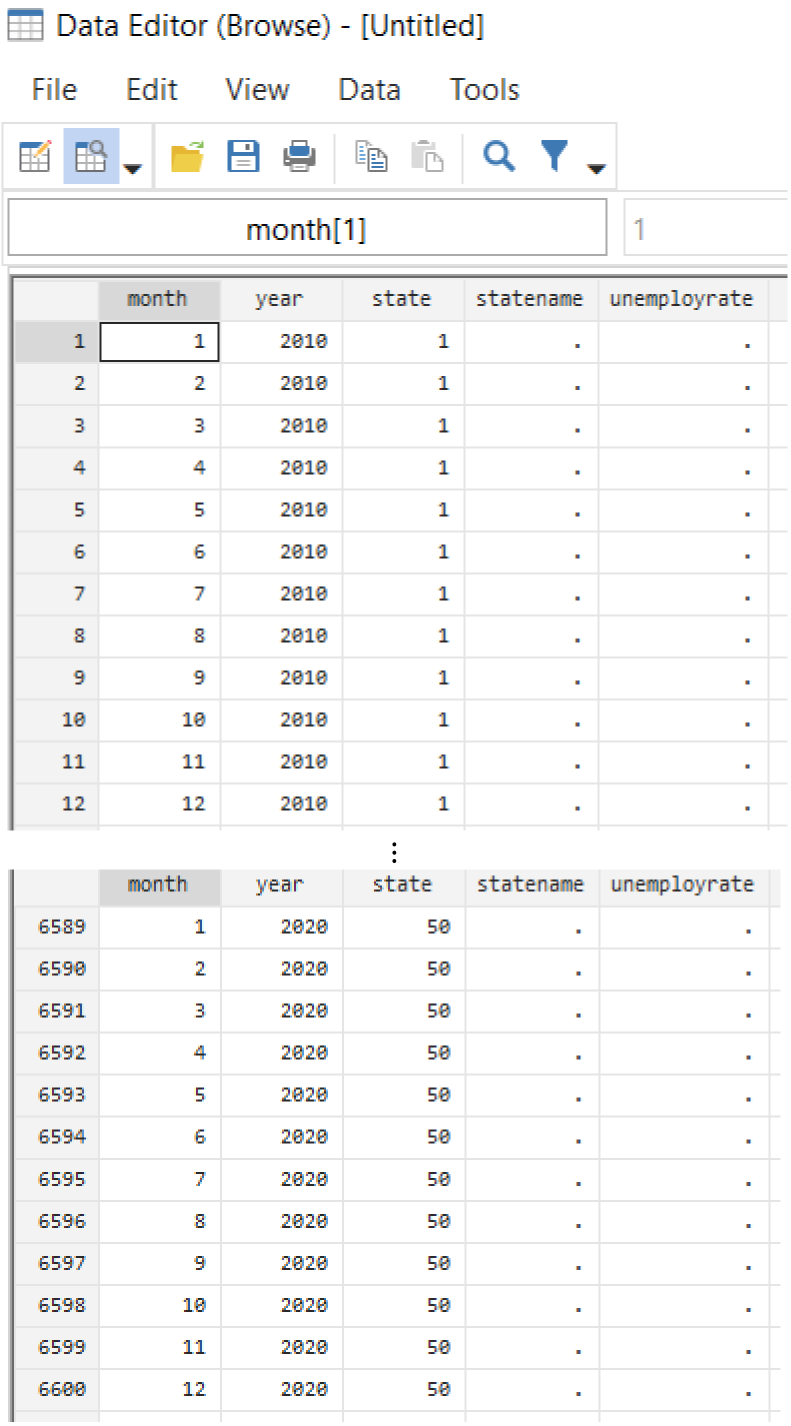
A base dataset is the theoretically possible dataset given your research question, theory, and research design. For example, what if I wanted to analyze the unemployment rate of all 50 U.S. states over the last 10 years by month. What type of dataset do I need? And how many observations (rows of data) could I possibly have?

The answer is a panel dataset that contains 6,600 observations (50 U.S states times 11 years times 12 months). You may be asking: "Why 11 years instead of 10 years, didn't you say over 10 years?" Yes, but if you count 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, and 2020, that is eleven unique years.

How can we create this base dataset in Stata? The answer may be to point-click-drag in Microsoft Excel or Google Sheets, or you can use Stata and the following code:

```
clear
set obs 6600
egen float month = seq(), from(1) to(12) block(1)
egen float year = seq(), from(2010) to(2020) block(600)
egen float state = seq(), from(1) to(50) block(12)
gen statename = .
gen unemployrate = .
```

These commands generate the following empty dataset of 6,600 observations of state-year-months:



	month	year	state	statename	unemployrate
1	1	2010	1	.	.
2	2	2010	1	.	.
3	3	2010	1	.	.
4	4	2010	1	.	.
5	5	2010	1	.	.
6	6	2010	1	.	.
7	7	2010	1	.	.
8	8	2010	1	.	.
9	9	2010	1	.	.
10	10	2010	1	.	.
11	11	2010	1	.	.
12	12	2010	1	.	.

⋮

	month	year	state	statename	unemployrate
6589	1	2020	50	.	.
6590	2	2020	50	.	.
6591	3	2020	50	.	.
6592	4	2020	50	.	.
6593	5	2020	50	.	.
6594	6	2020	50	.	.
6595	7	2020	50	.	.
6596	8	2020	50	.	.
6597	9	2020	50	.	.
6598	10	2020	50	.	.
6599	11	2020	50	.	.
6600	12	2020	50	.	.

Figure 4-7: Base dataset of 6600 observations of state-year-months

Taking the "Base dataset" approach helps you think of the data magnitude of the research question you are trying to answer. I do not think one should shy from large datasets, since that is partly the strength of data analysis today, but this approach allows you to be honest with yourself about the work that lies ahead.

Variable Data

Variable data are the columns of data that you want to have in your Analysis Ready dataset. However, going from idea to reality will take work, especially if you are creating a dataset from scratch. Let me demonstrate with an existing dataset compared to a new dataset.

Existing Dataset

Existing datasets may simply be a spreadsheet to you, but to the person or team who created it and populated the cells with data, it is proud achievement. While computers have eased this process in some ways, it can be as tedious as a cell-by-cell hand entry. Therefore, you should not sneeze at any dataset, because the amount of time and effort could be in the hundreds or thousands of hours.

Existing datasets already have variable data in them. Let us return to the Public Policy Institute of California (PPIC)'s data that I shared in Chapter 2. I visited PPIC Statewide Survey Data - 2020 - Public Policy Institute of California, downloaded the January 2020 Survey Data, unzipped the folder, and uploaded the .sav file onto the *Introduction to Political Science Research Methods* website. Now, I open Stata software and type the following command:

```
import spss using "https://www.ipsrm.com/stata/2020.01.15.release.sav"
```

After a few moments, 1,707 observations (rows of data) and 73 variables (columns of data) are loaded into Stata and essentially ready for analysis.

Creating a New Dataset

Creating new datasets is a lot of work. It takes time, planning, and perseverance. I will share a short story about my dissertation *Judicial Pork: The Congressional Allocation of Districts, Seats, Meeting Places, and Courthouses to the U.S. District Courts*.

Below are the lines of code from the section "Generate Datasets" in my Master.do file for my dissertation. Lines that begin with the asterisk (*) symbol are not processed by Stata, so they serve as an informative note.

See that I created the base dataset on Sunday, March 12, 2017. Through that month, I was importing variables left and right, then matters slowed down.

Another batch of variables were imported in June 2017, and then again February 2018, with the final variable imported on Sunday, April 15, 2018, a full year later.

```
*      3) Generate Datasets
* Create Base Dataset and Variables
run ".\Do\Stata Create Database of States Years.do" // Created 01, Completed: Sunday, March 12,
2017; UPDATED Friday, July 7, 2017 to add years 1787 and 1788

* Populate Variables into Base Dataset
run ".\Do\DpV_JDt.do" // Created 02, Completed: Sunday, March 12, 2017
run ".\Do\DpV_JSt.do" // Created 03, Completed: Monday March 13, 2017
run ".\Do\DpV_JMP.do" // Created 04, Completed: Saturday, March 18, 2017
run ".\Do\DpV_JCt.do" // Created 05, Completed: Saturday, March 18, 2017, Update: 3/8/18 added
GAO list
run ".\Do\IdV_S_MajLdr.do" // Created 06, Completed: Sunday, March 19, 2017
run ".\Do\IdV_S_MinLdr.do" // Created 07, Completed: Sunday, March 19, 2017
run ".\Do\IdV_S_JChair.do" // Created 08, Completed: Sunday, March 19, 2017 (from Judgeships
project)
*run ".\Do\IdV_S_JRkMbr.do" // Skipped, Data Not Readily Available, would require archival
research
run ".\Do\IdV_S_JMbr.do" // Created 10, Completed: Thursday, March 30, 2017, Updated 10/21/17
*run ".\Do\IdV_S_JMaj.do" // 11, Skip
*run ".\Do\IdV_S_JMin.do" // 12, Skip
*run ".\Do\IdV_S_Chrs.do" // 13, Skipped, Maybe refine to Appropriations Full/Sub Committee
Chairs ONLY 3/23/17
*run ".\Do\IdV_S_RkMbrs.do" // 14, Skipped, Data Not Readily Available, would require archival
research
run ".\Do\IdV_HR_Spkr.do" // Created 15, Completed: Thursday, March 30, 2017
run ".\Do\IdV_HR_MajLdr.do" // Created 16, Completed: Thursday, March 30, 2017
run ".\Do\IdV_HR_MinLdr.do" // Created 17, Completed: Thursday, March 30, 2017
run ".\Do\IdV_HR_JChair.do" // Created 18, Completed: Thursday, March 30, 2017
*run ".\Do\IdV_HR_JRkMbr.do" // 19, Skipped, Data Not Readily Available, would require archival
research
run ".\Do\IdV_HR_JMbr.do" // Created 20, Completed: Thursday, March 30, 2017
*run ".\Do\IdV_HR_JMaj.do" // 21, Skip
*run ".\Do\IdV_HR_JMin.do" // 22, Skip
run ".\Do\IdV_HR_Rules.do" // Created 23, Completed: Thursday, March 30, 2017
*run ".\Do\IdV_HR_Chrs.do" // 24, Skipped, Maybe refine to Appropriations Full/Sub Committee
Chairs ONLY 3/23/17
*run ".\Do\IdV_HR_RkMbrs.do" // 25, Skipped, Data Not Readily Available, would require archival
research
run ".\Do\Ctrl_JVac.do" // Created 26, Completed, Thursday, April 13, 2017
run ".\Do\Ctrl_StPop.do" // 27, Completed: Thursday, April 13, 2017
*run ".\Do\Ctrl_StPopChange_DpV_JDt.do" // 28, Skip
*run ".\Do\Ctrl_StPopChange_DpV_JSt.do" // 29, Skip
*run ".\Do\Ctrl_StPopChange_DpV_JMP.do" // 30, Skip
*run ".\Do\Ctrl_StPopChange_DpV_JCt.do" // 31, Skip
run ".\Do\Ctrl_POTUS.do" // Created 32, Completed: Thursday, April 13, 2017
*run ".\Do\Ctrl_VPOTUS.do" // 33, Skip
run ".\Do\Ctrl_JDtBalance.do" // 34, Completed: Friday, February 16, 2018
* 35-38 renumbered to 51-54 and relocated
run ".\Do\Other_StatehoodYear.do" // Created 39, Completed: Thursday, April 13, 2017
run ".\Do\Other_StateGeoSizeSqMi.do" // Created 40, Completed: Saturday, May 27, 2017
run ".\Do\Ctrl_UnifiedGov.do" // Created 41, Completed: Thursday, April 13, 2017
run ".\Do\Ctrl_S_ApChr.do" // Created 42, Completed: Sunday, April 16, 2017
run ".\Do\Ctrl_HR_ApChr.do" // Created 43, Completed: Sunday, April 16, 2017
run ".\Do\Ctrl_HR_WMChr.do" // Created 44, Completed: Tuesday, April 18, 2017
run ".\Do\Ctrl_S_PWChr.do" // Created 45, Completed: Thursday, May 11, 2017
run ".\Do\Ctrl_HR_PWChr.do" // Created 46, Completed: Tuesday, May 9, 2017
run ".\Do\DpV_JDt_Dummy.do" // Created 47, Completed: Sunday, June 11, 2017
run ".\Do\DpV_JSt_Dummy.do" // Created 48, Completed: Sunday, June 11, 2017
run ".\Do\DpV_JMP_Dummy.do" // Created 49, Completed: Sunday, June 11, 2017
run ".\Do\DpV_JCt_Dummy.do" // Created 50, Completed: Sunday, June 11, 2017
run ".\Do\Ctrl_TimeSinceLast_DpV_JDt.do" // Created 51, Completed: Sunday, February 18, 2018
run ".\Do\Ctrl_TimeSinceLast_DpV_JSt.do" // Created 52, Completed: Sunday, February 18, 2018
run ".\Do\Ctrl_TimeSinceLast_DpV_JMP.do" // Created 53, Completed: Sunday, February 18, 2018
run ".\Do\Ctrl_TimeSinceLast_DpV_JCt.do" // Created 54, Completed: Sunday, February 18, 2018
run ".\Do\Ctrl_JudConf_JSt.do" // Created 55, Completed: Sunday, April 15, 2018
```

```
*modern area post-1945
* new state emergence, population growth, workload changes, courts (Judicial Conference)

run ".\Do\AR_01.do" // Completed: Thursday, May 11, 2017
```

You will also notice that there are many asterisks, which means I skipped collecting and eventually importing that data. At some point, what you have planned for your research does not pan out, so you need to choose how you are going to spend your limited time.

I want to draw your attention to the following lines of code, because I had to personally collect the data related to these variables.

```
run ".\Do\DpV_JDt.do" // Created 02, Completed: Sunday, March 12, 2017
run ".\Do\DpV_JSt.do" // Created 03, Completed: Monday March 13, 2017
run ".\Do\DpV_JMP.do" // Created 04, Completed: Saturday, March 18, 2017
run ".\Do\DpV_JCt.do" // Created 05, Completed: Saturday, March 18, 2017, Update: 3/8/18 added
GAO list
```

My research question was: How does Congress structure the Judiciary, specifically the organization of the lower District Courts? To answer this question, I needed variable data for judicial districts, judicial seats, judicial meeting places, and judicial courthouses.

Let me show you what just one of these .do files included. Below are the lines of code for the "DpV_JCt.do" file, which stands for Dependent Variable – Judicial Courthouses.

```
* DO File for DpV_JCt
* History:
* Created: Saturday, March 18, 2017
* Data Collection Process and Data Coding Instructions
* I collected data on Judicial Courthouses from the Federal Judiciary Center (FJC)'s website.
* First, I went to http://www.fjc.gov/history/courthouses.nsf.
* Second, I selected a state from the drop-down menu on the left hand navigation bar.
* Third, I reviewed the list of courthouse locations
* Forth, I clicked on the link of the courthouse location and collected the following
information: city and state of courthouse, year completed, supervising architect, year extension
completed, and status of courthouse.
* I then inputted this data into a Microsoft Excel spreadsheet and subsequently imported into
Stata.

clear
import excel "C:\Users\joshf\OneDrive\Dissertation\Data\Judicial Courthouses\USDC Courthouses
01.xlsx", sheet("for Stata") firstrow
duplicates tag, generate(dup)
duplicates list
duplicates list id_icpsr_statenamelower year
duplicates drop id_icpsr_statenamelower year, force
drop dup
drop if year<1789 // 2 Observations deleted
save ".\DTAs\DpV_JCt.dta", replace

clear
use ".\DTAs\04.dta"
drop DpV_JCt
merge 1:1 id_icpsr_statenamelower year using
"C:\Users\joshf\OneDrive\Dissertation\Data\Stata\DTAs\DpV_JCt.dta"
label variable DpV_JCt "Judicial Courthouses"

* Note: I do not have a DpV_JCt_Tot variable: January 18, 2018
drop _merge
save ".\DTAs\05.dta", replace
```

```
* Completed: Saturday, March 18, 2017
```

Now, the Microsoft Excel spreadsheet (“`USDC Courthouses 01.xlsx`”) I imported was originally created on March 2, 2017 and then over the next two weeks, I collected the data from the Federal Judiciary Center website and prepared the data so it could be imported into Stata on March 18, 2017.

Mini-Assignment #1: Instructions

Step 1: Declare a research question.

Step 2: Declare an independent variable (aka explanatory variable aka cause) that is derived from your research question.

Step 3: Declare a dependent variable (aka outcome variable aka effect) that is derived from your research question.

Step 4: Declare at least 3 search terms.

List search terms you would type in a Google search to try to find a dataset related to your research question.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Research question declared	Yes Missing	25 0
Independent variable declared	Yes Missing	25 0
Dependent variable declared	Yes Missing	25 0
Google search terms: #	3 2 1 Missing	75 50 25 0

Chapter 5 - Descriptive Statistics

About

Descriptive statistics include the number of observations, the number of variables, the mean, median, and mode of each variable, and charts and graphs that help to visualize these statistics.

Descriptive statistics are under the umbrella of exploratory analysis. Exploratory analysis is exploring and examining the data, partly to feed your curiosity, but also to search for outliers and other unexpected features of the data.

Estimated Time

An estimated 90-120 minutes is needed to complete this activity.

How do I produce descriptive statistics in Stata?

Producing descriptive statistics consists of a 4-part process: import, describe, summarize, and tabulate. Below is a basic bending diagram to visualize the production of descriptive statistics in Stata.

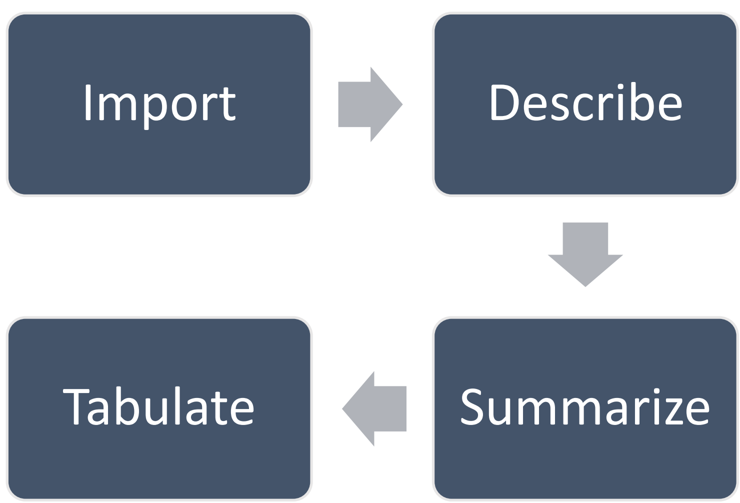


Figure 5-1: Basic bending process diagram to visualize the production of descriptive statistics.

Part 1: Import dataset into Stata

Let us use the Public Policy Institute of California's (PPIC)'s Statewide Survey Data for January 2020 Survey Data that I introduced in prior chapters.

Open Stata. Once in Stata, type in the following command:

```
import spss using "https://www.ipsrm.com/stata/2020.01.15.release.sav"
```

After this command is executed, you will see the following output:

```
(73 vars, 1,707 obs)
```

This output means there are 73 variables (columns) and 1,707 observations (rows) of data in the dataset.

Part 2: Describe the dataset

Next, let us type the following command:

```
describe
```

After this command is executed, you will begin to see the following output. I truncated the output in the image below because it goes on for about 75 more lines.

```
. describe

Contains data
  obs:         1,707
 vars:            73
--------------------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------------------------------
id              long    %8.2f
version         byte    %1.0f      labels0    Interview Version
county          byte    %2.0f      labels1    S2c. In which California county do you live?
q1              byte    %2.0f      labels2    Q1. First, which one issue facing California today do you think is the most impo
q2              byte    %1.0f      labels3    Q2. Overall, do you approve or disapprove of the way that Gavin Newsom is handl
q2a             byte    %1.0f      labels4    Q2a. Do you approve or disapprove of the way that Governor Newsom is handling t
q3              byte    %1.0f      labels5    Q3. Overall, do you approve or disapprove of the way that the California Legisla
q4              byte    %1.0f      labels6    Q4. Do you think that Governor Newsom and the state legislature will be able to
q5              byte    %1.0f      labels7    Q5. Do you think things in the California are generally going in the right direc
q6              byte    %1.0f      labels8    Q6. Turning to economic conditions in California, do you think that during the n
q7              byte    %1.0f      labels9    Q7. Next, some people are registered to vote and others are not. Are you absolut
```

Figure 5-2: Output from the "describe" command in Stata.

The describe command produces a table with five columns: variable name, storage type, display format, value label, and variable label.

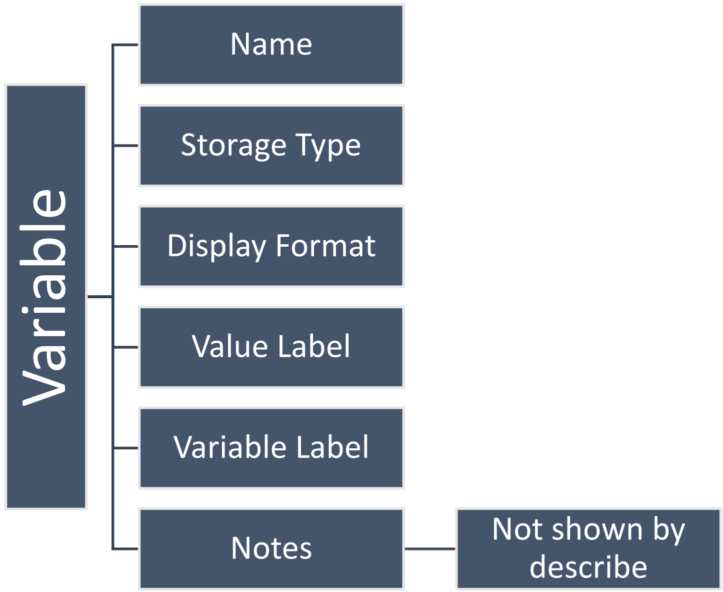


Figure 5-3: Variable properties

Variable name is the name of the variable, and it is the term you use when inputting variables in the Command field at the Center Bottom Panel or typing the variable name in .do file.

Storage type indicates how the variable is stored in the dataset. Types include byte, int, long, float, double, and str. The first five types store numeric only variables, while str stores alpha or alphanumeric.

The display format specifies how values are displayed. For example, for numeric variables, you can specify whether to include a leading minus sign and how many digits you want displayed after the decimal.

Value label is how the values of the variable are presented. For example, Question 5 asks: *Do you think things in California are generally going in the right direction or the wrong direction?* And there are four answer choices: 1 = right direction; 2 = wrong direction; 3 = (vol) don't know; and 4 = (vol) refuse.

In the Data Editor of Stata, you see the following:

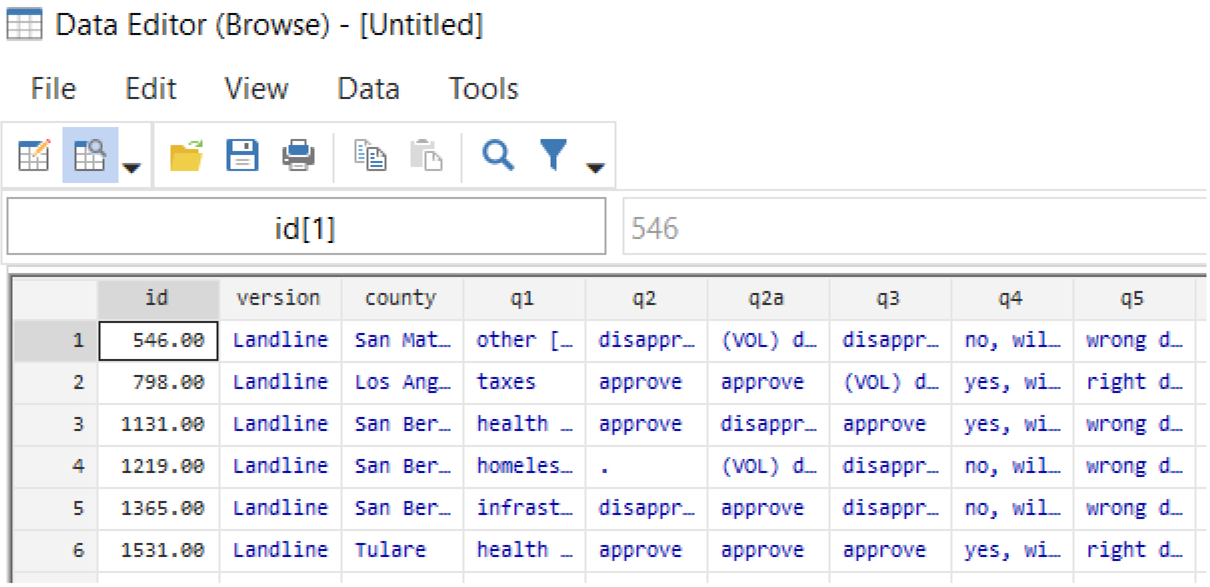


Figure 5-4: View of Data Editor in Stata

Notice that for *q5* (the last column in the figure above), we see that the survey respondent in row 1, with identification number 546.00, respond "wrong direction" when answering Question 5. We see "wrong direction" instead of the number "2" because of the value label applied to variable *q5*.

Lastly, variable label, which is not the same as value label, should be an informative, but short, description of the variable. For q5, the variable label is: *Q5. Do you think things in the California are generally going in the right direc.*

Part 3: Summarize all variables

Type the following command:

`summarize`

After this command is executed, you will begin to see the following output. Again, I truncated the output in the image below because it goes on for about 50 more lines.

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	1,707	2.15e+07	1.85e+07	546	5.00e+07
version	1,707	1.738137	.4397773	1	2
county	1,707	27.56942	13.48306	1	58
q1	1,698	26.15665	28.85894	1	98
q2	1,683	2.341652	2.331166	1	8
q2a	1,685	3.354896	2.988077	1	8
q3	1,683	2.278669	2.193698	1	8
q4	1,693	2.076787	2.111829	1	8
q5	1,695	1.739233	1.343299	1	8
q6	1,694	2.014168	1.912384	1	8
q7	1,706	1.271981	.8884402	1	8

Figure 5-5: Output from the "summarize" command in Stata.

The **`summarize`** command produces a table with 6 columns, the rows equivalent to the number of variables in the datasets. The 6 columns are: variable name, observations, mean, standard deviation, minimum, and maximum values.

Variable name and number of observations are straightforward. However, for a budding data analyst, the other four columns could use some explanation.

Let us inspect variable `q3`. According to the PPIC's Survey Codebook, `q3` stands for Question 3 of the survey. This question asks respondents (aka people taking the survey) the following question: *Overall, do you approve or disapprove of the way that the California Legislature is handling its job?*

There are four possible answer choices: approve, disapprove, and don't know. These alpha choices are coded numerically so they can be processed by statistical analysis software, like Stata. Therefore, approve = 1, disapprove = 2, and don't know = 8. The reason you see "approve" instead of the "1" in the spreadsheet is because of a value label that has been applied to Question 3 variable.

Mean is the average value of the variable across all observations. For q3, the mean is 2.278669. If the mean was 1, that means all 1,683 respondents approved of the California Legislature. And if the mean was 2, that means all 1,683 respondents disapproved of the California Legislature. And if the mean was 8, that means all 1,683 respondents don't know of the California Legislature handling its job. And so, you can conclude what the mean would be if all respondents refused to answer Question 3. Unfortunately, with the numeric coding of don't know and refuse as 8 and 9, respectively, these high numbers can skew the mean in one direction or another.

According to Standard deviation - Wikipedia, "In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range."

For q3, the standard deviation is 2.19. Relying on the 68–95–99.7 rule, this standard deviation suggests that 68% of observations have q3 values ranging between 0.08 and 4.46. However, with the numeric coding of "don't know" as 8, these high numbers can skew the standard deviation as well.

Minimum and maximum refer to the minimum value of the variable and the maximum value of the variable, respectively. Thus, for q3, the lowest numeric value is 1 (aka approve) and the highest numeric value is 8 (aka don't know).

At this point, you may be telling yourself: "The **summarize** command tells me nothing!" That is a strong statement, but I understand the sentiment you are expressing. However, the output of the `summarize` command is informative because it is suggesting that you need to try another command to get more useful output.

Part 4: Tabulate variables one-by-one

Given that we are disappointed with the output from the **summarize** command, we need to use another command to help us explore our data. There is where **tab1** can help since it produces a one-way table of frequencies.

Type the following command:

```
tab1 q3
```

After this command is executed, you will begin to see the following output.

```
. tab1 q3

-> tabulation of q3

 Q3. Overall, do |
  you approve or |
  disapprove of  |
the way that the |
      California |
         Legisla |      Freq.     Percent        Cum.
-----------------+-----------------------------------
         approve |        773       45.93       45.93
      disapprove |        703       41.77       87.70
(VOL) don't know |        207       12.30      100.00
-----------------+-----------------------------------
           Total |      1,683      100.00
```

Figure 5-6: Output from the "tab1" command in Stata

Recall Question 3 of the survey. This question asks respondents (aka people taking the survey) the following question: *Overall, do you approve or disapprove of the way that the California Legislature is handling its job?*

tab1 command produces a table with 4 columns: variable label, frequency, precent, and cumulative for q3 and its values. Column 1 restates the survey question, although it cuts it off after a certain number of characters, and the value labels of the responses provided. The three value labels are *approve*, *disapprove*, and *(VOL) don't know*. Interestingly, no one refused to answer this survey question, that is why we don't see a row for *"(VOL) refuse)"*. Side note: VOL means volunteered, since *don't know* and *refuse* are not formal answer choices and must be volunteered by the survey respondent with answering the question.

Column 2, labeled "Freq." stands for frequency, or the number of times that value appears in the dataset. In this case, we see that 773 respondents approve of the California Legislature, 703 disapprove, and 207 don't know.

Column 3, labeled "Percent" means percentage, or the percent of respondents who answered a particular value. We see the 45.93% of respondents approve, 41.77% disapprove, and 12.30% don't know how the California Legislature is handling its job.

Finally, Column 4, labeled "Cum." stands for cumulative, or the total cumulative percent of respondents who answered the survey question.

As you see, the `tab1` is more informative than the summarize command when it comes to this dataset and its variables.

Mini-Assignment #1: Instructions

Step 1: Select two parts of the four-part descriptive statistics process.

Step 2: Explain in 2 or more sentences how the two parts are related to each other.

Step 3: Explain in 2 or more sentences what is unclear to you about one of the two parts you selected.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Parts Selected: #	2 1 0	50 25 0
How Related: # sentences	2 1 0	50 25 0
What Unclear: # sentences	2 1 0	50 25 0

Chapter 6 - Model Selection

About

This chapter on model selection initiates our exploration of statistical and data analysis models that professors, researchers, scientists, data analysts, and students use to empirically examine the relationship between at least two variables of interest: an explanatory (aka independent) variable and an outcome (aka dependent) variable.

Estimated Time

An estimated 90-120 minutes is needed to complete this activity.

The Research Process

The research process includes question, theory, hypothesis, research design, empirical analysis, and results. These concepts, and more, are covered in Introduction to Political Science Research Methods – An Open Education Resource Textbook (ipsrm.com), which this Workbook is a companion to.

As a friendly refresher, below is a complex radial diagram that shows the six main branches of the research process: question, theory, hypothesis, research design, empirical analysis, and result. Each branch has at least one informative stem.

Figure 6-1: Complex radial diagram of the research process

As you will notice, Model Selection is a stem of the Empirical Analysis branch. And it should be clearly stated that empirical analysis does not appear out of thin air. It is a product of the question, theory, hypothesis, and research design.

What is Model Selection?

There is a plethora of models to select from: linear, binary, ordinal, categorical, count, fractional, generalized linear, choice, time series, panel, survival, endogenous covariates, structural equation modeling, and so on. The question inevitably is "Which model do I select?"

Model selection is your choice of statistical or data analysis model based on the nature of your dependent variable, and to some extent the nature of your independent variable(s), and the type of dataset (cross-sectional, time series, or panel).

Nature of Dependent Variable

Dependent variables can either be discrete or continuous. Discrete variables are typically non-negative integers, while continuous variables can be range from negative infinity to positive infinity. If your dependent variable is continuous, saying amount lost or gained in the stock market during a 12-month period, you will likely select a linear model and use the **`regress`** command in Stata.

However, if your dependent variable is discrete, then we need to consider what type of discrete variable is it. There are at least four types of discrete variables: binary, ordinal, categorical, and count.

Binary variables have only two values (say 0 for off and 1 for on). In Stata, we would use the **`logit`** or **`probit`** command.

Ordinal variables have values that can be ordered from least to most (say 0 for indifferent, 1 for some feelings, and 2 for strong feelings). And Stata has **`ologit`** and **`oprobit`** commands, among other ordinal models.

Categorical variables have values that have no logical ordering (say 1 for walking, 2 for biking, 3 for swimming, and 4 for running); for fitting models with these types of dependent variables you can use **`mlogit`**, or another one of Stata's commands for categorical outcomes.

Finally, count variables have at least 3 values (say 1 child, 2 children, 3 children, and so on) and Stata has **`poisson`** and **`nbreg`** commands for fitting models with these types of dependent variables. Below is a horizontal multi-level hierarchy diagram visualizing dependent variable, its nature, models, and corresponding Stata command(s).

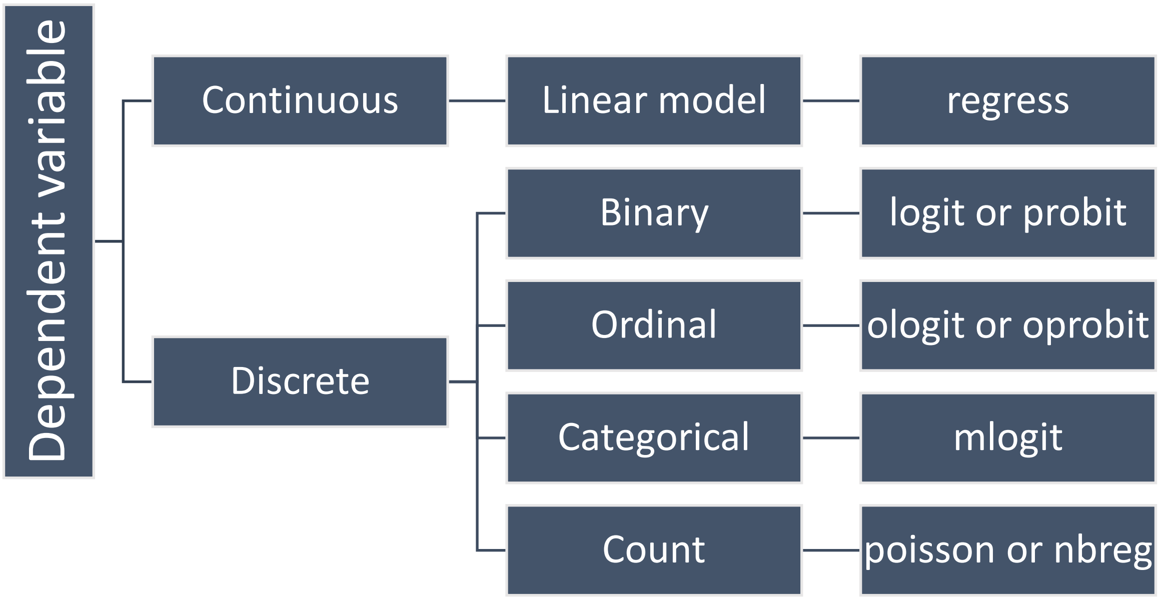


Figure 6-2: Horizontal multi-level hierarchy diagram visualizing dependent variable, its nature, the model, and the Stata command.

Type of Dataset

Recall from Chapter 3 on Datasets, there are three types: cross-sectional, time series, and panel. Along with the nature of the dependent variable, the type of dataset will inform you model selection as well.

Table 6-1: Stata Command by Dataset Type and Dependent Variable Type

	Linear	Binary	Ordinal	Categorical	Count
Cross-sectional	`regress`	`logit` `probit`	`ologit` `oprobit`	`mlogit`	`poisson` `nbreg`
Time series	*	*	*	*	*
Panel	`xtreg`	`xtlogit`	`xtologit`	`cmxtmixlogit`	`xtpoisson` `xtnbreg`

* I do not have the time series row populated since I have not yet meaningfully used time series models.

Mini-Assignment #1: Instructions

Step 1: Select one of the 1 of the 10 model options.

The 10 options are:

- **`regress`**
- **`logit/probit`**
- **`ologit/oprobit`**
- **`mlogit`**
- **`poisson/nbreg`**

- **`xtreg`**
- **`xtlogit`**
- **`xtologit`**
- **`cmxtmixlogit`**
- **`xtpoisson`**
- **`xtnbreg`**

Step 2: Explain in 2 or more sentences why you selected the model you chose.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Model selected	Yes Missing	50 0
Model selected: Why: # sentences	2 1 0	100 50 0

Chapter 7 - Linear Models

About

Linear models are used when your dependent (aka outcome) variable has values that range from negative infinity to positive infinity. Below is a list of real-world examples of continuous dependent variables:

- Amount of money lost or gained on the stock market.
- Number of jobs lost or gained in an economy.
- Amount of territory lost or gained in a conflict.

It is common in political science to use linear models for dependent variables that range from zero to a large positive number as well.

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a linear model in Stata using political science data?

For this walkthrough, we will use the Cooperative Congressional Election Study 2019 data and guide. According to the CCES website:

- "The CCES is a 50,000+ person national stratified sample survey administered by YouGov. Half of the questionnaire consists of Common Content asked of all 50,000+ people, and half of the questionnaire consists of Team Content designed by each individual participating team and asked of a subset of 1,000 people. In addition, several teams may pool their resources to create Group Content."
- "The survey consists of two waves in election years. In the pre-election wave, respondents answer two-thirds of the questionnaire. This segment of the survey asks about general political attitudes, various demographic factors, assessment of roll call voting choices, political information, and vote intentions. The pre-election wave is in the field from late September to late October. In the post-election wave, respondents answer the other third of the questionnaire, mostly consisting of items related to the election that just occurred. The post-election wave is administered in November."
- "In non-election years, the survey consists of a single wave conducted in the fall."

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/CCES19_Common_OUTPUT.dta"
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command.

Step 5. Type the following command:

```
summarize
```

Step 6. Review the output of the summarize command.

Step 7. In your review of the output, pay attention to the "Max" column to find a variable that has a large number. There are variables that will not work and variables that will work to serve as our dependent variable.

Step 8. Variables that will not work include:

a. `caseid` = unique number given to the survey respondent (the person taking the survey)

b. `birthyr` = a survey respondent's birth year is used to determine their age, but it is not an interesting dependent variable

c. `inputzip` = a survey respondent's zip code

d. There are others, but I just want to point out that the "Max" number is useful, but then you need to think of the variable itself.

Step 9. Variables that will work include:

a. `faminc_new` = family income ranges from 1 (less than $10,000) to 16 ($500,000 or more)

   i. Well, the max value is actually 97, but we see that this value represents a declined response.

b. `child18num` = number of children 18 years or younger the respondent has. The range of values for this variable is 1 to 20.

Step 10. Let us assume our research question is: What is the relationship between party identification and family income level?

Step 11. Type the following command:

```
tab1 faminc_new pid7
```

Step 12. Review the output.

Step 13. Type the following command:

```
twoway (lfit faminc_new pid7 if faminc_new<97 & pid7<8)
```

Step 14. Review the graph.

Step 15. Type the following command:

```
regress faminc_new pid7 if faminc_new<97 & pid7<8
```

Step 16. Let us review the output together:

```
. regress faminc_new pid7 if faminc_new<97 & pid7<8

      Source |       SS           df       MS      Number of obs   =    15,449
-------------+----------------------------------   F(1, 15447)     =     31.79
       Model |  362.491975         1  362.491975   Prob > F        =    0.0000
    Residual |  176150.715    15,447  11.4035551   R-squared       =    0.0021
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  176513.207    15,448  11.4262822   Root MSE        =    3.3769

------------------------------------------------------------------------------
  faminc_new |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pid7 |   .0682119   .0120985     5.64   0.000     .0444975    .0919264
       _cons |   5.942893   .0523558   113.51   0.000      5.84027    6.045517
------------------------------------------------------------------------------
```

Figure 7-1: Result of the regress command

A properly executed **regress** command, which follow the convention `regress` *`depvar indepvars`*, (where `depvar` = dependent variable name and `indepvars` = independent variable names) will produce the output seen above. There is a lot in this output table, but I want to focus your attention on the following elements on the lower table:

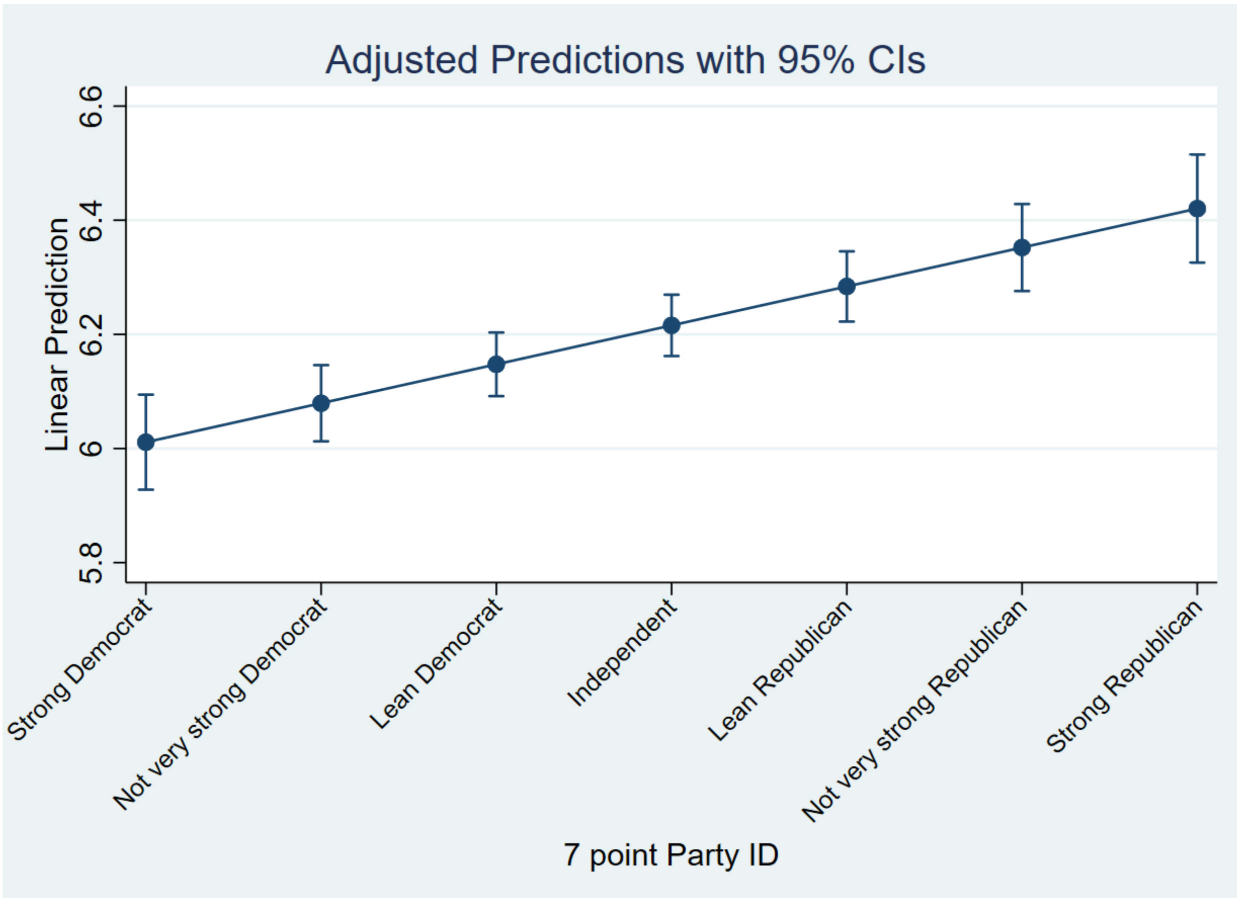
- `Number of obs` = number of observations include in the subset of the dataset analyzed. I say subset because the "**if faminc_new<97 & pid7<8**" excluded some survey respondents given how they answered those variable questions.
- `R-squared` = one measure of the statistical relationship between the dependent variable and independent variable(s). A higher R-square indicates a stronger relationship, while a lower R-square indicates a weaker relationship.
- `Coef.` Column = Coefficient estimate for the independent variable(s). `pid7` variable's coefficient of 0.0682119 indicates, that when there is 1-unit increase in `pid7`, there is a 0.0682119 increase in `faminc_new`. In other words, as we move from Strong Democrat to Strong Republican, there seems to be a positive increase family income level.
- `Std. Err.` Column = standard error estimate for the independent variable(s). Recall the [68–95–99.7 rule](), which means that 68% of our data falls within a -0.0120985 and +0.0120985 of the coefficient estimate 0.0682119; 95% of our data falls within a -0.024197 and +0.024197 of the coefficient estimate 0.0682119; and so on.
- `t` column = t-value. A larger t-value generally means the independent variable matters more, while a lower t-value generally means the independent variable matters less.
- `P>|t|` column = P-values compared to absolute value of t-value. For our introductory purposes, if a p-value is less than or equal to 0.10, and preferably 0.05, then the variable is "statistically significant". If the P-value is greater than 0.10, then the variable is "not statistically significant".

- `[95% Conf. Interval]` column = According to [Regression Analysis | Stata Annotated Output (ucla.edu)](), "This shows a 95% confidence interval for the coefficient. This is very useful as it helps you understand how high and how low the actual population value of the parameter might be."

Step 17. Type the following command:

```
margins, at(pid7=(1 2 3 4 5 6 7)) plot(xlabel(, labsize(small) angle(45)))
```

Step 18. Let us review the graph together:



- On the y-axis (vertical axis) of the graph, we see Linear Prediction range from 5.8 to 6.6.
- On the x-axis (horizontal axis) of the graph, from left to right, we observe the following values of Party Identification: Strong Democrat, Not very strong Democrat, Lean Democrat, Independent, Lean Republican, Not very strong Republican, and Strong Republic.
- You may be asking: "What is the `margins`" command? According to the Stata Help, "Margins are statistics calculated from predictions of a previously fit model at fixed values of some covariates and averaging or otherwise integrating over the remaining covariates." In other words, we can use -margins- after fitting our model to explore the results. For example, since we fit a linear regression model above, -margins- computed fitted values. So what we see are the average predicted values of faminc_new for each of the values we specified for pid7.

- One way to interpret the graph is to say as respondents move from Strong Democrat to Strong Republican, the average predicted family income level increases from 6.0 to 6.4.

Step 19. Visit Regression Analysis | Stata Annotated Output (ucla.edu) to learn more about `regress`.

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 19-step process above that you find most interesting.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 8 - Binary Outcome Models

About

Binary outcome models are used when your dependent (aka outcome) variable has two, and only two, values. Below is a list of real-world examples of binary dependent variables:

- Did a person vote or not?
- Will a person run for elected office or not?
- Does a person support a policy position or not?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a binary outcome model in Stata using political science data?

For this walkthrough, we will continue to use the Cooperative Congressional Election Study 2019 data and guide. This will help us further familiarize ourselves with this political science dataset.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/CCES19_Common_OUTPUT.dta", clear
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command.

Step 5. Type the following command:

```
summarize
```

Step 6. Review the output of the summarize command.

Step 7. In your review of the output, pay attention to the "Min" and "Max" column to find a variable which only has two numbers (say 0 and 1, or 1 and 2, respectively). There are variables that will not work and variables that will work to serve as our dependent variable.

Step 8. Variables that will not work include:

a. gender = gender is unlikely to be affected by politics

b. hispanic = whether some is or identifies as Hispanic is unlikely to be affected by politics

c. cit1 = whether someone is a U.S. citizen or not is unlikely to be affected by politics

d. Most demographic variables are unlikely to be dependent variables of interest to political scientists.

Step 9. Variables that will work include:

- CC19_300_1 = Read a blog in the past 24 hours
- CC19_300_2 = Watched TV news in past 24 hours
- CC19_300_3 = Read a newspaper in print or online

Step 10. Let us assume our research question is: What is the relationship between party identification and reading blogs?

Step 11. Type the following commands:

```
gen blog = CC19_300_1
replace blog = 0 if blog==2
```

Step 12. The reason we generated (**gen**) the variable blog is because CC19_300_1 has values of 1 and 2. However, binary outcome models need the values to be 0 and 1. While we can recode the values of CC19_300_1, it is best practice to generate a new variable, to leave the original data intact.

a. For your information, Step 11 could be completed with the following command: **recode CC19_300_1 (2=0), generate(blog)**

Step 13. Type the following command:

```
tab1 blog pid7
```

Step 14. Review the output.

Step 15. Type the following command:

```
twoway (lfit blog pid7 if pid7<8)
```

Step 16. Review the graph.

Step 17. Type the following command:

```
logit blog pid7 if pid7<8
```

Step 18. Let us review the output together:

```
. logit blog pid7 if pid7<8

Iteration 0:   log likelihood = -7695.7324
Iteration 1:   log likelihood = -7687.0846
Iteration 2:   log likelihood =  -7687.077
Iteration 3:   log likelihood =  -7687.077

Logistic regression                             Number of obs     =     17,433
                                                LR chi2(1)        =      17.31
                                                Prob > chi2       =     0.0000
Log likelihood =  -7687.077                     Pseudo R2         =     0.0011

------------------------------------------------------------------------------
        blog |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pid7 |  -.0383256   .0092365    -4.15   0.000    -.0564288   -.0202225
       _cons |  -1.509651   .0391501   -38.56   0.000    -1.586384   -1.432918
------------------------------------------------------------------------------
```

Figure 8-1: Result of the logit command

A properly executed **logit** command, which follow the convention `logit` *`depvar indepvars`*, (where `depvar` = dependent variable name and only has values of 0 and 1, and `indepvars` = independent variable names) will produce the output seen above. There is a lot in this output table, but I want to focus your attention on the following elements on the lower table:

- `Number of obs` = number of observations include in the subset of the dataset analyzed. I say subset because the "**`if pid7<8`**" excluded some survey respondents given how they answered the `pid7` variable question.
- `Pseudo R2` = A higher pseudo R-square indicates a stronger relationship, while a lower R-square indicates a weaker relationship
- `Coef.` Column = According to [Logistic Regression | Stata Data Analysis Examples (ucla.edu)](), "The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the independent variable." In this case, the log odds of reading a blog in the past 24 hours decreases as `pid7` increases from Strong Democrat to Strong Republican.
- `Std. Err.` Column = standard error estimate for the independent variable(s). Recall the [68–95–99.7 rule]() , which means that 68% of our data falls within a -0.0092365 and +0.0092365 of the coefficient estimate -0.0383256; 95% of our data falls within a -0.018473 and +0.018473 of the coefficient estimate -0.0383256; and so on.
- `z` column = z-value. A larger z-value generally means the independent variable matters more, while a lower z-value generally means the independent variable matters less.
- `P>|z|` column = P-values compared to absolute value of z-value. For our introductory purposes, if a p-value is less than or equal to 0.10, and preferably 0.05, then the variable is

"statistically significant". If the P-value is greater than 0.10, then the variable is "not statistically significant".

- `[95% Conf. Interval]` column = According to [Regression Analysis | Stata Annotated Output (ucla.edu)](), "This shows a 95% confidence interval for the coefficient. This is very useful as it helps you understand how high and how low the actual population value of the parameter might be."

Step 19. Type the following command:

```
margins, at(pid7=(1 2 3 4 5 6 7)) plot(xlabel(, labsize(small) angle(ninety)))
```

Step 20. Review the graph

Step 21. Visit Logistic Regression | Stata Data Analysis Examples (ucla.edu) to learn more about **`logit`**.

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 21-step process above that you find most interesting.

- The 3-step sequence should be sufficiently different from the prior Chapter's mini-assignment.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 9 - Ordinal Outcome Models

About

Ordinal outcome models are used when your dependent (aka outcome) variable has two or more values that can be logically ordered. Below is a list of real-world examples of ordinal dependent variables:

- On a scale of 1 to 3, with 1 being low and 3 being high, how much do you support a particular candidate for public office?
- Order a set of local policy issues from least important to most important.
- On a scale from Strongly agree to Strongly disagree, what do you think of the following statement?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run an ordinal outcome model in Stata using political science data?

For this walkthrough, we will continue to use the Cooperative Congressional Election Study 2019 data and guide. This will help us further familiarize ourselves with this political science dataset.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/CCES19_Common_OUTPUT.dta", clear
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command.

Step 5. There are four “Agreement” variables that lend themselves well to being considered ordinal outcomes:

- `CC19_343a` = Agreement - White people have certain advantages
- `CC19_343b` = Agreement - Racial problems are rare, isolated
- `CC19_343c` = Agreement - Women complain about being discriminated

- `CC19_343d` = Agreement - Feminists make reasonable demands

Step 6. Let us assume our research question is: What is the relationship between party identification and agreement issues related to race and/or gender?

Step 7. Type the following command:

```
tab1 CC19_343a CC19_343b CC19_343c CC19_343d
```

Step 8. Review the output and notice the values range from:

a. 1 – Strongly agree
b. 2 – Somewhat agree
c. 3 – Neither agree nor disagree
d. 4 – Somewhat disagree
e. 5 – Strongly disagree

Step 9. Select one of the four "Agreement" variables that interests you as your dependent variable.

Step 10. Type the following command:

```
twoway (lfit CC19_343a pid7 if pid7<8)
```

Step 11. Let us review the graph:

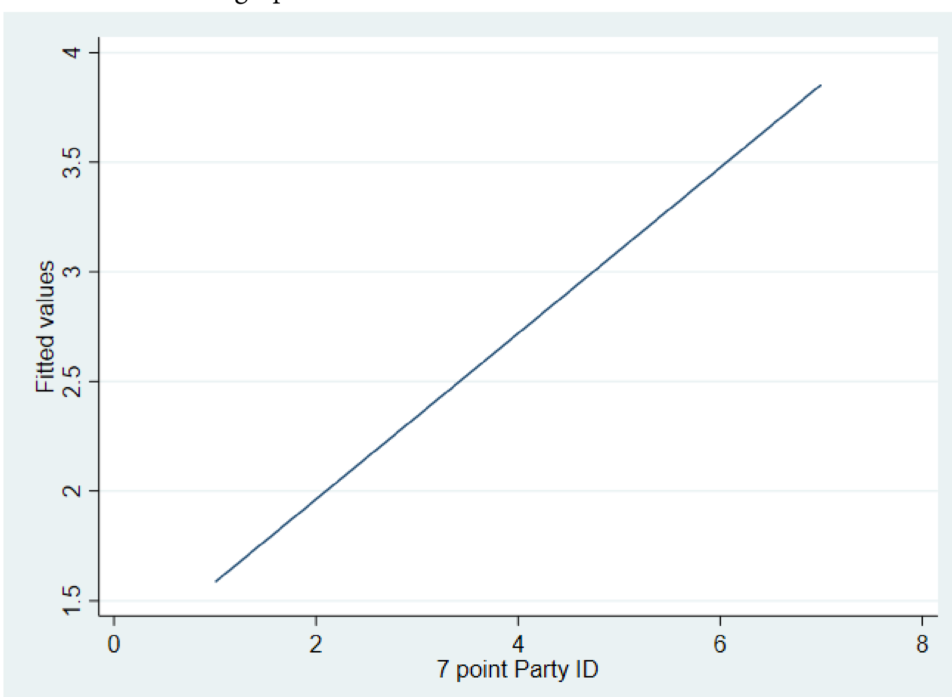


Figure 9-1: Two-way linear fit line plot

- The figure shows that has we move on the `pid7` variable from Strong Democrat to Strong Republican, that disagreement with the statement "White people have certain advantages" increases.

Step 12. Type the following command:

```
ologit CC19_343a i.pid7 if pid7<8, or
```

Step 13. Let us review the output together:

```
. ologit CC19_343a i.pid7 if pid7<8, or

Iteration 0:   log likelihood =  -26827.42
Iteration 1:   log likelihood = -23234.452
Iteration 2:   log likelihood = -23138.397
Iteration 3:   log likelihood = -23138.289
Iteration 4:   log likelihood = -23138.289

Ordered logistic regression                     Number of obs     =     17,413
                                                LR chi2(6)        =    7378.26
                                                Prob > chi2       =     0.0000
Log likelihood = -23138.289                     Pseudo R2         =     0.1375
```

CC19_343a	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
pid7						
Not very strong Democrat	2.265066	.1164948	15.90	0.000	2.047872	2.505297
Lean Democrat	1.332694	.0720384	5.31	0.000	1.198723	1.481637
Independent	7.373517	.3523783	41.81	0.000	6.714227	8.097545
Lean Republican	24.41567	1.416525	55.07	0.000	21.79136	27.35601
Not very strong Republican	13.20005	.7405373	45.99	0.000	11.82557	14.73428
Strong Republican	23.77502	1.129162	66.72	0.000	21.66179	26.09442
/cut1	.5084542	.0298132			.4500214	.566887
/cut2	1.797562	.0335951			1.731716	1.863407
/cut3	2.790824	.0370844			2.71814	2.863508
/cut4	3.479715	.0395324			3.402232	3.557197

Note: Estimates are transformed only in the first equation.

Figure 9-2: Result of the ologit command

A properly executed **ologit** command, which follows the convention `ologit` *`depvar indepvars`*, (where `depvar` = dependent variable name and has at least two ordered values and `indepvars` = independent variable names) will produce the output seen above. There is a lot in this output table, but I want to focus your attention on the following elements on the lower table:

- `Number of obs` = number of observations include in the subset of the dataset analyzed. I say subset because the "**`if pid7<8`**" excluded some survey respondents given how they answered the `pid7` variable question.
- `Pseudo R2` = A higher pseudo R-square indicates a stronger relationship, while a lower R-square indicates a weaker relationship
- `Odds Ratio` Column
    - Odds Ratio is not the same as Coefficient from the prior two chapters examples. Odds Ratio appears because we used the **`, or`** option after the **`ologit`** command sequence.
    - An odds ratio greater than 1 means the odds are higher.
    - An odds ratio lesser than 1 means the odds are lower.

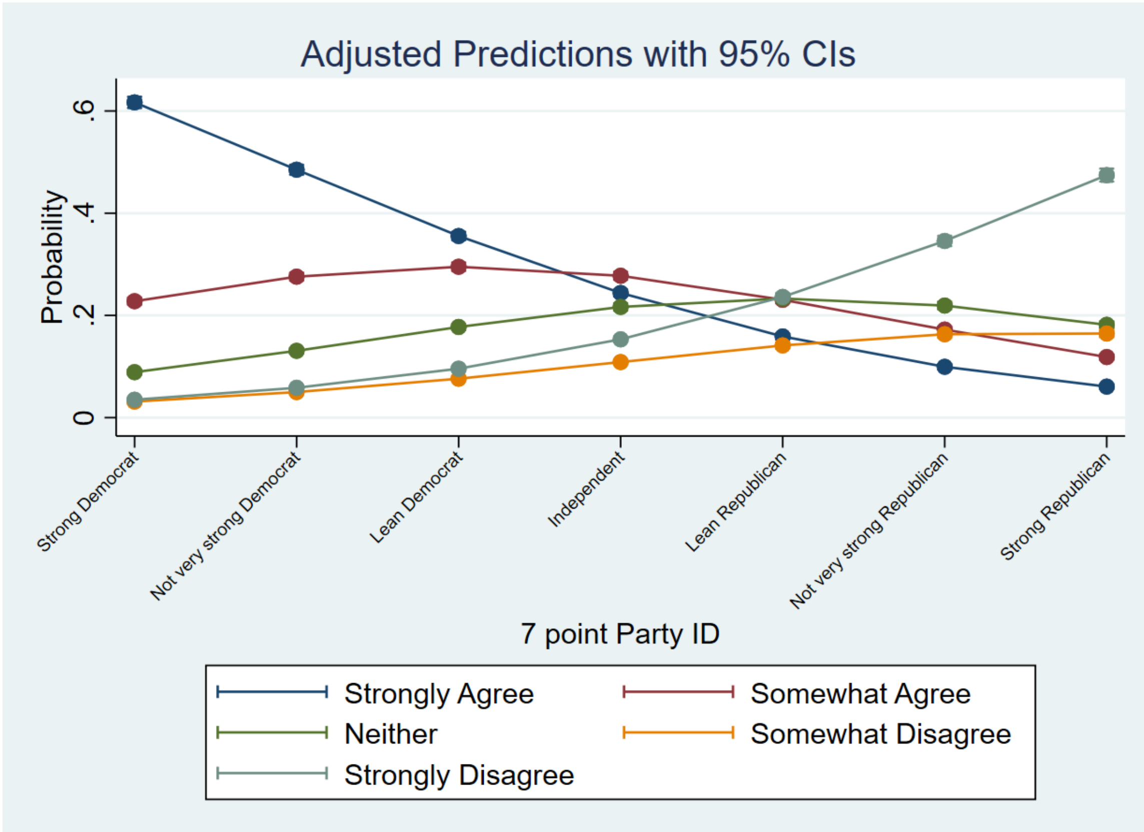
- In this example, we see the odds ratio varies by pid7 level. Therefore, someone who identifies as Lean Democrat has 1.3 greater odds of disagreeing with the statement: “White people have certain advantages”, while someone who identifies as Lean Republicans is 24.4 greater odds of disagreement with the same statement.

- `Std. Err.` Column = standard error estimate for the independent variable(s). Recall the [68–95–99.7 rule](#).
- `z` column = z-value. A larger z-value generally means the independent variable matters more, while a lower z-value generally means the independent variable matters less.
- `P>|z|` column = P-values compared to absolute value of z-value. For our introductory purposes, if a p-value is less than or equal to 0.10, and preferably 0.05, then the variable is “statistically significant”. If the P-value is greater than 0.10, then the variable is “not statistically significant”.
- `[95% Conf. Interval]` column = According to [Regression Analysis | Stata Annotated Output (ucla.edu)](#), “This shows a 95% confidence interval for the coefficient. This is very useful as it helps you understand how high and how low the actual population value of the parameter might be.”

Step 14. Type the following command:

```
margins, at(pid7=(1 2 3 4 5 6 7)) plot(xlabel(, labsize(vsmall)
angle(forty_five)) legend(order(1 "Strongly Agree" 2 "Somewhat Agree" 3
"Neither" 4 "Somewhat Disagree" 5 "Strongly Disagree")))
```

Step 15. Let us review the graph together.



- This margins graph shows us the predicted probability of Strongly Agreeing to Strongly Disagreeing with the following statement "White people have certain advantages" given political party identification.
- On the x-axis are the seven-values of pid7, ranging from "Strong Democrat" on the left to "Strong Republican" on the right.
- On the y-axis is the probability of offering a specific "Agreement" response to the statement: "White people have certain advantages."
- There are 5 "Agreement" response options: Strongly Agree, Somewhat Agree, Neither Agree or Disagree, Somewhat Disagree, and Strongly Disagree.
    - We observe that someone who identifies "Strong Democrat" has a predicted probability of 60% of ranking "Strongly Agree" with the statement: "White people have certain advantages".
    - While someone who identifies "Strong Republican" has a predicted probability near 0% of ranking "Strong Agree" with the same statement.

Step 16. Visit Ordered Logistic Regression | Stata Data Analysis Examples (ucla.edu) to learn more about **`ologit`.**

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 16-step process above that you find most interesting.

- The 3-step sequence should be sufficiently different from the prior Chapter's mini-assignment.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 10 - Categorical Outcome Models

About

Categorical outcome models are used when your dependent (aka outcome) variable has three or more values that are not naturally ordered. Below is a list of real-world examples of categorical dependent variables:

- What news sources do you read on regular basis?
- Which of the following issues are important for the government to address?
- Which of the following primary election candidates would you vote for?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a categorical outcome model in Stata using political science data?

For this walkthrough, we return to the Public Policy Institute of California's (PPIC)'s Statewide Survey Data for January 2020 Survey Data.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
import spss using "https://www.ipsrm.com/stata/2020.01.15.release.sav"
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command.

Step 5. There are two "choice" variables that lend themselves well to being considered categorical outcomes:

- `q17` = Thinking about these four areas of state spending…
- `q19` = The state is projected to have a budget surplus…

Step 6. Let us assume our research question is: What is the relationship between homeownership status and public policy choice?

Step 7. Type the following command to inspect our independent variable of homeownership status:

```
tab1 d2
```

Step 8. Let us review the output together:

```
. tab1 d2

-> tabulation of d2
```

D2. Do you own or rent your current residence?	Freq.	Percent	Cum.
Own	913	54.28	54.28
Rent	704	41.85	96.14
[VOL] Neither	65	3.86	100.00
Total	1,682	100.00	

Figure 10-1: Output from tab1 d2 command

- d2 variable asks: "Do you own or rent your current residence?" And the answer choices are "Own", "Rent", and the respondent can volunteer "Neither".
    - Own is coded as 1.
    - Rent is coded as 2.
    - [VOL] Neither is coded as 8.
- We see that 54.28% of respondents own their residence, while 41.85% of respondents rent their residence.

Step 9. Type the following command to review our two possible dependent variables:

```
tab1 q17 q19
```

Step 10. Let us review the output together:

Q17. Thinking about these four areas of state spending, I'd like you to name the	Freq.	Percent	Cum.
K-to-12 public education	606	36.29	36.29
higher education	194	11.62	47.90
health and human services	699	41.86	89.76
prisons and corrections	124	7.43	97.19
[VOL] don't know	47	2.81	100.00
Total	1,670	100.00	

-> tabulation of q19

Q19. The state is projected to have a budget surplus of several billion dollars.	Freq.	Percent	Cum.
pay down debt and build up the reserve	427	25.28	25.28
increase state funding for education, h	763	45.17	70.46
one-time state spending for transportat	435	25.75	96.21
(VOL) other	41	2.43	98.64
(VOL) don't know	23	1.36	100.00
Total	1,689	100.00	

Figure 10-2: Output from tab1 q17 q19 command

- Q17 asks: "Thinking about these four areas of state spending, I'd like you to name the one you think should have the highest priority when it comes to state government spending…" and list four choices:
  - K-12 public education
  - higher education
  - health and human services
  - prisons and corrections
- Q19 asks: "The state is projected to have a budget surplus of several billion dollars. In general, how would you prefer to use this extra money?" and lists three choices:
  - pay down the debt and build up the reserve.
  - increase state funding for education, and health and human services.
  - one-time state spending for transportation, water, infrastructure

Step 11. Select one of the two dependent variables from above that interest you.

Step 12. Type the following command:

```
twoway (lfit q17 d2 if d2<3)
```

Step 13. Review the graph.

Step 14. Type the following command:

```
mlogit q17 i.d2 if q17<8 & d2<3
```

Step 15. Let us review the output together:

```
. mlogit q17 i.d2 if q17<8 & d2<3

Iteration 0:   log likelihood = -1817.1748
Iteration 1:   log likelihood = -1802.0094
Iteration 2:   log likelihood = -1801.8664
Iteration 3:   log likelihood = -1801.8662

Multinomial logistic regression                 Number of obs     =      1,537
                                                LR chi2(3)        =      30.62
                                                Prob > chi2       =     0.0000
Log likelihood = -1801.8662                     Pseudo R2         =     0.0084

------------------------------------------------------------------------------------
                       q17 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------------+--------------------------------------------------------
K_to_12_public_education   |
                        d2 |
                      Rent |  -.4543413   .1155861    -3.93   0.000    -.6808859   -.2277967
                     _cons |   .0563112   .0770106     0.73   0.465    -.0946267    .2072491
---------------------------+--------------------------------------------------------
higher_education           |
                        d2 |
                      Rent |    -.06508   .1673992    -0.39   0.697    -.3931765    .2630164
                     _cons |  -1.260414   .1174797   -10.73   0.000     -1.49067   -1.030158
---------------------------+--------------------------------------------------------
health_and_human_services  |  (base outcome)
---------------------------+--------------------------------------------------------
prisons_and_corrections    |
                        d2 |
                      Rent |  -.9491002   .2169641    -4.37   0.000    -1.374342   -.5238584
                     _cons |  -1.338666   .1211474   -11.05   0.000    -1.576111   -1.101222
------------------------------------------------------------------------------------
```

Figure 10-3: Result of the mlogit command

A properly executed **`mlogit`** command, which follows the convention `mlogit` *`depvar indepvars`*, (where `depvar` = dependent variable name and has at least three unordered values and `indepvars` = independent variable names) will produce the output seen above. There is a lot in this output table, but I want to focus your attention on the following elements on the lower table:

- `Number of obs` = number of observations include in the subset of the dataset analyzed. I say subset because the "**`if q17<8 & d2<3`**" excluded some survey respondents given how they answered the `d2` variable question.
- `Pseudo R2` = A higher pseudo R-square indicates a stronger relationship, while a lower R-square indicates a weaker relationship
- `Coef.` Column

    - The numbers in this column are relative log-odds, so their interpretation is complicated.
    - The numbers are produced relative to the `(base outcome)` which is **health_and_human_services**
    - For example, in the **K_to_12_public_education** row, we see Coef. -0.4543413 for `Rent`. This can be roughly interpreted as the following: Being a renter is associated with a .454 decrease in the relative log odds of choosing **K_to_12_public_education** vs. **health_and_human_services**. In other words, it appears that a renter would prioritize health and human services over K-12 public education.
- `Std. Err.` Column = standard error estimate for the independent variable(s). Recall the [68–95–99.7 rule](https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule).
- `z` column = z-value. A larger z-value generally means the independent variable matters more, while a lower z-value generally means the independent variable matters less.
- `P>|z|` column = P-values compared to absolute value of z-value. For our introductory purposes, if a p-value is less than or equal to 0.10, and preferably 0.05, then the variable is "statistically significant". If the P-value is greater than 0.10, then the variable is "not statistically significant".
- `[95% Conf. Interval]` column = According to [Regression Analysis | Stata Annotated Output (ucla.edu)](https://stats.oarc.ucla.edu/stata/output/regression-analysis/), "This shows a 95% confidence interval for the coefficient. This is very useful as it helps you understand how high and how low the actual population value of the parameter might be."

Step 16. Type the following command:

```
margins, at(d2=(1 2)) plot(xlabel(, labsize(vsmall) angle(forty_five))
legend(order(1 "K-to-12 public education" 2 "higher education" 3 "health and
human services" 4 "prisons and corrections")))
```

Step 17. Let us review the graph together.

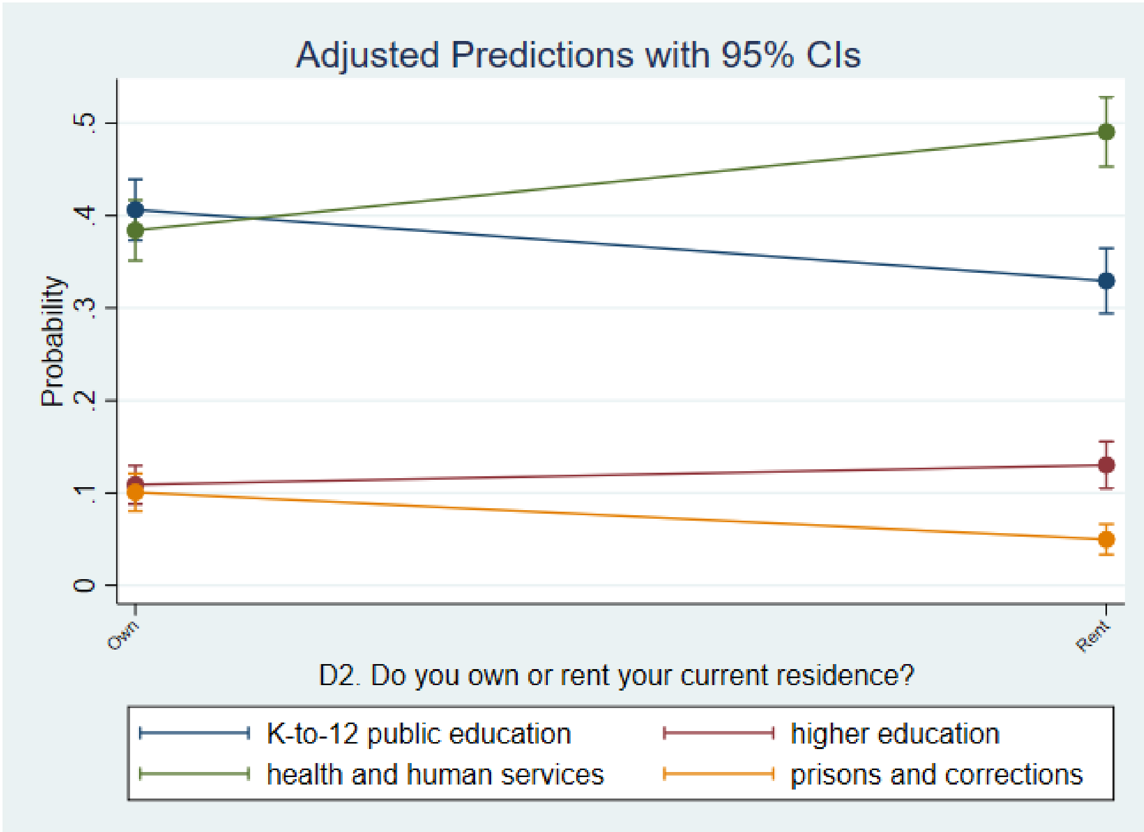


Figure 10-4: Adjusted Predictions with 95% CIs

- This margins graph shows us the predicted probability of choosing one of four policy priority choices given homeownership or renter status.
- On the x-axis are the two-values of `d2`: Homeowner or Renter.
- On the y-axis is the probability of choosing [1] K-to-12 public education, [2] higher education, [3] health and human services, or [4] prisons and corrections.
- We observe homeowners have the following:
    - ~40% predicted probability of choosing K-to-12 public education as their highest priority for state spending.
    - ~38% predicted probability of choosing health and human services as their highest priority for state spending.
    - ~10% predicted probability of choosing higher education as their highest priority for state spending.
    - ~10% predicted probability of choosing prisons and corrections as their highest priority for state spending.
- We observe renters have the following:
    - ~32% predicted probability of choosing K-to-12 public education as their highest priority for state spending.

- ~49% predicted probability of choosing health and human services as their highest priority for state spending.
- ~12% predicted probability of choosing higher education as their highest priority for state spending.
- ~6% predicted probability of choosing prisons and corrections as their highest priority for state spending.

Step 18. Visit [Multinomial Logistic Regression | Stata Annotated Output (ucla.edu)]() to learn more about `mlogit`.

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 18-step process above that you find most interesting.

- The 3-step sequence should be sufficiently different from the prior Chapter's mini-assignment.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 11 - Count Outcome Models

About

Count models are used when your dependent (aka outcome) variable represents a count of some object or actions and ranges from 0 to positive infinity. Below is a list of real-world examples of count dependent variables:

- How many courthouses did Congress authorize for a specific state?
- How many hearings did a state legislative committee hold in a specific legislative session?
- How many times did a U.S. citizen donate to political candidates in a campaign election cycle?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a count outcome model in Stata using political science data?

For this walkthrough, we will use a dataset that I am thoroughly familiar with because I collected the data for my dissertation: Judicial Pork: The Congressional Allocation of Districts, Seats, Meeting Places, and Courthouses to the U.S. District Courts.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/Franco_Judicial_Pork_July_3_2018.dta”
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command.

Step 5. There are four variables that lend themselves well to being considered count outcomes:

- `DpV_JDt` = Count of Judicial Districts
- `DpV_JSt` = Count of Judicial Seats
- `DpV_JMP` = Count of Judicial Meeting Places
- `DpV_JCt` = Count of Judicial Courthouses

Step 6. Let us assume our research question is: What is the relationship between committee membership and securing judicial seats?

Step 7. Type the following command to inspect our independent variable of homeownership status:

`tab1 IdV_HR_JMbr if year==1961`

Step 8. Let us review the output together:

```
. tab IdV_HR_JMbr if year==1961
```

House Judiciary Member	Freq.	Percent	Cum.
0	26	52.00	52.00
1	16	32.00	84.00
2	7	14.00	98.00
6	1	2.00	100.00
Total	50	100.00	

Figure 11-1: Output from tab1 IdV_HR_JMbr if year==1961 command

- The independent variable IdV_HR_JMbr tells us how many representatives a state had serving on the House Judiciary Committee.
- This a panel dataset. To work with a cross section of this data, we can use the **`if`** qualifier. Here, I restricted the sample to the observations from 1961.
- We see that 26 states had 0 representation on the committee, 16 states had 1 member, 7 states had 2 members, and 1 state has 6 members.

Step 9. Type the following command to review our possible dependent variable:

`tab1 DpV_JSt if year==1961`

Step 10. Let us review the output together:

```
. tab1 DpV_JSt if year==1961

-> tabulation of DpV_JSt if year==1961

   Judicial |
      Seats |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         20       40.00       40.00
          1 |         15       30.00       70.00
          2 |          7       14.00       84.00
          3 |          3        6.00       90.00
          4 |          3        6.00       96.00
          6 |          1        2.00       98.00
          8 |          1        2.00      100.00
------------+-----------------------------------
      Total |         50      100.00
```

Figure 11-2: Output from tab1 DpV_JSt command

- The dependent variable DpV_JSt tells us how many judicial seats a state was given during the year 1961.
- We see that 20 states were allocated no seats, 15 states received 1 seat, 7 states obtained 2 seats and so forth.

Step 11. Type the following command:

twoway (lfit DpV_JSt IdV_HR_JMbr if year==1961)

Step 12. Review the graph.

Step 13. Type the following command:

poisson DpV_JSt IdV_HR_JMbr if year==1961

Step 14. Let us review the output together:

```
. poisson DpV_JSt IdV_HR_JMbr if year==1961

Iteration 0:   log likelihood = -76.682465
Iteration 1:   log likelihood = -72.752568
Iteration 2:   log likelihood = -72.727308
Iteration 3:   log likelihood = -72.727285
Iteration 4:   log likelihood = -72.727285

Poisson regression                              Number of obs     =         50
                                                LR chi2(1)        =      24.84
                                                Prob > chi2       =     0.0000
Log likelihood = -72.727285                     Pseudo R2         =     0.1459

------------------------------------------------------------------------------
     DpV_JSt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 IdV_HR_JMbr |   .3899785   .0650178     6.00   0.000      .262546    .5174111
       _cons |  -.1623412   .1604376    -1.01   0.312     -.476793    .1521106
------------------------------------------------------------------------------
```

Figure 11-3: Result of the poisson command

A properly executed **poisson** command, which follows the convention `poisson depvar indepvars`, (where `depvar` = dependent variable name and has at least three unordered values and `indepvars` = independent variable names) will produce the output seen above. There is a lot in this output table, but I want to focus your attention on the following elements on the lower table:

- `Number of obs` = 50 since there are 50 states in the year 1961
- `Pseudo R2` = 0.1459 demonstrates there is some relationship between these two variables
- `Coef.` Column
    - The numbers in this column are relative log-odds, so their interpretation is complicated.
- `Std. Err.` Column = standard error estimate for the independent variable(s). Recall the [68–95–99.7 rule](#).
- `z` column = z-value. A larger z-value generally means the independent variable matters more, while a lower z-value generally means the independent variable matters less.
    - 6.00 is a large number.
- `P>|z|` column = P-values compared to absolute value of z-value. For our introductory purposes, if a p-value is less than or equal to 0.10, and preferably 0.05, then the variable is "statistically significant". If the P-value is greater than 0.10, then the variable is "not statistically significant".
    - In this case, since 0.000 demonstrates a statistically significant relationship.
- `[95% Conf. Interval]` column = According to [Regression Analysis | Stata Annotated Output (ucla.edu)](#), "This shows a 95% confidence interval for the coefficient. This is very useful

as it helps you understand how high and how low the actual population value of the parameter might be."

Step 15. Type the following command:

```
margins, at(IdV_HR_JMbr=(0 1 2 3 4 5 6)) plot(xlabel(, labsize(vsmall)
angle(forty_five)))
```

Step 16. Let us review the graph together.

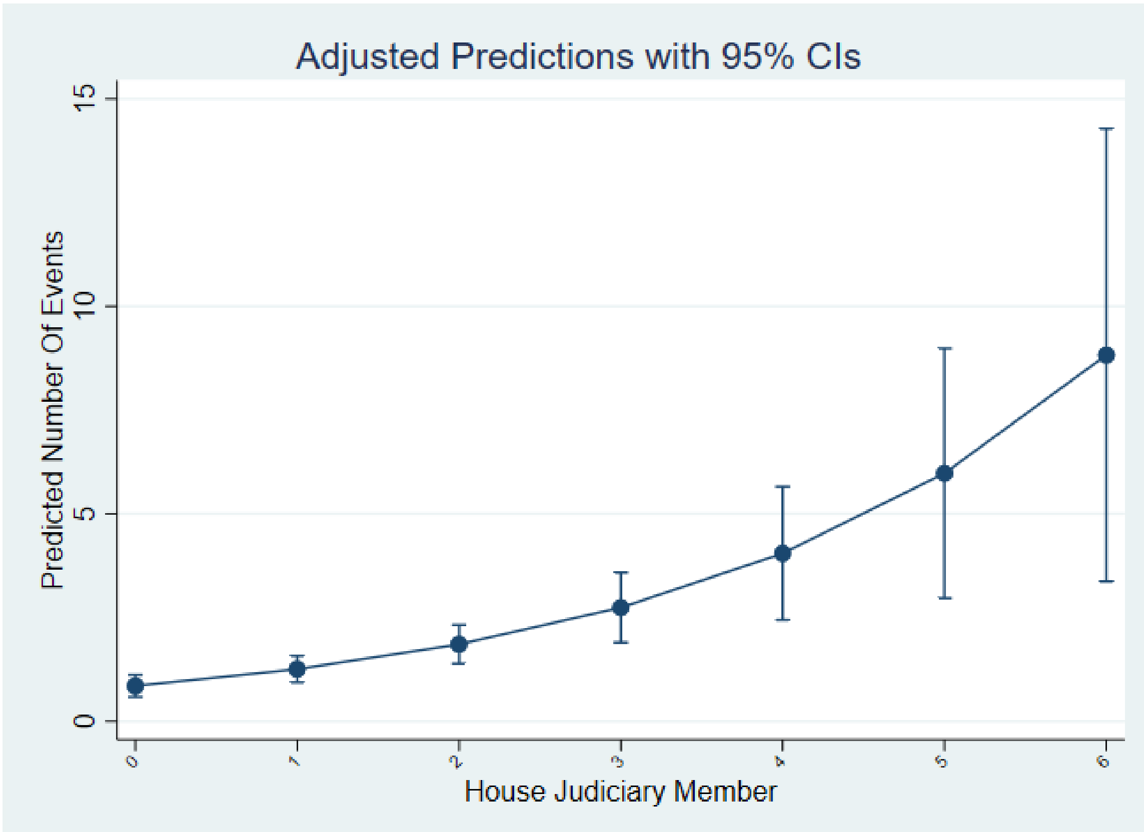


Figure 11-4: Adjusted Predictions with 95% CIs

- This margins graph shows us the predicted number of judicial seats allocated to a state given the number representatives a state has serving on the House Judiciary Committee.

Step 17. Visit [Poisson Regression | Stata Data Analysis Examples (ucla.edu)] to learn more about **poisson**.

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 17-step process above that you find most interesting.

- The 3-step sequence should be sufficiently different from the prior Chapter's mini-assignment.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 12 - Panel Data Linear Models

About

Panel data linear models are used when your dependent (aka outcome) variable has values that range from negative infinity to positive infinity, *and* you are using a panel dataset. Recall a list of real-world examples of continuous dependent variables:

- Amount of money lost or gained on the stock market.
- Number of jobs lost or gained in an economy.
- Amount of territory lost or gained in a conflict.

Panel Data: Multiple Objects across Multiple Time Periods

A panel dataset is when you are observing multiple objects across multiple time periods. Consider the following example. Say we are interested in the social, economic, and political demographics of counties in a single state over a 30-year period. Specifically, we observe the State of California and its 58 counties from 1990 to 2020. If we were just observing these 58 counties in 1990, we would have a cross-sectional dataset of 58 counties in year 1990. However, the moment we observe these same counties in 1991, we have just created a panel dataset.

$$\begin{array}{ccccc} C_1 & O_{1990} & O_{1991} & \ldots & O_{2020} \\ C_2 & O_{1990} & O_{1991} & \ldots & O_{2020} \\ & & \vdots & & \\ C_{58} & O_{1990} & O_{1991} & \ldots & O_{2020} \end{array}$$

Time-variant and time-invariant

The example above will help illustrate some key concepts of panel data. First, is that within Counties, there are variables that will and will not change over time. For example, variables that will change over time are the population, number of small businesses, registered Democratic versus Republican, and so on. However, there are variables that will not change over time: geographic square miles of the county, the type of county government, and the number of seats on the county Board of Supervisors.

Second, is that within time periods, there are variables that will and will not change. For example, in the year 1990, all counties may have experienced a reduction in state funding due to an economic recession. Therefore, a variable labeled “reduction in state dollars” may be labeled “Yes” for all 58 counties in year

1990. However, a related variable, say "amount of reduced state dollars" can be different between counties: County 1 experienced $1,000,000 reduction, while County 58 had a $5,000,000 reduction.

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a panel data linear model in Stata using political science data?

For this walkthrough, we will be reproducing results from an Appendix and Figure 5 of The Countervailing Effects of Competition on Public Goods Provision: When Bargaining Inefficiencies Lead to Bad Outcomes by Dr. Jessica Gottlieb and Dr. Katrina Kosec. The Abstract, or summary of the article, follows:

> *"Political competition is widely recognized as a mediator of public goods provision through its salutary effect on incumbents' electoral incentives. We argue that political competition additionally mediates public goods provision by reducing the efficiency of legislative bargaining. These countervailing forces may produce a net negative effect in places with weak parties and low transparency—typical of many young democracies. We provide evidence of a robust negative relationship between political competition and local public goods using panel data from Mali. Tests of mechanisms corroborate our interpretation of this relationship as evidence of legislative bargaining inefficiencies. To explore the generalizability of these findings, we analyze cross-country panel data and show that political competition leads to better (worse) public goods provision under high (low) levels of party system institutionalization. The paper sheds light on why political competition is only selectively beneficial and underscores the importance of considering both the electoral and legislative arenas."*

In the Supplementary Materials of the article, the authors provided several statistical analyses that were not included in the final article itself. Supplementary Materials are useful for the deeply interested reader. Additionally, the Replication Data for: The Countervailing Effects of Competition on Public Goods Provision: When Bargaining Inefficiencies Lead to Bad Outcomes is available on Dataverse website.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/gottlieb2019.dta”
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command. With 259 variables in this dataset, you may feel overwhelmed with the number of variables and their labels. However, just keep calm and continue.

Step 5. Type the following command:

```
summarize
```

Step 6. Review the output of the summarize command.

Step 7. How do the authors declare their independent variable and dependent variable? On page 104, Dr. Gottlieb and Dr. Kosec describe these variables:

> *"To test these predictions outside the Mali case, we constructed a panel dataset of 164 countries spanning the period 1975–2015. Since we are interested in legislative competition, we use a Herfindahl index (HHI) for legislative elections coded from the DPI dataset (Keefer 2005) as our independent variable. For our dependent variable, we use data on both public expenditures from the Statistics on Public Expenditures for Economic Development (SPEED) database (IFPRI 2017) and on development outcomes related to public expenditures from the World Development Indicators (WDI) database (World Bank 2017) as proxies for public goods provision."*

Recall that earlier in the article, on page 93, they explain what the HHI is:

> *We measure electoral competitiveness in two ways: with a Herfindahl-Hirschman index (HHI) and the winning party's margin of victory. The first better captures the idea that increasing political competitiveness exacerbates the complexity of coalition formation. The second corresponds more closely to the idea that the relative strength of the plurality party matters for its ability to form durable coalitions. Because of the way each measure is constructed, larger values indicate less political competitiveness.*

Step 8. Type the following command:

```
xtreg gdpeducation_ppp i.partysys2#c.herf herf t_init_gdpeducation_ppp i.year
population, fe cluster(ison)
```

- Let us breakdown the prior command:
    - **xtreg** = command for fitting linear regression model with panel data
    - **gdpeducation_ppp** = percentage of education expenditure in total GDP
    - **i.partysys2#c.herf** = interaction between party system institutionalization and the Herfindahl Index
    - **herf** = Herfindahl Index
    - **t_init_gdpeducation_ppp** = Initial period value of gdpeducation_ppp interacted with a time trend
    - **i.year** = year variable
    - **population** = Population of country (100,000s)

- **, fe cluster(ison)** = an option after the xtreg command that specifies fixed effects with clustering on country

Step 9. Let us review the command output together:

```
. xtreg gdpeducation_ppp i.partysys2#c.herf herf t_init_gdpeducation_ppp i.year population, fe cluster(ison)

Fixed-effects (within) regression               Number of obs     =      2,815
Group variable: ison                            Number of groups  =        126

R-sq:                                           Obs per group:
     within  = 0.1731                                         min =          1
     between = 0.1911                                         avg =       22.3
     overall = 0.0608                                         max =         33

                                                F(36,125)         =       5.29
corr(u_i, Xb)  = -0.7316                        Prob > F          =     0.0000

                                 (Std. Err. adjusted for 126 clusters in ison)
------------------------------------------------------------------------------
                       |               Robust
      gdpeducation_ppp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------+------------------------------------------------------
      partysys2#c.herf |
                     2 |  -2.077928   .7375346    -2.82   0.006    -3.537601   -.6182558
                       |
                  herf |   1.215566   .5541937     2.19   0.030     .1187479    2.312384
t_init_gdpeducation_ppp|  -.0176638   .0038305    -4.61   0.000    -.0252448   -.0100827
```

Figure 12-1: Result of the xtreg command

- Focus on the `Coef.` Column and the values for `partysys2#c.herf` and `herf`, which are -2.077928 and 1.215566, respectively

Step 10. Now, let us compare the results from above with how it appears in the Appendix of the article.

Table A.20: Effect of HHI on Public Goods Provision by Party System Institutionalization

	(1) Education share GDP	(2) Health share GDP	(3) Education per capita	(4) Health per capita	(5) Primary completion	(6) Immunization (measles)
Panel A: By Party System Institutionalization (PSI)						
HHI	1.216**	0.947**	260.497***	224.670***	9.863***	6.109**
	(0.554)	(0.363)	(68.004)	(64.938)	(3.003)	(2.894)
HHI × high PSI	−2.078***	−1.630***	−270.844**	−305.474***	−11.443**	−4.644
	(0.738)	(0.531)	(116.932)	(105.521)	(5.554)	(4.320)
Observations	2815	2750	2815	2750	3072	4617

Figure 12-2: Screenshot of Table A.20: Effect of HHI on Public Goods Provision by Party System Institutionalization

- Looking just at column (1) Education share GDP, we see that HHI is 1.216 and HHI*high PSI is -2.078
- These numbers match those we saw in the output from the previous Stata command.

Step 11. Next, we want to reproduce the top-left panel of Figure 5 in the article itself.

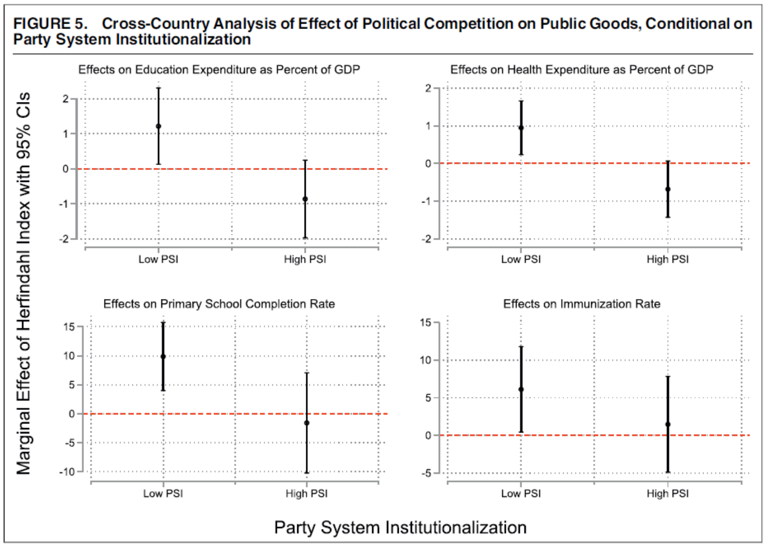


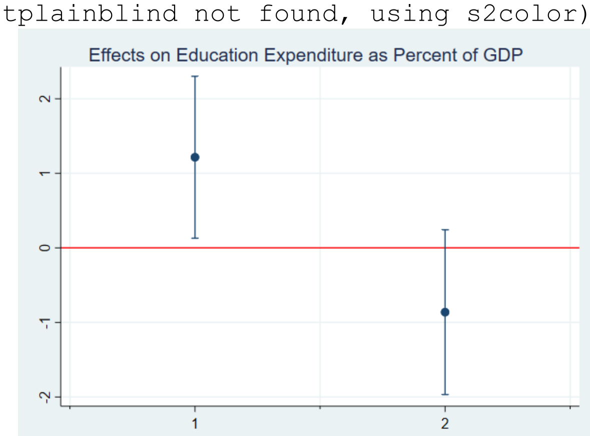
Figure 12-3: Screenshot of Figure 5: Cross-Country Analysis of Effect of Political Competition on Public Goods, Conditional on Party System Institutionalization

Step 12. Type the following command to attempt to reproduce the top-left panel of Figure 5:

```
margins, dydx(herf) by(partysys2)
marginsplot, plotopts(connect(none)) name(gdped, replace) title(Effects on
Education Expenditure as Percent of GDP) scheme(plotplainblind) yline(0,
lcolor(red)) xlabel(, valuelabel) xmtick(.5(1)2.5, grid gmin gmax)
xsc(range(.5(1)2.5)) ytitle("") xtitle("")
```

- After you run this command, the following text and graph will appear:

```
Variables that uniquely identify margins: partysys2
(note: scheme plotplainblind not found, using s2color)
```



- This text appears because the commands in Step 12 use a community-contributed command named plotplainblind. To install this community-contributed command, type the following text in the Command window:

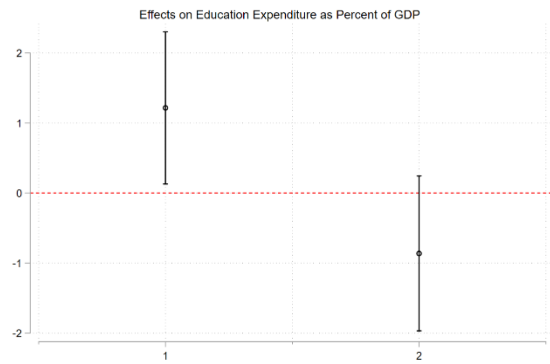
```
ssc install blindschemes, replace
net install blindschemes_fix, from(http://digital.cgdev.org/doc/stata/MO/Misc)
```

Step 13. Now, type the following command to reproduce the top-left panel of Figure 5:

```
margins, dydx(herf) by(partysys2)
marginsplot, plotopts(connect(none)) name(gdped, replace) title(Effects on
Education Expenditure as Percent of GDP) scheme(plotplainblind) yline(0,
```

```
lcolor(red)) xlabel(, valuelabel) xmtick(.5(1)2.5, grid gmin gmax)
xsc(range(.5(1)2.5)) ytitle("") xtitle("")
```

Step 14. Let us review the graph together:



- Note that the graph we created has "1" instead of Low PSI and "2" instead of "High PSI". To correct for this, we could include the following command in Step 12

```
label define partyof2 1 “Low PSI" 2 "High PSI"
label values partysys2 partyof2
```

Step 15. How do Drs. Gottlieb and Kosec interpret the results from Table A.20 and Figure 5?

> *"Our empirical specification takes advantage of overtime changes in both legislative electoral competitiveness and public goods outcomes, running a two-way (country and year) fixed effects regression.*
>
> *We interact the time varying independent variable of competitiveness with a time–invariant country-level indicator for high party system institutionalization, which allows us to estimate differential slopes for the two sets of countries. Results appear in Appendix Table A.20 and key outcomes are depicted visually in Figure 5.*
>
> *We find strong support for our theory outside the Mali case, whether looking at public expenditures (inputs) or citizens' access to services (outputs). In countries with low party system institutionalization, there is a positive and statistically significant relationship between the HHI and several outcomes—education and health expenditures as a share of GDP, primary school completion rates, and immunization rates for measles—indicating a negative relationship between political competition and public goods provision.*
>
> *By contrast, in countries with high party system institutionalization, this relationship either attenuates or reverses to obtain the negative relationship between the Herfindahl index and public goods outcomes (or the positive relationship between competition and public goods) that is predicted by much of the existing literature."*

Step 16. Visit Using xtreg (ucla.edu) and What is the difference between xtreg, re and xtreg, fe? | Stata FAQ (ucla.edu) to learn more about **xtreg.**

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 16-step process above that you find most interesting.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 13 - Panel Data Binary Outcome Models

About

Binary outcome models are used when your dependent (aka outcome) variable has only two values. Previously, we fit a model with a binary outcome with cross-sectional data. Now, we will fit a binary outcome model with panel data. Below is a list of real-world examples of binary dependent variables:

- Did a person vote or not?
- Will a person run for elected office or not?
- Does a person support a policy position or not?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a panel data binary outcome model in Stata using political science data?

For this walkthrough, we will continue to use Gottlieb 2019 dataset from the chapter on Panel Data Linear Models.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/gottlieb2019.dta”
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command. With 259 variables in this dataset, you may feel overwhelmed with the number of variables and their labels. However, just keep calm and continue.

Step 5. Type the following command:

```
summarize
```

Step 6. Review the output of the summarize command.

Step 7. Let us assume our research question is: What is the relationship between country-legislative specific variables and whether a chief executive can serve multiple terms or not?

Step 8. Our dependent variable is:

a. Can chief executive serve multiple terms? = **multpl**

Step 9. Our independent variables are:

a. Total seats in legislature = **totalseats**

b. Legislative elections held = **legelec**

c. Legislative electoral competitiveness = **liec**

Step 10. Type the following command:

```
xtlogit multpl totalseats legelec liec if multpl>=0 & legelec>=0 & liec>=0, fe
```

- Let us examine the prior command before reviewing the results:
  - The **if multpl>=0 & legelec>=0 & liec>=0** is needed because these variables have a -999 value to denote missing or incomplete information

Step 11. Let us review the command output together:

```
. xtlogit multpl totalseats legelec liec if multpl>=0 & legelec>=0 & liec>=0, fe
note: multiple positive outcomes within groups encountered.
note: 153 groups (4,743 obs) dropped because of all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -255.05206
Iteration 1:   log likelihood =  -250.7424
Iteration 2:   log likelihood = -250.68856
Iteration 3:   log likelihood = -250.68853
Iteration 4:   log likelihood = -250.68853

Conditional fixed-effects logistic regression   Number of obs      =        537
Group variable: ison                            Number of groups   =         15

                                                Obs per group:
                                                             min =         21
                                                             avg =       35.8
                                                             max =         41

                                                LR chi2(3)         =      90.48
Log likelihood  = -250.68853                    Prob > chi2        =     0.0000

------------------------------------------------------------------------------
      multpl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  totalseats |  -.0063398    .003649    -1.74   0.082    -.0134917    .0008121
     legelec |  -.0592558   .2419897    -0.24   0.807    -.5335469    .4150353
        liec |  -.9580906   .1705042    -5.62   0.000    -1.292273   -.6239085
------------------------------------------------------------------------------
```

Figure 13-1: Result of the xtlogit command

  - Notice the two "Notes":

- The first note is saying that within country, there are years when the chief executive can serve multiple terms and years when the chief executive cannot serve multiple terms. Thus, for some panels there are multiple observations for which the dependent variable is equal to one.
- The second note is saying 153 countries were dropped from the analysis because every year the chief executive cannot serve multiple terms, or every year the chief executive can serve multiple terms. In other words, there is no variation on the dependent variable of interest.

- Focus on the `Coef.` Column:
  - `totalseats` is -0.0063398 with P>|z| of 0.082.
  - `legelec` is -0.0592558 with P>|z| of 0.807.
  - `liec` is -0.9580906 with P>|z| of 0.000.

Step 12. Type the following command to produce a graph:

```
margins, at(liec=(1 2 3 4 5 6 7)) plot
```

Step 13. Let us review the graph together:

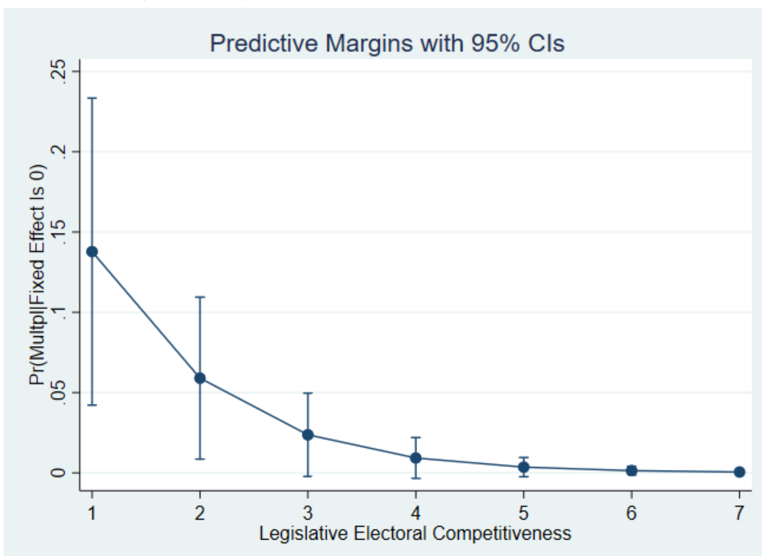


Figure 13-2: Predictive Margins with 95% CIs

- As legislative electoral competitiveness increases, the probability of a chief executive being able to serve multiple terms decreases.

Step 14. Visit [Longitudinal-Data/Panel-Data Reference Manual | Stata Press (stata-press.com)] to learn more about **`xtlogit.`**

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 14-step process above that you find most interesting.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 14 - Panel Data Ordinal Outcome Models

About

Ordinal outcome models are used when your dependent (aka outcome) variable has two or more values that can be logically ordered. Previously, we fit a model with an ordinal outcome with cross-sectional data. Now we'll fit an ordinal outcome model with panel data. Below is a list of real-world examples of ordinal dependent variables:

- On a scale of 1 to 3, with 1 being low and 3 being high, how much do you support a particular candidate for public office?
- Order a set of local policy issues from least important to most important.
- On a scale from Strongly agree to Strongly disagree, what do you think of the following statement?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a panel data ordinal outcome model in Stata using political science data?

For this walkthrough, we will continue to use Gottlieb 2019 dataset from the chapter on Panel Data Linear Models and Binary Outcome Models.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/gottlieb2019.dta”
```

Step 3. Type the following command:

```
describe
```

Step 4. Review the output of the describe command. With 259 variables in this dataset, you may feel overwhelmed with the number of variables and their labels. However, just keep calm and continue.

Step 5. Type the following command:

```
summarize
```

Step 6. Review the output of the summarize command.

Step 7. Let us assume our research question is: What is the relationship between country-legislative specific variables and the level of legislative electoral competitiveness?

Step 8. Our dependent variable is:

a. Legislative electoral competitiveness = **liec**

Step 9. Our independent variables are:

a. Total seats in legislature = **totalseats**

b. Proportional representation system = **pr**

c. Media bias = **v2mebias**

Step 10. We need to replace `liec` values that are not whole numbers (3.5, 5.5, and 6.5) because there is a reporting error when we try to **margins** plot the estimation results:

```
replace liec=4 if liec==3.5
replace liec=6 if liec==5.5
replace liec=7 if liec==6.5
```

Step 11. Type the following command:

```
xtologit liec totalseats pr v2mebias if liec>=0 & pr>=0
```

- Let us examine the prior command before reviewing the results:
  - The **if liec>=0 & pr>=0** is needed because these variables have a -999 value to denote missing or incomplete information

Step 12. Let us review the command output together:

```
Random-effects ordered logistic regression      Number of obs     =      4,401
Group variable: ison                            Number of groups  =        157

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =          3
                                                              avg =       28.0
                                                              max =         41

Integration method: mvaghermite                 Integration pts.  =         12

                                                Wald chi2(3)      =     337.70
Log likelihood  = -1767.8579                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        liec |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  totalseats |  -.0010904   .0011074    -0.98   0.325    -.0032609    .0010801
          pr |   1.499522   .2430959     6.17   0.000     1.023062    1.975981
    v2mebias |   1.531166   .0945023    16.20   0.000     1.345945    1.716387
-------------+----------------------------------------------------------------
       /cut1 |  -8.861536   .5151242                     -9.871161   -7.851911
       /cut2 |  -8.059862   .4852102                     -9.010856   -7.108867
       /cut3 |   -5.71674   .4362924                     -6.571857   -4.861622
       /cut4 |  -4.206457   .4174544                     -5.024653   -3.388261
       /cut5 |  -1.200217   .4048179                     -1.993646   -.4067888
-------------+----------------------------------------------------------------
    /sigma2_u |   11.00525   2.078085                      7.600996    15.93418
------------------------------------------------------------------------------
LR test vs. ologit model: chibar2(01) = 1404.22       Prob >= chibar2 = 0.0000
```

Figure 14-1: Result of the xtologit command

- Focus on the `Coef.` Column:
  - `totalseats` is -0.0010 with P>|z| of 0.325, which is not statistically significant.
  - `pr` is +1.4995 with P>|z| of 0.000, which suggests countries with proportional representation electoral systems are more competitive.
  - `v2mebias` is +1.5311 with P>|z| of 0.000, which suggests that as media bias shifts from left to right, legislative competitiveness increases.

Step 13. Type the following command to produce a graph:

```
margins, at(v2mebias=(-3 -2 -1 0 1 2 3)) plot
```

Step 14. Let us review the graph together:

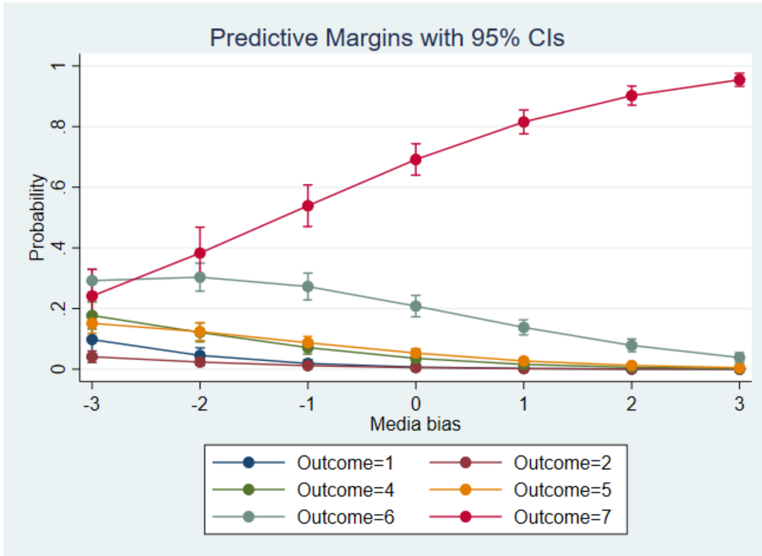


Figure 14-2: Predictive margins with 95% CI

- Given that there are 7 outcomes, or levels, of legislative electoral competitiveness, we need to observe each line:
  - Outcomes 1-6 demonstrate that as media bias shifts from left to right, the probability of legislative electoral competitiveness decreases.
    - Note that Outcome 3 is not included in the predictive margins above. To check this for yourself, you can type the following command: **`tabulate liec if pr>=0`**
  - Outcome 7 demonstrates that as media bias shifts from left to right, the probability of legislative electoral competitiveness increases.

Step 15. Visit [Longitudinal-Data/Panel-Data Reference Manual | Stata Press (stata-press.com)] to learn more about **`xtologit`.**

Step 16. You can also watch [Ordered logistic and probit for panel data in Stata® - YouTube]

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 16-step process above that you find most interesting.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points

Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

1988

Chapter 15 - Panel Data Categorical Outcome Models

About

Categorical outcome models are used when your dependent (aka outcome) variable has three or more values that are not naturally ordered. In Chapter 10, we fit a categorical outcome model with cross-sectional data. Below, we will discuss how to fit a categorical outcome model with panel data. Below is a list of real-world examples of categorical dependent variables:

- What news sources do you read on regular basis?
- Which of the following issues are important for the government to address?
- Which of the following primary election candidates would you vote for?

In Stata, these types of models are called panel-data mixed logit choice models and can be fit with the **`cmxtmixlogit`** command.

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I create a panel dataset to fit a mixed logit choice model?

Unlike the prior chapters, panel data for mixed logit choice models are more complicated. After scouring open access journal articles and associated datasets, searching Dataverse, trying to access walled-off data at ICPSR and ANES, and a couple of other apolitical sources, I reached the conclusion that there is not suitable panel dataset to conduct a panel-data mixed logit choice model for walkthrough purposes.

Instead of accepting that I fruitlessly spent about 2 hours searching through articles and datasets, I am going to make lemonade out of lemons and use Stata Manual's example for the **`cmxtmixlogit`** to analogize a potential political science-oriented dataset.

Step-by-Step Walkthrough

Step 1. Visit [Panel-data mixed logit | New in Stata 16]

Step 2. Watch [Choice models - YouTube]

Step 3. Download the Stata Manual for **[CM] cmxtmixlogit**.

a. Scroll to page 6 and let us observe Example 1: Panel-data mixed logit model with alternative- and case-specific covariates.

Step 4. Let us observe how the Transportation choice data is organized:

```
. use https://www.stata-press.com/data/r16/transport
(Transportation choice data)

. list in 1/12, sepby(t)

     +--------------------------------------------------------------------------+
     | id   t       alt   choice   trcost   trtime   age   income    parttime |
     |--------------------------------------------------------------------------|
  1. |  1   1       Car        1     4.14     0.13   3.0        3   Full-time |
  2. |  1   1    Public        0     4.74     0.42   3.0        3   Full-time |
  3. |  1   1   Bicycle        0     2.76     0.36   3.0        3   Full-time |
  4. |  1   1      Walk        0     0.92     0.13   3.0        3   Full-time |
     |--------------------------------------------------------------------------|
  5. |  1   2       Car        1     8.00     0.14   3.2        5   Full-time |
  6. |  1   2    Public        0     3.14     0.12   3.2        5   Full-time |
  7. |  1   2   Bicycle        0     2.56     0.18   3.2        5   Full-time |
  8. |  1   2      Walk        0     0.64     0.39   3.2        5   Full-time |
     |--------------------------------------------------------------------------|
  9. |  1   3       Car        1     1.76     0.18   3.4        5   Part-time |
 10. |  1   3    Public        0     2.25     0.50   3.4        5   Part-time |
 11. |  1   3   Bicycle        0     0.92     1.05   3.4        5   Part-time |
 12. |  1   3      Walk        0     0.58     0.59   3.4        5   Part-time |
     +--------------------------------------------------------------------------+
```

Figure 15-1: Output of use https://www.stata-press.com/data/r16/transport command

Step 5. Now, let us create an analogous dataset, but instead of transportation choice, let us consider choice of primary presidential candidate.

id	t	alt	choice	cand_ideo	cand_age	cand_race	cand_gender	resp_age	resp_race	resp_gender	resp_edu
1	1	Biden	0	0	78	White	male	35	Latino	male	College
1	1	Buttigieg	0	-1	38	White	male	35	Latino	male	College
1	1	Sanders	1	-3	79	White	male	35	Latino	male	College
1	1	Warren	0	-2	71	White	female	35	Latino	male	College
1	2	Biden	0	0	78	White	male	35	Latino	male	College
1	2	Buttigieg	0	-1	38	White	male	35	Latino	male	College
1	2	Sanders	0	-3	79	White	male	35	Latino	male	College
1	2	Warren	1	-2	71	White	female	35	Latino	male	College
1	3	Biden	0	0	78	White	male	35	Latino	male	College
1	3	Buttigieg	0	-1	38	White	male	35	Latino	male	College
1	3	Sanders	0	-3	79	White	male	35	Latino	male	College
1	3	Warren	1	-2	71	White	female	35	Latino	male	College

Figure 15-2: Screenshot of Choice of Primary Presidential Candidate dataset

- `id` = unique id for the respondent of a survey that occurs over a 3-month period
- `t` = time period in months
- `alt` = alternatives or choices available to the respondent

- `choice` = is a 0/1 indicator of the candidate the respondent chose. Only 1 of the 4 choices can be marked with 1, the other three choices are marked 0.
- `cand_ideo` = The candidate's ideology on a scale from -3 (most liberal) to 0 (moderate)
- `cand_age` = Age of the candidate
- `cand_race` = Race of the candidate
- `cand_gender` = Gender of the candidate
- `resp_age` = Age of respondent
- `resp_race` = Race of respondent
- `resp_gender` = Gender of respondent
- `resp_edu` = Education level of the respondent

Step 6. Instead of Primary President Candidate choices, we could consider California primary statewide elected candidate choices for U.S. Senator, governor, lieutenant governor, attorney general, insurance commission, secretary of state, state controller, state treasurer, and superintendent of public instruction.

Step 7. Both prior examples are "high profile"; however, what could be other political choices?

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 7-step process above that you find most interesting.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 16 - Panel Data Count Outcome Models

About

Count models are used when your dependent (aka outcome) variable represents a count of some object or actions and ranges from 0 to positive infinity. Previously, we fit a count outcome model with cross-sectional data in Chapter 11. Now we will fit a count outcome model with panel data. Below is a list of real-world examples of count dependent variables:

- How many courthouses did Congress authorize for a specific state?
- How many hearings did a state legislative committee hold in a specific legislative session?
- How many times did a U.S. citizen donate to political candidates in a campaign election cycle?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I run a panel data count outcome model in Stata using political science data?

For this walkthrough, we will return to a [Judicial Pork: The Congressional Allocation of Districts, Seats, Meeting Places, and Courthouses to the U.S. District Courts] dataset that I am thoroughly familiar with because I collected the data for my dissertation.

Step-by-Step Walkthrough

Step 1. Open Stata

Step 2. Type the following command:

```
use "https://www.ipsrm.com/stata/Franco_Judicial_Pork_July_3_2018.dta"
```

Step 3. Let us assume our research question is: What is the relationship between committee leadership and majority and minority committee members and securing judicial seats?

Step 4. Our dependent variable is:

- `DpV_JSt` = Count of Judicial Seats

Step 5. Our independent variables are:

- Senate Judiciary Chair= `IdV_S_JChair`
- Senate Judiciary Majority Member = `IdV_S_JMbr_Maj`

- Senate Judiciary Minority Member = `IdV_S_JMbr_Min`
- House Judiciary Chair = `IdV_HR_JChair`
- House Judiciary Majority Member = `IdV_HR_JMbr_Maj`
- House Judiciary Minority Member = `IdV_HR_JMbr_Min`
- Judicial Vacancies = `Ctrl_JVac`

Step 6. Type the following command:

```
xtnbreg DpV_JSt IdV_S_JChair IdV_S_JMbr_Maj IdV_S_JMbr_Min IdV_HR_JChair
IdV_HR_JMbr_Maj IdV_HR_JMbr_Min Ctrl_JVac if DpV_JSt>=0, fe irr
```

Step 7. Let us review the output together:

```
. xtnbreg DpV_JSt IdV_S_JChair IdV_S_JMbr_Maj IdV_S_JMbr_Min IdV_HR_JChair IdV_HR_JMbr_Maj IdV_HR_JMbr_Min Ctrl
> _JVac if DpV_JSt>=0, fe irr

Iteration 0:   log likelihood = -2063.1075
Iteration 1:   log likelihood = -1865.8168
Iteration 2:   log likelihood = -1857.7928
Iteration 3:   log likelihood = -1857.6737
Iteration 4:   log likelihood = -1857.6734
Iteration 5:   log likelihood = -1857.6734

Conditional FE negative binomial regression     Number of obs     =      8,705
Group variable: id_icpsr_st~de                  Number of groups  =         50

                                                Obs per group:
                                                              min =         56
                                                              avg =      174.1
                                                              max =        226

                                                Wald chi2(7)      =      27.37
Log likelihood  = -1857.6734                    Prob > chi2       =     0.0003

------------------------------------------------------------------------------
        DpV_JSt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
   IdV_S_JChair |   .5531653   .2569059    -1.27   0.202     .2226037    1.374603
 IdV_S_JMbr_Maj |   1.180312   .1543307     1.27   0.205     .9134794    1.525088
 IdV_S_JMbr_Min |   1.183964   .1710839     1.17   0.243     .8919468    1.571585
  IdV_HR_JChair |   .7149959   .2127242    -1.13   0.259     .3990761    1.281006
IdV_HR_JMbr_Maj |   1.257846   .0861945     3.35   0.001     1.099762    1.438654
IdV_HR_JMbr_Min |   1.167565   .0698733     2.59   0.010     1.038343     1.31287
      Ctrl_JVac |   .6383333   .0989621    -2.90   0.004     .4710683    .8649899
          _cons |   .0635517   .0067916   -25.79   0.000      .051542    .0783597
------------------------------------------------------------------------------
```

Figure 16-1: Result of the xtnbreg command

- `IRR` Column
  - IRR stands for incidence rate ratio. <1 means a % decrease in incidence rate of being allocated judicial pork while >1 means a % increase in incidence rate of being allocated judicial pork.
- `P>|z|`: The following three independent variables have a P>|z| less than 0.10.
  - House Judiciary Committee majority member
  - House Judiciary Committee minority member
  - Judicial vacancies

Step 8. Visit Negative Binomial Regression | Stata Data Analysis Examples (ucla.edu) to learn more about **`nbreg`.**

Mini-Assignment #1: Instructions

Step 1: Select a 3-step sequence subset from the 8-step process above that you find most interesting.

Step 2: Explain in 6 or more sentences why you selected this specific 3-step sequence.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Selected 3-step sequence	Yes Missing	10 0
Explained selected 3-step sequence: # sentences	6 5 4 3 2 1 0	90 75 60 45 30 15 0

Chapter 17 - Survival Models

About

Survival models, which are also known as duration models or event history models, are used when your dependent (aka outcome) variable represents a time-to-event which ranges from 0 to some large positive number. Below is a list of real-world examples of survival dependent variables:

- How long will a candidate for president stay in the race before dropping out?
- How long does it a take for a bill to become law?
- How long does a Cabinet-level appointee stay in their position before resigning?

Estimated Time

An estimated 120-180 minutes is needed to complete this activity.

How do I interpret the statistical output of a survival model?

Normally, we explore how to run an empirical model in Stata using political science data. However, for survival models, I want you to read the following excerpt from a draft of my dissertation that explains the output of two types of survival models.

> Binary outcome models determine what effect covariates have on the allocation of judicial pork to a state. However, judicial pork is a rare event, like discovering gold or riding in a helicopter, so we may be interested in how covariates affect the time until Congress allocates a state judicial pork.
>
> Survival models are known as event history models, duration models, or time-to-event models (Allison, 2014; Box-Steffensmeier & Jones, 2004). Survival models can be used to answer the question: *given a set of covariates, how long will a state survive without obtaining judicial pork?*
>
> There are two types of survival models to consider: Panel parametric survival model and Cox semi-parametric survival model. The primary difference is that parametric models assume a distribution (exponential or Weibull) to determine the hazard rate, while semi-parametric models do not make a distributional assumption.

Depending on the nature of the event(s) of interest (Metzger & Jones, 2016), it matters which survival model specification is used to estimate results. Given that states can repeatedly experience the allocation of a type of judicial pork, the use of a semi-parametric Cox model specified as clustered or shared frailty is most appropriate (Box-Steffensmeier, Linn, & Smidt, 2014; Cleves, 2017).

The cluster specification accounts for intra-state correlation through the standard errors while the shared frailty specification accounts for intra-state correlation through the hazard function (Cleves, 2017). It is possible to use parametric survival models, but this would impose restrictive assumptions (Box-Steffensmeier & De Boef, 2006; Box-Steffensmeier et al., 2014; Box-Steffensmeier & Zorn, 2002).

The table below shows the results of Clustered (CL) versus Shared Frailty (SF) Cox Models by Judicial Pork Type. While both model specifications are present in the table, the analysis for Districts will be focused on clustered (CL) version, while my analysis for Seats, Meeting Places, and Courthouses will focus on the shared frailty (SF) version. The reason is that shared frailty cannot be ruled out for the latter three, while it can be ruled out for Districts.[2]

This table reports hazard rates instead of standard beta coefficients. A hazard rate above one means that a state's rate of obtaining judicial pork increases, while a rate below one means that a state's rate decreases. First, we find that Senate Judiciary Committee Chairmanship is not statistically significant across any of the models.

This result corresponds with the fixed-effect logit models. Second, a state having a Senator on the Judiciary Committees or holding the House Judiciary Committee Chairmanship increases the rate of securing courthouses by 28% and 69%, respectively. Third, states with rank-and-file representatives on the House Judiciary Committee increases the rate of securing a Judicial District by 65%, a Meeting Place by 16%, and a Courthouse by 17%. However, this covariate is not statistically significant for Judicial Seats within the shared frailty (SF) model.

A one-unit change in Judicial Vacancies decreases the rate of which a state is allocated a Seat and Courthouse by 34% and 74%, respectively. Unlike the across-the-board positive effect of Unified Government, in the survival models we find that it only has a positive effect for Seats and Meeting Places, but not Districts or Courthouses.

---

[2] This is based on a likelihood-ratio test that H_0: $\Theta=0$

Table 17-1: Clustered (CL) versus Shared Frailty (SF) Cox Models by Judicial Pork Type

	Hypothesized Value	Districts CL	Districts SF	Seats CL	Seats SF	Meeting Places CL	Meeting Places SF	Courthouses CL	Courthouses SF
Senate Judiciary Chair	>1	2.85e-20	1.71e-15	0.746	0.864	1.359	1.346	0.967	0.963
		(.)	(-0.00)	(-0.71)	(-0.31)	(1.06)	(0.91)	(-0.09)	(-0.13)
Senate Judiciary Member	>1	1.431	1.431	1.085	1.110	1.183	1.169	1.288*	1.285*
		(1.41)	(1.28)	(0.79)	(0.97)	(1.36)	(1.36)	(2.18)	(2.46)
House Judiciary Chair	>1	0.727	0.727	0.921	0.861	1.165	1.033	1.832*	1.685*
		(-0.49)	(-0.29)	(-0.37)	(-0.47)	(0.51)	(0.10)	(1.98)	(2.03)
House Judiciary Member	>1	1.648**	1.648*	1.283***	1.102	1.222*	1.163*	1.248*	1.165*
		(2.60)	(2.03)	(5.35)	(1.67)	(2.50)	(2.03)	(2.40)	(2.15)
Judicial Vacancies	<1	6.56e-20	1.07e-14	0.733*	0.638**	0.813	0.811	0.259**	0.263**
		(.)	(-0.00)	(-2.09)	(-2.87)	(-0.77)	(-1.07)	(-2.94)	(-2.79)
Unified Government	>1	1.598	1.598	1.573***	1.718***	1.398**	1.391**	1.165	1.161
		(1.80)	(1.63)	(4.26)	(4.40)	(2.91)	(2.69)	(1.64)	(1.43)
Senate Majority Leader		5.55e-19	1.65e-15	1.536	1.633	0.819	0.764	1.023	1.039
		(.)	(-0.00)	(1.13)	(1.22)	(-0.35)	(-0.45)	(0.04)	(0.08)
Senate Minority Leader		2.30e-19	1.87e-15	1.203	1.050	1.251	1.120	0.848	0.799
		(.)	(-0.00)	(0.47)	(0.11)	(0.48)	(0.22)	(-0.52)	(-0.48)
House Speaker		1.018	1.018	0.853	0.888	0.761	0.749	0.397***	0.370**
		(0.04)	(0.02)	(-0.61)	(-0.36)	(-1.12)	(-0.86)	(-3.53)	(-2.66)
House Majority Leader		4.114	4.114	2.191**	2.134*	2.018*	1.719	2.216*	1.935
		(1.89)	(1.62)	(3.06)	(2.55)	(2.13)	(1.42)	(2.46)	(1.89)
House Rules Chair		2.927	2.927	1.634**	1.468	2.463***	2.042*	2.751***	2.256**
		(1.07)	(0.96)	(2.64)	(1.36)	(4.18)	(2.29)	(3.44)	(2.74)
House Minority Leader		1.85e-18	2.03e-15	0.970	0.993	1.268	1.260	1.000	0.933
		(.)	(-0.00)	(-0.09)	(-0.02)	(0.57)	(0.53)	(-0.00)	(-0.16)
Senate Appropriations Chair		2.980*	2.980	1.374	1.361	1.354	1.350	2.185**	2.080**
		(2.16)	(1.43)	(1.07)	(0.82)	(0.63)	(0.80)	(2.73)	(2.63)
House Appropriations Chair		1.70e-19	1.38e-15	1.431	1.311	1.104	0.922	1.077	0.941
		(.)	(-0.00)	(1.26)	(0.91)	(0.23)	(-0.20)	(0.22)	(-0.18)
House Ways and Means Chair		0.885	0.885	0.939	0.878	0.746	0.700	0.908	0.843
		(-0.29)	(-0.15)	(-0.25)	(-0.43)	(-0.98)	(-0.99)	(-0.27)	(-0.59)
Senate Public Works Chair or Equivalent		1.98e-19	4.35e-15	0.528	0.562	1.173	1.216	0.955	0.890
		(.)	(-0.00)	(-1.55)	(-1.23)	(0.43)	(0.54)	(-0.20)	(-0.37)
House Public Works Chair or Equivalent		4.453**	4.453*	1.186	1.120	1.657	1.636	1.127	0.996
		(2.75)	(2.27)	(0.59)	(0.36)	(1.69)	(1.58)	(0.51)	(-0.01)
President		0.808	0.808	1.254	1.117	0.914	0.896	0.733	0.723
		(-0.27)	(-0.28)	(1.59)	(0.43)	(-0.28)	(-0.34)	(-0.96)	(-1.04)
N		8751	8751	8751	8751	8751	8751	8751	8751
chi2		41.63	20.73	182.6	46.47	149.7	33.07	106.1	53.57
aic		627.2	641.2	2936.1	2912.1	2956.3	2940.1	3491.3	3473.8
bic		705.1	768.6	3063.5	3039.5	3083.7	3067.5	3618.7	3601.2

Mini-Assignment #1: Instructions

Step 1: Select at least three objects (word, paragraph, or concept) from the reading above that you are found interesting or perplexing.

Step 2: For each object, explain why you found it interesting or perplexing.

Mini-Assignment #1: Rubric

Criteria	Ratings	Points
Object 1 selected	Yes No	10 0
Object 2 selected	Yes No	10 0
Object 3 selected	Yes No	10 0
Explain Why Object 1	Yes No	30 0
Explain Why Object 2	Yes No	30 0
Explain Why Object 3	Yes No	30 0

Chapter 18 - Share

About

Share is an opportunity for you to share with your peers any of your Polimetrics Chapter Assignments.

Estimated Time

An estimated 90-120 minutes is needed to complete this activity.

Instructions

Post

- Specify which Polimetrics Chapter Assignment you are sharing with the class.
- In 3 or more sentences, explain why you wanted to share this assignment, compared to other, chapter assignments.
- Ask a specific question that you would like a peer to reply to. Examples of questions include:
    - What do you think of my submission for a specific assignment?
    - How is my explanation similar to what you wrote for the same assignment?
    - How is my explanation different from what you wrote for the same assignment?
    - What is a strength of my submission for the specific assignment?
    - What is an area I could expand upon for the specific assignment?

Reply to a Peer’s Post

- In 3 or more sentences, respond to the question your peer asked in their original post.

Rubric

Criteria	Ratings	Points
Post: Assignment specified	Yes No	20 0
Post: Why You Chose to Share this Assignment	3 sentences 2 sentences 1 sentence Missing	30 20 10 0

Post: Included Question for Peer to Respond To	Yes No	20 0
Post Quality: Subjective evaluation by Professor	01 – Superb 02 – Excellent 03 – Great 04 – Good 05 – Insufficient	0 0 0 0 0
Reply: # sentences	3 sentences 2 sentences 1 sentence Missing	30 20 10 0
Reply Quality: Subjective evaluation by Professor	01 – Superb 02 – Excellent 03 – Great 04 – Good 05 – Insufficient	0 0 0 0 0

2224

Chapter 19 - Reflection

About

Reflection is an opportunity for you share with me, your professor, your thoughts about the Polimetrics Chapter Assignments. No other student will read your reflection.

Estimated Time

An estimated 60-120 minutes is needed to complete this activity.

Instructions

Please write at least 5 sentences reflecting on the Polimetrics Chapter Assignments. To be clear, this reflection should focus on the Polimetrics Chapter Assignments as a whole. This reflection should not be about a specific chapter or another assignment.

Sentence #1: Your 1st Sentence should be a question. Examples of questions include:

- How can I apply the Polimetrics Chapter Assignments to my daily life or academic studies?
- What did you find most interesting about the Polimetrics Chapter Assignments? Why did you find this the most interesting?
- What did you find most relevant to your daily life about the Polimetrics Chapter Assignments? Why did you find this the most relevant?
- You are welcome to ask and answer your own question.

Sentence #2-5: Sentences 2 through 5 should be your response to the question you posed in sentence #1.

Rubric

Criteria	Ratings	Points
1st Sentence a Question	Yes No	25 0
Quantity: # Sentences	4 3 2 1 0	75 60 45 30 0
Quality: Subjective evaluation by Professor	01 – Superb 02 – Excellent	0

	03 – Great 04 – Good 05 – Insufficient	

2248

References

Allison, P. D. (2014). *Event history and survival analysis* (Second edition. ed.). Los Angeles: SAGE.

Box-Steffensmeier, J. M., & De Boef, S. (2006). Repeated events survival models: the conditional frailty model. *Statistics in Medicine, 25*(20), 3518-3533. doi:10.1002/sim.2434

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event history modeling : a guide for social scientists*. Cambridge ; New York: Cambridge University Press.

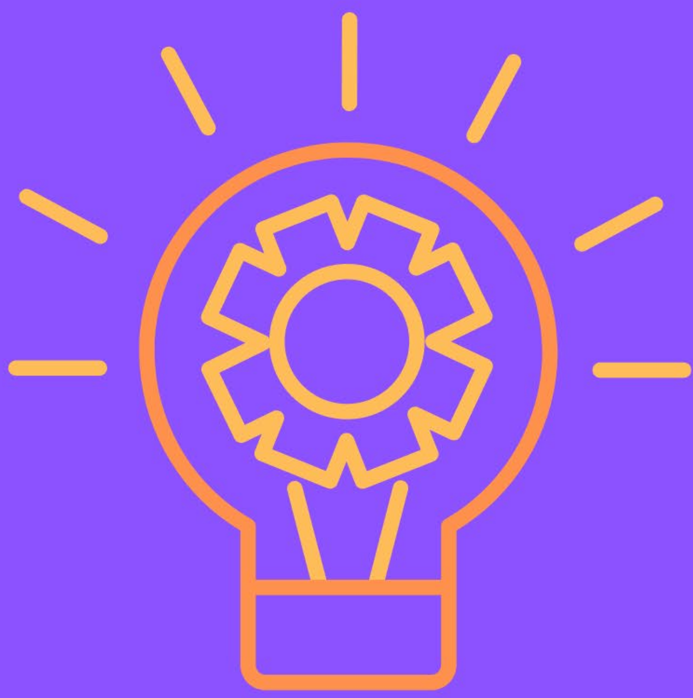
Box-Steffensmeier, J. M., Linn, S., & Smidt, C. D. (2014). Analyzing the Robustness of Semi-Parametric Duration Models for the Study of Repeated Events. *Political Analysis, 22*(02), 183-204. doi:10.1093/pan/mpt015

Box-Steffensmeier, J. M., & Zorn, C. (2002). Duration models for repeated events. *Journal of Politics, 64*(4), 1069-1094. doi:10.1111/1468-2508.00163

Cleves, M. (2017). How do I analyze multiple failure-time data using Stata? Retrieved from https://www.stata.com/support/faqs/statistics/multiple-failure-time-data/

Fish, P. G. (1973). *The politics of Federal judicial administration*. Princeton, N.J.,: Princeton University Press.

Metzger, S. K., & Jones, B. T. (2016). Surviving Phases: Introducing Multistate Survival Models. *Political Analysis, 24*(4), 457-477. Retrieved from <Go to ISI>://CCC:000386939300004



AN OPEN EDUCATION RESOURCE

ISBN: PENDING