

Matthew Loveless

POLITICAL ANALYSIS

A Guide to Data & Statistics

Second Edition

 **Sage**

Contents

AN INTRODUCTION.....	3
About the Datasets	3
<i>More Manageable</i>	4
<i>Data Manipulation</i>	4
From Practice to Research.....	6
Embrace the Software.....	6
An Introduction to STATA.....	7
An Introduction to SPSS.....	13
An Introduction to R.....	17
An Introduction to Python.....	27
QUESTIONS FOR CHAPTER 1: The Scientific Method and Statistics.....	33
QUESTIONS FOR CHAPTER 2: Theory and Hypotheses	36
QUESTIONS FOR CHAPTER 3: Data and Variables	41
QUESTIONS FOR CHAPTER 4: Research Design and the Scientific Study of Politics	44
QUESTIONS FOR CHAPTER 5: The Ethics of Data Analysis	49
QUESTIONS FOR CHAPTER 6: Univariate Descriptive Statistics.....	53
DATASETS.....	58
QUESTIONS FOR CHAPTER 7: Measures of Association I: Nominal- and Ordinal-Level Variables.....	59
DATASETS:	63
QUESTIONS FOR CHAPTER 8: Measures of Association II: Means Comparison and Correlation.....	65
Datasets	73
QUESTIONS FOR CHAPTER 9: Measures of Association III: (Bivariate) Regression.....	74
DATASETS:	79
QUESTIONS FOR CHAPTER 10: An Introduction to Inference	80
QUESTIONS FOR CHAPTER 11: Inference for Nominal- and Ordinal-Level Variables.....	84
Datasets	89
QUESTIONS FOR CHAPTER 12: The Central Limit Theorem	90
QUESTIONS FOR CHAPTER 13: Inference for Interval-level Variables	96
Datasets	102
QUESTIONS FOR CHAPTER 14: An Introduction to Multiple Regression.....	103
Datasets	112
QUESTIONS FOR CHAPTER 15: Extensions to Multiple Regression	113

Political Analysis Workbook

Datasets	120
QUESTIONS FOR CHAPTER 16: Issues with Multiple Regression.....	121
Datasets	128
QUESTIONS FOR CHAPTER 17: Binary Logistic Regression	129
Datasets	134
QUESTIONS FOR CHAPTER 18: Categorical and Limited Dependent Variables	135
Datasets	139

AN INTRODUCTION

This workbook is designed to help you move from reading about statistical techniques to actively using them. While the textbook introduces the concepts, this workbook is a place where you can develop and improve your statistical literacy, statistical abilities, and research skills. Here you can wade into the data and get your hands dirty practicing with the concepts and techniques found in *Political Analysis: A Guide to Data & Statistics*. This is data analysis in practice and will bring you nearer the world of empirical social science research.

Across all disciplines, but especially in the social sciences, the ability to correctly apply techniques to real data is nearly as essential as knowing how to use theory. Think of this workbook as your personal creativity room, your private dojo, your isolated recording studio. It offers space to experiment, question, and develop fluency—not just in syntax or statistical logic, but in the broader process of generating meaningful results from empirical data.

The workbook contains three types of questions, each reinforcing a different aspect of statistical learning:

1. **Theoretical Questions:** These questions ask you to reflect more deeply on the conceptual and philosophical underpinnings of the techniques. What assumptions are being made? What are the limitations of the method? These questions link statistical work to broader methodological debates in social science.
2. **Practice Questions:** These questions build directly on examples from the textbook. These are designed to help you apply methods you've already seen in familiar contexts, reinforcing the mechanics of each statistical tool.
3. **Expansion Questions:** These questions introduce new variations or challenges. They might ask you to modify a model, work with a different dataset, or consider what happens when the assumptions of a method are violated. These questions help extend your skills beyond the core examples.

Taken together, these three types of questions are intended to support both competence and curiosity. Don't be afraid or frustrated. Confronting this trifold variety of questions is meant to strengthen your grasp of techniques while also pushing you to think like a Social Science researcher.

About the Datasets

The datasets provided with *Political Analysis: A Guide to Data & Statistics* are also for this workbook. As I noted in the introduction of the textbook, there are a number of reasons I wanted you to use 'real data' that boils down to: **because you can**. What you are learning is not a facsimile of quantitative analysis, it is quantitative analysis. So, why should the data be anything else than real data?

The datasets for *Political Analysis* are drawn from several large, respected public sources. They represent real survey and political data from across the world, collected using transparent and ethical methods. These cleaned datasets have been curated to allow you to focus on analytical questions rather than spending hours sorting out technical issues.

Here is an example of what you will find below at the end of each chapter, including links to their original sources, codebooks, and documentation:

UCDP/PRIO Armed Conflict dataset

- Davies, Shawn, Garoun Engström, Therese Pettersson & Magnus Öberg (2024). Organized violence 1989-2023, and the prevalence of organized crime groups. *Journal of Peace Research* 61(4). [Version 24.1]
 - Available here: <https://ucdp.uu.se/downloads/>.
 - *Political Analysis* dataset: "Armed Conflict"

Feel free to explore the original data and look at the codebooks. Perhaps something will catch your eye.

I have included the citations to the original datasets so that you can see for yourself that the included, smaller datasets here are derivations that deviate only in two ways from the original data.

More Manageable

One, I deleted some of the variables from the original datasets for the sake of manageability. Some of the available datasets are very large and can be a bit cumbersome to search through for specific variables (especially when many variable names are inconveniently similar). Large datasets also demand a lot of computing power which is a testament to the seriousness of these data.

Data Manipulation

Two, we need to talk about a dirty word. The second reason that we are using smaller, derived datasets from the originals is 'data manipulation.' Data manipulation makes it sound like we are 'altering the data.' In fact we are - but we are not *changing* the data. What data manipulation refers to is the *cleaning* and *recoding* of data to make working with these data easier, more intuitive, and thus less subject to misunderstanding or misinterpretation.

I'll give you two examples of data manipulation that alter the data but do not change them: data cleaning and recoding.

Data Cleaning

Data cleaning (or even occasionally: 'data wrangling') is just that, preparing data for a greater ease and intuition of use. For the purposes of our research, we might need to *clean* the data to make them more useable. It is not glamorous work but it's one of the most important parts of any research project. Let's take an example from survey data and research.

From survey data on citizens, we will are often very interested in – and very often work with - **voting**. This survey question is usually some form of: 'did you vote in [your country's] last [parliamentary/presidential/local, etc...] election'.

Nearly all the time, what we want to know is 'Yes, I did' or 'No, I did not.' However, most surveys offer a number of additional response categories that may not be directly relevant to our research. Here are some common alternative response categories:

- *I tried to vote but couldn't*: Captures logistical or personal barriers to voting.
- *I am not eligible to vote*: Common in surveys that include non-citizens, minors, or disenfranchised individuals.
- *I don't remember*: Useful in longer recall periods to acknowledge memory limitations.
- *Prefer not to say*: Allows respondents to opt out of answering a sensitive question.

The reason for this is that these other answers, even while usually only having a single percentage point of observations, can be interesting in elections that were highly polarized or at a crucial time. It also minimizes *social desirability bias* in which people say ‘yes, I did’ when they did not because they feel pressured – even unconsciously – from the interviewer to be socially responsible. Including these alternative responses has been found to reduce the overreporting of voting - making the data more accurate. Always a plus.

For our purposes here, these alternatives can be interesting questions. However, they can also be a hassle for straightforward questions needed for learning statistics (and not data manipulation). Therefore, a few of the variables we will work with have been ‘cleaned’ to make them easier to work with. You can do this just as much as I can but we do so for analytical reasons (and of course always report it!).

Recoding Data

Another reason for data manipulation is that some data can be ‘unintuitive.’ Sometimes, surveys will ask about how interested a respondent might be in politics. The response codes may be, for convenience:

1. Very interested
2. Somewhat interested
3. Hardly interested
4. Not interested at all

This is fine. However, when we go to use these data, we naturally tend to think of variables having an intuitive scale. That is, if we know the question asks about a respondent’s level of interest in politics, we might assume that a higher number, say, on this scale above from 1 to 4, to be 4. And if the respondent doesn’t have a high level of interest in politics, we might expect the response to be low, such as 1. However, in the coding above, we would be wrong, in fact, we would be exactly backwards. Here ‘high interest’ is 1 and ‘not interested at all’ is 4.

The most common action is simply to ‘recode’ the variable. That is, we reverse the coding [1=4, 2=3, 3=2, and 4=1]. Doing so puts a higher score for what the question ostensibly is asking for. It does not, however, change the data. Those who responded ‘very interested’/‘somewhat interested’/‘hardly interested’/‘not interested’ at all are still in the same groups but now 1 represents ‘low interest’ and 4 represents ‘high interest.’

Without changing the underlying nature of the data and improving the intuition of the variable, this practice is both widespread and – frankly - expected. Therefore, for the smaller datasets we use here, some variables have been recoded.

I note that, more generally, public datasets are rarely “analysis-ready.” You will have to put in the time to clean them up as they often include inconsistent variable names, missing values, complex coding schemes, and documentation that is difficult to interpret. Transforming these datasets involves recoding variables, checking for errors, simplifying categories, and creating versions that align with the analytical goals of a project.

Yes, this is also ‘doing statistics’. Understanding how to move from a messy dataset to an analysis-ready one is a fundamental research skill—and one that you will almost certainly use in your own work. Hello, research skills and statistical literacy.

From Practice to Research

This brings us to a sidenote for the textbook and workbook. Working with real datasets using statistical software is not just an academic exercise. It is the foundation of original research. Every chart you create, every model you estimate, and every transformation you perform is a small act of inquiry. By the time you reach the end of this workbook, you'll have not only improved your technical skills but also sharpened your ability to design, critique, and carry out research projects of your own.

As you progress, I also encourage you to revisit the textbook's chapter on research ethics. Data analysis is never neutral. From how data is collected to how it is presented, every stage involves choices that carry ethical implications. Be thoughtful about how you handle missing data, how you interpret outliers, and how you communicate uncertainty. Respect for data is part of respect for the people and communities behind it. Make yourself – and your research – impossible to ignore.

This workbook is more than a set of drills. It's a step toward becoming a more independent, critical, and capable researcher. Take your time. Ask questions. Experiment. And remember that every error message is just an invitation to learn something new.

Embrace the Software

This workbook is available in four versions, tailored for STATA, SPSS, R, and Python. If you're new to statistical software, don't worry. It can be a bit intimidating at first but know that no one is born knowing how to interact with statistical software. The learning curve is real, but it's also manageable – and it pays off.

Statistical software is not just a tool for getting results; it's a language for asking questions and exploring relationships in data. The more you use it, the more confident and flexible you'll become – not just replicating textbook examples and working on new problems but also adapting methods to your own research needs.

Make the software your ally. It may look intimidating at first—a wall of commands, errors, or unfamiliar syntax—but that's true of any tool worth learning. Like a pair of nunchucks, it feels impossible in your hands at first. You fumble them, drop things, maybe even bang your shin (or your face!). But give it time. I'm pretty sure not every DJ matched beats on day 1.

Statistical software is not an obstacle to statistics, it is a tool we can use to ask questions, test theories, and advance our work (our world?) with data. Give yourself space to explore, experiment, and occasionally fail. That's how fluency is built. And imagine how amazed other people will be when you whip that software out on your desktop and just go!

An Introduction to STATA

The goal and utility of a workbook is to offer students the opportunity to engage the concepts introduced in the textbook and practice generating and interpreting statistical techniques. Together, this is the process of ‘analysing the data’ or ‘doing statistics.’

This workbook is associated with the Stata statistical software. SPSS uses a ‘code driven’ approach, you have to tell Stata what to do and how to do it.

Let's introduce ourselves to the Stata environment. Like other statistical software, Stata is based on the $N \times K$ design where 'N' refers the number of observations and 'K' refers to the number of variables. Most often, datasets order observations and variables in the form of rows (observations) and columns (variables).

FIGURE 1: N x K Design

[illegible]

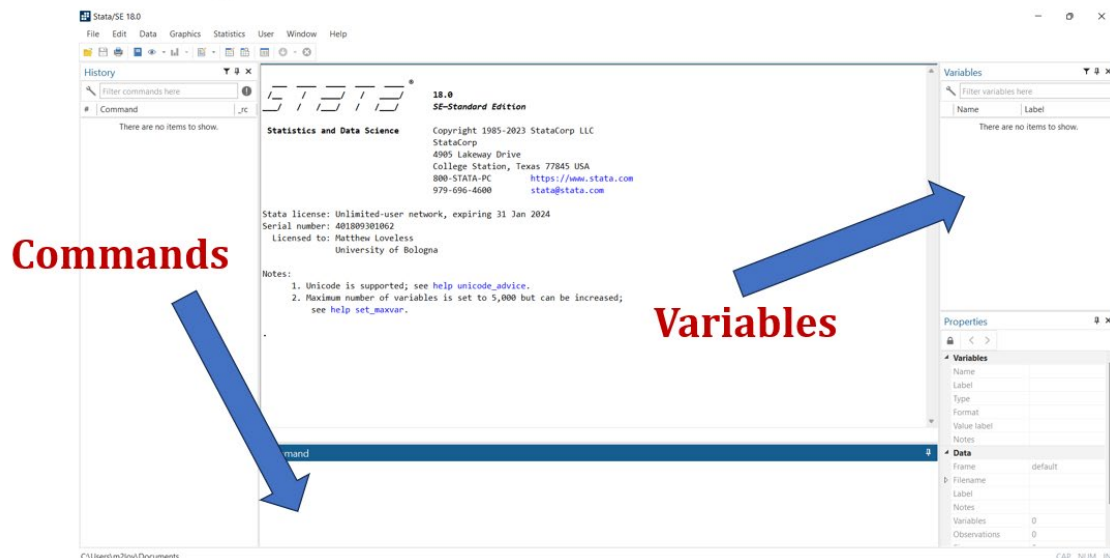
For STATA, however, the opening window looks a little different (Figure 2a):

FIGURE 2a The Opening Window in STATA



However, this opening window is designed to maximize our working with our data. In Figure 2b, when we open a dataset (as we will below), there is a place for us to put our commands – that ask STATA to do certain statistical techniques – and view the variables from our dataset (see Figure 2b).

FIGURE 2b The Opening Window in STATA: Commands and Variables



In order to get the the N x K Desgin, we need to ask STATA simply to show it to us (see Figure 3).

FIGURE 3a: Getting to the N x K Design in STATA

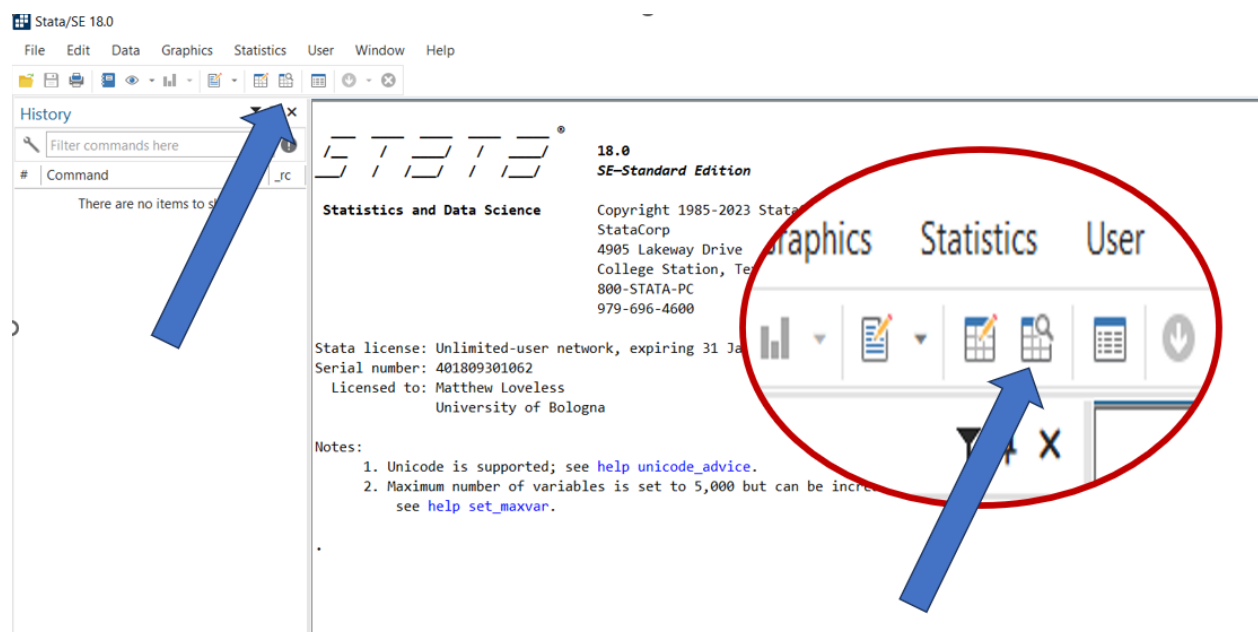
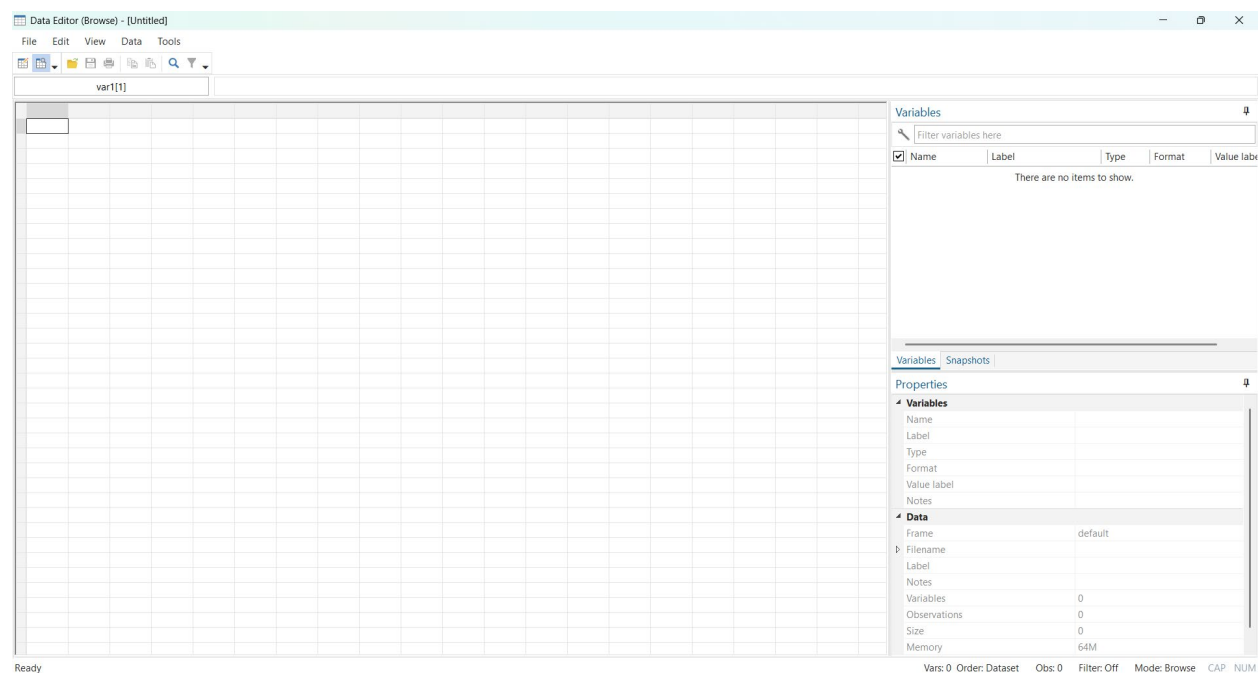


FIGURE 3b: N x K Design in STATA



Returning to the opening window (simply close the N x K window), we can open the “DS - Levels of Measurement.dta” STATA dataset. We use the ‘open files’ icon (Figure 4a) and get the following (Figure 4b).

FIGURE 4a: Using ‘Open Files’ to get the ‘DS - Levels of Measurement.dta’ Dataset

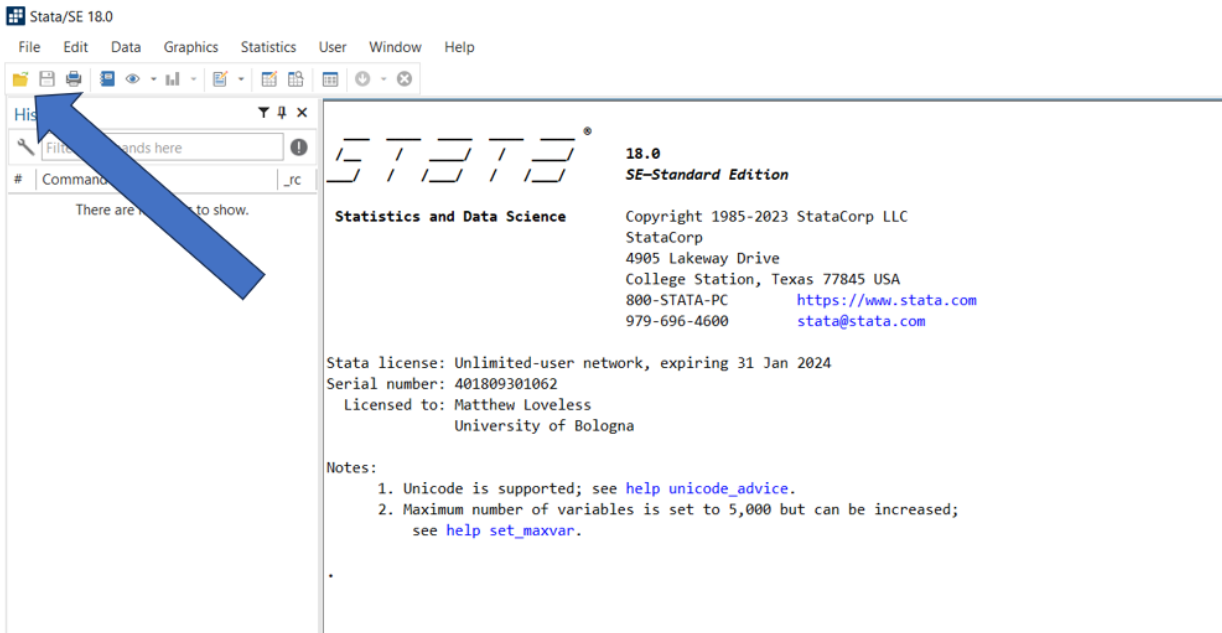
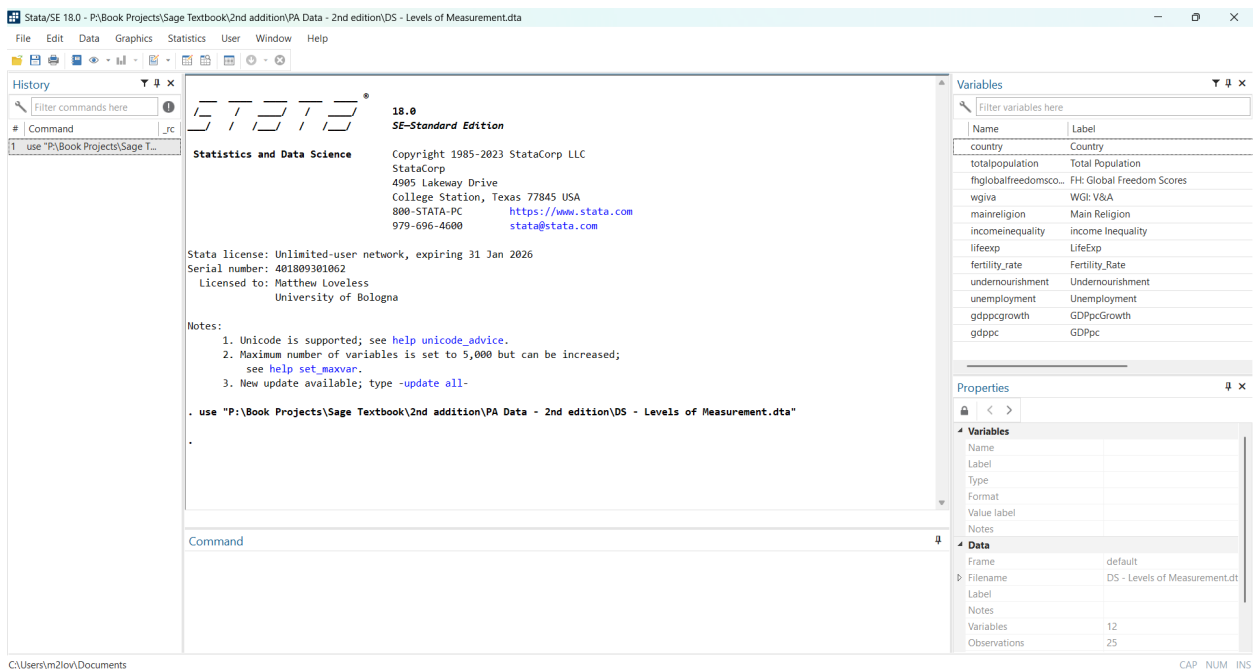


FIGURE 4b: STATA Displaying the ‘DS - Levels of Measurement.dta’ Dataset



Here we can see a number of things. Recall Figure 2b. At the top right, we have the list of variables in our dataset and the command window (center bottom) awaits our instructions. For a fuller look at our data, we can return to the N x K Design page (recall Figure 3a above).

FIGURE 5: Viewing the Variables in STATA (using the 'DS - Levels of Measurement.dta Dataset)

The screenshot shows the STATA Data Editor interface. The main window displays a list of 25 countries with various socio-economic and political variables. The right-hand pane shows the 'Variables' list and the 'Properties' window for the 'country' variable.

country	totalpop...	fhglobalfre...	wgiva	mainreligion	income...	lifeexp	fertility_r...	undem...	unempl...	gdppcg...	gdppc
1 Viet Nam	97338.6	Not Free	-1.38	Folk Religions	35.7	75.4	27.4	6.4	2.3	2	2785.7
2 Democratic Republic of the Co...	89561.4	Not Free	-1.37	Christianity	42.1	60.7	121	.	4.6	-2.3	556.8
3 Egypt	102334	Not Free	-1.43	Islam	31.5	72	52.4	4.7	10.5	1.6	3547.9
4 Ethiopia	114964	Not Free	-1.3	Christianity	35	66.6	63.4	19.7	2.8	3.4	936.3
5 China	1.4e+06	Not Free	-1.61	Unaffiliated Religions	38.5	76.9	7.7	2.5	5	2	10500.4
6 Iran	83993	Not Free	-1.37	Islam	42	76.7	40.8	4.7	11	4	2252.6
7 Thailand	69800	Not Free	-83	Buddhism	34.9	77.2	44.5	9.3	1	6.3	7189
8 Myanmar	54409.9	Not Free	-84	Buddhism	30.7	67.1	27.9	14.1	1.8	-10.6	1400.2
9 Turkey	84339.1	Not Free	-81	Islam	41.9	77.7	24.6	2.5	13.9	-7	8538.2
10 Russian Federation	145935	Not Free	-1.1	Christianity	37.5	73.1	18.5	2.5	5.7	-2.8	10126.7
11 Nigeria	206140	Partly Free	-41	Christianity	35.1	54.7	103.6	12.8	9	-4.3	2097.1
12 Mexico	128933	Partly Free	.02	Christianity	45.4	75.1	58.5	7.1	4.7	-9.2	8346.7
13 Philippines	109581	Partly Free	.03	Christianity	42.3	71.2	55.4	14.5	3.4	-10.8	3298.8
14 Tanzania	59734.2	Partly Free	-5	Christianity	40.5	65.5	115.5	25	2.2	-9	1076.5
15 India	1.4e+06	Partly Free	29	Hinduism	35.7	69.7	11	14	7.1	-8.9	1900.7
16 Indonesia	273524	Partly Free	16	Islam	38.2	71.7	46.4	9	4.1	-3.1	3869.6
17 Bangladesh	164689	Partly Free	-72	Islam	32.4	72.6	81.7	13	5.3	1.4	1968.8
18 Italy	60461.8	Free	97	Christianity	35.9	83.2	5	2.5	9.3	-8.6	31676.2
19 Germany	83784	Free	1.34	Christianity	31.9	80.9	7.6	2.5	4.3	-5.1	45723.6
20 France	65273.5	Free	1.14	Christianity	32.4	82.6	4.7	2.5	8.6	-8.3	38625.1
21 United Kingdom	67886	Free	1.26	Christianity	35.1	81.2	11.9	2.5	4.3	-10.3	40284.6
22 Brazil	212659	Free	34	Christianity	53.4	75.9	56.7	2.5	13.7	-4.7	6796.8
23 South Africa	59308.7	Free	67	Christianity	63	64.1	67.8	5.7	28.7	-8.1	5080.7
24 Japan	126477	Free	96	Unaffiliated Religions	32.9	84.4	3.5	2.5	3	5	40113.1
25 United States of America	331003	Free	97	Christianity	41.4	78.8	17.3	2.5	8.3	-3.8	63543.6

Variables

Name	Label	Type	Format	Value l
country	Country	str32	%32s	
totalpopulation	Total Population	float	%8.0g	
fhglobalfreedo...	FH: Global Freedom Sco...	byte	%11.0g	ggg
wgiva	WGI: V&A	float	%8.0g	
mainreligion	Main Religion	str22	%22s	
incomeinequality	Income Inequality	float	%8.0g	
lifeexp	LifeExp	float	%8.0g	
fertility_rate	Fertility_Rate	float	%8.0g	
undemourishme...	Undemourishment	float	%8.0g	
unemployment	Unemployment	float	%8.0g	

Properties

Variables

- Name: country
- Label: Country
- Type: str32
- Format: %32s
- Value label: Notes

Data

- Frame: default
- Filename: DS - Levels of Measurement.dta
- Label: Notes
- Variables: 12
- Observations: 25
- Size: 2.22K
- Memory: 64M

Returning to the opening window (simply closing the N x K window), we can begin to consider the techniques and approaches with which we are familiar in our study of statistics: Descriptive Statistics, Means Comparison, Correlation, Multiple Regression, and Logistic Regression, *inter alia*. We will put our instructions to STATA in the command window (see Figure 2b).

Or, if we want to do a lot of them, we can use what is called a 'do-file'. A do-file is essentially a text function that allows us to input, change, and update our commands and – crucially – save them. I recommend using the do-file as a 'all in one' place to save the code for the various techniques we are learning. You can write notes to yourself and remind yourself of what each one does and how to interpret them. You can open a do-file this way (Figure 6a) and it looks like this (Figure 6b).

FIGURE 6a: Opening a STATA Do-File

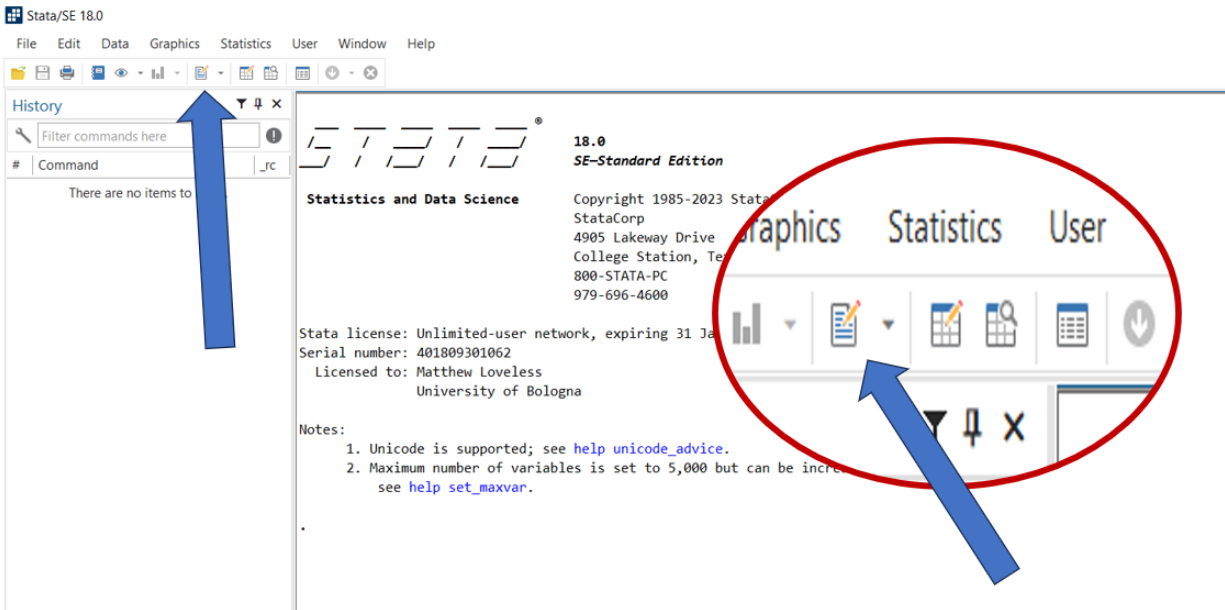
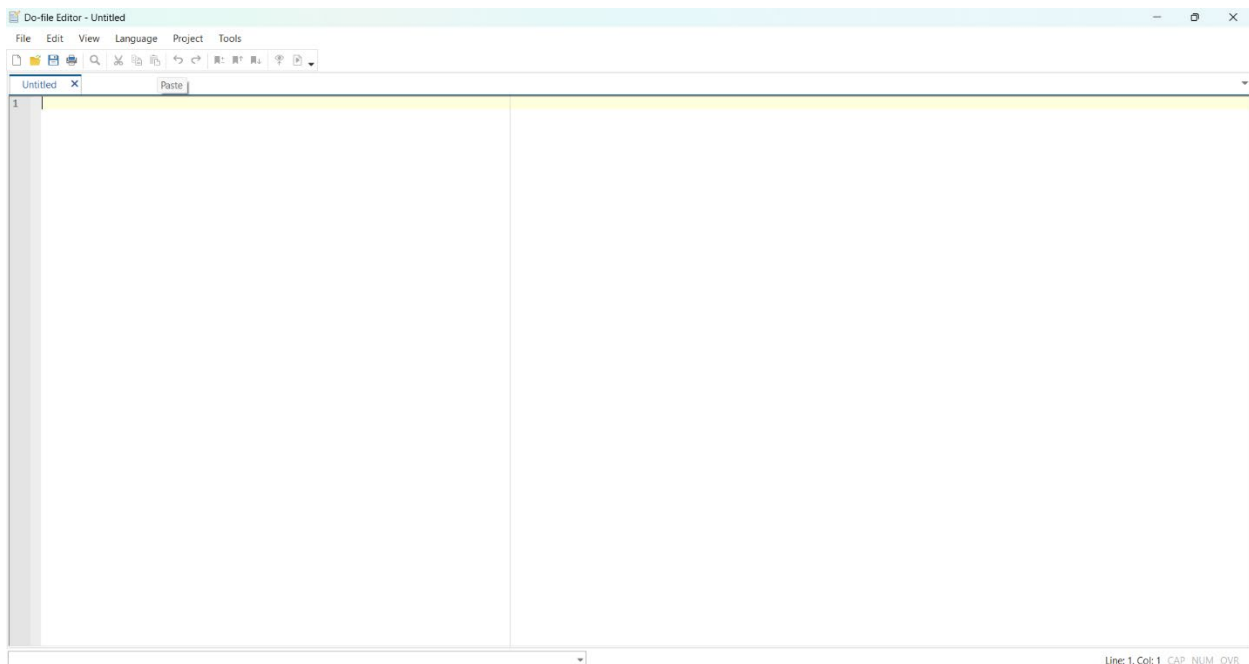


FIGURE 6b: A STATA Do-File



Specific instructions on which code to use and how to further use STATA will be given by your instructor.

An Introduction to SPSS

The goal and utility of a workbook is to offer students the opportunity to engage the concepts introduced in the textbook and practice generating and interpreting statistical techniques. Together, this is the process of ‘analysing the data’ or ‘doing statistics.’

This workbook is associated with the SPSS statistical software. SPSS uses the less common but perfectly reasonable approach of using a graphic user interface (GUI) rather than a strictly ‘code driven’ approach. This simply means that the statistical techniques will be produced from an interface window using buttons, many with drop-down options.

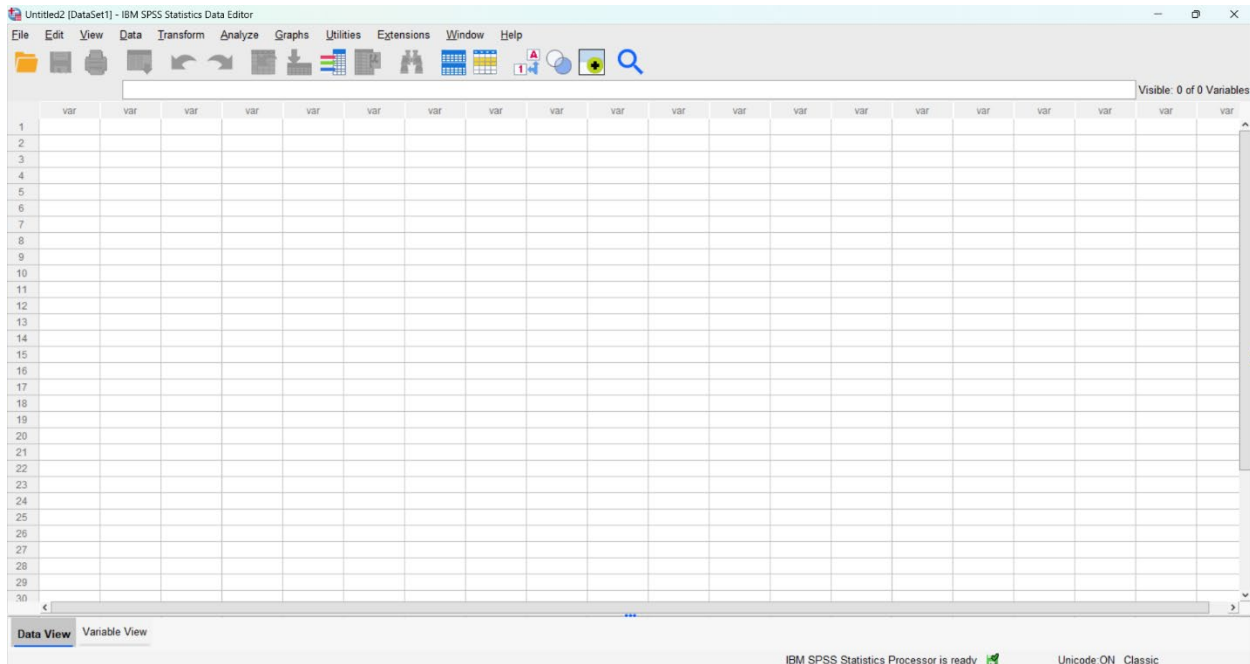
Let's introduce ourselves to the SPSS environment. Like other statistical software, SPSS is based on the $N \times K$ design where 'N' refers the number of observations and 'K' refers to the number of variables. Most often, datasets order observations and variables in the form of rows (observations) and columns (variables).

FIGURE 1: N x K Design

[illegible]

For SPSS, that looks like this (Figure 2):

FIGURE 2: N x K Design in SPSS



So, for example, if we open the “DS - Levels of Measurement.sav” SPSS dataset, we get the following (Figure 3).

FIGURE 3: SPSS Displaying the ‘DS - Levels of Measurement.sav’ Dataset

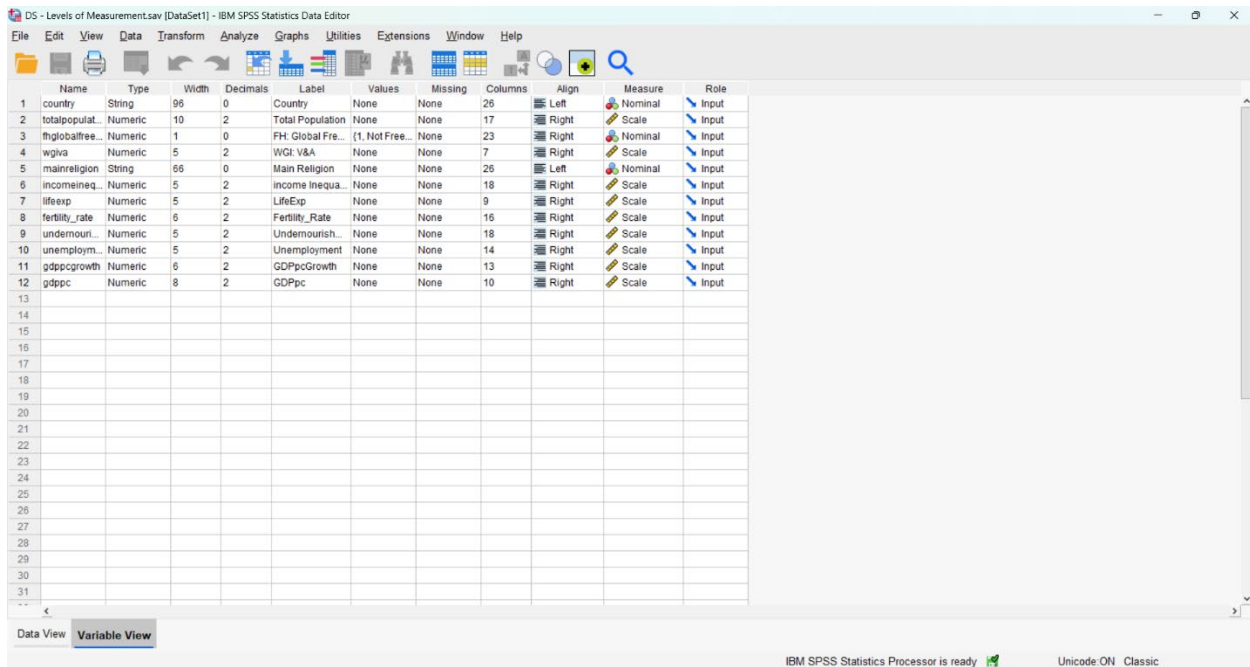
The screenshot shows the IBM SPSS Statistics Data Editor window titled 'DS - Levels of Measurement.sav [DataSet2] - IBM SPSS Statistics Data Editor'. The menu bar and toolbar are the same as in Figure 2. The main data grid displays 25 rows of data with 12 columns. The status bar at the bottom indicates 'IBM SPSS Statistics Processor is ready' and 'Unicode: ON Classic'.

	country	totalpopulation	thglobalfreedomcores	wgkva	mainreligion	incomeinequality	lifeexp	fertility_rate	undemourishment	unemployment	gdpccgrowth	sdpy
1	Viet Nam	97338.59	1	-1.38	Folk Religions	35.70	75.40	27.40	6.40	2.30	2.00	2785
2	Democratic Republic of the Con...	89561.41	1	-1.37	Christianity	42.10	60.70	121.00	-	4.60	-2.30	556
3	Egypt	102334.40	1	-1.43	Islam	31.50	72.00	52.40	4.70	10.50	1.50	3547
4	Ethiopia	114963.60	1	-1.30	Christianity	35.00	66.60	63.40	19.70	2.80	3.40	936
5	China	1439324.00	1	-1.61	Unaffiliated Religions	38.50	76.90	7.70	2.50	5.00	2.00	10500
6	Iran	83992.95	1	-1.37	Islam	42.00	76.70	40.80	4.70	11.00	-6.30	2282
7	Thailand	69799.98	1	-.83	Buddhism	34.90	77.20	44.50	9.30	1.00	-6.30	7185
8	Myanmar	54409.79	1	-.84	Buddhism	30.70	67.10	27.90	14.10	1.80	-10.60	1400
9	Turkey	84339.07	1	-.81	Islam	41.90	77.70	24.60	2.50	13.90	.70	8538
10	Russian Federation	145934.50	1	-1.10	Christianity	37.50	73.10	18.50	2.50	5.70	-2.80	10126
11	Nigeria	206139.59	2	-.41	Christianity	35.10	54.70	103.60	12.60	9.00	-4.30	2097
12	Mexico	128932.80	2	.02	Christianity	45.40	75.10	58.50	7.10	4.70	-9.20	8346
13	Philippines	109581.10	2	.03	Christianity	42.30	71.20	55.40	14.50	3.40	-10.80	3298
14	Tanzania	59734.21	2	-.50	Christianity	40.50	65.50	115.50	25.00	2.20	-.90	1078
15	India	1380004.00	2	.29	Hinduism	35.70	69.70	11.00	14.00	7.10	-8.90	1900
16	Indonesia	273523.59	2	.16	Islam	38.20	71.70	46.40	9.00	4.10	-3.10	3865
17	Bangladesh	164689.41	2	-.72	Islam	32.40	72.60	81.70	13.00	5.30	1.40	1968
18	Italy	60461.83	3	.97	Christianity	35.90	83.20	5.00	2.50	9.30	-8.60	31676
19	Germany	83783.95	3	1.34	Christianity	31.90	80.90	7.60	2.50	4.30	-5.10	45723
20	France	65273.51	3	1.14	Christianity	32.40	82.60	4.70	2.50	8.60	-8.30	38625
21	United Kingdom	67886.01	3	1.26	Christianity	35.10	81.20	11.90	2.50	4.30	-10.30	40284
22	Brazil	212559.41	3	.34	Christianity	53.40	75.90	56.70	2.50	13.70	-4.70	6796
23	South Africa	59308.69	3	.67	Christianity	63.00	64.10	67.80	5.70	28.70	-8.10	5090
24	Japan	126476.50	3	.96	Unaffiliated Religions	32.90	84.40	3.50	2.50	3.00	50	40113
25	United States of America	331002.69	3	.97	Christianity	41.40	78.80	17.30	2.50	8.30	-3.80	63543

Political Analysis Workbook

Note that – as we can see in the bottom left-hand corner, this is the ‘Data View’. The Data View is just that, a display of the data so that we can view them. If we click on ‘Variable View’, we get a look at the variables themselves and their characteristics (see Figure 4).

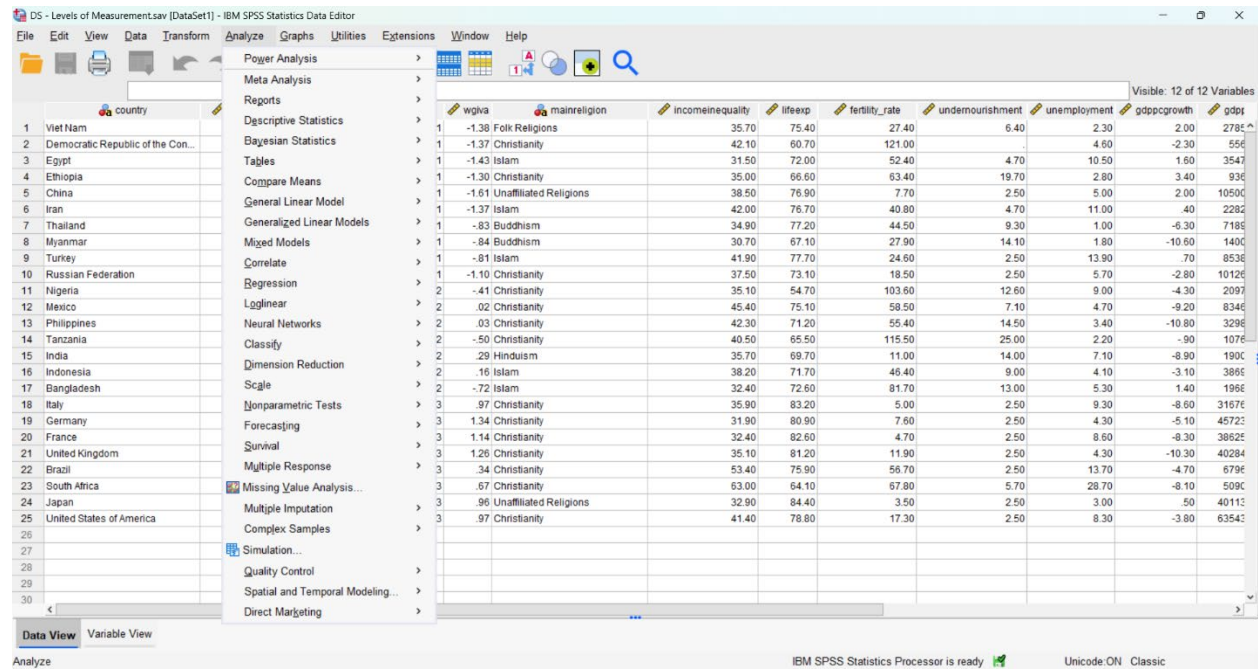
FIGURE 4: Variable View in SPSS (using the ‘DS - Levels of Measurement.sav’ Dataset)



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	country	String	96	0	Country	None	None	26	Left	Nominal	Input
2	totalpopulat...	Numeric	10	2	Total Population	None	None	17	Right	Scale	Input
3	fhglobalfree...	Numeric	1	0	FH: Global Fre...	{1, Not Free...	None	23	Right	Nominal	Input
4	wgiwa	Numeric	5	2	WGI: V&A	None	None	7	Right	Scale	Input
5	mainreligion	String	66	0	Main Religion	None	None	26	Left	Nominal	Input
6	incomeineq...	Numeric	5	2	Income Inequa...	None	None	18	Right	Scale	Input
7	lifeexp	Numeric	5	2	LifeExp	None	None	9	Right	Scale	Input
8	fertility_rate	Numeric	6	2	Fertility_Rate	None	None	16	Right	Scale	Input
9	undernourish...	Numeric	5	2	Undernourish...	None	None	18	Right	Scale	Input
10	unemployment...	Numeric	5	2	Unemployment	None	None	14	Right	Scale	Input
11	gdppcgrowth	Numeric	6	2	GDPPcGrowth	None	None	13	Right	Scale	Input
12	gdppc	Numeric	8	2	GDPPc	None	None	10	Right	Scale	Input
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											

Going back to the Data View, we will be using the ‘Analyze’ button to produce nearly all of the techniques found in *Political Analysis* (see Figure 5).

FIGURE 5: The ‘Analyze’ Button and Its Functionalities



In this drop-down menu, we can see many of the techniques and approaches with which we are familiar in our study of statistics: Descriptive Statistics, Compare Means (a.k.a.: ‘Means Comparison’), Correlate, Regression (Linear, Binary, Ordinal, Multinomial!). As I am sure you also noticed, there are a lot of other buttons for a lot of other statistical techniques. I remind you that what you have learned serves as an essential foundation for myriad other types of statistical techniques (often suited for fewer, more highly specified analytical needs). So, no fear.

An Introduction to R

The goal and utility of a workbook is to offer students the opportunity to engage the concepts introduced in the textbook and practice generating and interpreting statistical techniques. Together, this is the process of ‘analysing the data’ or ‘doing statistics.’

This workbook is associated with the statistical software R. R is a common software used by many different fields for statistical analysis, data management and visualization. Being open-source and free, R is steadily extended by practioners to conduct high quality and reproducible research.

One relevant distinction should be made right at the beginning. While R is the actual programming language, Rstudio is the Graphical User Interface (GUI) for R. RStudio offers a more straightforward and intuitive application of the programming language and is now commonly used by people to code in R. In the following, R and Rstudio might be used interchangeable which is why it is important to keep the difference in mind.

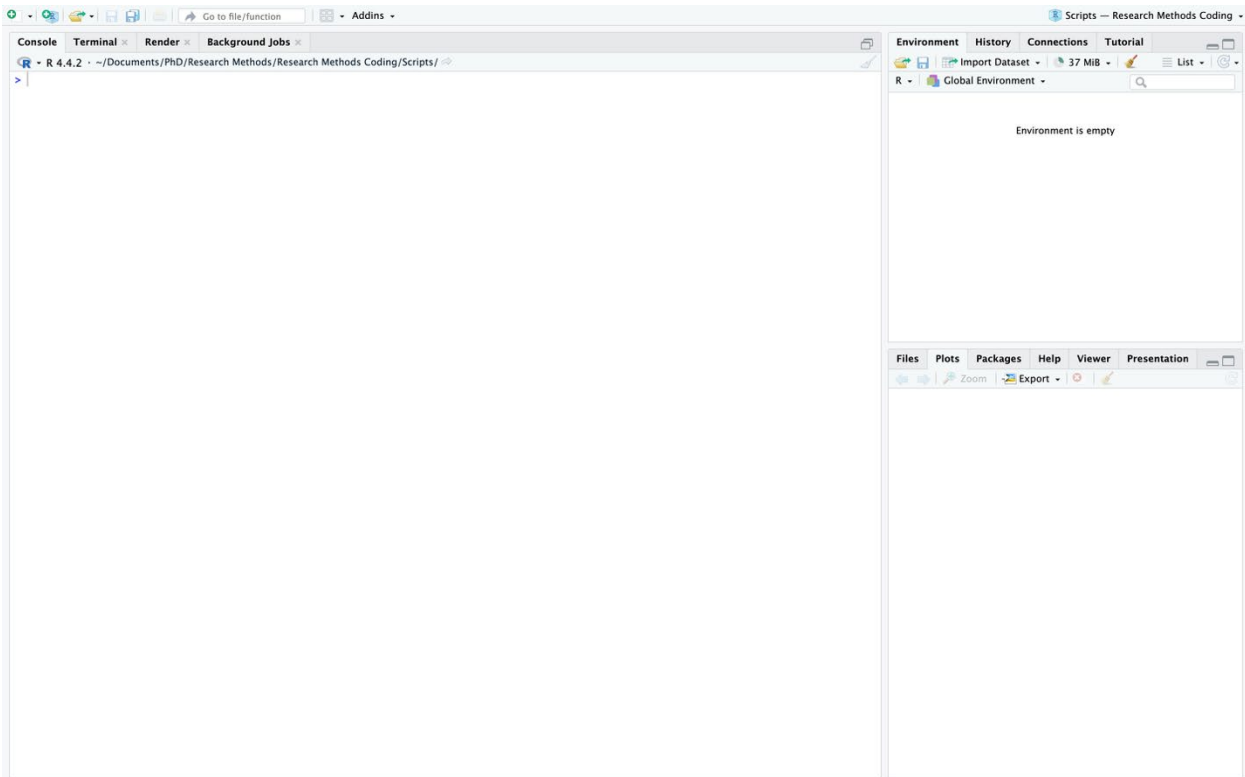
Let's introduce ourselves to the R environment. Like other statistical software, R can be based on the $N \times K$ design where 'N' refers the number of observations and 'K' refers to the number of variables. Most often, datasets order observations and variables in the form of rows (observations) and columns (variables).

FIGURE 1: N x K Design

[illegible]

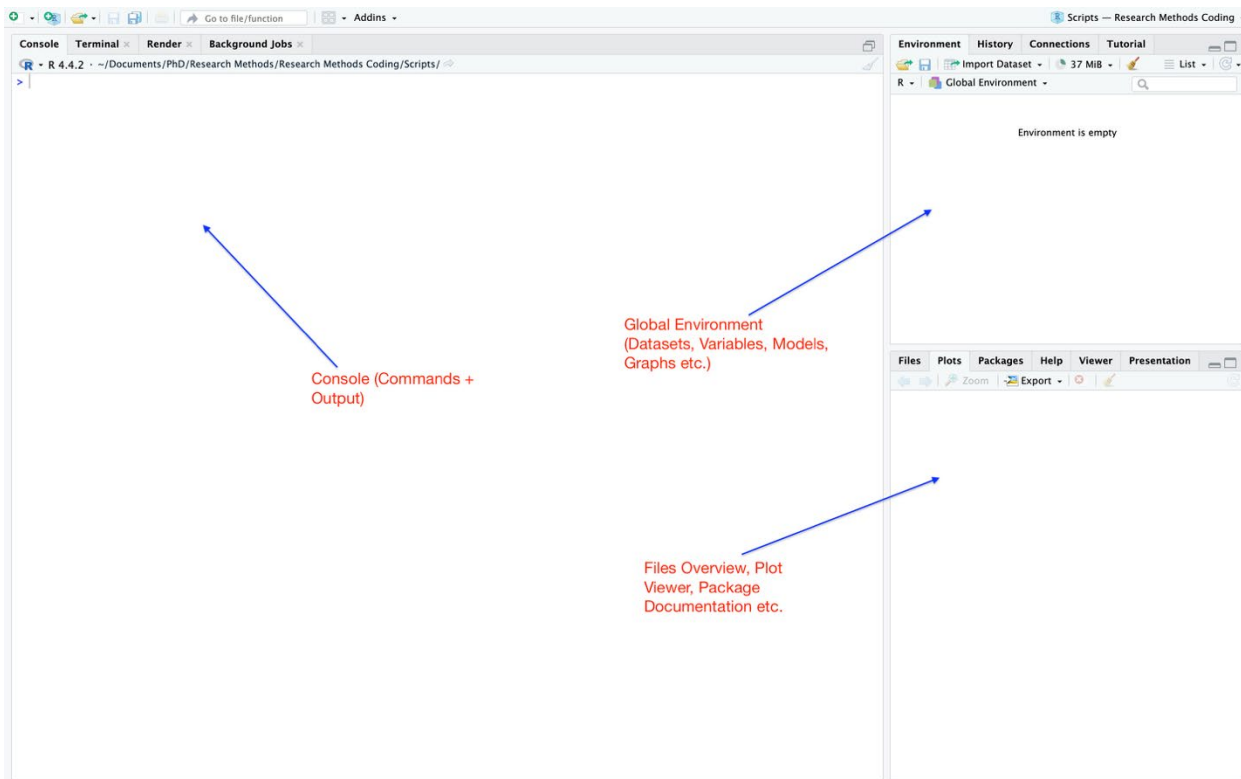
For R, however, the opening window looks a little different (Figure 2a):

FIGURE 2a The Opening Window in R



However, this opening window is designed to maximize our working with our data. In Figure 2b, when we open a dataset (as we will below), there is a place for us to put our commands – that ask R to do certain statistical techniques – and view the variables from our dataset (see Figure 2b).

FIGURE 2b The Opening Window in R: Commands and Variables



In order to get the the N x K Desgin, we need to ask R simply to show it to us. One way is to click on the button indicated in Figure 3a. Alternatively, you could also use the command 'View()' by inserting the name of the dataset in the brackets of the 'View()' command in the console (See Figure 3b). Note however, that the N x K design is not the only way in which R can represent data. While it is useful to open datasets and their variables in the N x K design, there are other types in which R stores data and allows to inspect it. This is what makes R a powerful tool for analysing data of various types.

Political Analysis Workbook

FIGURE 3a: Getting to the N x K Design in R

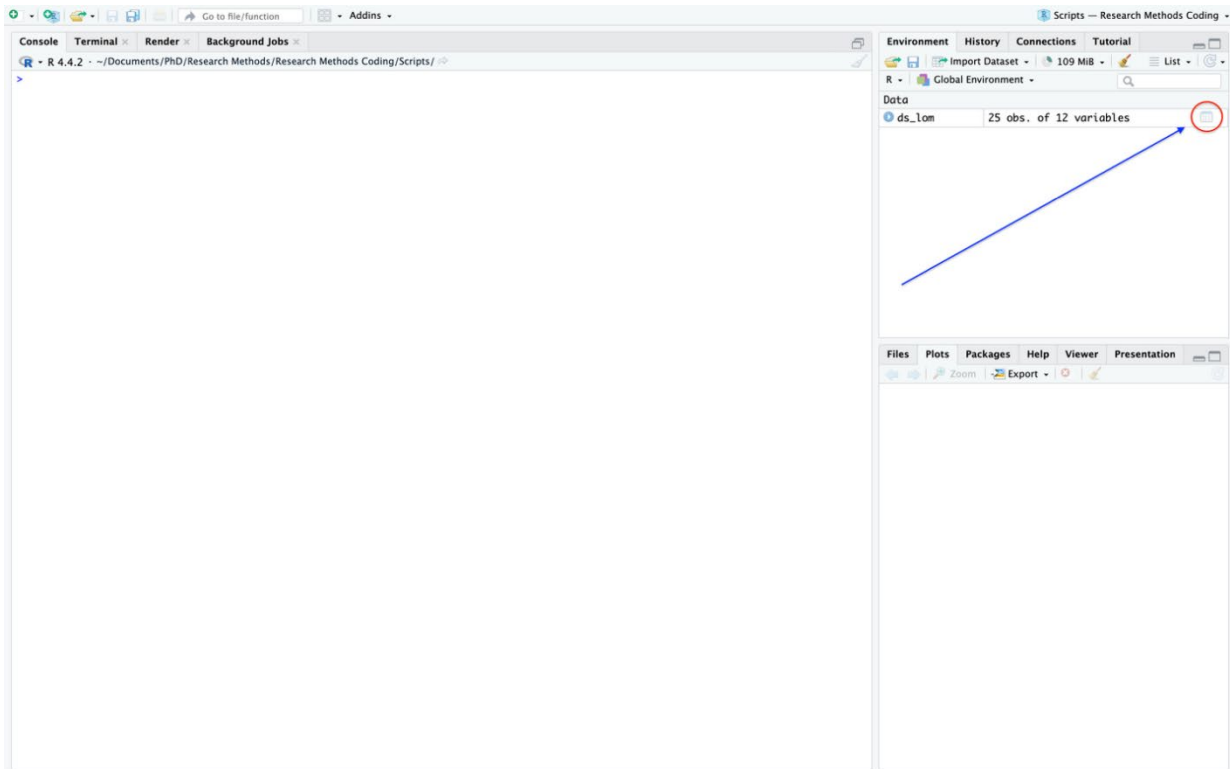
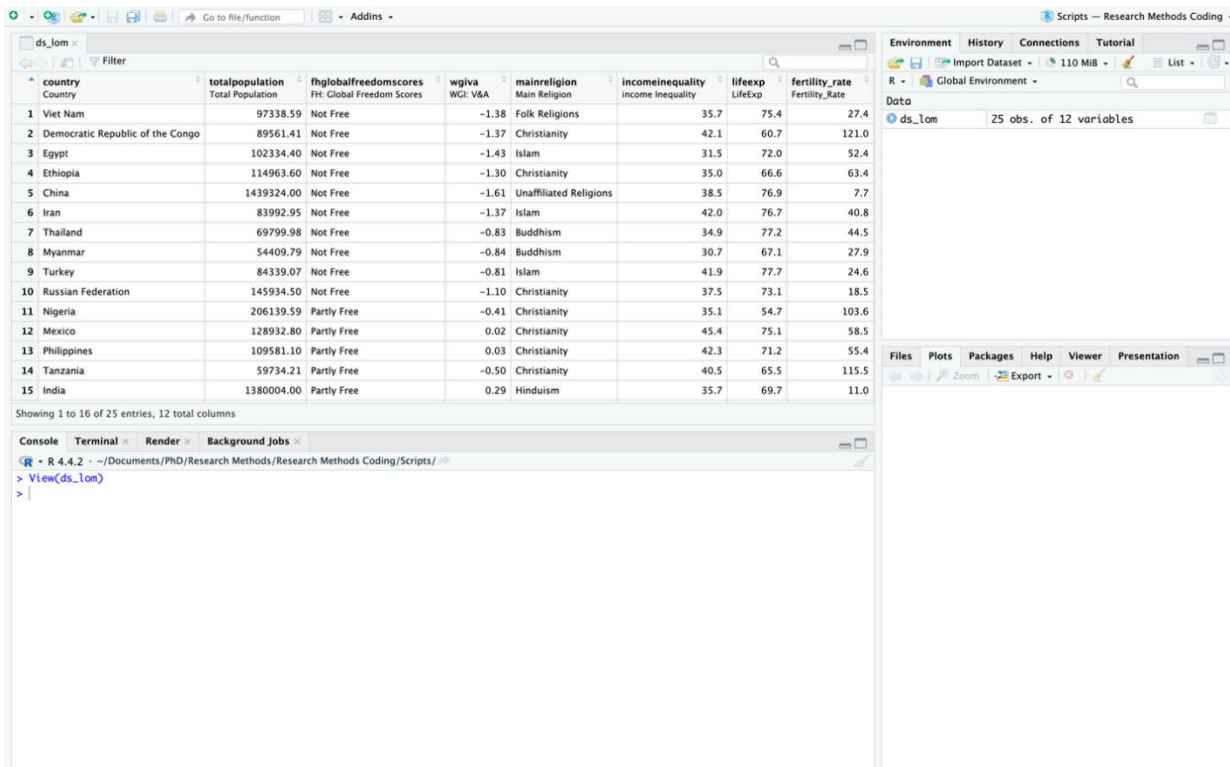
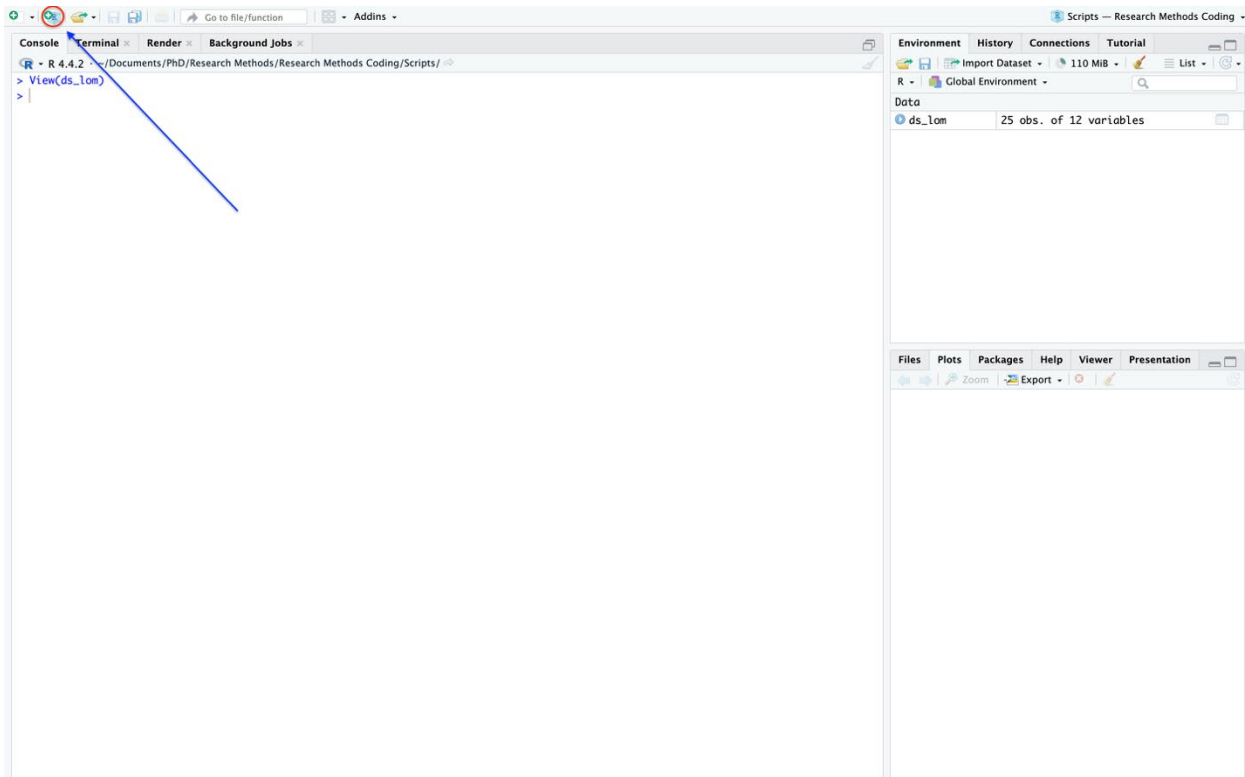


FIGURE 3b: N x K Design in R



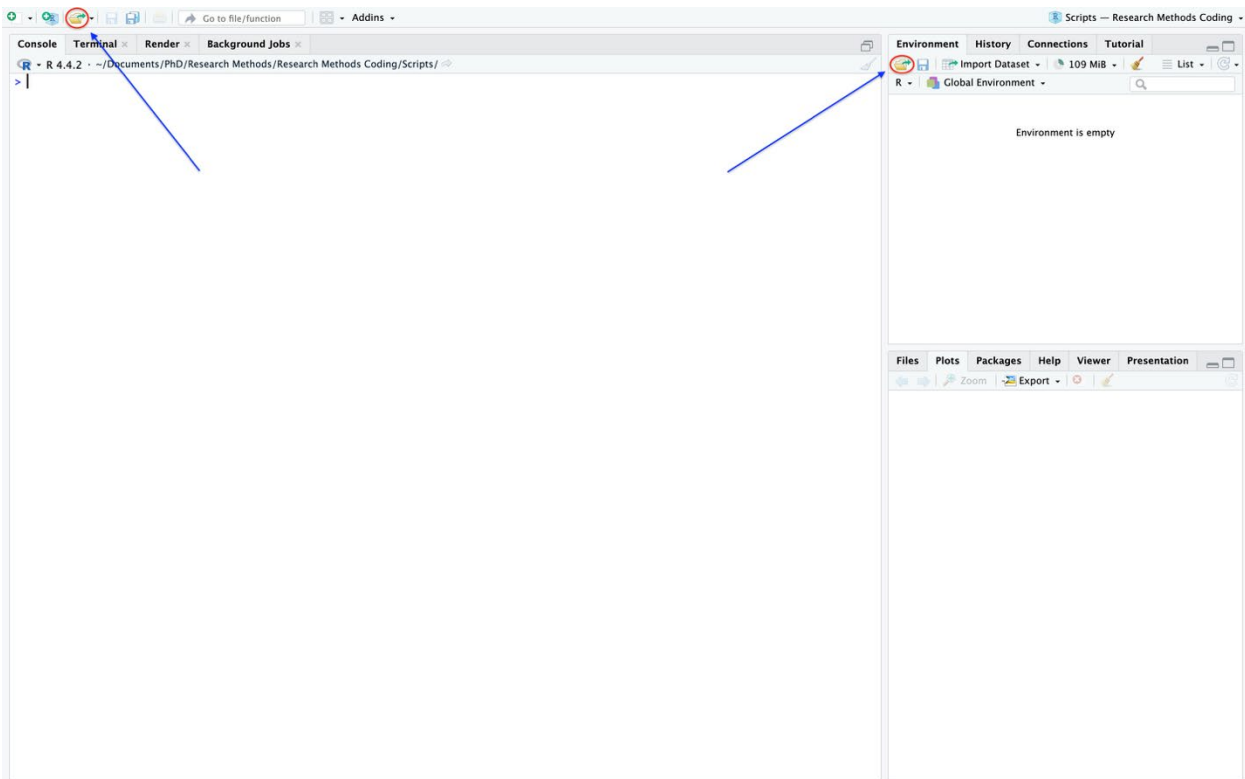
In order to return to the opening window, you can either close the tab that has been opened. Or you choose to leave it open in case you would like to return it at a later point. The tab-system within R allows you to work on different windows and projects at the same time. For instance, here we are working in one R environment but you could easily open a new project in parallel by creating a new R environment (See Figure 4).

Figure 4: Opening a new R environment



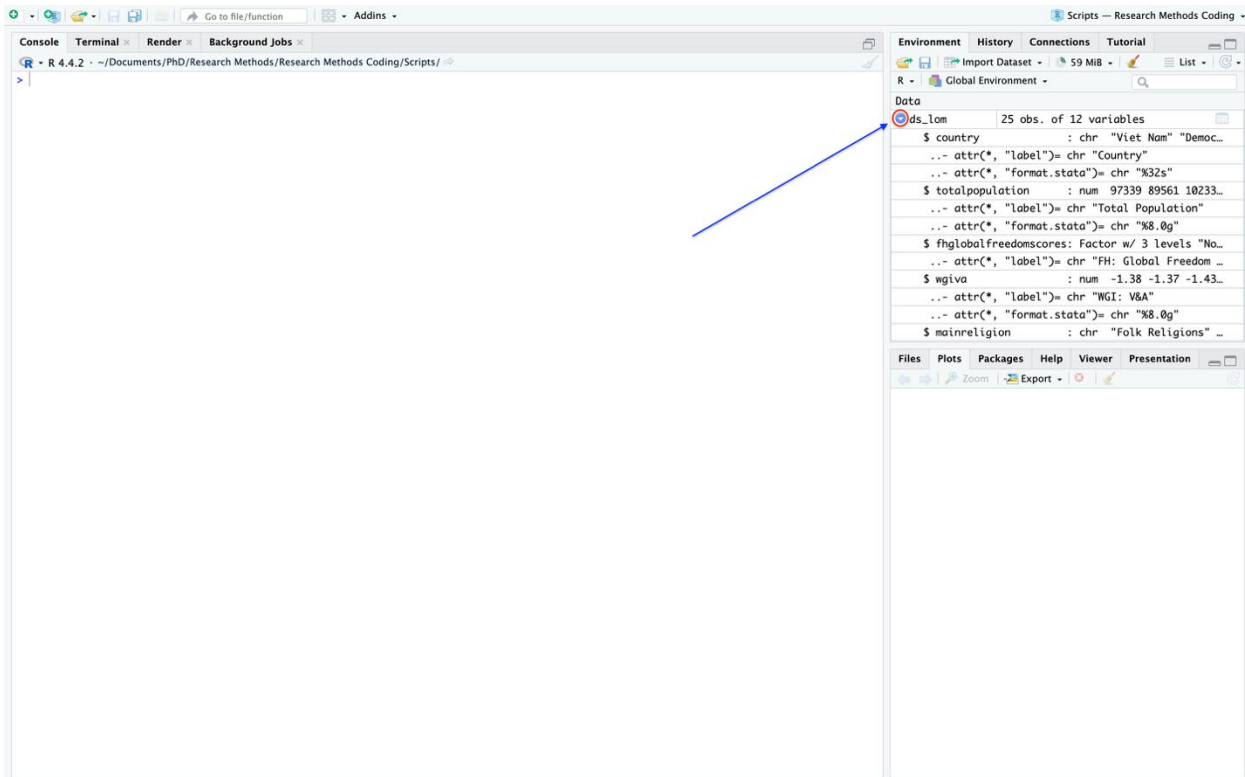
Let's open the "DS - Levels of Measurement.RData" dataset. To do so, we use one of the 'open files' icons (Figure 5a) and get the following (Figure 5b).

FIGURE 5a: Using ‘Open an existing file’ to get the ‘DS - Levels of Measurement.dta’ Dataset



Here we can see a number of things. Recall Figure 2b. At the top right, we have the list of objects in our R working environment. In order to have a look at the variables in the dataset in the Environment window, we can click on the arrow positioned in front of the dataset.

Figure 6: Viewing variables in R in the Environment window.



The console window (on the left side) awaits our instructions. For a fuller look at our data, we can return to the N x K Design page (recall Figure 3a above).

We can begin to consider the techniques and approaches with which we are familiar in our study of statistics: Descriptive Statistics, Means Comparison, Correlation, Multiple Regression, and Logistic Regression, *inter alia*. We will put our instructions to R in the console (see Figure 2b).

Or, if we want to do a lot of them, we can use what is called a 'R script'. A R Script is essentially a text function that allows us to input, change, and update our commands and – crucially – save them. I recommend using the R Script as a 'all in one' place to save the code for the various techniques we are learning. You can write notes to yourself and remind yourself of what each one does and how to interpret them. You can open a R Script this way (Figure 7a) and it looks like this (Figure 7b). From Figure 7a you can already see that there are not only R-scripts you can create but numerous different files. For now, let's concentrate on the R script and not be put off by the many options.

FIGURE 7a: Opening a R Script

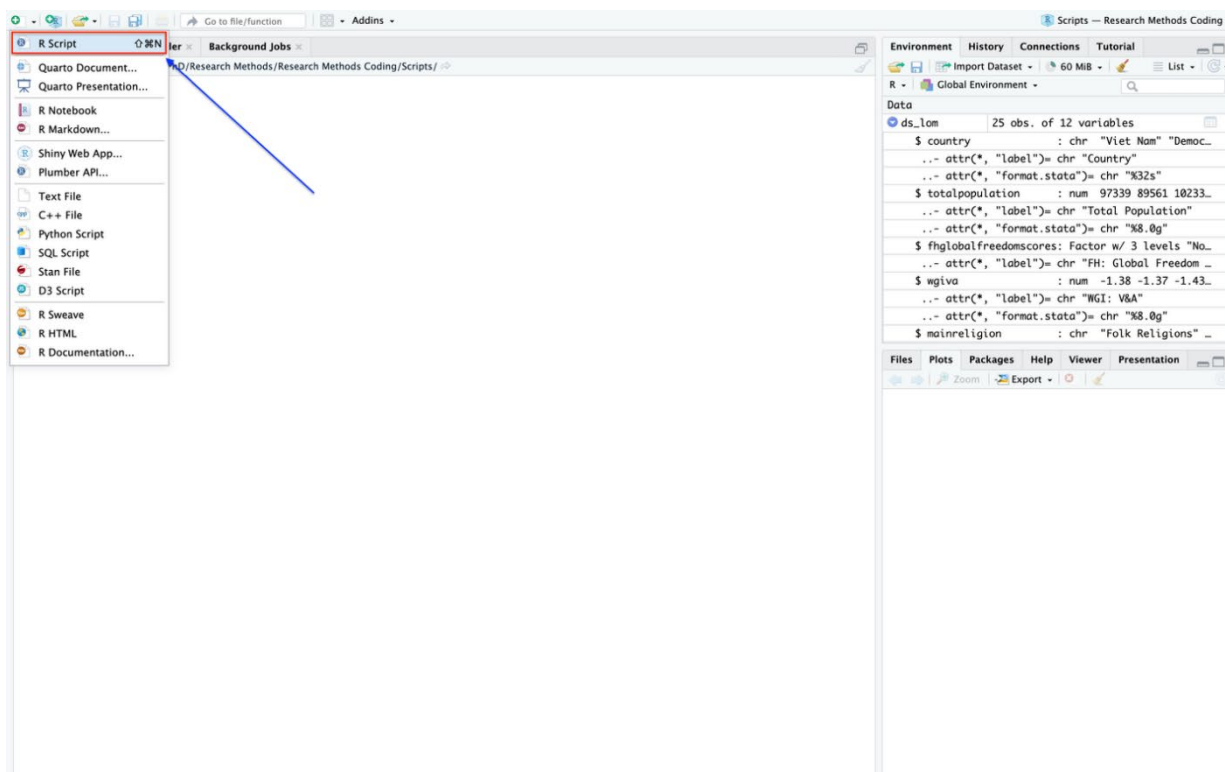
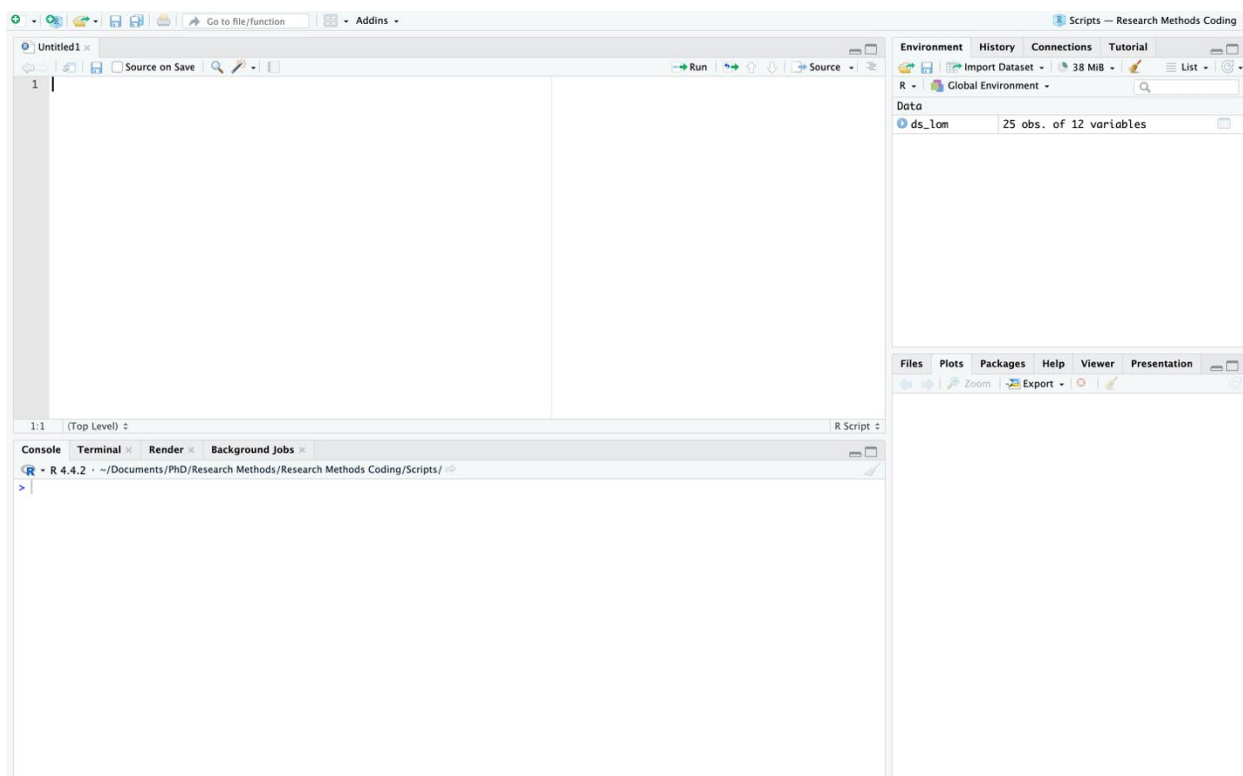


FIGURE 7b: A R-Script



The commands that you use in R often stem from packages that you need to load before you can apply them. The only exemption are those functions which R is equipped with as standard. This is called 'Base-R' or the 'R base package'. While these basic functions already allow you to apply a series of different operations, you will need to install and load other packages to exploit the full potential of the R programming language. To install packages, you can type the following command (replacing *'packagename'* with the name of the corresponding package you would like to use):

```
install.packages("packagename")
```

Afterwards you need to load the package to be able to use it by typing:

```
library(packagename)
```

At the beginning of each chapter you will find a list of packages which you need to complete the exercises. Make sure to have installed and loaded the packages as shown above.

As mentioned at the beginning, R is continuously developed by the community of R users. Therefore, it is important to regularly update the packages which you have installed to ensure a proper functioning. This can be done by clicking on the 'Update' button in the 'Packages' tab in the bottom right (See Figure 8a). This will open a window, in which you can select all or some packages you would like to update (See Figure 8b).

Figure 8a: Update packages in R

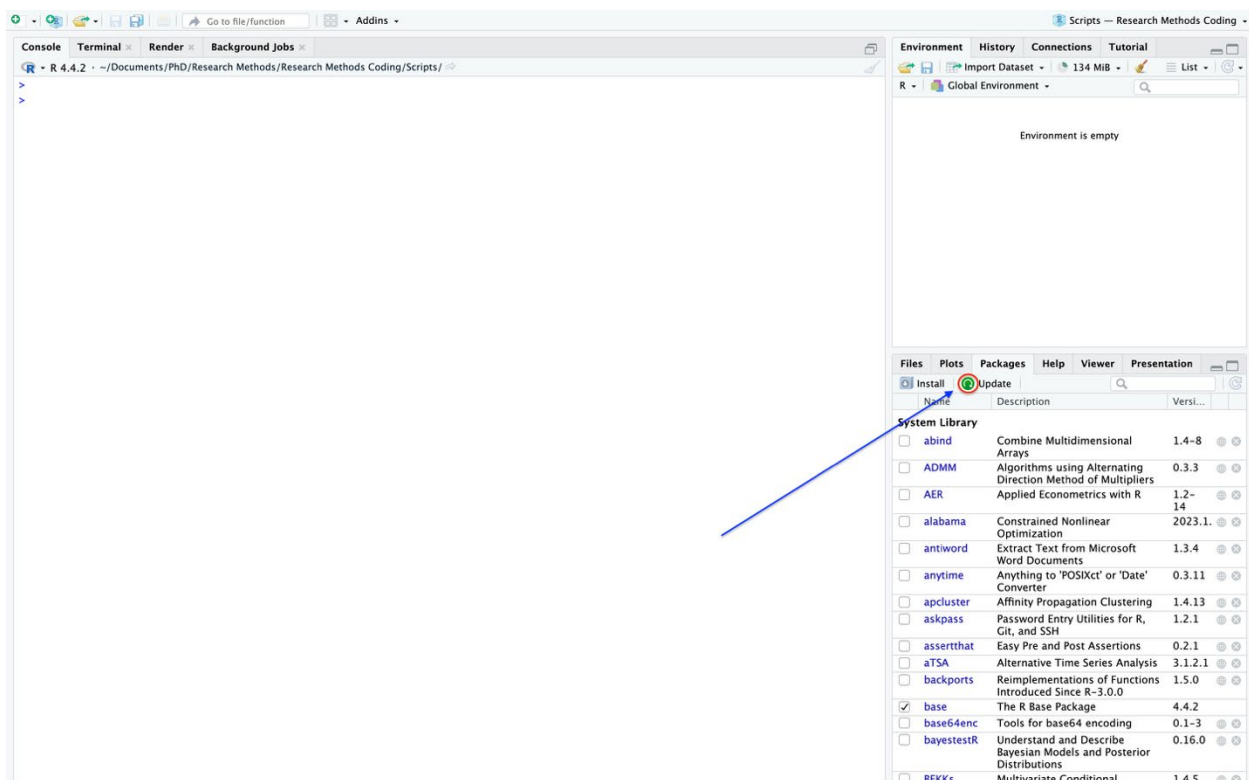
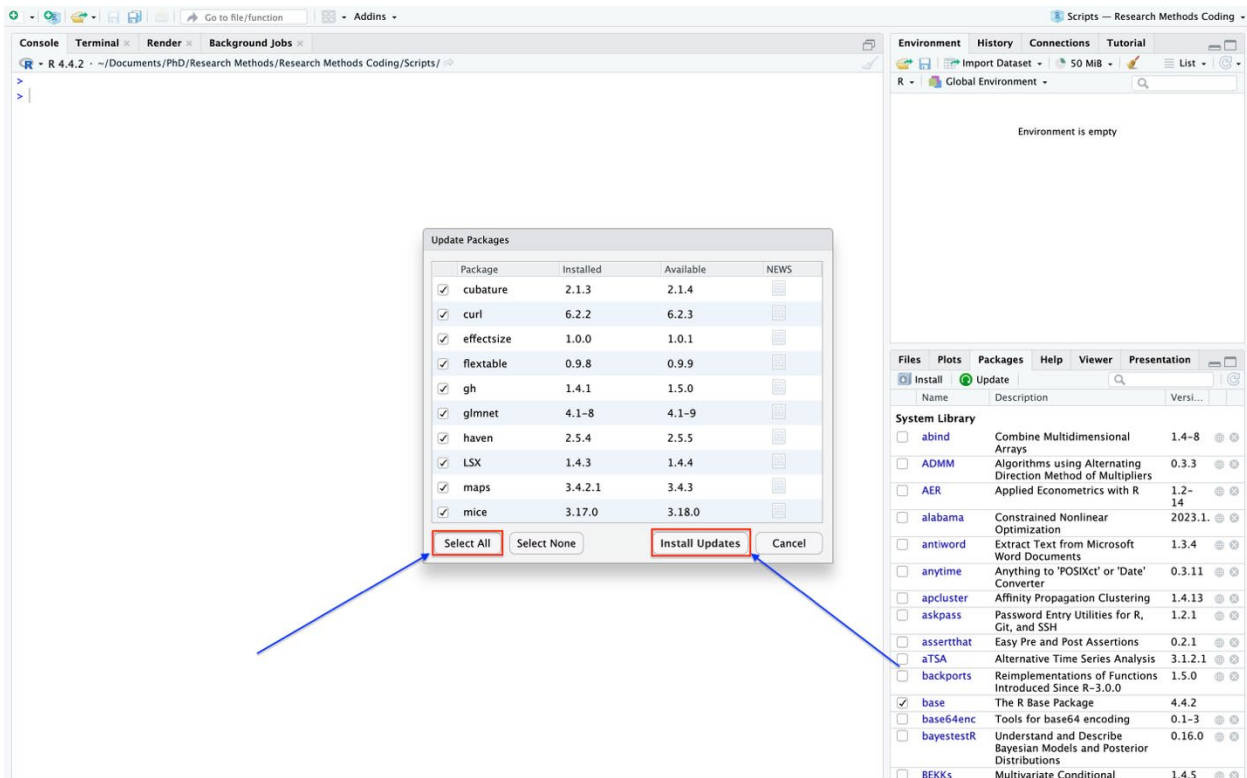


Figure 8b: Choose packages to update in R

Packages are being ongoingly optimized and updated. This also means that the code presented in this workbook might be outdated at one point. In fact, being deprecated is the fate of most code. If deprecated functions are used, R then indicates this with a warning. It is helpful to pay attention to this when implementing the code and to generally keep in mind that the code presented here is one of many ways to arrive at the right solution. Specific instructions on which code to use and how to further manipulate R will be given by your instructor.

An Introduction to Python

The goal and utility of a workbook is to offer students the opportunity to engage the concepts introduced in the textbook and practice generating and interpreting statistical techniques. Together, this is the process of ‘analysing the data’ or ‘doing statistics.’

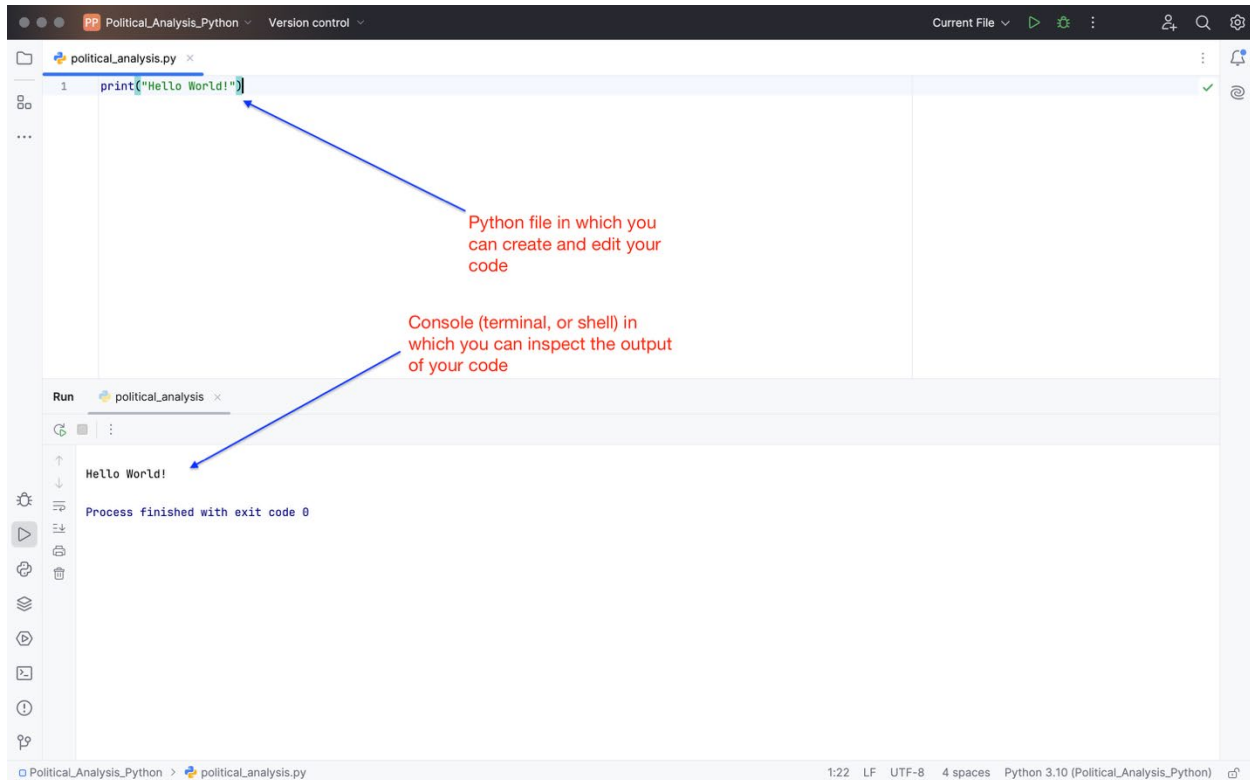
This workbook is associated with the general-purpose programming language Python. As such, it is designed to build software for a variety of domains and applications. In other words, other than Stata, SPSS, or R (although to a lesser degree), Python hasn’t been created for the sole purpose of conducting statistical analysis. It is a highly ‘code driven’ approach, you have to tell Python what to do and how to do it. However, when used in the right way, it can also be very efficient tool for statistical analysis. Particularly, Python integrability with other software provides you with a breadth of new possibilities to expand your toolkit.

What also distinguishes Python from other statistic software is that it is not associated with one standard editing program. There are a number of programs to code in Python, commonly referred to as Integrated Development Environment (IDE). There is no space here to present all the alternatives, but each comes with its advantages and disadvantages and should therefore be chosen carefully. Popular Python IDEs are ‘PyCharm’, ‘Spyder’, or ‘Visual Studio Code’. Alternatively, there is ‘Jupyter Notebook’ or ‘GoogleColab’, a cloud code version of Jupyter Notebook, which combines the creation and editing of Python code with more input of text.

All of these programs allow you to create, edit, and modify Python code in a so-called ‘.py file’, a text file which contains Python code. In many ways, especially when doing statistics with Python, a py file is similar is to Stata’s ‘do-file’ or Rstudio’s ‘R script’. It is essentially a text function that allows us to input, change, and update our commands and – crucially – save them. The Python file will become your ‘all in one’ place to save the code for the various techniques we are learning. You can write notes to yourself and remind yourself of what each one does and how to interpret them. This will form the basis for the application of techniques and approaches with which we are familiar in our study of statistics: Descriptive Statistics, Means Comparison, Correlation, Multiple Regression, and Logistic Regression, *inter alia*.

While IDEs differ from one another, however, all typically contain a window in which you can edit your code, subsequently run it, and view the output of the code in another window (depending on the IDE, this is called console, terminal, or shell). Figure 1 shows an example of this basic two-windows setup (Note that the window in Figure 1 is taken from ‘PyCharm’).

Figure 1: Code editor and console



There is one further relevant point which differentiates Python from other languages. There is no environment (like in R) or global storage of variables of the dataset you loaded (like in Stata). A Python file is mostly intended to run through in one go. This means that you always have to tell Python explicitly which data and which packages you will need to execute the entirety of your code. This is typically done at the beginning of every Python file.

Jupyter Notebook (as well as the web-based Google Colab) deviates from this in that it allows you to edit your code in separate cells and display the output right next to it. Therefore, it has become a useful tool for 'debugging', the process of identifying and removing errors in code. This workbook is conceptualised such that each solution is given with Python files that are designed to run through in one go. You can copy each chunk of code, insert it in your Python IDE, and run it. This ensures that the code has all the necessary statements to work while giving you the opportunity to customize the code, depending on which IDE you will end up using.

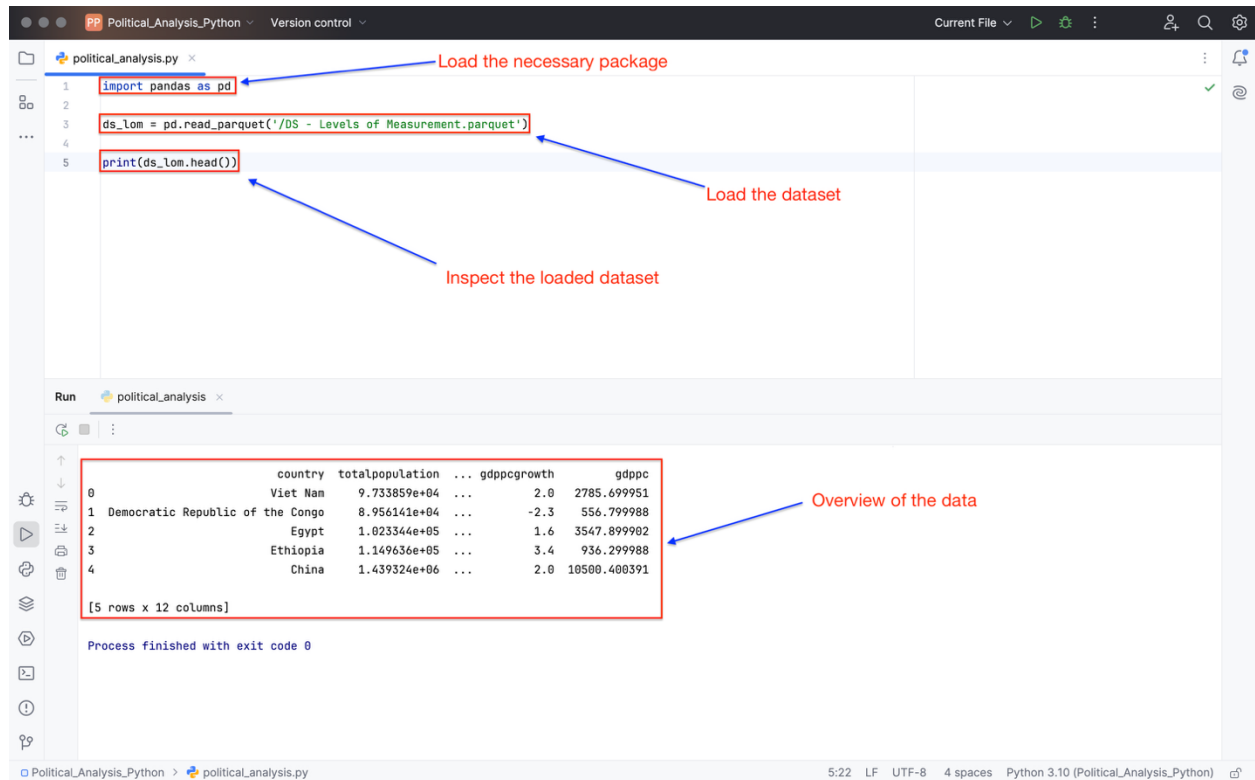
Let's introduce ourselves to the Python syntax. Like other statistical software, Python can also be based on the $N \times K$ design where 'N' refers to the number of observations and 'K' refers to the number of variables. Most often, datasets order observations and variables in the form of rows (observations) and columns (variables). However, rather than being able to view the $N \times K$ design in a separate window, one would call specific functions in order to inspect the dataset in the console (Figure 3).

Political Analysis Workbook

FIGURE 2: N x K Design

[illegible]

Figure 3: Data management with Python



There is a lot happening in Figure 3, so let's unpack this one by one. The first two lines of code are telling Python which package (In the example, pandas, a very common data manipulation library) and which dataset we want to use. For very frequently used packages, one often adds an abbreviation, like 'pd' for 'pandas' in our example. This has the advantage that when you are calling a function later in the script, you don't need to always write the full name of the package.

Python can deal with datasets of all kind of formats, such as csv or json. Throughout this workbook, datasets are stored in the so-called 'parquet' format because it allows for a straightforward and fast processing of datasets while preserving the data types we use during statistical analyses. As in Figure 3, you will often see that instead of a path, only the name of the Political analysis dataset is given. Please make sure that this string matches the path which leads to the parquet file on your local machine.

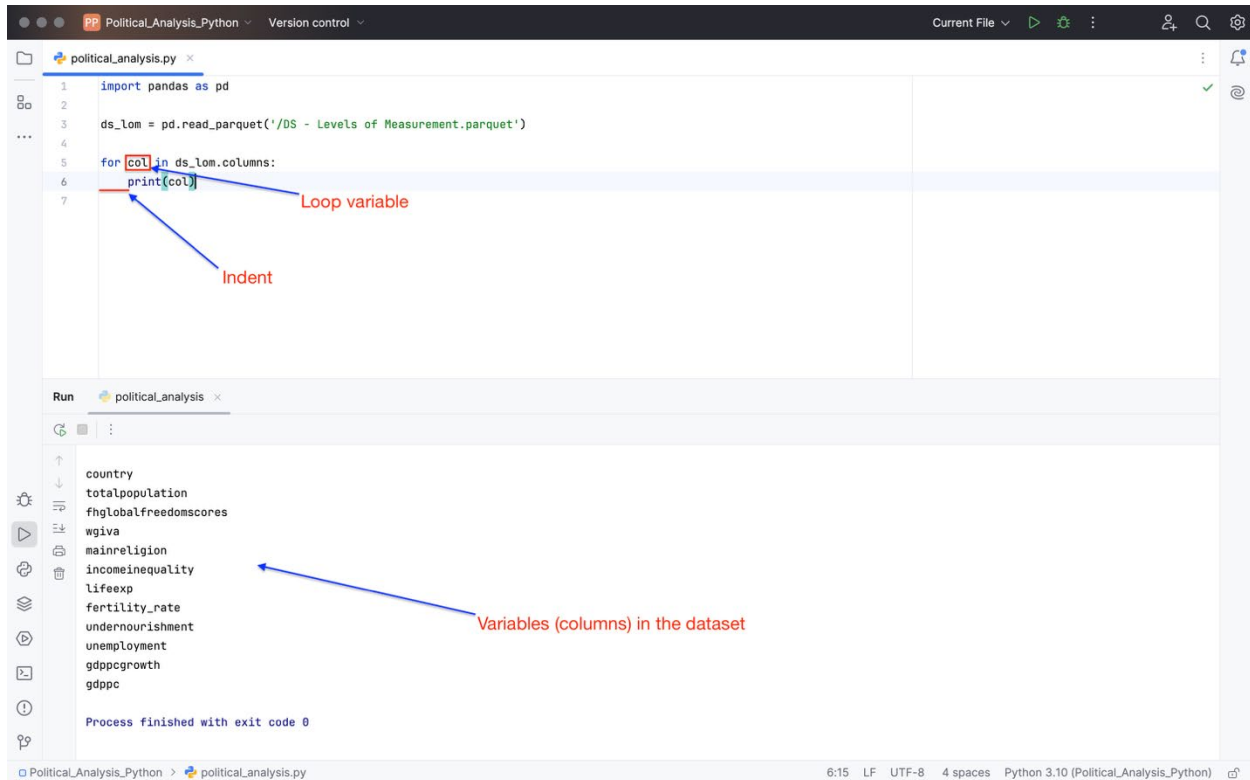
In line 3 in the py file in Figure 3, you can see that we call the function 'read_parquet' from pandas by typing 'pd.read_parquet'. This tells Python nothing more than to use the function 'read_parquet' from the pandas library which we have abbreviated with 'pd'. During this workbook, you will also see syntax like 'from *packagename* import *function*'. As you can already guess, we are telling Python here to only import one specific function from a package rather than all of its functions. As stated in the beginning, if you don't call the package or the dataset, Python will fail to execute the commands afterwards.

In the lower part of Figure 3, you can see a preview of the imported dataset in the console which we called with the 'head()' function. In addition to giving you a small preview of the data, the function also tells you how many variables (columns) and how many observations (rows) it is currently displaying (the '...' in the header line is telling you that it is leaving out some variables in the display). Due to the limited size of the console, it is rarely the case that you view the entire data set at once, but instead use certain commands to display exactly what part of the data you would like inspect. This is the peculiarity of a 'code driven' language as it demands you to have always a precise idea of what it is you would like to see. Importantly, also note that Python starts counting from 0 and not from 1, like other programs. The first observation is typically the 0th observation.

One of the most characteristic feature of the Python syntax is indentation which assumes a highly important function for the code. An indent refers to the space at the beginning of a code. Since a picture is worth more than a thousand words, let's make the use of indentation clearer with the help of another coding example. Figure 4 includes the code to print each variable (column) of our 'DS – Levels of Measurement' dataset individually.

Line 5 of the py file contains a for loop that is structured as follows: 'for *loop variable* in *data frame*:". In words, we ask Python to apply a function for each element of the specified dataset, one by one – this is what looping through an object means. This is where the indent comes into play which is positioned in front of the print() command in line 6 of the py file. The indent basically states that the command we are using – print() – is in reference to the line of code above. If the print() command would be at the beginning of the line, that is, we would omit the indent, the code would not work.

Note also that we have decided to call the loop variable col – because it is one single item (one column, more precisely) which belongs to the larger object (columns) we are looping through – but we could have given it a different name like x or y, for example. In the console you can now see the variables included in 'DS – levels of Measurement'. This is also a good point to remind you that you will be using the print() command very often because without it, nothing will be shown in the console.

Figure 4: Python syntax

This code is another example of how to engage with a dataset in Python without visualising it in a dedicated window. This is one aspect in which Python differs from the statistical software you may be used to, and you will find more examples where this is the case throughout the rest of this workbook. For instance, being a more numerically based language, Python is very sensitive to missing values. Therefore, often in the code you will see that we will have to remove missing cases before we can apply our statistical methods.

A further note on the packages used in Python. Python comes with a set of pre-installed packages which include basic operations that are frequently used. Other packages which include more specialised functions need to be installed before you can use them. Always make sure to have installed the packages that are being stated at the beginning of each chunk of code. How to install packages depends on the IDE you are using and in which environment you have installed it. However, most packages are being installed with the help of 'pip', a package management system written in Python. Here, you normally just need to use the line of code: 'pip install *packagename*'. I recommend that you take a look at the Python package index 'PyPi' from which pip also installs packages. This is a very useful resource to obtain instruction on how to install a package and to familiarize yourself with the documentation of packages.

Python packages are being developed and continuously updated by researchers, open source communities, or companies and they usually maintain their packages on 'PyPi'. Therefore, it is important that you update your packages from time to time to ensure that you are not using deprecated versions. But don't worry, if this is the case, Python prints a warning message to the console to inform you about this.

Political Analysis Workbook

Be aware that after updating the packages, in some cases, there can be conflicts between versions of different packages or between package versions and the version of Python you are using. While these conflicts are being resolved with time by developers, in some cases, you will need to install an older version of a package to ensure that your code is working. It is also relevant to keep in mind which version of Python you are using. For example, this workbook is written in Python 3.10.

As a general purpose programming language, Python can be used for a myriad of applications. You can scrape a webpage, edit videos, program a computer game – or do statistics. This workbook offers you one solution for carrying out statistical analyses in Python, but there are many alternative ways to achieve the same results. Specific instructions on which code to use and how to further use Python will be given by your instructor.

QUESTIONS FOR CHAPTER 1: The Scientific Method and Statistics

In chapter 1, we learned that science doesn't describe *what* we study, it describes *how* we study. To make the most of statistics, it is imperative that we provide just the right material, in just the right way, in order to get an answer that can be applied to our problem. If we have designed our analysis to answer a well-developed question and have the most appropriate data for our question, statistics can provide a result in which we can have a great deal of confidence. Inasmuch as each of these steps leading up to the use of statistics break down or is poorly implemented, we will have less confidence in our ability to provide an answer our question. The implementation is the scientific method. That is, formally, statistics is an empirical technique that allows us to collect, analyse, interpret, and present data. Statistics cannot however be the arbiter of the question of 'truth.'

1. What does it mean, 'Politics are loaded (or value-laden)'? Why is this a potential problem for the scientific study of politics?
2. Isn't it strange that so many people are studying politics and still there is plenty on which to do research. Why are there still many questions to answer in Political Science?
3. The following article started off a long line of research into why those who voted for parties that won elections reported being more satisfied with the way that democracy works in their country than those who voted for parties that lost. While this may seem initially intuitive, the literature has grown to understand that it is mostly 'losing' that lowers satisfaction rather than 'winning' raising satisfaction. A troubling discovery as many democracies are showing signs of waning popularity and legitimacy.

Anderson, Christopher and Christine Guillory. 1997. "Political Institutions and Satisfaction with Democracy: A Cross-National Analysis of Consensus and Majoritarian Systems," *American Political Science Review* 91(1): 66-81. <https://doi.org/10.2307/2952259>

- a. What is the research question being studied by the authors? Is the research topic normative or non-normative? Explain your answer.
 - b. Does the article provide empirical evidence? You do not need to understand this evidence, but list the page numbers of the article where such evidence is presented and explained.
 - c. Is the knowledge presented in the article transmissible? In other words, do(es) the author(s) provide enough information to allow one to replicate the study? List the page numbers and sections of the articles where this information is provided. Also indicate whether additional information on the methodology is presented in footnotes (list footnotes) or appendices (list page numbers).
 - d. Take a closer look at the end of the article and address these questions. Do the authors include a normative discussion in the conclusion or the final few paragraphs? Do the authors seek to explain a larger set of events than the one under investigation? If so, briefly summarize this discussion.
4. Explain the difference between descriptive statistics and inferential statistics

Political Analysis Workbook

5. What is the goal of Principle Number 1: The scientific method requires a transparent and replicable description of the research design and analysis?
6. What is the goal of Principle Number 2: The scientific method attempts to identify, isolate and explain the relationship under investigation.
7. What is the goal of Principle Number 3: The scientific method seeks to derive and make appropriate inferences from the results of our research.
8. How does the correct replication of results improve our confidence in the results of an analysis?
9. Statistics and quantitative methods more broadly are referred to as positivist approaches. What does this mean? And how does it differ from other approaches to understand?
10. Connect Principle Number 1 to one of the three key analytical tools of statistics – description, inference and control.
12. Connect Principle Number 2 to one of the three key analytical tools of statistics – description, inference and control.
13. Connect Principle Number 3 to one of the three key analytical tools of statistics – description, inference and control.
14. Generally, epistemology is a question of about how we formulate questions and answer those questions. Science is one of those ways. What are two others mentioned in the chapter?
15. If science is not necessarily in a laboratory conducted by a serious person (inevitably wearing glasses) in a white lab coat, are there other ways of being scientific but not statistical?
16. What is meant by the phrase: ‘it is important that we can distinguish between a statement that is true and a statement that we have a great deal of confidence in?’
17. Generalization is part of the process of theory building, determining how far what we have learned can be extended to explain a class of events. What, then, is inference?
18. If we observe in our analysis that parties that focus on one issue tend to do more poorly in elections than parties that give attention to many issues, give an example of descriptive inference.
19. If we explain what we have observed in our analysis that parties that focus on one issue tend to do more poorly in elections than parties that give attention to many issues, saying that the narrow focus gives voters the impression that the party is over-attentive to only one of the many challenges voters face, give an example of causal inference.
20. Science is said to be objective. Science is not objective; scientists only try to be. Why is trying to be objective important?
21. When we say that science is a method. What does this mean?
22. Science is a method to be less wrong. What does this mean?
23. Uncertainty is not a limitation. Knowing how much we are uncertain tells us how certain we can be. How does this approach map onto probabilistic confirmation?
24. Prediction and causation are different in important ways. What is one important way?

Political Analysis Workbook

- Causation – as a theoretical construct – is grounded in reality.
- Causation – unlike prediction – can be proven.
- Prediction – unlike causation – can be proven.
- Prediction – as an empirical construct – is grounded in reality.

25. What's wrong with common sense?

QUESTIONS FOR CHAPTER 2: Theory and Hypotheses

Theories are not right or wrong. Theories are weak or strong. We use theories to explain how the phenomena in which we are interested are related. We also don't test theories directly but rather through the use of declarative statements of the expected theoretical relationship between the independent variable and dependent variable (i.e.: hypotheses). The ability to use theories and hypotheses correctly lays the foundation on which we build our research design and where statistics can do their most impressive work.

1. The theory of Priming states that media shape the standards or criteria by which citizens evaluate political leaders and candidates. Which of the following could be subjects in a study of the theory of Priming? *Click all that apply*
 - a. Citizens of the UK
 - b. Citizens of Italy
 - c. Candidates for election
 - d. Members of Parliament
 - e. Taylor Swift
 - f. Ordinary people
 - g. None of these
 - h. All of these
2. Good theories are falsifiable. How does this shape traditional deductive research?
3. In the field of political communication, there are a number of exciting theories. However, one well-known and long-standing theory has demonstrated itself to be strong: Agenda Setting (McCombs and Shaw, 1972). The theory states that the issues most prominent in the news are also the issues most important in public opinion. Or – and, yes, you can use this at parties to impress others – ‘Media do not tell people what to think, but what to think about.’ Which of the following represents a potential hypothesis of the agenda-setting theory? *Click all that apply*
 - a. Surveys show that Germans’ support for or opposition to Covid restrictions is related to whether they watch television or listen to the radio.
 - b. In a year, BBC viewers will post (or re-post) on social media topics that reflect the topics given the largest amount of time in the BBC news.
 - c. According to many who read *Financial Times*, Christopher Wylie, Cambridge Analytica whistleblower, is either a patriot or a traitor.
 - d. There have been several interviews with top corporate and financial representatives on CNN International who have stated clearly and consistently that the economic crises in 2008 and again now are related to governmental interference in the market. Many viewers agree.

- f. Over a one-month period in [your country], greater diversity in the headline news stories of each night broadcast will correspond to less convergence in the opinions of viewers on the most important problem facing the country.
4. One might extend the agenda-setting theory to explain electoral behaviour such that news media carry increasing number of mentions of election – making people think about it and ultimately act. One could see if in countries in which the media carry increasing number of mentions of election, people think about it more and ultimately act (i.e. vote). Which of the following are possible hypotheses that allow him to test this theory? *Click all that apply*
- a. Voting in the upcoming election will be higher in counties in which there is more news coverage of the issues pertaining to the election
 - b. Men, who sit in front of the television in the evening, are more likely to vote than women
 - c. People who did not vote in the last election will not vote in this election, even if they read a newspaper
 - d. People who watch television channels that carry information about the election are more likely to vote than those who watch television channels that carry little or no information about the elections
 - e. If people had watched more news in the last election, voting levels would have been higher
5. In the Agenda-Setting theory, what is theorized to cause what?
- a. Do 'the media' form viewers' opinions?
 - b. Does the issues most reported in the news change viewers' values and morals?
 - c. Does the news cause viewers' opinions to change?
 - d. Does the prominence of issues in the news cause viewers' reported salience given to these same issues?
6. The cultivation theory states that media tend to offer uniform and relatively consensual versions of social reality and their audiences are 'acculturated' accordingly to this agreed upon consensual, centrist view. Which of these correctly falsifies cultivation theory? *Click all that apply*
- a. Respondents to a survey in Italy were separated by whether they watched RAI, the public channel or the Berlusconi-owned MediaSet. These groups showed strong, contrasting views on politics.
 - b. In the early 1950s, television offered few channels and it could be expected that most viewers would be watching very much the same small offering of programs and news. Despite this, people surveyed at that time displayed a surprising wide variety of views about the probability of success of the US war in Korea.
 - c. Over the past 30 years, sitcoms and entertainment television have offered a formulaic set of storylines and characters. Entertainment television consumption worldwide over the same period has increased every single year.

- d. Liberals also read the Economist magazine.
- e. Pop stars are nearly perfectly interchangeable. The only thing that distinguishes them are their inevitable – and big selling – paths to self-destruction.

7. Create four plausible *hypotheses* from the following list of concepts:

- a. Gender
- b. Age
- c. Voter turnout
- d. Party identification (an individual's preference for one party over the other)
- e. Congressional election
- f. Political Interest
- g. Income

Recall that a hypothesis is a declarative statement of the *expected theoretical relationship* between the independent variable and the dependent variable. Produce two plausible hypotheses for the following theories:

- 8. A simplified version of the median voter theory sets out that in a two-party system, candidates and parties will converge on the voter 'in the middle' to secure the majority.
- 9. The congruence postulate sets out the countries that have relatively high levels of democratic culture will have higher performing democratic institutions and thus be more stable, long lived and legitimate.
- 10. The democratic peace theory suggests that liberal democracies will try to avoid coming into open conflict with other democracies.
- 11. The Iron Law of Oligarchy is a subset of the broader set of elite theories that states political – and to some extent economic – elites not only hold a disproportionate amount of power in democracies and do so despite (ostensibly democratic) elections.
- 12. Duverger's law sets out that majoritarian – first past the post – electoral systems are more likely to produce strictly two-party systems than proportional representational system, which are more likely to produce multi-party systems.
- 13. Unlike more recent developments that focus on individuals' attitudes and psychological profiles, the Michigan model of voting bases its explanation on socio-demographic, socio-economic and socialization factors.
- 14. The principle-policy paradox describes the lack of synchronicity between public opinion ('support for the *principle*') and support for a *policy* related to that principle.
- 15. The spiral of silence identifies the increasing reticence of people to express opinions that are less commonly or widely held.

16. World-systems theory states that the unit of analysis or central subject of study for international relations is not necessarily the nation state but rather transnational relationship among core and *semi-peripheral*, and peripheral countries.
17. Relative deprivation – the perceived exclusion from social, economic or political goods relative to others – serves as an individual source of social mobilization and even political violence.
18. What is intended by the author's statement, '*This, my dear reader, is the central load bearing column for the scientific method*'?
19. If the concept of parsimony is to maximize leverage, how does this improve theory?

Read this abstract of a proposed research project and answer the following questions:

Abstract: What factors explain the amount of national newspaper coverage that political candidates for the Congress of Deputies receive in Spain? We hypothesize that the gender of a candidate will be an important variable for predicting the level of news coverage, with male candidates receiving more attention than female candidates. In addition to gender, we examine the effect of the candidate's party identification and the candidate's previous political experience. We test our hypothesis by analyzing the amount of El País coverage received by political candidates in the upcoming election.

20. What is the dependent variable in this study?
21. What is the independent variable that is of *primary* interest to the researchers?
22. What hypothesis is proposed for the relationship of primary interest to the researchers?
23. What are the *other* independent variables in this study?
24. What is the unit of analysis in this study?
25. Recall Figure 2.1 A Model of the Role of Theory from the textbook. Explain how deduction connects a theory to potential data.
26. Recall Figure 2.1 A Model of the Role of Theory from the textbook. Explain how induction connects observed data to theory.

Recall that a hypothesis is a declarative statement of the *expected theoretical relationship* between the independent variable and the dependent variable. It should be general, non-normative that are plausible, directional, consistent with the data and testable. What are potential problems with the following hypotheses?

27. Income inequality and taxes are related.
28. When the economy is performing well, the incumbent president should be re-elected.
29. Using data on elections, individuals with growing incomes are more likely to vote for conservative parties.
30. Countries that are more democratic have taller than average populations.
31. You are investigating the causes of genocide. Identify three independent variables and specify how they are related to your dependent variable, 'the likelihood of genocide'.

Political Analysis Workbook

32. You are interested in studying the consequences of the drop in journalistic freedom in Western democracies. Develop three hypotheses you might investigate.

QUESTIONS FOR CHAPTER 3: Data and Variables

Data is not your concept. Data do not belong in your title. What you are examining is the focus of the project, not the evidence used to draw conclusions and inferences. Like theory and hypotheses in the previous chapter, the thrust of this is not only the importance of conceptualization and operationalization but the impact of both will have on what one will find.

Ultimately, the variables that we choose to represent the actual thing in reality that we can't really measure and the calculated nature of their relationships each remove us from the actual relationships about which we would like to be able to say something. Be wary and sceptical of clean, easy fitting results. It is easy to be critical of poor performing results - but less so when results align with what you are interested in demonstrating with your analysis. Recall, we are wrong, we are only trying to minimize how much so.

1. Robert Putman popularized the concept of 'social capital' in his books *Making Democracy Work* (1993) and *Bowling Alone* (2000). At some point, the definition included four conceptual elements: (1) trust in people and institutions, (2) norms of reciprocity, (3) networks and (4) membership in voluntary associations. Many researchers tried various empirical approaches but ultimately none were able to achieve a measure of all 4 simultaneously. A common - if greatly diminished - approach was the use of the simple survey question on whether the respondent trusted others in his\her society ('horizontal trust' or 'social trust'). This was aggregated to the national level to compare countries' various levels of 'social capital'. This actually worked quite well. Given the 4 original dimensions and the resulting single question used, briefly discuss what may have been lost and what may have been retained.

Let's look at this article and answer the following questions: Granato, Jim, Ronald Inglehart and David Leblang. 1996. "The Effect of Cultural Values on Economic Development: Theory, Hypotheses, and Some Empirical Tests," *American Journal of Political Science* 40(3): 606-631.

2. What is the unit of analysis for the research?
3. What is the central theory being tested?
4. What hypotheses does the article test?
5. What are the dependent and independent variables?
6. How do the authors conceptualize culture?

Three studies are described below. One of these describes a potential reliability problem and the other two potential validity problem. Indicate what is the MAJOR potential problem for each underlined measure.

7. We would expect that citizens who have negative attitudes towards feminists would be supportive of right-wing candidates for prime minister. A researcher measured attitudes towards women and found no relationship with attitudes towards right-wing candidates for prime minister.
8. A few months before the 2024 European Parliamentary election, the BBC news produced a documentary about the members of the EU parliament who had shown the least support for the

positions of Roberta Metsola, current president of the EP. As their indicator of support for the positions of Roberta Metsola, they used the representatives' votes on one issue – support for taking the side of Ukraine.

9. A study finds that during a one-week period, 40% of Europeans changed their opinions on taxing corporate windfalls.

10. Which of the following are data?

- a. In this class, I collect the name of the party everyone voted for in the last parliamentary election.
- b. Rhianna has sold more albums than The Beatles.
- c. The average final mark for this class last year was 23 (out of 30).
- d. The final marks for all the students from last year's class.
- e. Every day, my friend records the length (in centimeters) of column space on the front page of The Times given to non-economic issues.
- f. Most of my Facebook friends post political messages on their walls at least once a week.

Choose the best answer for the following sentences from the word bank below.

Independent	Induction	Variable	Normative
Dependent	Deduction	Theory	Empirical
Concept	Hypothesis	Validity	Operationalization
Sample	Population	Systematic measurement error	Non-systematic measurement error

- 11. Variables that are thought to be influenced, affected or caused by some other variable are ____ variables
- 12. Observing (dis-)similarities of actions or events and formulating a generalization to cover them is the process of ____.
- 13. ____: A concept which is assigned a number of quantitative values or qualitative categories.
- 14. Abstract, generalized statements linking relationship in a unified and explanatory structure constitute a ____.
- 15. Variables that influence the value of other variables are ____ variables.
- 16. ____: The process of selecting observable phenomena to represent abstract concepts.
- 17. Statements that explicitly or implicitly incorporate evaluations of 'good' or 'bad' are ____ statements.
- 18. ____: A declarative statement, meant for testing data, that depicts the expected relationships between phenomena.

Political Analysis Workbook

19. ____ consists of applying previous generalizations to some specific situation to arrive at a conclusion.
20. ____: An analytical term that is used to abstract, summarize or refer to a particular type of event or behaviour.
21. A ____ is the entire set of what you wish to draw conclusions about.
22. A ____ is a subset of units in the population of interest.
23. ____ is a deviation in our data collection that affects all of the observations in more or less the same way.
24. ____ are measurements perturbed unpredictably.

Identify the dependent variable (DV), independent variable (IV) and unit of analysis (UA) in the following hypotheses.

25. Subnational regions with higher median incomes spend a larger proportion of their budgets on education than subnational regions with lower median incomes.
26. Corporations that receive national subsidies enjoy higher profit margins.
27. Countries with unequal distributions are less likely to be democracies than nations with more equal income distributions.
28. Individuals with higher levels of education are more likely to vote.
29. Explain how units of analysis and levels of analysis are related.

QUESTIONS FOR CHAPTER 4: Research Design and the Scientific Study of Politics

This chapter is more of an overview of the quantitative approaches in Political Science that reflect the 'classic' training. Research Design is difficult and complicated work. This chapter has provided a skeleton on ideas about what it entails. Consider the treatment in the chapter as a brief introduction to an important and complex topic.

While specific methods and tools vary between disciplines (while potentially a wild weekend, what would a Political Scientist do with the Hubble Space Telescope?), the aims are to adhere as closely as possible to the core characteristics of the scientific method. Hence, our attention to them here. While the practical implementation of the scientific method varies by discipline (Astrophysicists don't survey stars for their preferred political party), each are discipline-specific means to achieve the rigorous evaluation of observations to test and improve our knowledge of the processes of the world. The overall goal is to provide a highlights reel of various research approaches in Political Science.

1. Social scientists often refer to and rank approaches by the concepts of 'internal validity' and 'external validity' when they discuss research designs. What do these terms mean? What are the relative advantages and disadvantages of experimental and non-experimental research designs with regard to validity issues?

Read the following article and answer the subsequent questions:

Downes, James, Matthew Loveless and Andrew Lam. 2021. "The Looming Refugee Crisis in the EU: Right-Wing Party Competition and Strategic Positioning," *Journal of Common Market Studies* 59(5): 1103–1123. <https://doi.org/10.1111/jcms.13167>

2. What is the research question being studied by the authors?
 - a. How have mainstream European centre right been able to offset a potential surge of far-right party contenders during a crisis of immigration?
 - b. How can far-right parties defeat centre-right parties during an immigration crisis in Europe?
 - c. Did the European Refugee crisis create an opportunity for far-right parties to exploit immigration for electoral gain?
 - d. Why do far-right parties and centre-right parties use immigration as an electoral issue?
3. What is the 'gap' the authors identify?
 - a. The literature on far-right parties often ignores 'right party' competition – that is, other right parties – as a factor in explaining their electoral success.
 - b. The literature ignores the gap between centre- and far-right parties.
 - c. The literature doesn't explain the European Refugee crisis.
 - d. The literature doesn't explain the impact of the European Refugee crisis on right party competition.

4. Do the authors seek to explain a larger set of events than the one under investigation?
 - a. Yes, they argue that the results help us understand the electoral strategies of centre-right parties to undermine challenges from far-right parties during crisis.
 - b. No, the authors are interested in only explaining the elections and countries included in the analysis.
 - c. Yes, they want to show that right party competition is the same all the time and not conditioned on periods of crisis.
 - d. Yes, the *strategic positioning* theory is applicable to left party competition as well.
5. Which of the following is a hypothesis that the article tests?
 - a. Centre-right parties that adopt more hard-line immigration positions will perform better electorally in the context of the refugee crisis.
 - b. Centre-right parties that adopt more hard-line campaign rhetoric salience will perform better electorally in the context of the refugee crisis.
 - c. Centre-right parties that adopt more hard-line stances on the EU will perform better electorally in the context of the refugee crisis.
 - d. Centre-right parties that adopt more 'accommodationist' stances on immigration will perform worse electorally in the context of the refugee crisis.
6. The authors include a normative discussion in the conclusion. Which best summarizes this discussion?
 - a. By adopting 'hard-line' ('accommodationist') strategies, centre-right might be able to weather far-right electoral challenges. However, this mainstreams such hard-line immigration positions making them both increasingly acceptable and polarizing.
 - b. Centre-right should fight far-right parties because their positions are not mainstream and too hard line.
 - c. It is polarizing for centre-right parties to contest far-right electoral challenges. Centre-right parties should let them use hard-line immigration positions but not do so themselves.
 - d. 'Hard-line' ('accommodationist') strategies are the only ones that work for centre-right to defeat far-right electoral challenges. Even though it is polarizing, this makes these attitudes more mainstream and thus ok.
7. According to the textbook, what are possible roles of a research question?
8. Apart from the role of theories in the scientific method, what does the 'theory section' of a research design do?
9. How can a literature review help the reader or audience of your work?
10. Our choice of analytical approach is informed by what elements of the research project?

To make your research question 'researchable', you very often will need to 'narrow it down' by being more specific. How would you 'narrow down' these statements into research questions?

12. There are a lot of extreme parties in Europe.

13. Economic development leads to democracy.

14. Income inequality is bad for politics.

15. Europeans distrust the EU.

16. Conceptualization and Operationalization: Income inequality

a. What does 'income inequality' mean? Which of the following best fits with the notion of 'income inequality'?

- Income inequality is the sum of individuals' assets, wealth or income.
- Income inequality is the difference between highest paid member of society and the lowest.
- Income inequality is the difference between individuals in the distribution of their assets, wealth or income.
- Income inequality is an intrinsically multidimensional concept, which captures disparities along a number of dimensions that matter for the lives of individuals and the societies where they live.

17. How are you going to measure this?

- Subtract the lowest income from the highest
- Ask citizens what they think the level of income inequality is and take the average.
- Make a queue with everyone in the country from lowest to highest income and choose the income of the person in the middle.
- With a measures of the distribution of all incomes among a country's residents

18. There is a perfect research design if you think about it enough.

- True
- False

19. Recent critics of quantitative methods have pointed out that the increasing sophistication of statistics over the past two or three decades has led to more research with less long-term value to political science. Is this just sour grapes – meaning, they dismiss the sophistication as undesirable simply because they can't do it - or is there a meaningful criticism here?

20. Experiments are no longer limited to the laboratory. Field experiments such as randomized experiments, can extend the external validity of analysis. See the following article:

Kalla, J.L. and D.E. Broockman. 2016. "Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment," *American Journal of Political Science* 60: 545–558. <https://doi.org/10.1111/ajps.12180>

Explain how they design the experiment and what they find.

21. Other examples of experiments include natural experiments in which history or nature creates an opportunity for quasi-experimental control. See the following article:

Holger Lutz Kern and Jens Hainmueller. 2009. "Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes," *Political Analysis* 17: 377–399.
<http://www.jstor.org/stable/25791984>

How do they exploit nature to conduct an experiment?

22. Women in the EU earn around 16% less per hour than men. What do you suppose 'causes' that difference? Describe the analytical procedures by which you might test your conjectures.
23. Why is the statement, 'democracy is the best form of government' so fraught with difficulty for scientific testing? Are there ways to test the validity of this statement?
24. Do we impose bias where there is none? How can we find out? Below are three different experiments. Which *one* would *best* provide a preliminary answer this question?
- a. Participants evaluate the content of a news report about the 2022 Brazilian presidential election. The news report comes from a news station commonly associated with a left-leaning news station, a right-leaning news station or a fictional TV station. Comparing randomly populated groups' viewing and subsequent responses to the report, the results suggest that individuals not only distinguish between the reputations of the media outlets as ideologically left or right which heavily influences the perceptions of bias in content.
 - b. Participants evaluate the content of a news report about the 2022 Brazilian presidential election. The news report comes from a news station with no bias or another one with bias. Comparing randomly populated groups' viewing and subsequent responses to the report, the results suggest that individuals can distinguish between media outlets and agree that there is bias in the biased news report.
 - c. Participants evaluate the content of a news report about the 2022 Brazilian presidential election. Regular watchers of the left-leaning news station watch a right-leaning news station. Similarly, regular watchers of the right-leaning news station watch a left-leaning news station. Both report more bias in the content about the election than there is in each of the respective reports.
25. If causal inference is the crucial methodological problem for social scientific research, then experiments are the ideal solution. Which of the below are counterarguments to this statement? (*check all that apply*)
- a. Experiments have trouble with 'big process' such as 'culture' as it is hard to isolate such mechanisms.
 - b. Experiments are good at direct and 'observable' mechanisms and indicators (e.g. behaviours) but less good with latent or gradual changes (i.e. learning).

- c. Because experiments must focus on singular causal tests, they often struggle to deal effectively with more complex processes.
 - d. Experiments are strong on internal validity but not on external validity. In other words, while we might observe a process in the experiment, this same process may have trouble being generalized to non-experimental settings.
 - e. Experiments – as an approach – do not resolve measurement issues
26. In the article, ‘Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes’ (2009, *Political Analysis* 17: 377–399: doi:10.1093/pan/mpp017), Holger Lutz Kern and Jens Hainmueller, study the impact of West German television on public support for the East German communist regime during the Cold War. One big problem with studying media use – particularly at that time in a Communist country – was identifying the effect of media. However, they noticed the topographical oddity that allowed some cities to receive West German over-the-air television broadcasts and others not to (such, as they note, the Dresden district). They find that ‘contrary to conventional wisdom, East Germans exposed to West German television were more satisfied with life in East Germany and more supportive of the East German regime’ (2009: 377). Aside from the counter-intuitive finding, why is this a natural experiment and how does this affect the quality of their findings?
27. One uncontroversial observation about Big Data is its focus on prediction. How is a focus on prediction different than a focus on explanation?
28. Using Mahdavi and Ishiyama’s (2020) article as an example, describe the potential advantage of Data Science to get at novel questions.
- Mahdavi, Paasha and John Ishiyama. 2020. “Dynamics of the Inner Elite in Dictatorships: Evidence from North Korea,” *Comparative Politics* 52(2): 221–240.
<https://www.jstor.org/stable/26867765>
29. What does it mean when someone describes Big Data as data-driven?

QUESTIONS FOR CHAPTER 5: The Ethics of Data Analysis

As stated and re-stated in the book, statistical – and scientific – work is largely subjective. As such the onus of understanding what it is what we are doing, what we are producing, and thus what we are advancing is on us both individually and collectively. To the extent that we practice and advocate for an ethical code is, fundamentally, support for taking one's work - and the work of their discipline - seriously. You will be challenged to make this distinction in your work and your life. How you decide to handle that challenge will say a great deal about how you view yourself, your place in the world, and how the world should ultimately work. That's heavy stuff.

1. Ethical research is crucial for the resulting research to be taken seriously. At the same time, Social Scientists must investigate phenomena that are not 'value-free'. Politics, as one example, is extremely value-laden as acts and thoughts are motivated and directed and used to change outcomes that affect others. Social scientists, one might argue, face more difficulty in conducting objectivity-driven science given the topics and units of analysis (most commonly, people and countries). How does an attention to ethical research help resolve this problem?
2. Contrast the deontological and consequentialist approaches. Are there instances in which Social Scientists might defer to the consequentialist approach?
3. It is said that ethical issues are paramount in research for the simple reason that another human is submitting him/herself to the manipulation by another person. While this is true for investigations and analysis on people, how does this apply to other units of analysis, such as countries or ethnic groups or parties?
4. Protecting the subjects of an investigation is called beneficence. However, if we are interested in studying the effect of tension and fear, this can limit the study of situations in which respondents express or feel some tension or even fear. Even though such situations can be generated as hypothetical situations, is it right to do so? What are the limits to the potential for harm?
5. Explain the importance of anonymity (right to privacy) for producing reliable results.
6. Justify beneficence as a cornerstone of ethical research
7. Justify respect for persons as a cornerstone of ethical research
8. Justify justice as a cornerstone of ethical research
9. Why is this statement not correct: Informed consent is a verbal agreement between the research and the participant.
10. What is the difference between anonymity and confidentiality?
11. Even though informed consent is a concept familiar to many social investigators, the standard practice is fairly recent. How does the following example demonstrate how the lack of informed consent violates ethical research?

The Obedience Studies. In the 1960s, Prof. Stanley Milgram asked participants to give electric shocks to another person – whom they could not see – for wrong answers. What he did not reveal to the participants is that the shocks were not real despite the ‘respondents’ yelling, crying, screaming and in some extreme cases, failing to respond at all (as if they were unable). He wanted to see that, if in the presence of an ‘authority figure’ (the experimenter, in a white lab coat), they would continue to administer the shocks to the increasingly distressed respondents (actors who volunteered to be part of the experiment) under the direction of authority figure. Participants were not informed about the nature of the experiment nor debriefed about its falsity afterward.

12. As above, how does the following example demonstrate how the lack of informed consent violates ethical research?

The Tearoom Trade Study: Laud Humphreys, then a PhD student in sociology observed men engaged in homosexual sex in public restrooms in St. Louis, Missouri, in the 1960s. He never revealed himself as conducting research and subsequently went to the men’s homes (he had recorded the license plates from the parking lot) and interviewed them under the guise of a general health study.

13. Look at the website for the General Data Protection Regulation (GDPR): <https://gdpr.eu/what-is-gdpr/>. What are the *privacy rights of individuals* articulated by this legislation?
14. Look again at the website for the General Data Protection Regulation (GDPR): <https://gdpr.eu/what-is-gdpr/>. Among the *Data Protection Principles*, which ones refer directly to the researcher’s responsibility with the data?
15. Universities are increasingly evaluating academics’ plans for research before the research is conducted to mitigate any ethical concerns before they might arise. Why are there diverse members from different disciplines?
16. There are many reasons to insist the published research is replicable. Why is this important?
17. What are the widely dispersed and generally agreed upon university of disciplinary penalties for research that is not replicable?

Which of the following are potential *conflicts of interest*, which are *ethical dilemmas*, and which are *neither*?

18. The outputs of the research will determine if a new experimental laboratory will be allocated to the researcher’s department.
19. Publishing in one of the top five journals in one’s discipline results in a cash reward to you from your department.
20. You have been told that one of your favourite podcasts is considering you for an interview. You also hear that the interviewers are lukewarm about your participation as your research only somewhat aligns with the interviewers’ viewpoints. You send them a ‘new research project you have just started’ that aligns closely with the views of the interviewers although you have neither planned nor conducted this research.

21. You are asked to review an article for a journal and even though the author's name has been removed you know it to be a colleague who is struggling with publishing so that she/he can be promoted.
22. There is a rise in the number and variety of 'pay to publish' journals and although their prestige is very low (if any), you consider encouraging a student entering the job market (in which a publication can really help) to publish in the journal using all their (meagre) research money. A job for your student – in this tough market – would be a feather in your cap around the department.
23. In some disciplines, winning grants is not the free and open competition of the best ideas. You know that being able to 'become known' to reviewers involved in the evaluation process can be helpful in making sure your grant application is not immediately eliminated. Someone mentions a way to identify the reviewers and this would allow you to meet them at conferences and other professional opportunities.
24. The collection of our internet-produced data are being used by platforms to improve 'user experience' and increase 'engagement'. This is novel as the product of many websites (certainly free ones) is you – or more specifically – your data. Your visits and behaviours and choices online are increasingly used to develop a profile of you so as to better respond to your internet use. That is, what you engage online is increasingly less neutral to you. Your environment is becoming a motivated space in which decisions and choices are limited rather than expanded as a function of your profile. To what extent is this massive and passive data collection of your data are potentially violating the cornerstones of ethics: beneficence, respect for persons and justice?
25. In 1971, Prof. Philip Zimbardo designed and conducted an experiment that quickly became famous and is now referred to as the 'Stanford Prison Experiment'. As a psychologist, he wanted to examine individuals' behaviour when placed into roles of either prisoner or guard, particularly whether they would assume these 'roles' partially or fully including the exhibition of in-group norms and out-group rejection. Using 24 undergraduate volunteers from Stanford University who responded to a newspaper ad, he gave all of them preliminary psychological exams and briefed them on what was going to happen.

After telling them that they were free to leave at any time, he randomly assigned them the role of guard or prisoner in a fake campus-based prison he had built in the basement of an on campus building. Neither set of subjects were given specific instructions or training other than their role assignment. It was designed to run for two weeks. Within 6 days, he had to stop the entire experiment. After an initial uncertainty about what they were to do, the 'guards' began to exhibit sadistic tendencies. On the second day of the experiment, there was a prison 'rebellion' which was brought under control by the 'guards' who then more fully took on the profiles of prison guards. They began a privilege system used to interfere with the solidarity of prisoners, cultivating distrust. They reported becoming increasingly paranoid about the prisoners and their intentions and as a result extended the privilege system to all aspects of prisoners' days including going to the restroom and eating.

Unlike the guards, the prisoners were put into situations purposely meant to cause disorientation, degradation and depersonalization. Prisoners began to experience emotional disturbances, depression and learned helplessness. Within days of the outset, two prisoners had

to be removed due to emotional trauma. The 'prisoners' used their prison number instead of their names to address the visiting 'chaplain' and when asked how they planned to leave the prison, they became confused. Like the obedience experiments (1963), they show how individuals assume roles under the right circumstances, even if the roles are arbitrarily assigned. This experiment seems fraught with potential ethical issues and many have called this one of the most unethical experiments conducted.

Given key elements of ethical considerations above [*informed consent; right to privacy; right to service*], this controversial experiment is often introduced as 'an experiment gone bad.'

Therefore, to what extent did this experiment violate informed consent, voluntary participation, minimizing harm to participants, deception, anonymity and justice?

QUESTIONS FOR CHAPTER 6: Univariate Descriptive Statistics

Description – the summary process – is both doable and valuable. But why do we want to do this? What are the advantages and disadvantages of doing so (i.e.: summarizing data)? What do measures of central tendency tell us? And why are measures of dispersion invaluable? These fundamental activities of summarizing what we are interested in is a showcase of what statistics do. Regardless the sophistication of the following chapters and beyond this book, at the heart of statistics is summarizing. So, wrapping our heads around this process is not ‘baby statistics’ or ‘first steps’ but rather the planting of the cornerstone, breaking of ground, the raising of the first beam in the construction of an understanding of what statistics can do.

The statistical commands that you will learn and use in this chapter are summarizing commands. They will tell the statistical software to summarize the variables and report various elements about those summaries.

1. Why is it important to distinguish between nominal, ordinal, and interval-level variables when choosing statistical techniques? What might go wrong if we ignore these distinctions?
2. The mean is often seen as the most precise measure of central tendency. However, in what situations might using the mean be misleading or even incorrect? Can you think of real-world examples where another measure (mode or median) would be more appropriate?
3. Two variables have the same mean but very different standard deviations. What does this tell us about the representativeness of the mean in each case? Why should we care about dispersion rather than just the central tendency?
4. When people hear "average," they usually think of the mean. But in different contexts—like income, home prices, or test scores—why might the median be a better summary statistic? Can you give an example where using the mean would create a misleading impression?
5. Outliers can sometimes distort summary statistics but can also reveal important insights. How do we decide when an outlier is an error to be ignored versus an important case that tells us something meaningful? Think of an example from politics where an outlier changed our understanding of an issue.
6. The mode is the best measure of central tendency for nominal variables, but it often provides very little insight. Why is this the case? Can you think of a scenario where knowing the mode would be useful, and one where it would be nearly meaningless?
7. All summary statistics lose information. When we condense a dataset into a single number (like a mean or median), what trade-offs are we making? How do we balance the need for simplification with the risk of oversimplification?
8. In the textbook, there is a table of three variables on 25 of the most populated countries (TABLE 6.1). There are also some follow up questions:
 - In the 25 most populous countries in the world, what is the most common religion?

Political Analysis Workbook

- In the 25 most populous countries in the world, how free are they according to the Freedom House measures?
- In the 25 most populous countries in the world, how responsive are the governments?
- And finally, how good are each of these summary measures at describing these countries?

Using our knowledge of level of measurement and measures of central tendency and dispersion, let's answer these questions. Using the following table, insert the level of measurement and **calculate** (using the 'DS - Levels of Measurement' dataset) **or determine by hand** both the most appropriate measure of central tendency and measure of dispersion (i.e. one answer per cell).

	FH: Global Freedom Scores	WGI: Voice & Accountability	Main Religion
Central Tendency			
Dispersion			

9. What are the levels of measurement for the following:

Populations of nations	
Race (White, Asian, Hispanic, etc.)	
Level of education (high school, college, etc.)	
Order of finish in a race (first, second, third, etc.)	
Marital status (married, single, divorced, etc.)	
Percent of eligible voters registered to vote	
Ideological orientation (very liberal, somewhat liberal, somewhat conservative and very conservative)	
Region of your country where you were born	
Whether your country requires annual safety inspections of automobiles	
Attitudes towards nuclear energy (strong approve, approve, neither, disagree and strongly disagree)	
Amount of attention respondent paid to campaign commercials (none, some or a lot)	
Type of voluntary groups respondent is a member	
Respondent's income	

10. In the table above, how does the level of measurement change if we change the information we are looking for? Look at the last five variables. Does the level of measurement change and if so, what do level of measurement do they change to?

Percentage of cars in your country that have annual safety inspections	
Attitudes towards nuclear energy (pro-nuclear or anti-nuclear)	
Amount of attention respondent paid to campaign commercials (by minutes of attention)	
Number of voluntary groups respondent is a member	
Respondent's income bracket [<15,000; 15,000-45,000; 45,000-100,000; 100,000+]	

11. There is a measure of Income Inequality in which variable goes from 0 ('no income inequality') to 100 ('complete income inequality'). We'll call this the '**Solt Index**' (from its author).
- What is the level of measurement of the data?
 - Very high
 - Interval level
 - Nominal level
 - Ordinal level
 - Which of the following descriptive statistics can be used to describe the data?
 - Mode, variation ratio
 - Median, range
 - Mean, standard deviation
 - Actually, all of these can be used
12. Using the most recent Solt inequality data on European countries, what is the *best* measure of central tendency?
- The mode, which is 28.1
 - The median, which is 28.9
 - The mean, which is 29.2
 - The degree of Communism
13. Using the most recent Solt inequality data on European countries, what is the *best* measure of dispersion?
- The variation ratio, which is 0.935
 - The range, which is 22.6–35.6
 - The standard deviation, which is 3.72
 - The standard measure, which is 1
14. What is the crucial difference between an interval level variable and an ordinal level variable?
15. What is the crucial difference between an ordinal level variable and a nominal level variable?
16. What level of measurement is represented in the following examples?
- Whether a newspaper story portrayed a candidate in a negative, neutral or positive manner.
 - The percentage of each country's population that is non-native born.
 - Whether a country is in North America, South America, Europe, Africa or Asia.

Table: Various Scores of Democracy and Freedom

	Freedom House	WB WGI
--	---------------	--------

Political Analysis Workbook

	PR	CL	Status	Rule of Law
Albania	3	3	PF	0.21
Armenia	5	4	PF	0.31
Azerbaijan	6	5	NF	0.16
Belarus	7	6	NF	-0.10
Bulgaria	2	2	F	0.54
Croatia	1	2	F	0.44
Czech Republic	1	1	F	1.06
Estonia	1	1	F	1.40
Georgia	3	3	PF	0.68
Hungary	1	2	F	0.97
Kyrgyzstan	5	5	PF	-0.20
Latvia	2	2	F	1.00
Lithuania	1	1	F	1.10
Poland	1	1	F	0.96
Romania	2	2	F	0.54
Russia	6	5	NF	-0.36

17. Using the table above, calculate by hand the most appropriate measure of central tendency and dispersion for the following four variables:

	Measure of central tendency	Measure of dispersion
FH: Political Rights		
FH: Civil Liberties		
FH: Status		
WGI: Rule of Law		

18. Calculate the mean and standard deviation for both of the Freedom House: political rights and civil liberties.

	Measure of central tendency	Measure of dispersion
FH: Political Rights		
FH: Civil Liberties		

19. In the context of the most appropriate measures (median and range), explain whether using the mean and standard deviation are also an appropriate summary of these variables. Does this help or hurt our understanding of those variables?
20. Look up more information about the *Anscombe Quartet* dataset. Anscombe (1973) created variables with identical measures of central tendency and dispersion but vastly different distributions that reveal themselves only by graphing [e.g.: <https://youtu.be/Kd--Q-aTwpM/>].
21. In surveys, we often ask people to provide basic demographics which are often important when we want to know why people vote for one party or another, how much they might

participate, and what values they might hold. Below are three common questions. Please identify the variables' *the most* appropriate measures of central tendency and dispersion (i.e.: one answer per cell).

Survey Question	How old are you? ____ years old	Which class do you belong to? 1 working class 2 lower middle class 3 middle class 4 upper middle class 5 upper class	What is your current work situation? 1 self-employed 2 employed 3 in school 4 working in the household 5 retired 6 unemployed 7 other
Measure of Central Tendency			
Measure of Dispersion			

22. Give an example of a variable that might have a bimodal distribution. Draw the distribution.
23. Give an example of a variable that might have a positively skewed distribution. Draw the distribution.
24. In the 'DS - Levels of Measurement' dataset, let's look at some interval-level variables: totalpopulation [total population], incomeinequality [income inequality], and lifeexp (life expectancy). Produce the most appropriate measures of central tendency and dispersion for each.
25. In the 'CPD mini' dataset, let's look at some ordinal-level variables: cabinet composition [gov_party] and type of electoral system [prop]. Produce the most appropriate measures of central tendency and dispersion for each.
26. In the 'ESS mini' dataset, let's look at some nominal-level variables: badge and PolPart (a variable we will see again soon). Produce the most appropriate measures of central tendency and dispersion for each.

Datasets

DS - Levels of Measurement' dataset

- Prepared for *Political Analysis*

'Solt' Gini Index: Solt, Frederick. 2020. "Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database," *Social Science Quarterly* 101(3): 1183–1199. SWIID Version 9.3, June 2022.

Anscombe Quartet dataset

- Prepared for *Political Analysis*

Comparative Political Data Set

- Armingeon, Klaus, Sarah Engler, Lucas Leemann and David Weisstanner. 2024. *Comparative Political Data Set 1960-2022*. Zurich/Lueneburg/Lucerne: University of Zurich, Leuphana University Lueneburg, and University of Lucerne
 - Available here: www.cpbs-data.org/
 - *Political Analysis* dataset: "CPD mini"

European Social Survey

ESS Round 9: European Social Survey Round 9 Data (2018). Data file edition 3.1. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data Archive and distributor of ESS data for ESS ERIC. [doi:10.21338/NSD-ESS9-2018](https://doi.org/10.21338/NSD-ESS9-2018).

- Available here: <https://www.europeansocialsurvey.org/data/>
- *Political Analysis* dataset: "ESS mini"

QUESTIONS FOR CHAPTER 7: Measures of Association I: Nominal- and Ordinal-Level Variables

Expanding the distribution of a single variable (which, for example, we may have summarised in the previous chapter) across a new variable can be an enlightening for students' appreciation of statistical analysis. The notion that our description of a variable can help us understand elements of the variable itself can be improved by knowing another is the fundamental nature of (measures of) association. Seeing the overall distribution of a single variable change into separate distributions across various outcomes of another variable underscores the nature of 'a relationship' and more specifically, what an association means. To the extent that one variable changes with another is to say that they are associated. What this means, what we are trying to get at, is that this association – this relationship – is helpful in that knowing one can tell us a lot about the other. That is ultimately the value of knowing the association between two variables.

Second a PRE (Proportional Reduction of Error) statistic is about minimizing our errors by knowing the nature of the relationship between two variables. More importantly, reducing our errors merges with the narrative about scientific study – we are assumed to always be wrong or at a very minimum asymptotically uncertain. So, rather than thinking of our analysis as getting better, we can think of it as getting less worse. Reducing our errors we make in guessing outcomes is to our advantage. Finally, controlled comparisons introduce the notion of 'control,' as in controlling for other effects.

The statistical commands that you will learn and use in this chapter are summarizing commands. They will tell the statistical software to summarize the relationship between variables and report various elements about those summaries.

1. Measures of association quantify the strength and, in some cases, the direction of relationships between variables. What do the difference in Lambda vs. Gamma shape the conclusions we can draw about the nature of a relationship?
2. Chapter 7 discusses how different levels of measurement (nominal, ordinal) dictate which measure of association is appropriate. How does this constraint impact the way researchers construct and interpret statistical relationships in the social sciences?
3. Gamma provides information on the direction of relationships between ordinal-level variables. Is directionality important – or at least helpful - in social science research? Is this ordinal directionality made more challenging when categories are not evenly spaced conceptually?
4. How do controlled comparisons strengthen or limit the claims we can make about the relationship in which we are interested?
5. In Chapter 7, we moved from a binary measure of democracy to an ordinal measure (not free, partly free, free). How does this change impact the interpretation of associations with

internet access, and what does it reveal about the challenges of operationalizing political concepts?

6. In the following questions, state the associational analysis that you would use to test the relationship between the variables (choose from: Yule's Q, Gamma or Lambda)? Explain your choice.
 - a. You are testing whether citizens' ideologies (liberal, moderate or conservative) are related to their trust in parliament (low, moderate or high).
 - b. You are comparing whether graduating from college is related to whether people vote in local elections.
 - c. You want to see whether countries with more than 5% annual change in unemployment also have various levels of crime (low, medium, high) than countries that have less than 5% annual change in unemployment.

From the "Armed Conflict" dataset, we can determine whether the type of conflict [type_of_conflict] helps us anticipate whether a conflict since its onset has exceeded 1,000 battle-related deaths (prior to 1946) [cumulative_intensity].

7. Calculate the appropriate measure of association and interpret the finding.

Perhaps the type of conflict is better at explaining the main conflict issue in regards to the incompatibility about territory and/or government [incompatibility].

8. Calculate the appropriate Measure of Association and interpret the finding.

Party systems can shape the subsequent distribution of the people that inhabit them. Let's look at the DPI2020 Database of Political Institutions (see citation at end of chapter, you can use the 'DPI mini' dataset).

9. Produce a Measure of Association (with the software or by hand) so that we can see the joint distribution of the type of political system [system] with the religious status of the chief executive party [execrel] (NB: the non-religious are excluded from this analysis). And interpret the result
10. This is quite strong. Perhaps there are other features of national governments that shape the subsequent nature of the governing parties. For example, we can determine to what extent a proportional representation system [pr] affects the religious status of the chief executive party [execrel]. Produce a measure of association and interpret the result.
11. Related to our research on democracy and the internet in textbook, let's also confront a change we might make in the internet access variable. Freedom house also publishes an 'Internet Freedom' variable. That is, instead of Table 7.9, we could have chosen to present the cross-tab of Freedom House Global Freedom Scores and Freedom House Internet

Freedom Scores (from 'DS - Gamma Lambda' dataset). With the Global Freedom scores as the Dependent Variable, produce a measure of association and interpret the result.

12. What are the problems of using the Freedom House 'Internet Freedom' variable
13. We were interested in studying how the electoral system of a country is related to the cabinet composition. From the Comparative Political Dataset ('CPD mini'), we can test this by examining European countries' electoral systems (i.e. Proportional Representation or Single-Member Districts: 'prop') and cabinet compositions as measured by the Schmidt index (gov_party). Produce a cross-tab and an appropriate measure of association:
14. Fully interpret by characterizing the magnitude or strength of the relationship. Does this relationship have a direction? If so, how would you interpret this? If not, why not? How do the results speak to our initial question?

Let's reconsider the relationship of cabinet composition and electoral systems by controlling for whether the countries in the sample come from Western or Eastern Europe. Using a controlled comparison, we arrive at the following two tables.

15. How do we update the interpretation of our initial results?
16. We are interested in finding what explains higher levels of political activity. By hand, calculate the most appropriate Measure of Association for the following data displayed in the contingency table below and interpret this result.

Table Gender and Political Activity I

	Female	Male	Total
Not politically active	224	670	894
Politically active	776	330	1,106
Total	1,000	1,000	2,000

Hmm, interesting. From the findings above, maybe there is something to this relationship. You decide to partition the level of activity to see if the relationship suggested above can be further parsed out.

Table Gender and Political Activity II

	Female	Male	Total
Low political activity	260	680	940
Medium political activity	305	220	525
High political activity	435	100	535
Total	1,000	1,000	2,000

17. Calculate the most appropriate measure of association for gender and a scale of increasing levels of political activity and interpret the result.

18. What, then, might we say about our original hunch in the previous question?

Maybe something else is involved in shaping individuals' political activity. Let's examine political activity and social class.

Table: Social Class and Political Activity

	Low	Middle	High	Total
Low political activity	222	186	148	556
Medium political activity	148	297	295	740
High political activity	111	223	370	704
Total	481	706	813	2,000

19. In the table above, Calculate the most appropriate measure of association and interpret the results.

Let's use some real data. Using the 'ESS mini' dataset across 19 countries and keeping in mind that gender is the independent variable, produce a cross-tab of gender [Gender] and having voted in the most recent, national parliamentary elections [Vote].

20. Produce a Measure of Association appropriate to this 2×2 relationship and interpret the result.

Perhaps the confidence to participate more broadly in politics – rather than voting *per se* – is shaped by gender. Keeping in mind that gender is still the independent variable, produce a cross-tab of gender [Gender] and confidence to participate in politics [cptppola].

21. Produce a Measure of Association appropriate to this relationship and interpret the result.

With an eye on the direction of the proposed relationship, produce a cross-tab of interest in politics [Pollnt] and confidence in participating [cptppola].

22. Produce a Measure of Association appropriate to this relationship and interpret the result paying careful attention to the direction.

23. One big question about political participation is whether there is a type of person that does 'all of them' and a person that does none. One way to find out is to see how associated some political participation activities are.

Using the ESS mini dataset, produce measures of association between the following variables.

- pblmn: Taken part in lawful public demonstration last 12 months
- pstplonl: Posted or shared anything about politics online last 12 months
- sgnptit: Signed petition last 12 months
- wrkprty: worked in political party or action group last 12 months

24. Which of the two ideas - a type of person that does 'all of them' and a person that does none - is supported by these data? Why?

25. From the V-Dem mini dataset, look at the two variables:

- v2clacfree_ord: Freedom of academic and cultural expression
 - v2mecrit_ord: Print/broadcast media critic
- a. Using Freedom of academic and cultural expression as the dependent variable, what is a viable theory linking the extent of critical print/broadcast media to that freedom?
 - b. If we use these two variables to test an empirical hypothesis of this relationship, test this empirically by generating an appropriate measure of association.
 - c. Do your findings support or fail to support your theory?

26. From the V-Dem mini dataset, look at the two variables:

- v2mecrit_ord: Print/broadcast media critic
 - v2elfrfair_ord: Election free and fair
- a. Does the extent and power extent of critical print/broadcast media extend to have political impacts, particularly for elections?
 - b. If we use these two variables to test an empirical hypothesis of this relationship, test this empirically by generating an appropriate measure of association.
 - c. Do your findings support or fail to support your theory?

27. Are there any big regional differences between the level of individual-level reported happiness? Using the ESS mini dataset, look at the following two variables:

- Clsprty: Feel closer to a particular party than all other parties
 - regions: West/East/Med
- a. Does there seem to be any substantial regional differences in terms of how close individuals might feel to a political party? Generate a measure of association.
 - b. Given the output and an inspection of the cross-tab, what might you conclude?

Datasets

UCDP/PRIO Armed Conflict dataset

- Davies, Shawn, Garoun Engström, Therese Pettersson & Magnus Öberg (2024). Organized violence 1989-2023, and the prevalence of organized crime groups. *Journal of Peace Research* 61(4). [Version 24.1]
 - Available here: <https://ucdp.uu.se/downloads/>.
 - *Political Analysis* dataset: "Armed Conflict"

DPI2020 Database of Political Institutions

- Scartascini, C., Cruz, C., & Keefer, P. (2021). *The Database of Political Institutions 2020 (DPI2020)*.
 - Available here; <https://doi.org/10.18235/0003049>
 - *Political Analysis* dataset: "DPI mini"

Comparative Political Data Set

- Armingeon, Klaus, Sarah Engler, Lucas Leemann and David Weisstanner. 2024. *Comparative Political Data Set 1960-2022*. Zurich/Lueneburg/Lucerne: University of Zurich, Leuphana University Lueneburg, and University of Lucerne
 - Available here: www.cpbs-data.org/
 - *Political Analysis* dataset: "CPD mini"

European Social Survey

- ESS Round 9: European Social Survey Round 9 Data (2018). Data file edition 3.1. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data Archive and distributor of ESS data for ESS ERIC. [doi:10.21338/NSD-ESS9-2018](https://doi.org/10.21338/NSD-ESS9-2018).
 - Available here: <https://www.europeansocialsurvey.org/data/>
 - *Political Analysis* dataset: "ESS mini"

V-Dem

- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. "V-Dem [Country-Year/Country-Date] Dataset v6.2." Varieties of Democracy (V-Dem) Project.
 - Available here: <https://www.v-dem.net/vdemds.html>
 - *Political Analysis* dataset: "V-Dem mini"

QUESTIONS FOR CHAPTER 8: Measures of Association II: Means Comparison and Correlation

Here we move into more and less familiar territory simultaneously. At first glance, you may see this chapter as an island in the stream. 'Mean' and 'correlation' seem like familiar words, at least in colloquial usage. It can be expected that using means to compare outcome categories to (preliminarily) establish some type of pattern or association of interest starting to get pretty interesting and raises the empirical challenge. Correlation, despite its familiarity, is of course a new challenge.

It is crucial that we start to see these terms as statistical techniques with applications and limitations. This shouldn't be the process of formalizing what you 'already know' about these common terms but rather updating what you 'already know' to the formal use of these terms.

The statistical commands that you will learn and use in this chapter are still summarizing commands. They will tell the statistical software to summarize the relationship between variables and report various elements about those summaries.

1. Why does the choice of statistical measure for association depend on the level of measurement of the variables involved, and how does this impact our ability to compare different types of relationships?
2. The chapter suggests that categorizing interval-level variables (e.g., turning internet access into "<50%" and ">50%") leads to information loss. How does this trade-off between simplification and precision affect the interpretation of results in social science research?
3. In America, there is an old saying: 'If You Are Not a Liberal When You Are Young, You Have No Heart, and If You Are Not a Conservative When Old, You Have No Brain'

[Borrowed, as with many things, from the French: **François Guizot** (1787–1874): "*Not to be a republican at 20 is proof of want of heart; to be one at 30 is proof of want of head.*" Or more recently: **Georges Clemenceau** (1841–1929): "*If a man is not a radical at twenty, he has no heart; if he is not a conservative at forty, he has no head.*"]

If we take age to be the interval-level variable and ideological location (1 Very Left; 2 Left; 3 Moderate; 4 Right; and 5 Far Right) to be ordinal, which means are we comparing across which categories? And – given this Franco-American phrase, what do we expect to find?

4. Means comparisons allow us to summarize interval-level variables across categorical groups. Are there limitations to this method when the category distributions are highly skewed?
5. Pearson's correlation coefficient assumes a linear relationship between two interval-level variables. What are the implications of this assumption, and in what situations might correlation fail to capture meaningful associations?
6. Why can't we use cross-tabulation to look at two interval-level variables as we have been doing with nominal- and ordinal-level variables?

7. Now we are starting to understand the classic trope, ‘correlation is not causation,’ more profoundly. What is the difference between these two concepts?
8. The chapter discusses how different operational definitions of democracy (e.g., binary vs. ordinal vs. continuous) can lead to different conclusions. Recalling our detailed discussion in Chapter 3, in what way does the way we define and measure concepts shape our empirical findings in social science?
9. Democracy means a lot of things. Many times we think about elections and freedom of speech but it can also include secondary freedoms as well. For example, whether universities and other cultural institutions are not shaped by politics. Using the “V-Dem mini” data set, we would like to compare the *Liberal Democracy Index* by an ordinal-level measure of *Freedom of academic and cultural expression*. Using the following variables, let's look at this.

Liberal Democracy Index: v2x_libdem

Freedom of academic and cultural expression: v2clacfree_ord

- a. Given the nature of the variables we want to compare – with the liberal democracy index as an interval-level variable and academic and cultural freedom as an ordinal-level variable - perhaps we need to do a means comparison. What are the means of the *Liberal Democracy Index* for each category of *freedom of academic and cultural expression*?
 - b. What do we find? Please describe the results.
10. When we were examining European Parliamentary election turnout by country (Table 6.4), we summarized the interval-level variable, percent turnout to summarize these data. One thing that we noted is the apparent difference in turnout between the newer members in the East and the older members in the West. As percent turnout is an interval-level variable and East/West is a nominal-level variable, let's look at a means comparison.

Table: Means Comparison: EP Election Turnout

	EP election turnout	Number of observations [country/years]
Western Europe	57.5%	132
Eastern Europe	32.0%	43
Overall mean	51.2%	175

First, there are not 132 countries in Western Europe (or 43 in Eastern Europe). The 16 (11) countries each have participated in several elections. Second, of all, as a *description of these data*, it appears that there may be a higher level of turnout in the West (57.5% > 32.0%).

Perhaps, the East Europeans were excited by the prospect of being able to vote in the EP elections and quickly became disappointed. And West Europeans may have also been a little disappointed after Eastern Europe joined the EU by feeling that their ‘Western’ votes didn’t

count as much. Thus, the first full election of the expanded EU in 2004 brought on a post-expansion funk thus explaining what seems to be the disparity in voting levels between East and West.

One way to start to uncover this is to compare the means of pre-2004 West and East (pre-funk era) with the means of post-2004 East and West (post-funk era). Why 2004? In 2004, the countries Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia joined the European Union, the largest expansion in its history.

We can do this by integrating the controlled comparisons (from Chapter 7) with the means comparisons of this chapter. We use the same logic as before. Controlled Means Comparisons takes this one step further and subdivides the total group by a third variable to see if the original relationship is affected by this separation. That is, controlled means comparisons allow us to identify the potential influence of a third variable on bivariate relationships between interval-level variables and nominal-/ordinal-level variables.

Table: Controlled Means Comparison: EP Election Turnout

	2004 and previous (<i>pre-funk</i>)		After 2004 (<i>post-funk</i>)	
	EP election turnout	Number of observations [country/years]	EP election turnout	Number of observations [country/years]
Western Europe	59.4%	81	54.4%	51
Eastern Europe	31.2%	8	32.2%	35
Overall mean	56.9%	89	45.4%	86

Comment on two specific things that this means comparison and controlled means comparison potentially has shown us about the East/West difference in European Parliamentary turnout. [*Hint: 'no change' is also a result.*]

11. Using the 'V-Dem min' dataset, does the Egalitarian Democracy Index move with equality for gender ('power distributed by gender')?
 - Egalitarian democracy index: v2x_egalDEM
 - Power distributed by gender (Ordinal): v2pepwrGEN_ord
 - a. Given the nature of the variables we want to compare, perhaps we need to do a means comparison. What are the means of the *Egalitarian Democracy Index* for each category of *power distributed by gender*?
 - b. What do we find? Please describe the results.
12. Like above, we can use the concept of Controlled Means Comparison. Given what we observed above, perhaps there is something more than merely symbolic about a female Head of State on gender parity. Let's compare the relationship between *Egalitarian*

Democracy Index and power distributed by gender across two groups, one with a female head of state and one without. To control for Female Head of State, we can use the following:

- Egalitarian democracy index: v2x_egaldem
 - Power distributed by gender (ordinal): v2pepwrngen_ord
 - Female Head of State 'v2exfemhos'
- a. What do we find? Please describe the results.

13. Let's compare the World Governance Scores for countries at various levels of freedom, using Freedom House's freedom status measure: Not Free, Partly Free and Free. Recall that the World Governance Scores range from -2 (lowest score) to 2 (highest score).

	<i>FH: Not Free</i>	<i>FH: Partly Free</i>	<i>FH: Free</i>
Government effectiveness	-0.60	-0.50	0.55
Rule of law	-0.94	-0.46	0.50
Voice and accountability	-1.23	-0.12	0.77
Observations	35	72	239

Using this means comparison between measures of governmental performance and freedom status from Freedom House (measuring an average rating for Political Rights and Civil Liberties for each country), what do we learn about government performance on these three dimensions moving across the categories of Freedom House from Not Free to Free?

14. We can take a further look at what might be driving the difference in scores for governmental performance by separating the above cross-tab by the level of political corruption. At the top, we include only the countries that rank in the lower half of the Political Corruptions Index (i.e. less corruption). At the bottom, we include only the countries that rank in the upper half of the Political Corruptions Index (i.e. more corruption).

Political Corruption Index: Less than 50%			
	<i>FH: Not Free</i>	<i>FH: Partly Free</i>	<i>FH: Free</i>
Government Effectiveness	-0.51	0.47	0.73
Rule of Law	-0.87	0.36	0.69
Voice and Accountability	-1.40	0.48	0.90
Observations	7	2	184
Political Corruption Index: More than 50%			
	<i>FH: Not Free</i>	<i>FH: Partly Free</i>	<i>FH: Free</i>
Government Effectiveness	-0.62	-0.53	-0.06
Rule of Law	-0.96	-0.48	-0.14
Voice and Accountability	-1.19	-0.13	0.33
Observations	28	70	55

What do we learn about government performance on these three dimensions moving across the categories of Freedom House from Not Free to Free, controlling for political corruption?

15. How have things changed in the controlled comparison from the previous table of simple means comparison?
16. Do you see any limitations to this controlled means comparison?
17. Recall the discrepancy we found in the text that ‘partly free’ democracies have the lowest mean level of internet access [Table 8.5]. This roughly corresponds to the curious finding we stumbled upon in the previous chapter in the *Controlled Comparison* section. Recall, we found that the relationship between democracies and internet access is conditioned on economic performance. Specifically, countries in which unemployment has changed less than 1% between 2018-20, the relationship between democracy and internet is quite strong. At the same time, countries in which unemployment has changed more than 1% between 2018-20, the relationship between democracy and internet is weighted heavily toward higher internet access regardless of the level of democracy. This should lead us all to the same lightbulb moment: perhaps the ‘partly free’ discrepancy can be accounted for by economic performance. Only one way to find out.

Let’s take a look, Using the ‘DS - Gamma Lambda’ dataset, produce a controlled means comparison of the level of Internet access [Perc_Internet] over the Freedom House Global Freedom Index [FH_GF] over low and high unemployment [UNEMP]

Does this Controlled Means Comparison provide further insight into this relationship?

18. Which is more appropriate, means test or correlation?

a.	What is the difference of the percentage of women in parliament, in democracies and non-democracies?	
b.	Does the national percentage of women in parliament correspond to the amount of money spent on education at the national level?	
c.	Does amount of money spent on education at the regional level differ across the five regions of Denmark?	
d.	Are the total percentages of people who can vote in a country the same for each ethnic group?	
e.	Does the total percentages of people who can vote in a country reflect the economic performances of countries?	
f.	Do economic performances of countries tell us about the percentage of women in parliament?	

19. A research study reports a correlation of $r = -0.64$ between the individuals’ party preference (Green Party, Centre-Left Party, Centre-Right Party, Extreme Far Right Party) and the amount they donate annually to Amnesty International. Which of the following statements is *most correct*:
 - Party preference and donations to Amnesty International are negatively correlated.
 - Party preference and donations to Amnesty International are substantially and negatively correlated.
 - Pearson’s correlation coefficient is inappropriate to establish an association between these variables as individuals’ party preferences is not an interval-level variable.

- People who prefer far right parties do not donate to Amnesty International.

20. In the previous chapters, we used PRE Measures of Association. How does correlation differ in calculating and indicating a relationship between variables?

21. Using the 'QOGmini' dataset, let's look at the empirical relationships between Income inequality [wdi_gini] and other interval-level variables:

- the HDI [undp_hdi]
- Life expectancy [wdi_lifexp]
- GDP *per capita* [wdi_gdpcapcur]

Insert the correlations and the most appropriate description

- a. Income Inequality [wdi_gini] and Human Development Index [undp_hdi]
- b. Income Inequality [wdi_gini] and Life Expectancy [wdi_lifexp]
- c. Income Inequality [wdi_gini] and GDP *per capita* [wdi_gdpcapcur]

22. Which of the below provides an initial narrative?

- Overall, unequal countries only have high Human Development and Life Expectancy because, presumably, the economy works.
- Overall, things that we prefer – longer lives, better economic conditions, and capacity for higher quality of life are all negatively – if only slightly - correlated with income inequality.
- Overall, Income Inequality and Life Expectancy are highly correlated which should be true because living longer allows people more time to get ahead (or behind).
- Overall, income inequality is only strongly correlated with economic variables.

23. From the 'QoG mini' dataset, produce a scatterplot between the HDI [undp_hdi] and Life expectancy [wdi_lifexp]

- a. What is the correlation and strength?
- b. Which interpretation is correct:
 - This is a negative correlation, we can definitely state that when Life Expectancy increases, the HDI goes down.
 - Life expectancy causes the HDI to increase.
 - Life Expectancy is the most important element of the HDI
 - This is a substantial positive correlation. When Life Expectancy increases, the HDI as well increases.
 - Correlation is not the most appropriate measure of association.

24. From the 'QoG mini' dataset, produce a scatterplot between the HDI [undp_hdi] and GDP *per capita* [wdi_gdpcapcur]

a. What is the correlation and strength?

b. Which interpretation is correct:

- GDP *per capita* and the HDI are not strongly associated.
- While a correlation has been produced, it seems that the relationship might be better modeled without the assumption of linearity.
- GDP *per capita* causes the HDI to increase.
- Compared to Life expectancy, the GDP per capita produces 12% less change in HDI.
- Correlation is always appropriate between interval-level variables.

25. The World Bank produces the World Governance Indicators, which is a set of six measures of governmental performance. Let's look at their relationships.

- Voice and Accountability [wbgi_vae]: The extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association and a free media.
- Political Stability and Absence of Violence/Terrorism [wbgi_pve]: Reflects perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including politically motivated violence and terrorism.
- Government Effectiveness [wbgi_gee]: The quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation and the credibility of the government's commitment to such policies.
- Regulatory Quality [wbgi_rqe]: Reflects perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.
- Rule of Law [wbgi_rle]: Reflects perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police and the courts as well as the likelihood of crime and violence.
- Control of Corruption [wbgi_cce]: Reflects perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as 'capture' of the state by elites and private interests.

Produce a correlation table for all 6 indicators of governmental performance:

26. Which variables are correlated over $r = 0.90$.

27. Make an argument as to why these variables are so closely correlated.

28. What are the lowest two relationships?

29. What does this tell us about these relationships? That is, make an argument as to why these variables are not so closely correlated.

30. At the same time, from the same dataset, produce a scatterplot for a measure of media corruption [vdem_mecorrpt] and with World Bank's World Governance Score for Regulatory Quality [wbgi_rqe]. Produce and interpret the correlation
31. What is the (theoretical) implication of this correlation?

Datasets

DS - Gamma Lambda dataset

- Prepared for *Political Analysis*

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG mini"

V-Dem

- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. "V-Dem [Country-Year/Country-Date] Dataset v6.2." Varieties of Democracy (V-Dem) Project.
 - Available here: <https://www.v-dem.net/vdemds.html>
 - *Political Analysis* dataset: "V-Dem mini"

QUESTIONS FOR CHAPTER 9: Measures of Association III: (Bivariate) Regression

You may not realize this but introducing regression as a part of Descriptive Statistics is new. And it will help a great deal later. Coming on the heels of the notions of association, levels of measurement, and reducing errors, we set the foundation for what regression does as a quantitative representation of a relationship. Regression is a measure of association. Well, technically, it is not as measure of association as it is a formula rather than a number but for concept of substantive significance (i.e.: a measure of association), this provides initial insight into what we will eventually discover are the multiple capacities of multiple regression.

The statistical commands that you will learn and use in this chapter are summarizing commands. They will tell the statistical software to summarize the relationship between variables and report various elements about those summaries.

1. How does (bivariate) regression analysis differ from correlation in describing relationships between variables?
2. Why is the error term (ϵ) fundamental to regression analysis, and what does it represent?
3. What does the coefficient of determination (R^2) tell us about a regression model?
4. What is the significance of the regression coefficient (β), and how should it be interpreted?
5. Why is model fit important in regression, and what challenges arise when assessing it?
6. You are presented with a model with an adjusted- R^2 of .50. Which of the following statements is correct?
 - There is a 50% chance that the model is correct.
 - 50% of the cases in the data set are predicted by the model.
 - 50% of the coefficients are significant.
 - 50% of the variation in the dependent variable is explained by variation in the independent variable.
 - 50% of the variation in the independent variable is explained by variation in the dependent variable.
7. What is the difference between the systematic and random parts of the regression equation? *Click all that apply.*
 - In a regression equation, the systematic components refer to the intercept and the beta coefficient.
 - The systematic components always sum to zero.
 - The random components explain the variance of the independent variable.
 - The random parts of a regression equation refer to the error term.
 - The systematic components are the part of the regression equation that we control by introducing an independent variable.
 - The random components of the regression equation represent the random effects that explain the dependent variable but are not included.

8. In the chapter, there was a discussion of how OLS regression coefficients are BLUE, the Best Linear Unbiased Estimator. What does 'best' mean in this context?

In a study of the negative effects of rising unemployment on youth, a team of researchers regress the change in the percent of annual unemployment on the annual suicide rate of people under the age of 35 (per 100,000 people). The resulting regression equation looks like this:

$$\text{Suicide rate} = 11.3 + 0.04(\text{Change in Unemployment, \%}) + \varepsilon$$

9. What does the regression coefficient tell us?
10. What does the intercept tell us?
11. Before 2020, the change in unemployment was very slowly rising (a change of 0.5%). After the pandemic period, the change in unemployment grew to nearly 4%. Given these two values, what is the expected change in the suicide rate?
12. Given this study, what is a policy recommendation that we might make?

Recall from the textbook (Chapter 9) that Reporters Without Border's Press Freedom Index is strongly correlated with the World Bank's World Governance Indicator 'Voice and Accountability'. With the Press Freedom Index as the dependent variable, fully model this relationship in six steps. You can use the 'QoG mini' dataset:

13. *Step One: A Research Question*
14. *Step Two: Identify Variables*
15. *Step Three: Pearson's Product Moment Correlation Coefficient*
16. *Step Four: The Regression Equation*
17. *Step Five: Model Fit*
18. *Step Six: Interpretation*

Using the 'Polity V mini' dataset, we can try to determine if democracies are more long lasting than autocratic regimes. Let's model this relationship. The first two steps are the Research Question and Identifying the Variables.

Step One: The RQ is: As countries range from more autocratic states to more democratic states, we expect them to have longer periods of stability.

Step Two: Identify Variables

- **DV: Regime durability** ['durable']: The number of years since the most recent regime change or the end of transition period defined by the lack of stable political institutions since 1800. Mean: 20.93; sd: 24.58
- **IV: Polity score** ['polity']: The polity score is computed by subtracting the AUTOC score from the DEMOC score; the resulting unified polity scale ranges from -10 (strongly autocratic) to +10 (strongly democratic). Mean: -0.52; sd: 7.10

19. *Step Three: Pearson's Product Moment Correlation Coefficient*

20. *Step Four: The Regression Equation*

21. *Step Five: Model Fit*

22. *Step Six: Interpretation*

From the 'CPD mini' dataset, we can examine the relationship between national-level income inequality and national-level spending on families (subsidies and support for childcare and related expenditures).

23. *Step One: A Research Question*

24. *Step Two: Identify Variables*

- **DV: Expenditure for families** ['family_pmp']: Total public and mandatory private expenditure for families as a percentage of GDP. Range: 0.15% to 4.45%. Mean: 1.93; sd: 0.955
- **IV: National-level income inequality** ['postfisc_gini']: Gini index of disposable income (after taxes and transfers) among household members. Range: 18.3–36.5. Mean: 27.26; sd: 4.125

25. *Step Three: Pearson's Product Moment Correlation Coefficient*

26. *Step Four: The Regression Equation*

27. *Step Five: Model Fit*

28. *Step Six: Interpretation*

Let's go back to the impression that income inequality is not great for us – as we found in the previous chapter - by looking again at the 'QoG mini' dataset. Life Expectancy is the most highly correlated with Human Development so, let's regress the Gini Index for income inequality [wdi_gini] on Life Expectancy [wdi_lifexp] to see to what extent national-level income inequality is related to the longevity of its citizens.

29. *Step One: A Research Question:* Which of the following research questions most aligns with our inquiry?

- To what extent does life expectancy affect states' levels of income inequality?
- Why do states' levels of income inequality shape life expectancy?
- Do life expectancy and states' levels of income inequality move with the same variables?
- How can we get income inequality to shorten life expectancy?
- To what extent do states' levels of income inequality contribute to shorter life expectancy?

30. *Step Two: Identify Variables:* Which is the dependent variable, and which is the independent variable?

- The dependent variable is the GINI Index [wdi_gini] and the independent variable is Life Expectancy [wdi_lifexp].
- Life Expectancy [wdi_lifexp] and the GINI Index [wdi_gini] are both independent variables.
- The dependent variable is Life Expectancy [wdi_lifexp] and the independent variable is the GINI Index [wdi_gini].
- Life Expectancy [wdi_lifexp] and the GINI Index [wdi_gini] are both dependent variables.
- Neither, this is a trick question because in regression the variables are symmetrical. It doesn't matter

31. *Step Three: Pearson's Product Moment Correlation Coefficient:* Calculate r

- Lambda is actually the most appropriate measure of association for this relationship.
- $r=0.515$
- $r=-0.316$
- They are very strongly correlated.
- Correlation only works with two interval-level variables.

32. *Step Four: The Regression Equation:* Calculate the regression equation

- $LE = 83.02 - 0.268(GINI) + \epsilon$
- $LE = 18.17 - 2.26 (GINI) + \epsilon$
- $LE = 43.05 - 0.415(GINI) + \epsilon$
- $LE = -0.268 + 83.02(GINI) + \epsilon$
- $LE = -0.316 + 18.17(GINI) + \epsilon$
- Regression cannot be calculated for these variables.

33. What is the *expected* average life expectancy of a country with a GINI of...

- a. 25:
- b. 45:

34. *Step Five: Model Fit:* What is the most appropriate reporting for this model?

- None of these are appropriate.

Political Analysis Workbook

- 8.1% of the variation in national-level income inequality can be explained by variation in life expectancy.
- 10.0% of the variation in national-level income inequality can be explained by variation in life expectancy.
- National-level income inequality explains 8.1% of life expectancy.
- 8.1% of the variation in life expectancy can be explained by variation in national-level income inequality.

Datasets

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG mini"

Polity5 Project: Political Regime Characteristics and Transitions: 1800-2018

- Polity5 dataset, version 2018
 - Available here: <https://www.systemicpeace.org/inscrdata.html>
 - *Political Analysis*: 'Polity V mini'

Comparative Political Data Set

- Armingeon, Klaus, Sarah Engler, Lucas Leemann and David Weisstanner. 2024. *Comparative Political Data Set 1960-2022*. Zurich/Lueneburg/Lucerne: University of Zurich, Leuphana University Lueneburg, and University of Lucerne:
 - Available here: www.cpds-data.org/
 - *Political Analysis* dataset: "CPD mini"

QUESTIONS FOR CHAPTER 10: An Introduction to Inference

Inferential Statistics takes up the unique and separate process of linking samples to population. This is done separately to demonstrate its distinct contribution to what statistics offers us; that is, summaries and some degree of confidence of these summaries being, in a sense, 'real.' Thus, we learn that inference – or, Inferential Statistics - is the process of determining how confident we can be in our best guess of what we want to know but can never really know. Statistical significance is based on the notion of whether what we have observed in the sample is likely to have been a function of chance or not. This is included in the definition: Inferential Statistics involves making informed guesses about population values – or parameters - from a sample by estimating the probability that the result could be due to chance.

1. How does inferential statistics extend the power of descriptive statistics in scientific research?
2. Why is randomization essential for statistical inference, and what risks arise when it is not properly implemented?
3. How does probability theory underpin the logic of inference, and why is a larger sample preferred?
4. A simple random sample occurs when:
 - a. a minimum of 10% of the population is included in the sample.
 - b. the sample has a normal distribution.
 - c. each unit in the population has an equal chance of becoming a part of the sample.
 - d. each unit in the population has been selected for the sample.
 - e. None of the above
5. The key to being able to make proper inferences in statistical analysis is
 - a. Randomization and sufficient sample size
 - b. A famous professor leading the research
 - c. Collecting information on the entire population
 - d. Making zero mistakes measuring the variables in which we are interested
 - e. Using Big Data
 - f. None of the above
6. Whether we can infer the substantive relationship observed in the sample to the population from which the sample was drawn is _____. While the pattern we observe in the sample is the _____ of a relationship.
 1. substantive significance, statistical significance
 2. statistical significance, substantive significance
 3. statistic, parameter
 4. parameter, statistic
 5. causal, random
7. A _____ is a descriptive characteristic of a population whereas _____ refers to the estimate of a population parameter from sample data.

- a. substantive significance, statistical significance
 - b. statistical significance, substantive significance
 - c. statistic, parameter
 - d. parameter, statistic
 - e. causal, random
8. How is statistical significance different than substantive significance?
- a. Substantive significance describes the relationship in the sample. Statistical significance is whether these results can be inferred to the population from which the sample was drawn.
 - b. Substantive significance describes the relationship in the sample. Statistical significance is whether these results are the same as the population.
 - c. Substantive significance describes the relationship in the sample. Statistical significance is whether these results are true for all populations.
 - d. Substantive significance describes the relationship in the sample. Statistical significance is whether these results can be inferred to the rest of the sample which is random.
 - e. None of the above
9. Statistical significance means that
1. the substantive significance doesn't matter
 2. what we have observed in the sample are unlikely to be a function of chance
 3. we are sure that this is the only relationship
 4. we can generalize to other theories our explanation
 5. we have explained a class of events
10. What does statistical significance tell us about causation?
1. Nothing. Statistical significance only tells us when our sample is not a good representative of the population.
 2. Nothing. Only substantive significance can tell us anything about causation.
 3. Nothing. Causation is proven by other statistical techniques and tests.
 4. Nothing. Although statistical significance is necessary but not sufficient in order to identify a potential causal relationship.
 5. None of the above
11. Probability is a formal model of uncertainty that assigns a numerical measure of the chance or likelihood, that a particular event will occur. Distinguish the classical and the relative frequency methods on the basis of this assignment.
12. The intersection is the joint probability of two (or more) events. In the case of two events, the intersection contains the outcomes which belong to both E_i and E_j (i.e. $E_i \cap E_j$). The formula tells us to multiply the outcomes, but instead of getting large, this multiplication makes the intersection smaller than both the events. Why is that?
13. The union of events contain all outcomes in which we are interested ($E_i \cup E_j$). In the formula, why do we subtract the intersection $P(E_i \cap E_j)$?

14. The complement of an event is all the other outcomes that are not part of the event. Does this mean that the complement is always larger than the event?
15. When an event occurs and does not affect the probability of another event, these events are independent. Explain how the dependency of two events changes our assessment of their probability of occurring (hint: see *conditional probability*).
16. Using online data collection through a social media website, an online polling and survey company recently studied public opinion in several of the countries in the EU on the issue of the EU program itself. Aiming to sample the population of EU voices, name two of the potential challenges to the representativeness of this sample?
17. Which of the following events in Political Science are likely *best understood* as probabilities?
 1. GDP *per capita*
 2. Freedom House scores for democracy
 3. The percentage of voter turnout on election day
 4. *A country goes to war*
 5. Chance that a specific party wins a majority
 6. The number of votes a party gets
18. Inferential statistics involve making informed guesses about population parameters from a sample by estimating the probability that the result we observe in the sample could be due to ____.
19. Explain the importance of representativeness of the sample with the goal of inference.
20. Recall Table 10.1 from the textbook. Provide another sample for each population.

Population	Sample
Level of EU support among European voters	
Success of environmental legislation	
Effectiveness of crisis management in sub-Saharan Africa	
Ideological tenor of political posts on Twitter	
Impact of political consumerism in America	

21. Figure 10.1 confronts us with which review we think is the most accurate as to the actual quality of the product. Which one is the most useful and why?
22. Why is inferential statistics practically necessary?
23. Recall the example of early election calling in the textbook (see Table 10.2). Perhaps we can change our attention to the longest times. We are interested in the likelihood that there will be early elections called more than 3 years before the regularly scheduled election (i.e. >36 months). What is the probability this will occur?

24. Continuing from above, what is the probability that early elections will be called more than 2 years before the regularly scheduled election.
25. How are the frequentist and Bayesian approaches to statistics – including probability and uncertainty - different?
26. Use your favorite web browser to search, 'Applications of Bayes Theorem' and briefly describe its use in three 'real world' examples. They do not need to be examples in political science.
27. Recall the example of early election calling in the textbook (see Table 10.2). Let's move our attention to the longest times. We are interested in the likelihood that there will be early elections called more than three years before the regularly scheduled election (i.e.: >36 months). What is the probability this will occur?
- There is a 0% chance that, if an early election is called, it will be called more than 3 years before the regularly scheduled election.
 - There is a 12% chance that, if an early election is called, it will be called more than 3 years before the regularly scheduled election.
 - There is a 28% chance that, if an early election is called, it will be called more than 3 years before the regularly scheduled election.
 - There is a 7% chance that, if an early election is called, it will be called more than 3 years before the regularly scheduled election.
 - There is no way to know this from these probabilities.
28. Following from the previous question, what is the probability that early elections will be called more than 2 years before the regularly scheduled election.
- There is a 10% chance that, if an early election is called, it will be called more than 2 years before the regularly scheduled election
 - There is a 39% chance that, if an early election is called, it will be called more than 2 years before the regularly scheduled election.
 - There is a 17% chance that, if an early election is called, it will be called more than 2 years before the regularly scheduled election
 - There is a 72% chance that, if an early election is called, it will be called more than 2 years before the regularly scheduled election
 - There is no way to know this from these probabilities.

QUESTIONS FOR CHAPTER 11: Inference for Nominal- and Ordinal-Level Variables

In a broader scientific perspective, we aim to determine whether the substantive relationship we uncover in the sample data also achieves statistical significance. Merging the descriptive power of statistics with the ability to infer the observed relationship as an attribute of a larger set of observations – or not – is the power of inferential statistics. This chapter focuses on the test for statistical independence, χ^2 [Chi-squared].

However, in order to get to the significance testing, we first develop a rubric of Classical Hypothesis (or Significance) Testing. Again, like the modelling in the chapter on (bivariate) regression (Chapter 9), this five-step hypothesis testing is not limited to significance testing and has a more universal utility. In the case of determining statistical significance, however, this gives us a clear and reliable guide.

The statistical commands that you will learn and use in this chapter are a combination of descriptive and inferential commands. They will tell the statistical software to determine the substantive and statistical significance of relationships. It is becoming increasingly important for you to understand what the output is telling you about the variables under investigation.

1. What is the *conceptual* difference between Type I and Type II errors, and why does it matter in statistical inference?
2. How does the chi-square (χ^2) test determine whether an association between nominal- or ordinal-level variables is statistically significant?
3. In hypothesis testing, why is the critical region also called the rejection region?
4. What is the logic behind rejecting or failing to reject the null hypothesis?
5. When is χ^2 appropriate as a test of statistical significance?
 - When a nominal- and an ordinal-level variable interact with an interval-level variable.
 - When you need to know about the statistical significance of relationships among interval- and ordinal-level data.
 - When you want to make inferential claims about relationships among nominal- and ordinal-level variables.
 - Only when the substantive relationship between nominal- and ordinal-level variables is very low or very high.
 - It is always appropriate.
6. Gamma and Lambda are measures of association that confer ____ significance on the relationships between nominal- and ordinal-level variables. χ^2 , on the other hand, confers ____ significance.
7. Let's look at how people who place various levels of importance in living in a democratic society (not important, somewhat important, very important) rank the performance of

democracy in their own country (poor, satisfactory, very good). If we want to examine the joint distribution of these two variables, which variable goes across the top (columns) and which goes down the (left-hand) side (rows) in our cross-tab? Why?

8. Based on the level of measurement of these two variables, you get one of the following measures of association:
 - Gamma: -0.36
 - Lambda: 0.45
 - Yule's Q: -0.51
 - Correlation: $r = -0.48$
9. What does this Measure of Association tell you about the strength of the relationship?
 - It is moderately strong and negative, meaning that as respondents' preferences for democracy go up, their evaluations go down.
 - It is moderately strong and positive, meaning that as respondents' preferences for democracy go up, their evaluations as well go up.
 - It is moderately strong and direction cannot be implied.
 - They move together in a moderate, negative correlation, meaning that as respondents' preferences for democracy go up, their evaluations go down.
10. You produce a χ^2 test statistic from these data of 10.25. Using the table below, report whether you can be 95% confident that the relationship exists in the population.

Degrees of freedom	P-value = 0.05	P-value = 0.01	P-value = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46

11. Can you be 99% confident? Explain.

Our physical health is important to our state of mind. The better we feel physically, we often report feeling happier. Using data from the 'WVS mini' dataset, we can look at two variables that might help us understand this relationship.

Feeling of happiness [Happy]

State of health (subjective) [Health]

12. Generate a cross-tab

13. To investigate this relationship, we can produce the most appropriate measure of association.
14. Use Classical Hypothesis Testing to test the inferential nature of this relationship. What are the null and alternative hypotheses:

Null hypothesis:

Alternative hypothesis:

15. Generate a χ^2 test statistic with the statistical software or by hand.
16. Critical or rejection region [Using the χ^2 table in the textbook, start with 95% confidence.]
17. Can we be more confident? Report the highest level of confidence.
18. Interpretation and conclusion

Using the dataset "CPD mini", we have investigated whether the nature of the electoral system [prop] has any perceptible effect on the ideological balance of cabinet composition [gov_party] by generating a gamma of 0.43. IS this result significant?

19. What are the null and alternative hypotheses?

Null Hypothesis:

Alternative Hypothesis:

20. Generate a χ^2 test statistic
21. Which of the following is the most appropriate interpretation for these results? *Click all that apply*
 - The observed relationship ($\gamma=0.43$) between the type of electoral system and the ideological balance of cabinet composition exists in the population from which the sample was drawn.
 - At the 99.9% confidence level, we fail to reject the null hypothesis.
 - We are 100% confident that we have proven our theory.
 - At the 99.9% confidence level, we reject the null hypothesis.
 - The observed relationship ($\lambda=0.43$) between the type of electoral system and the ideological balance of cabinet composition exists in all of the samples of the population.
 - The observed relationship ($r=0.43$) between the type of electoral system and the ideological balance of cabinet composition only exists in the sample and not in the population from which the sample was drawn.
 - Substantively, there is no relationship between the electoral system and cabinet composition.
 - Substantively, as the electoral system moves toward proportional representation, left parties are more highly represented in cabinet composition.

We are interested in the level of internet censorship efforts in relation to the amount of internet access. That is, perhaps the level of access to the internet really doesn't tell us much about governments' efforts to censor it. Let's use the 'DS - Gamma Lambda' dataset.

22. What are the response categories for a variable about efforts of internet censorship [Internet_Censor]:

If you look in the codebook, they give a fuller description [the original variable 'v2mecenef' comes from the V-Dem dataset: <https://www.v-dem.net/documents/55/codebook.pdf>]:

- 'The government successfully blocks Internet access except to sites that are progovernment or devoid of political content' (1)
- 'The government attempts to block Internet access except to sites that are progovernment or devoid of political content, but many users are able to circumvent such controls' (2)
- 'The government allows Internet access, including to some sites that are critical of the government, but blocks selected sites that deal with especially politically sensitive issues' (3)
- 'The government allows Internet access that is unrestricted - with the exceptions mentioned in the description (4).'

So, one question is whether the level of aggregate internet access explain the efforts to control or constrain its use? Produce a cross-tabulation of internet censorship efforts and internet access (as a dichotomous variable <50%/>50%).

23. What are the null and alternative hypotheses?

Null Hypothesis (H_0):

Alternative Hypothesis (H_a):

24. Generate the appropriate measure of association.

25. Generate a χ^2 test statistic.

26. Given an χ^2 Critical Value $\alpha=0.05$; $df=3$]: 7.82. What is the interpretation?

[By hand] Are more populous countries more economically developed? A sample of countries produces the following joint distribution of population size (in millions of people) and economic development (low, medium, high).

If Gamma gives us 0.36, use Classical Hypothesis Testing to test the inferential nature of this relationship.

27. What are the null and alternative hypothesis?

Null hypothesis:

Alternative hypothesis:

28. Generate a χ^2 test statistic

Observed Frequencies:

Political Analysis Workbook

		Size of Country by Population (in millions)			
		<i>Small (under 1 m)</i>	<i>Moderate (1–29 m)</i>	<i>Large (30 m+)</i>	Total
Economic Development	<i>Low</i>	3	18	1	22
	<i>Medium</i>	3	35	16	54
	<i>High</i>	2	18	10	30
	Total	8	71	27	106

Expected Frequencies:

		Size of Country by Population (in millions)			
		<i>Small (under 1m)</i>	<i>Moderate (1–29m)</i>	<i>Large (30m+)</i>	Total
Economic Development	<i>Low</i>				
	<i>Medium</i>				
	<i>High</i>				
	Total				

29. Critical or rejection region [Using the χ^2 table in the textbook, start with 95% confidence.]

30. Interpretation and conclusion

Datasets

Comparative Political Data Set

- Armingeon, Klaus, Sarah Engler, Lucas Leemann and David Weisstanner. 2024. *Comparative Political Data Set 1960-2022*. Zurich/Lueneburg/Lucerne: University of Zurich, Leuphana University Lueneburg, and University of Lucerne:
 - Available here: www.cpbs-data.org/
 - *Political Analysis* dataset: "CPD mini"

DS - Gamma Lambda dataset

- Prepared for *Political Analysis*

V-Dem

- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. "V-Dem [Country-Year/Country-Date] Dataset v6.2." Varieties of Democracy (V-Dem) Project.
 - Available here: <https://www.v-dem.net/vdemds.html>
 - *Political Analysis* dataset: "V-Dem mini"

World Values Surveys

- WVS: Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 4.0*. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVS Secretariat. doi:10.14281/18241.18
 - Available here: <https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp>
 - *Political Analysis* dataset: "WVS mini"

QUESTIONS FOR CHAPTER 12: The Central Limit Theorem

The Central Limit Theorem is the link between the reality in which we are interested and the statistical techniques that we use to assess it. Many textbooks will mention the CLT and move on. In fairness to you, there are challenges to this reliance. As we confront its implications for making inferential claims – both theoretically and empirically – it is worth making clear what it is, what it does, and the limitations of using it.

We will not use statistical commands to produce the results for this chapter.

1. Why is the Central Limit Theorem (CLT) fundamental to inferential statistics?
2. How does the Standard Normal distribution serve as the foundation for many statistical tests?
3. What role does probability play in linking sample estimates to population parameters?
4. How does a Z-distribution and Z-scores standardize variables so that they can be directly compared?
5. What are the three useful properties of the Normal distribution (*click all that apply*)?
 - It is symmetric
 - We can infer from it
 - It is bell-shaped
 - The mean is 1
 - The total area under the curve is 1
 - It has fixed proportions under the curve.
6. Insert the correct level of confidence:
The 3.295 z-score creates a critical value for ____ confidence
The 1.96 z-score creates a critical value for ____ confidence
The 2.575 z-score creates a critical value for ____ confidence
7. The Central Limit Theorem makes a number of strong assumptions. Which of the following are specific assumptions of the CLT? *Click all that apply.*
 - Samples from the sampling distribution are actually samples of samples.
 - The distribution of the sample means approaches a normal distribution as the sample size increases, regardless of the population distribution, as long as the sample size is sufficiently large (usually more than 30 observations)
 - All samples from the population are normally distributed.
 - Large and random samples do not produce normally distributed distributions
8. Suppose you surveyed 1,500 Argentinian adults and asked them how they felt about tax cuts for the rich. We used what is known as a ‘feeling thermometer’. On a scale of

Political Analysis Workbook

1 to 100 in which 1 is 'extremely negative' and 100 'extremely positive.' We find that the average thermometer rating for the sample is 24.3 and the standard deviation for the sample is 12.9.

At 95% confidence, what is the margin of error?

- 1.962
- $12.9/\sqrt{1500}$
- $(2.575)(24.3/\sqrt{1500})$
- $(1.962)(12.9/\sqrt{1500})$
- none of these

9. What is our confidence interval at 95%

- 21.99 to 30.01
- 23.98 to 25.21
- 22.88 to 25.32
- 23.65 to 24.95
- 21.79 to 26.55

10. What is the correct interpretation of our confidence interval?

11. Below are five more countries and their measures of corruption. Standardize each (to z-scores) and compare each country's two different measures of corruption.

Country	Corruptions Perception Index	Political Corruption Index
France	30	0.063
Belarus	56	0.321
Nigeria	73	0.815
United Kingdom	18	0.039
United States	25	0.135

Recall:

- CPI: mean: 57.9 with a standard deviation of 18.5
- PCI: mean: 0.512 with a standard deviation of 0.286

	$z = \frac{(x_i - \bar{x})}{s}$	
	Corruptions Perceptions Index Z-score	Political Corruption Index z-score
France		
Belarus		
Nigeria		
UK		

US		

12. Which countries have similar scores? Which have dissimilar ones? Do you see a pattern between the countries in which the scores are similar and dissimilar?

13. Below are five countries and their measures of democracy. Standardize each (to z-scores) and compare each country's two different measures of democracy.

- Voice and Accountability*: mean: 0.299 with a standard deviation of 0.695. Range from -2 [lowest] to 2 [highest].
- Political Rights* (Freedom House): mean: 2.67 with a standard deviation of 1.88. Range from 1 [most free] to 7 [least free].

Country	Political Rights [Freedom House]	Voice and Accountability
Ukraine		
Slovenia		
Poland		
Bulgaria		
Czech Republic		

	Political Rights Z-score	V&A z-score
8. Ukraine		
9. Slovenia		
10. Poland		
11. Bulgaria		
12. Czech Republic		

14. Which countries have similar scores? Which have dissimilar ones? Do you see a pattern between the countries in which the scores are similar and dissimilar?

15. Below are five countries and their measures of effective number of parties at the votes and seats levels. Standardize each (to z-scores) and compare each country's conversion of votes to seats. Are they similar or different? Why do you think that is?

- Effective number of parties on the **votes** level*: mean: 5.41 with a standard deviation of 2.19. Range from 2.1 to 13.3.
- Effective number of parties on the **seats** level*: mean: 2.67 with a standard deviation of 1.88. Range from 1.97 to 8.14.

Country	Effective number of parties on the votes level	Effective number of parties on the seats level
---------	---	---

Political Analysis Workbook

Australia	3.69	2.44
Canada	3.33	2.50
Iceland	6.09	5.10
New Zealand	3.28	2.96
Switzerland	5.84	4.90

	Effective number of parties on the votes level Z-score	Effective number of parties on the seats level z-score
19. Australia		
20. Canada		
21. Iceland		
22. New Zealand		
23. Switzerland		

16. Which countries have similar scores? Which have dissimilar ones? Do you see a pattern between the countries in which the scores are similar and dissimilar?

17. The variable 'Conflict Intensity' from the Quality of Government dataset ranges from 1 – no conflict – to 10 – widespread violent conflict and is approximately normally distributed with a mean of 5.4 and a standard deviation of 2.1 (n = 54).

- Using what we know about the Z-distribution, what are the probabilities of randomly observing a country with conflict intensity less than 5.4?
- Greater than 7?
- Less than 5?
- Between 5.4 and 6?
- Between 4 and 7?

Less than 5.4		
Greater than 7		
Less than 5		
Between 5.4 and 6		
Between 4 and 7		

18. The variable 'Ethnic Fragmentation' – also from the Quality of Government dataset – ranges from 0 in which a country is perfectly homogeneous to 1 in which a country is highly fragmented. It is normally distributed with a mean of 0.460 and a standard deviation of 0.254 (n = 61).

Political Analysis Workbook

- a. Using what we know about the Z-distribution, what are the probabilities of randomly observing a country with ethnic fragmentation greater than 0.460?
- b. Less than 0.400?
- c. Greater than 0.500?
- d. Between 0.460 and 0.750?
- e. Between 0.400 and 0.750?

Greater than 0.460		
Less than 0.400		
Greater than 0.500		
Between 0.460 and 0.750		
Between 0.400 and 0.750		

19. Over the past decade, the mean Gini Index score is 31.82 with a standard deviation of 5.09 (with a lower score being less inequality).
 - a. If the United States reported a 38 on the Gini index, what is the probability of randomly observing a country with a higher Gini Index score?
 - b. What is the probability of randomly observing a country with a Gini Index score less than 30?
 - c. What is the probability of randomly observing a country with a Gini Index score greater than 40?
 - d. What is the probability of randomly observing a country with a Gini Index score less than 25?
 - e. What is the probability of randomly observing a country with a Gini Index score between 25 and 45?
 - f. What is the probability of randomly observing a country with a Gini Index score between 25 and 30?

Greater than 38		
Less than 30		
Greater than 40		
Less than 25		
Between 25 and 40		
Between 25 and 30		

20. In 2016, according to the Comparative Political Dataset, the vote share of the largest Green Party has a mean of 2.71 and a standard deviation of 3.43.
 - a. What is the probability of randomly observing a country 5% or more vote share?
 - b. What is the probability of randomly observing a country 2% or more vote share?

Political Analysis Workbook

- c. What is the probability of randomly observing a country 7.5% or less vote share?
- d. What is the probability of randomly observing a country with 2 and 5% vote share?
- e. What is the probability of randomly observing a country with 5 and 7.5% vote share?

Greater than 5		
Greater than 2		
Less than 7.5		
Between 2 and 5		
Between 5 and 7.5		

QUESTIONS FOR CHAPTER 13: Inference for Interval-level Variables

Student's *t* is the workhorse for inference. We will see *t*-tests and *t*-scores and *t*-values for the rest of our study. So, let this be an invitation to spend some time on this and get it right.

I do want to underscore one thing (well, three things about one thing). Our preferred point estimates should have three desirable properties.

- It would be *unbiased* meaning that the point estimate of the sample is 'equal' to the parameter.
- It would be *consistent*. That is, our sample estimate approaches the population parameter as the sample size increases.
- It would be *efficient*. Estimates with smaller distributions indicate more efficient estimates than other potential estimates.

To produce unbiased, consistent, and efficient point estimates of population parameters from our sample we combine the characteristics of large and random samples and exploit the known behaviours of probability (often with theoretical justification of the Central Limit Theorem). *Et voilà*.

The statistical commands that you will learn and use in this chapter are inferential commands. They will tell the statistical software to determine the statistical significance of relationships we determine using measures of association. It is becoming increasingly important for you to understand what the output is telling you about the substantive nature of the variables under investigation.

1. Why do we use the *t*-distribution instead of the Normal distribution for small samples?
2. What is the conceptual difference between a one-tailed and a two-tailed *t*-test?
3. Why do we often assume unequal variances in a difference of means *t*-test, and how does this affect the results?
4. What does the difference of means test actually test?
5. **[By hand]** We collected data on corruption in 28 democracies. Their mean level of corruption, as measured by Transparency International's Corruption Perceptions Index, is 6.48 with a standard deviation of 1.99 (with a higher score being less corrupt). What is our population estimate with a 95% confidence interval?
6. Another researcher comes along with more data and now our sample size is 65. The mean and standard deviation stay the same. Recalculate the confidence interval and describe any changes (recall that the sample size also changes the *t*-score).
7. Using another sample of 65 countries, we find the mean level of corruption is still the same but our measure of dispersion has gotten smaller (std.dev. = 1.69). Recalculate the confidence interval and describe any changes.

8. Returning to the previous study with a mean of 6.48 and a standard deviation of 1.99 (N = 65), let's calculate a 99% confidence interval.
9. Let's calculate a 99.9% confidence interval of these same data. Consistent with the previous question, what happens to the confidence interval?
10. **[By hand]** The difference in support for the EU between West and East citizens is interesting and important. A further question would be whether East Europeans' support has in fact dropped over the last decade. If we look back to the European Social Survey of 2008, we can again examine the question of whether they think European unification has gone too far (0) or could go further (10).

From the textbook, we know that the mean level for European unification for Eastern respondents (N = 9,821) is 5.09 with a standard deviation of 2.72. The East European countries included in the ESS 2008 include Bulgaria, Croatia, the Czech Republic, Estonia, Hungary, Latvia, Poland, Romania, Slovakia, and Slovenia. We find that the mean level for European unification for Eastern respondents (N = 15,365) is 5.84 with a standard deviation of 2.74.

Our first impression is that East Europeans in 2008 were substantially more supportive of EU unification (recall: European unification has gone too far (0) or could go further (10)). But there's only one way to find out.

Based on Classic Hypothesis testing, state the null and alternative hypotheses.

H₀:

H_a:

11. Generate the difference of means test statistic t-statistic using the formula:

$$t - statistic_{Diff\ of\ Means} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}}$$

12. What is the critical or rejection region (Table 13.2)?
13. What is the interpretation and conclusion?
14. Can you be more confident?
15. **[By hand]** Ok, East Europeans were more supportive in 2008 than 2018 (with 99.9% confidence) but less supportive than their Western counterparts in 2018 (with 99.9% confidence). Perhaps West Europeans felt the same – that is more supportive in 2008 than they did in 2018. Let's find out.
 - In 2018, the mean support was 5.251 with a standard deviation of 2.652 for Western respondents (N = 16,210).

Political Analysis Workbook

- In 2008, the mean support was 5.032 with a standard deviation of 2.500 for Western respondents (N = 21,622).

Based on Classic Hypothesis testing, state the null and alternative hypotheses.

H₀:

H_a:

16. Generate the difference of means test statistic t-statistic using the formula:

$$t - statistic_{Diff\ of\ Means} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}}$$

17. What is the critical or rejection region for 99.9% confidence?

18. What is the interpretation and conclusion?

19. **[By hand]** Of course, to be fully informed, we must ask if West Europeans had stronger support than East Europeans in 2008. Conduct a difference of means test to find out. What is the highest level of confidence that we can have?

- For West Europeans in 2008, the mean support was 5.032 with a standard deviation of 2.500 for Western respondents (N = 21,622).
- For East Europeans in 2008, the mean support was 5.837 with a standard deviation of 2.731 for Western respondents (N = 16,494).

Based on Classic Hypothesis testing, state the null and alternative hypotheses.

H₀:

H_a:

20. Generate the difference of means test statistic t-statistic using the formula:

$$t - statistic_{Diff\ of\ Means} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}}$$

21. What is the critical or rejection region for 99.9% confidence?

22. What is the interpretation and conclusion?

23. In a random sample from 2019, of 1,398 murders in the United States, 894 were gun related. This means that 64% of murders were perpetrated by someone with a gun.

a. Calculate a 95% confidence interval around this proportion.

$$BCI = \pi \pm (z - score_{Confidence\ Level}) \sqrt{\frac{\pi(1 - \pi)}{n}}$$

b. Interpret this result

24. Report this with maximum confidence, 99.9%. Interpret this result

25. Recall our 'survey' of 1,500 Argentinian adults from the previous chapter about how they felt about tax cuts for the rich in the previous chapter. We divided this group into those making an annual income less than the median income (roughly 40,094 USD in 2021) and those making more. We suspect that those making more will feel more warmly than those making less.

- For those making **less than the median**, the mean 'warmth' was 17.4 with a standard deviation of 8.6 (N = 955).
- For those making **more than the median**, the mean 'warmth' was 34.8 with a standard deviation of 30.1 (N = 545).

Use a difference of means test to determine whether there is a statistically significant difference between these groups on their warmth to tax cuts for the rich at 95% confidence.

a. Based on Classic Hypothesis testing, state the null and alternative hypotheses.

H₀:

H_a:

b. Generate the difference of means test statistic t-statistic using the formula:

$$t - statistic_{Diff\ of\ Means} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}}$$

c. What is the critical or rejection region for 95% confidence?

d. What is the interpretation and conclusion?

e. How confident can we be?

f. We are reminded by a colleague that the richer parts of society have always had 'more warmth' for tax cuts that are aimed at them than those who make less than the median income. Therefore, without calculating another test, would a one-tailed test increase or decrease your chance rejecting the null hypothesis? Explain.

26. It has long been assumed that democracies experience less income inequality [wdi_gini] than non-democracies [bmr_dem]. Using the 'QoG mini' dataset, calculate a difference of means test and interpret.
27. The following output refers to two variables in the 'QoG mini' dataset. One, is the country a democracy (or not) [bmr_dem] and two, the level of ethnic fractionalization ['fe_etfra'] (0 ethnically uniform -100 everyone is ethnically different). We decide to do a difference of means test on the level of ethnic fractionalization between the democracies and non-democracies. Generate a difference of means test and interpret.
28. Remember the question about whether there was any substantial regional differences in terms of how close individuals might feel to a political party? We used a nominal-level measure of closeness and a multinomial-measure of region: West, East, Mediterranean. What if we compare them using a difference of means t-test?
- Using the ESS mini dataset, test the difference in 'close to party' [close2: with 0 No, 1 Yes] with the three regional measures (recall, we can only directly test between two groups):
- a) WestEU
 - b) EastEU
 - c) MedEU
29. In Chapter 7, it didn't look like a big difference ["It doesn't look like there are big regional differences in whether feel close to a political party ($\lambda = 0.0367$). While in the West, there seems to be slightly fewer feeling close than not, the other regions are balanced the other way."]. Given the output above, how would you update that now?
30. What are the differences between our Chapter 7 Measure of Association (Lambda) and this set of difference of means t-tests? And does this affect the result?
31. Look at the Trust in European Parliament question in the 'ESS mini' data [trstep]. Using the statistical software, produce a...
- a. 95% confidence interval for the mean
 - b. 99% confidence interval for the mean
 - c. 99.9% confidence interval for the mean
 - d. Interpret the 99.9% confidence interval
32. Create 95%, 99%, and 99.9% confidence intervals for the mean of Electoral Democracy Index in the 'V-Dem mini' dataset [v2x_polyarchy] and interpret the 99% confidence interval.
33. Create 95%, 99%, and 99.9% confidence intervals for the mean of the percentage of adult citizens with right to vote in the 'V-Dem mini' dataset [v2eladltvt] and interpret the 95% confidence interval.

34. In the textbook, we saw that the role of a female head of state was important to gender equality. Look at the following variables in the 'V-Dem' dataset.

- v2x_genc1: Women civil liberties index
 - v2exfemhos: HOS female [with 0 No, 1 Yes]
-
- a. Are women's civil liberties higher in states with female heads of state? Generate a different of means t-test to find out.
 - b. How would you interpret this?
 - c. Do you see any potential problems?

Datasets

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG mini"

European Social Survey

- ESS Round 9: European Social Survey Round 9 Data (2018). Data file edition 3.1. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data Archive and distributor of ESS data for ESS ERIC. [doi:10.21338/NSD-ESS9-2018](https://doi.org/10.21338/NSD-ESS9-2018).
 - Available here: <https://www.europeansocialsurvey.org/data/>
 - *Political Analysis* dataset: "ESS mini"

V-Dem

- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. "V-Dem [Country-Year/Country-Date] Dataset v6.2." Varieties of Democracy (V-Dem) Project.
 - Available here: <https://www.v-dem.net/vdemds.html>
 - *Political Analysis* dataset: "V-Dem mini"

QUESTIONS FOR CHAPTER 14: An Introduction to Multiple Regression

Having previously been introduced to (bivariate) regression in Chapter 9, you are well prepared for multiple regression. The crucial difference is obviously the introduction multiple independent variables and the concept of statistical significance applied to multiple regression results. We can merge these with a simple observation. Using multiple regression makes statistical significance more meaningful as we are controlling for the effects of several independent variables simultaneously and both substantive and statistical significance are likely to be a more accurate portrayal of the relationship in which we are interested.

The statistical commands that you will learn and use in this chapter are a combination of descriptive and inferential commands. They will tell the statistical software to determine the substantive and statistical significance of relationships. You will need to understand what the output is telling you about the variables and relationships under investigation.

1. What are the three key elements of the regression equation?

$$Y = \alpha + \beta_1(X_1) + \beta_2(X_2) + \dots + \beta_n(X_n) + \varepsilon$$

2. What are the three key analytical abilities the multiple regression provides? Explain each.
3. Why do we not throw all the variables that we can into the multiple regression model?
4. What is meant by the phrase, 'holding all other variables constant' in our interpretation?
5. Why are multiple regression coefficients referred to as 'partials'?
6. How does the F-test differ from R^2 ?
7. Why does multiple regression not establish causation, even when controlling for multiple variables?
8. Why is adjusted R^2 often preferred over regular R^2 in multiple regression analysis?
9. What is the purpose of standardizing regression coefficients, and how does it relate to z-scores?
10. How do researchers decide which independent variables to include in a multiple regression model?
11. In the Box: 'Nested Models and the Coefficient of Determination (R^2)' in the textbook, calculate the overall explanatory improvement for Model 3 over Model 1 and 2. Also calculate the explanatory power of adding civic skills (comparing Model 3 to Model 2). Explain fully what this tells us.

A nation's GDP is not simply a function of industrial and agricultural output. Economic, societal and political factors constrain and encourage its growth as well. Below is a simple model to explain the level of a nation's GDP (in \$ thousands).

Dependent Variable:

GDP Range: 0.4–18.07 (in millions USD); Mean: 4.80; Standard deviation 4.90

Independent Variables:

Ethnic Heterogeneity: Provides an index of ethno-linguistic heterogeneity (a high score indicates a more heterogeneous society). Range: 0–0.9; Mean: 0.41; Standard deviation: 0.29

GINI Index: Income Inequality: Measures income inequality (0 corresponds to perfect income equality (i.e. everyone has the same income) and 1 corresponds to perfect income inequality (i.e. one person has all the income, while everyone else has zero income). Range: 24.4–62.9; Mean: 41.04; Standard deviation: 10.64

Human Development Index: Measures national levels of standard of living. Range: 0.275–0.944; Mean: 0.70; Standard deviation: 0.19

Urban Population: The percent of the population that lives in urban areas. Range: 6–100; Mean: 55.78; Standard deviation: 23.04

Let's examine elements of the analytic output.

GDP	Unstandardized Regression Coefficient	Standard Error	t	p > t
<i>Ethnic heterogeneity</i>	1.869	1.929	0.97	0.339
<i>Income inequality</i>	–0.168	0.049	–3.46	0.002
<i>Human development</i>	30.543	5.746	5.32	0.000
<i>% Population urban</i>	0.041	0.031	1.33	0.194
<i>Constant</i>	–11.671	5.471	–2.32	0.027
Number of observations	40			
F(4, 35)	38.41			
Prob. > F	0.0000			
R ²	0.8496			
Adjusted-R ²	0.8275			

12. Interpret the output for Ethnic Heterogeneity, including substantive and statistical significances.

13. Interpret the output for income inequality, including substantive and statistical significances.

14. Interpret the output for Human Development Index, including substantive and statistical significances.
15. Interpret the output for percent urban population, including substantive and statistical significances.
16. What does the F-test tell us?
17. What does the adjusted- R^2 tell us?
18. What is the regression equation?
19. You can make a prediction equation from the regression equation by inserting the means of all the values. Do this and predict the 'average' GDP.
20. Keep all of the variables at their mean except for Human Development. Make it the highest value in the range. What does this predict for GDP?
21. Some would argue to drop 'Government spending on education' from the analysis. There is some literature that says this could be important but clearly it is not here. What do you do?
22. What is the takeaway of this regression analysis?
23. Using the 'QoG mini' dataset, let's investigate the potential explanations for countries' levels of Government Effectiveness. We are interested to see whether the level of investment in making the government more 'electronic' is important to the effective function of government. To do so, we will regress "E-Government Index" [egov_egov] on the World Bank's World Governance Indicator, "Government Effectiveness" [wbgi_gee]. However, we have been investigating this relationship for some time and know of other potential explanations for the quality or level of government effectiveness.
 - One is the simple size of the country. Larger countries are harder to manage.
 - Another potential explanation is that richer countries are typically easier to manage.
 - Others still have argued that ethnically diverse countries are more difficult to manage.
 - Finally, we have long been exposed to the idea that a free press is good for governance by shining a light on its performance and accountability.

Using the following variables, let's consider the performance of our theorized relationship in the context of other competing explanations.

- Total population [wdi_pop]
- GDP per capita, PPP [wdi_gdpcappppcur]
- Ethnic Fractionalization [fe_etfra]
- Press Freedom Index [rsf_pfi]

Note: In order to make the results interpretable, it will be helpful to generate new variables that shrink the size of total population and GDP *per capita*. Use the following commands. Be sure to use the new variables [totalpop and GDPpc] in your analysis

gen totalpop=wdi_pop/1000000

gen GDPpc= wdi_gdpcappppcur/1000

Step One: Choose the most appropriate research question for our project.

- To what extent does having a more educated population produce better performing governance?
- Does investment in AI makes democracy stronger?
- To what extent does online governance produce better performing governance?
- How can we make governance more effective?
- Why do some countries perform better than others?
- Which is the best indicator of better performing governments?

24. *Step Two:* Identify Variables:

- a. Which of the following is not an independent variable?
 - Total population
 - GDP *per capita*
 - Ethnic Fractionalization
 - Press Freedom Index
 - Government Effectiveness
 - E-Government Index
- b. What are the minimum and maximum values for Government Effectiveness?
 - -1.77 to 2.22
 - -2.30 to 2.22
 - -2.48 to 2.15
 - -2 to 2
 - 0 to 2.25
- c. What are the mean and standard deviation of GDP *per capita*?
 - 20.74; 48.61
 - 21.36; 20.05
 - 43.58; 20.42

Political Analysis Workbook

- 12.30; 19.10
- 28.23; 48.68

d. What is the highest amount of e-governance?

- 23.94%
- 48.45%
- 61.73%
- 90.53%
- It is not possible to determine

e. According to the actual codebook, the Press Freedom Index is not technically an indicator of the quality of a country's media. Instead, it measures (*click all that apply*):

- all human rights violations that affect press freedom
- none of these
- the amount of freedom journalists and the media have in each country
- low scores reflect higher levels of press freedom.
- the efforts made by governments to see that press freedom is respected

25. Step Three: Pearson's Product Moment Correlation Coefficients:

a. How would you describe the correlation between Government Effectiveness and GDP *per capita*?

- Substantial negative correlation
- Moderate positive correlation
- Low negative correlation
- Very strong positive correlation
- Negligible

b. How would you describe the correlation between Government Effectiveness and E-Governance?

- Very strong positive correlation
- Substantial positive correlation
- Moderate positive correlation
- Low negative correlation
- Negligible

26. Step Four: Multiple Regression

a. Write out the correct regression equation

- b. Which independent variables are statistically significant at or above the 95% confidence level (*click all that apply*)?
- ☐ Government Effectiveness
 - ☐ Total population
 - ☐ GDP *per capita*
 - ☐ Ethnic Fractionalization
 - ☐ Press Freedom Index
 - ☐ E-Government Index
- c. Is e-government statistically significant and if so, at what confidence level?
- ☐ Yes, 95% confidence
 - ☐ Yes, 99% confidence
 - ☐ Yes, 99.9% confidence
 - ☐ No, 98% confidence
 - ☐ No.
27. *Step Five:*
- a. The F-test tells us (*click all that apply*):
- ☐ There are 108 degrees of freedom
 - ☐ There are 5 dependent variables
 - ☐ These independent variables are not simultaneously indistinguishable from zero.
 - ☐ We are 99.9% confident in the result of the F-test
 - ☐ That we must fail to reject the F-test null hypothesis that the betas are simultaneously equal to zero
 - ☐ That 5 of the 108 letters in the names of the independent variables are F's.
- b. What is the correct interpretation of the most appropriate coefficient of determination?
- ☐ 82.64% of the variation in Government Effectiveness can be explained by variation in these independent variables.
 - ☐ 81.83% of the variation in Government Effectiveness can be explained by variation in the statistically significant independent variables.
 - ☐ 82.64% of Government Effectiveness can be explained by these independent variables.
 - ☐ We only need to explain about 20% of the remaining variation in Government Effectiveness.
 - ☐ 80.2% of the variation in Government Effectiveness can be explained by variation in these independent variables.
 - ☐ The coefficient of determination is not presented for us to interpret.

28. *Step Six: Interpretation:*

- a. What is the correct interpretation for Press Freedom Index?
- ☐ For every increase in Press Freedom Index (indicating less press freedom), Government Effectiveness decreases 0.0012 units ($\alpha < 0.05$), controlling for all other variables.

- For every 1-unit increase in Press Freedom Index (indicating less press freedom), Government Effectiveness decreases 0.012 units ($\alpha < 0.01$), controlling for all other variables.
- For every 1-unit increase in Press Freedom Index (indicating less press freedom), Government Effectiveness decreases 0.012 units ($\alpha < 0.001$), controlling for all other variables.
- For every 1-unit increase in Press Freedom Index (indicating less press freedom), Government Effectiveness increases 0.012 units ($\alpha < 0.05$).
- For every 1-unit increase in Press Freedom Index (indicating less press freedom), we can causally predict Government Effectiveness decreases 1.2% ($\alpha < 0.05$), controlling for all other variables.

b. Given the results of our analysis, what can we say about our research question (*click all that apply*)?

- GDP *per capita* is the most important variable to explain government effectiveness.
- Not all of the independent variables were statistically significant so there is something wrong with our regression model
- Having more online government seems to be related to increased government effectiveness, even when controlling for other explanations.
- Ethnically diverse countries do not have high government effectiveness.
- Counter to our expectations, large countries can be managed effectively as there is no difference between small and large countries.
- The measures of model fit tell us that our model performs perfectly.
- We have found evidence of our expected or theorized relationship.
- There is no reason to continue to investigate this question as we have found a statistically significant relationship between our variables of interest.

What explains the amount of CO₂ emissions in a country? We look at the 'QoG max' dataset and use a multiple regression modelling approach.

29. *Step One: A Research Question.*

30. *Step Two: Identify Variables*

- a. *What is the range, mean and standard deviation of the dependent variable, Emissions?*

Independent Variables:

- *Renewable energy output:* [wdi_elerepower: Renewable electricity output (% of total electricity output)]
 - *Electricity from coal:* [wdi_elprodcoal: Electricity production from coal sources (% of total)]
 - *Government spending on education:* [wdi_expedu: Government expenditure on education, total (% of GDP)]
 - *Military expenditure:* [wdi_expmil: Military expenditure (% of GDP)]
 - *GDP per capita:* [wdi_gdpcapppcur: GDP *per capita*, Purchasing Price Parity (\$)]
- b. What are the minimum and maximum values for Renewable energy?
- c. What are the mean and standard deviation of Electricity from Coal?

- d. What is the highest amount of government expenditure on the military (% of GDP)
31. *Step Three*:
- Produce correlations for all the variables..
 - Which three of the independent variables looks to be the most likely candidates for explaining CO₂ emissions?
32. *Step Four* asks for the regression equation.
- Produce regression output.
 - What is the regression equation?
33. *Step Five* asks us for measures of Model Fit. Describe and interpret.
34. *Step Six* asks for a comprehensive interpretation. Please include one paragraph – a table is not necessary.

35. Given the state of the literature, Brexit and what may be described as democratic backsliding in both Poland and Hungary, we are interested in the relationship between EU unification and citizens' orientations to democracy as a system of government. Thus, a research question [*Step One*] might be: 'To what extent do citizens' preferences for democracy shape their views about EU unification?'

Step Two asks us to identify the variables. Using the 'EES mini' dataset, let's take a look at European citizens' support for the ongoing EU project. Using survey data from all 27 EU member states and the United Kingdom, we can propose a straightforward model of support with four key independent variables capturing citizens' personal ideological position, their pro-/anti-market stance, pro-/anti-democracy stance and pro-/anti-immigration stance. Let's use the following variables to investigate our research question.

Dependent Variable (DV): European Unification [EU Unification]:

- Some say European unification should be pushed further. Others say it already has gone too far. What is your opinion? [Q23 → EUU]; Range: 0–10, where 0 means unification 'has already gone too far' and 10 means it 'should be pushed further'.

Independent Variables (IVs):

- Ideological Self-Placement:** In political matters people talk of 'the left' and 'the right'. What is your position? [Q11 → LRscale]; Range: 0 means 'extreme left' and 10 means 'extreme right'
- Support for the Market:** What do you think of state regulation and control of the economy? [Q14_1 → MktSupport]; Range: 0 means 'You fully favour state intervention in the economy' and 10 means 'You fully oppose state intervention in the economy'.
- Support for Democracy:** How important is it for you to live in a country that is governed democratically? [Q16 → DemSupport]; Range: 0 means 'Not at all important' and 10 means 'Absolutely important'
- Immigration Position:** What do you think of immigration? [Q14_5 → IMPos]; Range: 0 means 'Fully in favour of a restrictive policy on immigration' and 10 means 'fully opposed to a restrictive policy on immigration'

Political Analysis Workbook

- a. Generate the correlations for all the variables.
 - b. Describe these correlations and provide an intuition of how they relate to one another.
36. *Step Four* asks for the regression equation.
- a. Produce regression output.
 - b. What is the regression equation?
37. *Step Five* asks us for measures of Model Fit. Fully report the measures of Model Fit.
38. *Step Six* asks for a comprehensive interpretation. Please include one paragraph – a table is not necessary.

Datasets

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG mini"
 - *Political Analysis* dataset: "QoG max" – with more countries

European Election Studies: European Parliament Election Study 2019, Voter Study

- Schmitt, Hermann, Hobolt, Sara B., Brug, Wouter van der, & Popa, Sebastian A. (2022). European Parliament Election Study 2019, Voter Study. GESIS, Cologne. ZA7581 Data file Version 2.0.1.
 - Available here: <https://doi.org/10.4232/1.13846>
 - *Political Analysis* dataset: "EES mini"

QUESTIONS FOR CHAPTER 15: Extensions to Multiple Regression

The technique of multiple regression seems quite powerful, widely useful, and adaptable to various conditions. But it is also available for an abundance of after-market accessories as extension to 'basic' multiple regression. Among the most common adaptations or extensions, dummy variables are uniquely valuable by allowing the incorporation of non-interval level data as independent variables. We also make the technical and conceptual leap of including the relationships among independent variables through interactions. Finally, we have considered the nature of including various forms of variables that do not lend themselves readily to obvious or extant measurement though proxies.

The statistical commands that you will learn and use in this chapter are a combination of descriptive and inferential commands. They will tell the statistical software to determine the substantive and statistical significance of relationships. You will need to understand what the output is telling you about the variables and relationships under investigation.

1. Why are dummy variables necessary for including nominal- and ordinal-level variables in multiple regression? And how do they work?
2. What is the importance of choosing a reference category when using dummy variables?
3. What is the difference between an additive index and a latent variable in statistical analysis?
4. Why is model specification critical when using multiple regression with dummy variables and interactions?
5. In a multiple regression with only one independent dummy variable, the output is the same as a difference of means t-test. Why do the regression coefficient and t-value of the dummy variable change when other independent variables are added?
6. Another form of latent variable analysis is factor analysis. Briefly describe factor analysis.
7. Interacting variables produces a contingent result in the regression. What does this mean?
8. Often in research, we have to rely on the respondents to tell us what they did. For example, if we want to know if people were exposed to a particular event or were paying attention during an important period, we can ask them about their media use habits.

For example, if they watched something on the television – or if they listened to radio or read a newspaper – about news/politics/current affairs on an average weekday. Let's assume that our current research question is not interested in the differences *between* these media but rather how much the respondent uses media *in general*.

Perhaps we can consider making a single variable called, the 'Media Consumption' variable, from these three. The response categories for television, radio and newspaper are (per day): (1) less than ½ an hour, (2) ½ an hour to 1 hour, (3) between 1 and 1½ hours, (4) 1½ hours to 2 hours, (5) between 2 and 2½ hours, (6) 2½ hours to 3 hours and (7) more than 3 hours. These the three variables in 'Media Consumption' have a very high Cronbach's alpha (e.g. >0.90).

- a. What are some advantages of using this new additive variable as an independent variable to explain EU support?
 - b. Disadvantages:
9. From the 'ESS mini' dataset, we want to look at how people view political objects. We are interested in how people feel about the key political institutions of politics. A common approach is to ask people about their trust in government. However, this asks respondents to clump a wide variety of things into 'government'. We want to articulate these disparate references into a panoramic view of government rather than simply ask one broad question.
- Trust in country's parliament [trstprl]
 - Trust in the legal system [trstlgl]
 - Trust in the police [trstplc]
 - Trust in politicians [trstplt]
 - Trust in political parties [trstprt]
- a. Generate a new variable by adding five variables and produce a resultant variable: 'trust in government' [TrustGovt]
 - b. Describe this variable:
 - c. What is the Cronbach's alpha of these variables? And what does this tell us?

These five variables move together very well and thus must be capturing similar facets or dimensions of government trust.

- d. What are the operational advantages of compiling this variable versus asking a generic 'trust in government' question?
10. Using the 'ESS mini' dataset, we might suspect that immigration positions are different between men and women and interact these positions with a simple gender variable.
- a. Produce the output of regressing Immigration position [imwbcnt], male [Gender: where 0 female, 1 male], and their interaction on EU unification [euftf]
 - b. What do the results suggest in terms of gender?
 - c. Produce a graph to show this:
11. Using the 'ESS mini' dataset, we can see if there is an inequality of incomes related to gender. Below is a simple model to study this question.

Dependent Variable: (Household) Income [hinctnta]: Household's total net income, all sources

Independent Variables:

- Gender [Gender]: 0 Female, 1 Male]
- Size of household [hhmmb]: Number of people living regularly as member of household];
- Education [EDUC_std]: Highest level of education, ES - ISCED
- Union Member [mbtru]: Yes, currently; Yes, previously; No.

- a. Produce the regression output (one note, make sure that the reference category for mbtru is 'no')
 - b. What is the regression equation?
12. Interpret the model
 - a. Interpret the output for gender (male), including substantive and statistical significances.
 - b. Interpret the output for size of household, including substantive and statistical significances.
 - c. Interpret the output for education, including substantive and statistical significances.
 - d. Interpret the output for the two dummy variables for union membership, including substantive and statistical significances.
13. You can make an (OLS) prediction equation from the regression equation. Insert the means for the non-dummy variables and let's predict income for a female, current union member.
14. Insert the means for the non-dummy variables and let's predict income for a male, current union member.
15. Insert the means for the non-dummy variables and let's predict income for a male, non-union member.
16. What is the takeaway of this regression analysis?
17. One thing you may not have considered is that members of unions tend to be supportive of institutions as mechanisms of enforcement and support. That is, once a member, always a member. Use the variable [mbtru] to take a look at this.
 - a. Generate a new, dichotomous variable for those who have ever been in a union or current are versus those who have never been a member.
 - b. Include this dummy variable in the regression for Household Income. Produce the output here
 - c. What do you find?
18. Based on the previous results, one might propose that education may be more beneficial to women in terms of income; that is, higher education provides a higher income pay-off for women than it does for men. We can model this by interacting the gender dummy variable with the education variable. We rerun the model and the output is below.
 - a. Produce a table of the output.
 - b. What is the regression equation?
 - c. How have the other variables changed?
19. Given that gender is a dummy variable, we can however interpret the output for education separately for men and women.

- a. Given that the main effects and interaction are all statistically significant, we can interpret the results for men and women, holding the other variables at their means (assume that both are current union members):
- b. Produce a graphical representation of this relationship that represents the effect of the interaction. What does this tell us about our proposed importance of education to women?

Education does not produce an effect that compensates for the gender disparity in income. These data further suggest that higher education is in fact more costly to women (evidenced by the fact that the interaction coefficient is positive, that is, it helps men i.e. 0.03). Over the range of education, women with higher education are penalized further.

20. We haven't looked at the size of household. The regression coefficient tells us that as the size of a household grows by 1 person, income should increase by 0.71 deciles. Compare a single person to a family of four. Holding education at its mean, let's assume we are talking about males who are currently union members.

21. Previously, we used the 'QoG mini' dataset to investigate the potential explanations for countries' levels of Government Effectiveness. We found that, while there are a number of potential explanations, the level of E-governance is important to the effective function of government. Let's confront a slightly larger dataset. Using the 'QoG max' dataset, look at the following variables in which:

Dependent Variable:

wbgi_gee: World Bank's World Governance Indicator, "Government Effectiveness"

Independent Variables:

wdi_expedu: Government expenditure on education, total (% of GDP)

wdi_pop: Total population

wdi_gdpcappppcur: GDP per capita, PPP

fe_etfra: Ethnic Fractionalization

rsf_pfi: Press Freedom Index

To simplify the output, we add the following two (2) transformations [*generate a new variables*]:

- `totalpop=wdi_pop/1000000`
- `GDPpc= wdi_gdpcappppcur/1000`

- a. Produce the regression output

We presented our results to a small group and someone asked, what now seems, a simple question. Does it matter if the country is a democracy or not?

Using the dichotomous variable [bmr_dem] that categories countries as democracies [1] or not [0], determine the answer to this question.

- b. Does being a democracy substantively impact Government Effectiveness in the context of the full model?
 - Yes
 - No
 - It is impossible to determine
 - The adjusted- R^2 tells us that it does
 - Only if the F-test is insignificant
- c. What happens to other parts of the model when we include the dichotomous democracy variable for the model?
 - Total population becomes statistically significant.
 - The regression coefficient for GDP per capita shrinks
 - The Press Freedom Index becomes statistically insignificant
 - The Adjusted- R^2 increases by 5 percentage points
 - The F-test improves

Perhaps we are not considering the full potential of being a democracy. Let's consider some new research in the literature that suggests democracies require a robust free media to have a substantive improvement in governing effectiveness.

22. One way to test this is to interaction the dichotomous democracy variable with the Press Freedom Index. In this way,

- we can determine how press freedom shapes being a democracy.
- we can determine the impact of being a democracy on press freedom and government effectiveness simultaneously.
- we can determine to what extent government effectiveness is shaped by having a high level of press freedom.
- we can determine the conditional impact of being a democracy and the level of press freedom on government effectiveness.
- we can finally determine the comprehensive impact of being a democracy on government effectiveness.

23. Produce the regression output

- a. Which independent variables are statistically significant in this interaction model? (*click all that apply*)
 - Government expenditure on education, total (% of GDP)
 - Total population
 - GDP *per capita*
 - Ethnic Fractionalization
 - Press Freedom Index
 - Democracy or not
 - Press Freedom and Democracy interaction

b. What does the inclusion of the interaction tell us?

- The interaction didn't work
- The interaction is more important than GDP *per capita*
- Nothing as the main effects should both be statistically significant for the interaction to be relevant
- There may be potential for a conditional solution for democracy and press freedom to the model.
- Only democracies can have a free press

c. One way to determine the impact of including an interaction term is to graph the marginal effects of the independent variables included in the interaction. Do this:

d. Recall that the Press Freedom Index is measured as [0] Free Press and [100] Not Free Press. Given the resulting graph, we can we say about the interaction term? (*click all that apply*)

- For non-democracies, the level of press freedom is inconsistently related to government effectiveness.
- For democracies, it appears to be important to have a high level of press freedom (indicated by lower PFI scores) in order to have high levels of government effectiveness.
- At both medium and low levels of press freedom (indicated by middling and higher PFI scores), there isn't a clear statistically significant difference between democracies and non-democracies in terms of its impact on government effectiveness.
- In order to be 'more free' than non-democracies, democracies must have a level of press freedom no less than 50.
- For both democracies and non-democracies, press freedom is important for the level of government effectiveness.

24. Using the 'ESS mini' dataset, run this regression (set the reference category for 'mbtru': yes, currently).

DV: Trstprl

IVs agea; hinctnta; lrscale; EDUC_std; Gender; mbtru

a. Describe what we are regressing

25. One thing that bothers you about the model of Trust in [country's] Parliament above is that, going from liberal to conservative [lrscale] is positively related to institutional trust. That is, conservatives are more likely to have a higher level of institutional trust than liberals. To you, maybe a bit cynically, that it has really to do with socio-economic statuses (i.e.: income [hinctnta]). You decide to interact these two interval-level variables.

a. Produce that regression output:

26. We find that the interaction between income and left-right scale to be negative suggesting that they interact in interesting ways. The best way to see this is graph it. Do this for 0) far-left, 5) center, and 10) far-right at four levels ('deciles') of income 1, 4, 7, and 10.
27. Keeping in mind that the x-axis is income and the lines represent left-right ideological location, what can we conclude about their interaction on institutional trust (y-axis) based on the output and the graph? (*click all that apply*)
- Men with greater income are more likely to trust institutions.
 - Income is positively related to institutional trust at 99.9% confidence
 - The more conservative respondents are, the more likely they are to have higher levels of institutional trust ($\alpha=0.001$).
 - The level of institutional trust for ideologically moderate respondents (those in the ideological center) with low income is low.
 - A statistically significant interaction effect suggests that the main effects of income and left-right location are not important.
 - Given the slopes of the lines, it seems that the impact of income on left-leaning respondents' levels of institutional trust is potentially greater than on right-leaning respondents' levels.
 - If both main effects of the interaction – income and left-right location – are statistically significant, the interaction cannot be statistically significant.
 - At the highest level of income, there is much less difference between left-leaning, center, and right-leaning respondents' levels of institutional trust than the levels of institutional trust of respondents at lower levels of income.

Datasets

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG max"

European Social Survey

- ESS Round 9: European Social Survey Round 9 Data (2018). Data file edition 3.1. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data Archive and distributor of ESS data for ESS ERIC. [doi:10.21338/NSD-ESS9-2018](https://doi.org/10.21338/NSD-ESS9-2018).
 - Available here: <https://www.europeansocialsurvey.org/data/>
 - *Political Analysis* dataset: "ESS mini"

QUESTIONS FOR CHAPTER 16: Issues with Multiple Regression

Multiple (linear) regression is one of the most widely used tools in statistical analysis, but it is not perfect. There are crucial assumptions—such as linearity, independence, homoscedasticity, and normality—to ensure that regression results are interpretable and reliable. Violating these assumptions can lead to biased estimates, incorrect inferences, or misleading predictions. This chapter provides exercises to help you identify, test, and even correct these assumptions in practice. By working through real-world data scenarios, you will develop a deeper understanding of when linear regression is appropriate and how to diagnose potential violations.

The statistical commands that you will learn and use in this chapter are diagnostic. They are there to evaluate the performance of your choice of regression model. You will need to understand what the output is telling you model assessment.

1. What is the most correct response to a high level of multicollinearity?
 - Ignore the problem
 - Combine the variables
 - Drop one of the variables
 - It depends on the question and the state of the literature.
2. One of the means to identify multicollinearity is simply to look at the correlations between all of the independent variables. While seemingly a good approach, why is this approach potentially not sufficient?
3. What are the five assumptions of the linear regression model?
4. Normal variance for the error term is assumed. What does this assumption mean for the model itself?
5. Another assumption is that the expected value of the error term is zero. What is this usually a sign of?
6. When we refer to optimizing the included variables and their performance, we are referring to:
 - Model Specification
 - Model Fit
 - Gauss–Markov Optimization
 - Model Unbiasedness
 - Model Optimizing
7. If error terms are correlated with one another, this can cause the problem of:
 - Autocorrelation
 - Serial correlation
 - Spatial autocorrelation
 - *All of these*

- None of these
8. When all five assumptions are reasonably met, we know that the estimating procedure of OLS is the best for linear regression. What does 'best' mean?
 9. Endogeneity refers to a problem for causal inference. What is the problem?
 10. Very often in social science we are unable to directly measure a variable of interest. One way to approach this problem is to use proxy variables. What are proxy variables?
 11. Instrumental variable estimation takes a similar approach to measuring variables; however, the motivation for using instrumental variables is different. Why would we choose to use an instrumental variable approach?

Instrumental variable estimation addresses the problem in which an independent variable is correlated with the error term by substituting a variable that is uncorrelated with the error term and highly correlated with the problematic independent variable.

12. The term 'well-behaved' is used in reference to the error term. What does well-behaved mean?
 - The error term produces a clear signal with little noise.
 - The error term is measured without error.
 - The error term is normally distributed, has a mean of zero, and has constant variance.
 - The error term has a mean of zero with a variance of $1/\pi$.
 - The error term is a straight line and is not correlated with the independent variables.
13. 'Fixed in repeated samples' or 'error-in-variables' – Assumption 5 – refers to which problem?
 - All of the independent variables are measured incorrectly.
 - The independent variables are correlated with the error term.
 - The error term has a constant variance.
 - The dependent variable goes up unexpectedly.
14. The linear regression requires a set of assumptions to be reasonably met to perform well. The first two assumptions refer to problems often found in the systematic part of the regression equation [$\alpha + \beta x$], while the last three assumptions refer to problems in the random [or 'stochastic'] part (ε). Why do you think the error term is so problematic?
15. Violations of the assumptions of the linear model can produce misleading estimates of the coefficients' standard errors. Why is this a problem specifically?
16. Do these assumptions tell us which variables to include in our model?

- Not at all. The included independent variables must be in the model regardless of any problems with the estimation procedure. That there is a statistical problem for the proper theoretical model is the problem with statistics not the model.
- To some extent, yes. The relationship between the independent variables is important as is their relationships with the dependent variable. Not to mention the things we exclude and assume away to the error term.
- Absolutely yes. We should immediately adjust the inclusion and exclusion of our independent variables strictly according to any and all challenges to the assumptions of the linear regression model and model constraints.

17. Why does it make sense to check the robustness of our model? In contrast to the old saying, 'if it ain't broke, don't fix it.' Why go looking for problems?

18. What is the primary problem of violating the assumptions of the linear regression model?

- The regression coefficients all become statistically insignificant
- Makes the intercept get larger
- The dependent variable cannot be measured
- The regression coefficients become unreliable
- The regression coefficients shrink

19. One way to assess the robustness of a model is to use a proxy variable in place of key independent variables and look to see how much the model solution changes. How will we know whether our model is robust?

- The regression coefficients change very little.
- The intercept will go to zero
- The regression coefficient of the proxy variable will get substantially larger
- Some of the regression coefficients change signs
- The measures of model fit improve.

20. What is heteroscedasticity?

- Heteroscedasticity is constant variance of the error term.
- Heteroscedasticity is non-constant variance of the error term.
- Heteroscedasticity is a mean of zero for the error term.
- Heteroscedasticity is normal distribution of the error term.
- Heteroscedasticity is not a problem for the error term.

21. What does heteroscedasticity mean for the model itself?

- We have included too many independent variables.
- The model is solving all of the model equally.
- We will need to include GDP per capita.
- Our regression coefficient estimates are biased toward a specific independent variable.
- The model is solving some parts of the relationship better than others.

22. What is model specification?

- Another name for Model Fit
- Optimizing the selection of included variables and their performance
- The process of Gauss-Markov Optimization

- Producing model unbiasedness
- A list of the included independent variables and their transformations.

23. Biased estimates create which problem(s) for the model?

- The model prefers – or is ‘biased’ – toward one independent variable.
- They misrepresent the relationship between the independent variables and the dependent variable.
- The model prefers – or is ‘biased’ – toward the current model as it is specified.
- The variances of our estimates are increased thus making the model solution unreliable.
- The model cannot find a stable solution.

24. Inefficient estimates create which problem(s) for the model?

- The model solution is biased toward omitted variables.
- None of the regression coefficients can be statistically significant.
- The F-test is not statistically significant.
- The regression coefficients all increase by a standard margin
- The regression coefficients become unreliable as they are poor estimates of the ‘true’ parameter.

25. Why is the error term central to diagnosing problems in multiple regression models?

26. You are interested in studying national-level income inequality. To understand why it differs across Europe, you decide to analyse cross-national differences in the Gini Index of income inequality by including five independent variables in a multiple regression model.

You have data from a random sample of European countries ($N = 37$) and include *GDP per capita*, total population, percent rural population, the unemployment rate and *GDP per hour worked* (from the Quality of Government dataset) [see the Table below]. Unfortunately, you discover that *GDP per capita* and *GDP per hour worked* are correlated at $r = 0.922$.

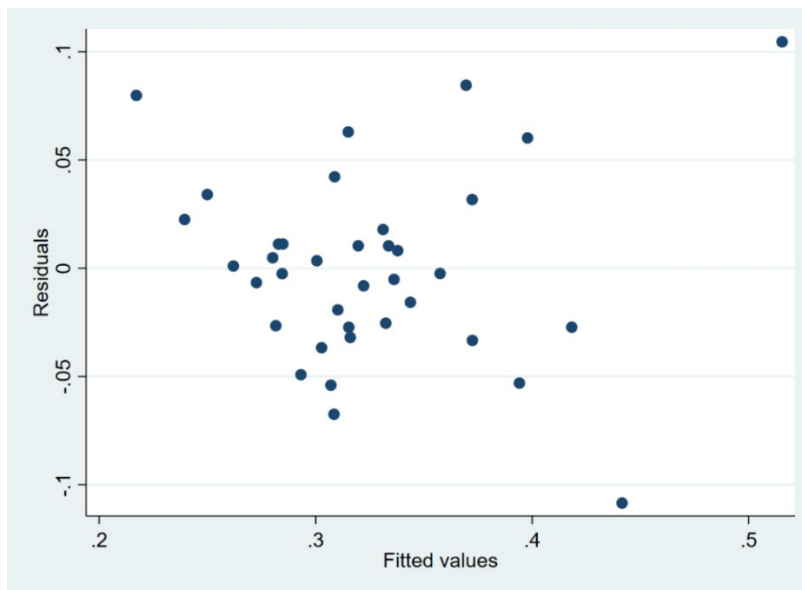
Table: Income Inequality in Europe

Income Inequality [Gini Index]	Unstandardized Regression Coefficient	Standard Error	t	p > t
<i>GDP per capita</i>	0.021	0.012	1.75	0.090
<i>Total Population</i>	0.001	0.000	3.42	0.002
<i>Rural Population</i>	-0.001	0.001	-1.30	0.203
<i>Unemployment Rate</i>	0.007	0.002	4.28	0.000
<i>GDP per hour worked</i>	-0.003	0.001	-3.21	0.003
<i>Constant</i>	0.364	0.037	9.72	0.000
Number of Observations	37			

F(5, 31)	10.84			
Prob. > F	0.0000			
R ²	0.6362			
Adjusted-R ²	0.5776			

Discuss your next steps of how to deal with this problem.

27. If the concept of GDP *per hour worked* is important in the literature as a means to understanding why income inequality varies cross-nationally, what are some strategies to retain this concept in the model? And what would you need to avoid to not create more problems?
28. Finally, just to be sure you also check the distribution of the error terms for your income inequality regression. In the Figure below, we can see the distribution of the error terms. It is not clear whether this is heteroscedastic or not so we run a test and find that we must reject the null hypothesis that the error terms are constant ($p < 0.001$). Now what should we do?



To look at some of the issues of multiple regression, we can often take a quick look at the assumptions. Using the 'QoG mini' dataset, let's consider the following empirical model

- RQ: To what extent does governmental regulation [wbgi_rqe] mitigate national-level CO2 emissions [wdi_co2]?
- Let's control for
 - Coal production [wdi_elprodcoal]
 - Percent of power from renewables [wdi_elerenew]
 - And national economic performance [wdi_gdpcapppcur]

- To make things a bit easier to see and interpret, you can transform the GDP *per capita* variable like this:
$$\text{GDPpc} = \text{wdi_gdpcappppcur} / 1000$$

29. Now, produce the regression outcome

30. Assumption one: Linearity

Looking at the first column of correlations, we can look for any non-linear relationships between the IVs and the DV. **Note:** this doesn't mean that any scattering of points that is not linear qualifies. Rather, we are looking to see if another functional form – other than a line – might better fit the points of the scatterplots (in the first column). What can you report?

31. Assumption Two: Uncorrelated Independent Variables.

To see if there is coordination among the Independent Variables, we can ask for a Variance Inflation Factor test:

- a. Produce a VIF estimation.
- b. Are there any obvious problems here?

32. Assumption Three: The Error Term has Constant Variance

- a. Compare the residuals (a.k.a. errors) versus the predicted values (a.k.a.: 'fitted') plot for for our regression on CO₂ emissions and report any potential problems.
- b. To be sure, we can run a 'Cook and Weisberg's test for heteroskedasticity'. If the null hypothesis is that there is homoscedacity (i.e.: constant variance), what can we conclude?
- c. What does this imply?

33. Assumption Four: Mean of Error Term is Zero

Once again, we find ourselves looking for omitted variables. We can test for omitted variable bias

- a. Do that now for our regression.
- b. If the null hypothesis is that the model has no omitted variables, how would you interpret this?

34. Assumption Five: Fixed Xs in Repeated Samples

Political Analysis Workbook

Although there is no direct test, one could sneak of peek at the relationship between the Independent Variables and the error terms to see if anything jumps out.

This requires two steps.

One is to create the errors (a new variable - called 'errors' - of the residuals (or errors, IOW: ε) and compare them to each IV looking for deviations from the midline (i.e.: zero).

While again this is a rudimentary way to consider the relationship between IVs and the error terms associated with each, what would you report that you see?

Datasets

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG mini"

QUESTIONS FOR CHAPTER 17: Binary Logistic Regression

Binary logistic regression is a powerful statistical method used to model relationships between a set of independent variables and a binary outcome—one that takes only two possible values. Unlike multiple (linear or OLS) regression, it estimates the probability of an outcome using the logistic function. Interpreting the coefficient and output is however a challenge. Here we apply these techniques to real-world data to develop practical skills for making informed predictions and decisions.

The statistical commands that you will learn and use in this chapter are inferential commands. They will tell the statistical software to determine the statistical significance of relationships we determine using measures of association. It is becoming increasingly important for you to understand what the output is telling you about the variables under investigation.

1. Why will using ordinary least squares (OLS) regression when the dependent variable is dichotomous be a problem?
2. Why is maximum likelihood estimation (MLE) used instead of least squares estimation in logistic regression?
3. How do odds ratios help in interpreting logistic regression results?
4. What is the difference between average marginal effects (AMEs) and predicted probabilities in logistic regression?
5. Why is pseudo- R^2 not as informative as R^2 in linear regression?
6. You run a binary logistic regression predicting whether an individual voted or not in the last election (coded 1 for voted). You have only one independent variable in your regression, sex, coded 1 for men and 0 for women. You obtain a regression coefficient of 0.58 for your sex variable. Express the coefficient as an odds ratio and interpret it.
7. What are Ideal Types?
8. What is the main limitation of predicted probabilities?
9. Why is graphing the predicted probabilities an effective way to present the results of binary logistic regression?
10. What are categorical dependent variables?
11. Logistic regression produces logits as regression coefficients. What are logits?
12. How is a Likelihood Ratio χ^2 test like the F-test?

13. Using the 'WVS mini' dataset, let's see if we can predict whether a respondents say that teaching tolerance to children is important or not by a set of responses to other question and attributes. Let's use the following variables:

Dependent Variable:

Tolerance [Q12 → Tolerance]: Important child qualities: tolerance and respect for other people; Range: 0 [No] to 1 [Yes]

Independent Variables:

Satisfaction with Life [Q49 → SWL]: Satisfaction with your life; Range: 1-10

Left-Right Ideology [Q240 → LR]: Left-right political scale; Range: 1 [Left] to 10 [Right]

Immigrant [Q263 → Immigrant]; Range: 0 [No] to 1 [Yes]

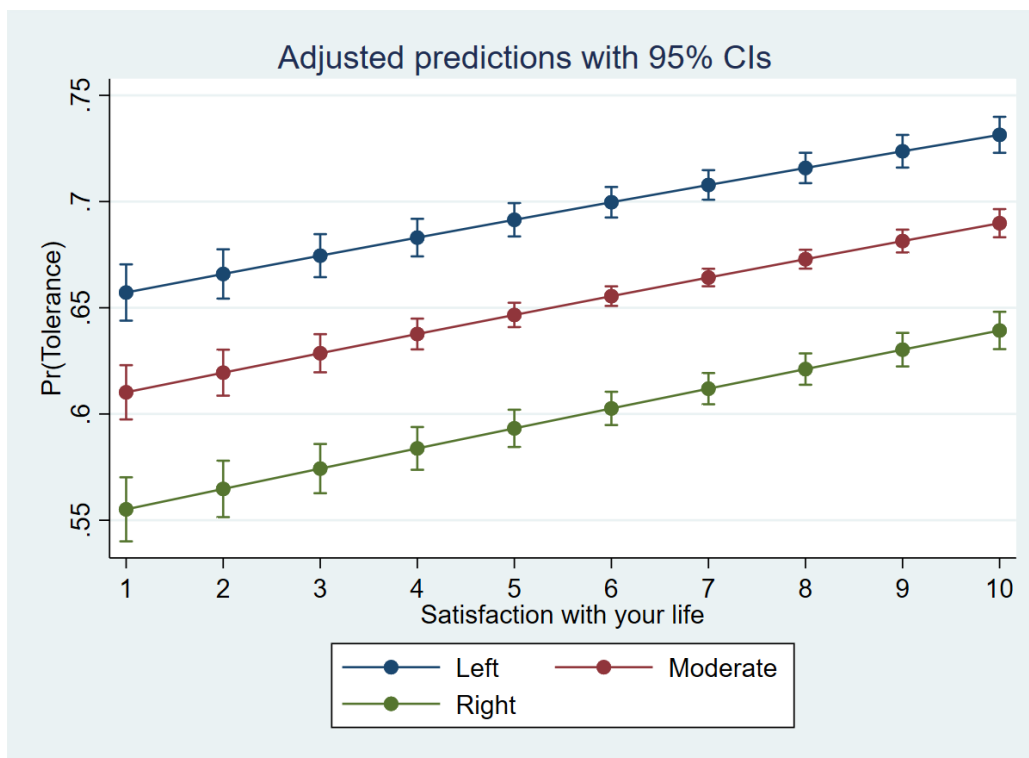
- d. Produce a table of the Logistic output:
 - e. What is the regression equation?
14. What does the Likelihood Ratio χ^2 test tell us?
15. What does the pseudo- R^2 tell us?
16. Let's look at the Odds Ratios
- a. Produce a table of the Odds Ratios Logistic output:
 - b. Interpret using the odds ratios
17. We would like to get a sense of the general performance of the model and compare the IVs. Let's look at the Average Marginal Effects
- a. Produce a table of the Average Marginal Effects:
 - b. Which of these appears to have the strongest average effect on Tolerance?
 - c. [*optional*] you can produce a graph for this
18. Let's consider the predicted probabilities.
- a. Generate predicted probabilities for SWL (1-10) and Immigrant (0,1)

- a. In the above figure, we can see that as life satisfaction increases, so does the probability that tolerance will be taught. But we can also see how this relationship between life satisfaction and teaching tolerance is different for immigrants. Explain.

19. Compare a native born and immigrant with a high level of life satisfaction using two Ideal Types.

20. Let's consider the interaction between ideology and life satisfaction on teaching tolerance. From the left-right ideology scale, ranging from 1 to 10, respondents were bundled into Left (scores: 1, 2, 3), Moderate (scores: 4, 5, 6, 7) and Right (scores: 8, 9, 10). Below we can see their interaction with life satisfaction on teaching tolerance.

We can see that as life satisfaction increases, so does the probability that tolerance will be taught. But we can also see how this relationship between life satisfaction and teaching tolerance is different for those of different ideological positions. Explain.



21. Let's look at some Ideal Types

- a. Compare a Left ideologue to a Moderate ideologue with a high level of life satisfaction using two Ideal Types.
- b. Compare Moderate ideologue to a Right ideologue with a moderate level of life satisfaction using two Ideal Types.

- c. Compare Left ideologue to a Right ideologue with a low level of life satisfaction using two Ideal Types.

22. In Political Psychology, we might be advised to add a measure of how people report feeling not merely in satisfaction but in mood. This may also partially explain their willingness to teach tolerance. Let's add the following independent variable

Average Happiness [wvs_hap → Happy]: Feeling of happiness; Range: 2.54 to 3.65; mean: 3.14, sd: 0.06

- a. Produce a table of the output.
- b. What is the regression equation?
- c. Given that the addition of 'Happiness' doesn't appear to make an impact. What are some of the reasons that we may return to the original model (that is, without 'Happiness')?
 - The Log Likelihood is lower in this model which means this model is better than the previous one.
 - It is not statistically significant
 - The estimations for the other independent variables are almost exactly the same as in the previous model.
 - 'Happiness' – despite what Political Psychologists may argue – seems to be a lot like 'Satisfaction with Life' and thus capturing a lot of the same conceptual and operational content and thus redundant.
 - Logit models cannot handle a large number of independent variables.
 - 'Average Happiness' is a nominal-level variable cannot be on the right-hand side of the equation.

23. Let's look at a new set of data to ask what determines whether Western election monitors need to be present during an election? Let's use the 'V-DEM mini" dataset

DV: Western election monitors present [v2elwestmon] 0 No 1 Yes

IVs

- Alternative sources of information index [v2xme_altinf]
 - Direct popular vote index [v2xdd_dd]
 - Election turnout [v2eltrnout]
 - Print/broadcast media critical [v2mecrit_ord]
- a. Produce a table of the logistic output (let's treat v2mecrit_ord as a series of dummies).

- b. What is the regression equation?
24. Interpret the results by reporting the Odds Ratios of the statistically significant estimates
- a. Produce a table of the output.
 - b. Interpret using the odds ratios
25. We would like to get a sense of the general performance of the model and compare the IVs. Let's look at the Average Marginal Effects
- a. Produce a table of the Average Marginal Effects:
 - b. Which of these appears to have the strongest average effect on explaining the need for Western election monitors?
 - a. **[optional]** you can produce a graph for this
26. Let's consider the predicted probabilities.
- a. Generate a graph plotting alternative sources of information index [v2xme_altnf]: Range:[0.02,0.99]
 - b. Generate a graph plotting *Election turnout* [v2eltrnout]: Range: [18.75,100]
27. For Measures of Model Fit and Performance, produce a table of the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC)
28. If you working at an NGO focused on democracy in the form of free and fair elections, how would present this to the members of your team?

Datasets

V-Dem

- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. "V-Dem [Country-Year/Country-Date] Dataset v6.2." Varieties of Democracy (V-Dem) Project.
 - Available here: <https://www.v-dem.net/vdemds.html>
 - *Political Analysis* dataset: "V-Dem mini"

World Values Surveys

- WVS: Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. World Values Survey: Round Seven - Country-Pooled Datafile Version 4.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WWSA Secretariat. doi:10.14281/18241.18
 - Available here: <https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp>
 - *Political Analysis* dataset: "WVS mini"

QUESTIONS FOR CHAPTER 18: Categorical and Limited Dependent Variables

Many real-world outcomes cannot be measured on a continuous scale—individual policy preferences, which legislation is adopted, or to which a country develops alliances are all examples of categorical or limited dependent variables. Neither OLS multiple regression nor Binary Logistic Regression can resolve these. However, we can exploit what we know about non-interval-level dependent variables to handle polynomial- and ordinal-level dependent variables. Here we focus on these types of models, interpreting their coefficients and assessing their model performance. Adding this to our statistical toolbox, we are able to confront complex social and political phenomena beyond simple linear relationships.

The statistical commands that you will learn and use in this chapter are inferential commands. They will tell the statistical software to determine the statistical significance of relationships we determine using measures of association. It is becoming increasingly important for you to understand what the output is telling you about the variables under investigation.

1. Why do we use the Ordinal Regression Model (ORM) instead of OLS for ordinal dependent variables?
2. What is the proportional odds assumption, and why is it important in ordinal logit models?
3. How does a multinomial logit model differ from an ordinal logit model?
4. Why is choosing a reference category not only important but crucial in multinomial logistic regression?
5. Does an (over-)reliance on statistical techniques shape the kinds of political science questions researchers choose to ask?
6. Why is the tension between explanation and prediction important in political science research?
7. How does this shift towards prediction challenge the scientific approach?
8. Despite a growing complexity in empirical analytical methods, what is one of the limitations of these empirical methods in 'explaining' political and social phenomena?
9. What is a fundamental problem with the classical hypothesis testing approach of comparing a null and alternative hypothesis?
10. Let's take a look at the dilemma of democratic political culture. The dilemma – the character – of democratic political culture is simply to what extent are individuals equally allowed individual liberty. Using the 'EVS mini' dataset, we can use the following variables:

Dependent Variable:

Freedom versus Equality: [E032: 1 Freedom above equality; 2 Equality above Freedom; 3 Neither]; Range: 1–3; Mode: 1

Independent Variables:

Trust [A165: Most people can be trusted]; Range: 0 [Can't be too careful] to 1 [Most people can be trusted]; Mode: 0; VR: 0.675

Household Financial Satisfaction [C006: Satisfaction with financial situation of household]; Range: 1 [Dissatisfied] to 10 [Satisfied]; Mean: 6.40; Standard deviation: 2.49

Fight for Country [E012: Willingness to fight for country]; Range: 0 [No] to 1 [Yes]; Mode: 1; VR: 0.315.

11. Given the dependent variable, it makes sense to look at the predicted probabilities in graphical form. Produce a graph that shows the joint effect of financial satisfaction [c006: 1-10] and trust [A165; 0/1] on the probability for 'Freedom above Equality' [1] (assuming that the respondent is willing to fight for her country [E021 = 1]).
12. How would you interpret this output?
 - a. Specifically compare a trusting but deeply financially dissatisfied respondent to a less trusting very financially satisfied (e.g. 10) respondent.
 - b. Specifically compare a trusting but averagely satisfied (e.g.: 5) respondent to similarly averagely satisfied cautious respondent.
13. How does this change if the respondent is not willing to fight for her country? Produce another graph.
14. What has changed from the previous results?
15. Produce a graph so that we can see the joint effect of financial satisfaction and trust on the probability for *Equality above Freedom* (assuming that the respondent is willing to fight for her country).
16. How would you describe this output?
 - a. Compare a trusting but dissatisfied respondent to a more cautious financially satisfied respondent.
 - b. Compare a trusting but averagely satisfied respondent to similarly averagely satisfied cautious respondent.
17. What explains the level of academic and cultural expression in a country? Are such cultural and social freedoms driven by political values such as the freedom of expression, or actual freedoms like voting, or even the level of political responsibility, given by the level of turnout at election time. Let's use the 'V-Dem mini' dataset to examine this question.

Dependent Variable:

Academic and Cultural Expression: [v2clacfree_ord: Freedom of academic and cultural expression]; Range: 0 [None], 1 [Weak], 2 [Somewhat], 3 [Mostly], 4 [Fully]

Independent Variables:

Freedom of Expression [v2x_freexp: Freedom of expression index]; Range: 0.01–0.99; Mean: 0.47; Standard deviation: 0.32

Universal Voting [v2elgvsufvl: Suffrage level]; Range: 1–12; Mean: 7.51; Standard deviation: 1.35.

Election Turnout [v2eltrnout: Election turnout]; Range: 18.75–100; Mean: 72.90; Standard deviation: 16.57

Produce the output for this ordinal regression

18. What does the Likelihood Ratio χ^2 test tell us?
19. Which independent variables are statistically significant?
20. What is the odds ratio for universal voting?
21. Produce a graphical display of the effect of electoral turnout on ‘fully academic and cultural freedom’.
22. How do you interpret this?
23. Given this graphical display of the effect of electoral turnout on ‘fully academic and cultural freedom’ across a trifold of popular suffrage. How do you interpret this?
24. Let’s do some quick comparisons
 - a. Compare a high suffrage country with 40% turnout with a low suffrage country with 40% turnout.
 - b. Compare a high suffrage country with 100% turnout with a low suffrage country with 100% turnout.
25. Using the ‘QoG mini’ dataset, let’s produce the output for an ordinal logit and interpret it.

If we want to explain the level of political violence and repression a state inflicts on its population, we could look at the variable for Political Terror Scale - Amnesty International [gd_ptsa].

It has five outcomes:

- 1 Countries under a secure rule of law
- 2 There is a limited amount of imprisonment for nonviolent political activity
- 3 There is extensive political imprisonment, or a recent history of such
- 4 Civil and political rights violations have expanded to large numbers of the population.

- 5 Terror has expanded to the whole population

Let's explain this level with the following independent variables.

- bmr_dem: Dichotomous democracy measure
- wbgi_pve: Political Stability and Absence of Violence/Terrorism, Estimate
- wdi_popden: Population density (people per sq. km of land area)
- Europe: QoG: European countries

a. Produce the output for an ordered logit

26. What independent variables are statistically significant (and at what level)?

- bmr_dem: Dichotomous democracy measure ($p < 0.05$)
- wbgi_pve: Political Stability and Absence of Violence/Terrorism, Estimate ($p < 0.001$)

1. Interpret the Europe dummy variable.
2. What does the Likelihood Ratio chi2 test tell us?

27. Produce the output with Odds Ratios

28. Produce an Average Marginal Effects table

29. Produce predicted probabilities for the World Bank's World Governance Indicators: Political Stability and Absence of Violence/Terrorism index to explain the outcome (5) Terror has expanded to the whole population. Interpret.

30. Produce the Measurements of Fit: Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Datasets

V-Dem

- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. "V-Dem [Country-Year/Country-Date] Dataset v6.2." Varieties of Democracy (V-Dem) Project.
 - Available here: <https://www.v-dem.net/vdemds.html>
 - *Political Analysis* dataset: "V-Dem mini"

Quality of Government

- Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute,
 - Available here: <http://www.qog.pol.gu.se>
 - *Political Analysis* dataset: "QoG mini"

European Values Study

- EVS (2022). European Values Study 2017: Integrated Dataset (EVS 2017). *GESIS, Cologne. ZA7500 Data file Version 5.0.0.*
 - Available here: <https://doi.org/10.4232/1.13897>
 - *Political Analysis* dataset: "EVS mini"