Econometrics for Policy Analysis

Jared Greathouse

2024-07-08

Table of contents

1 Syllabus: PMAP 4041, Fall 2024

> **Note**
>
> This is an ongoing project. *None* of the material is in its final form yet. Comments and suggestions are welcome. Jared Greathouse. Office Hours: By Request.

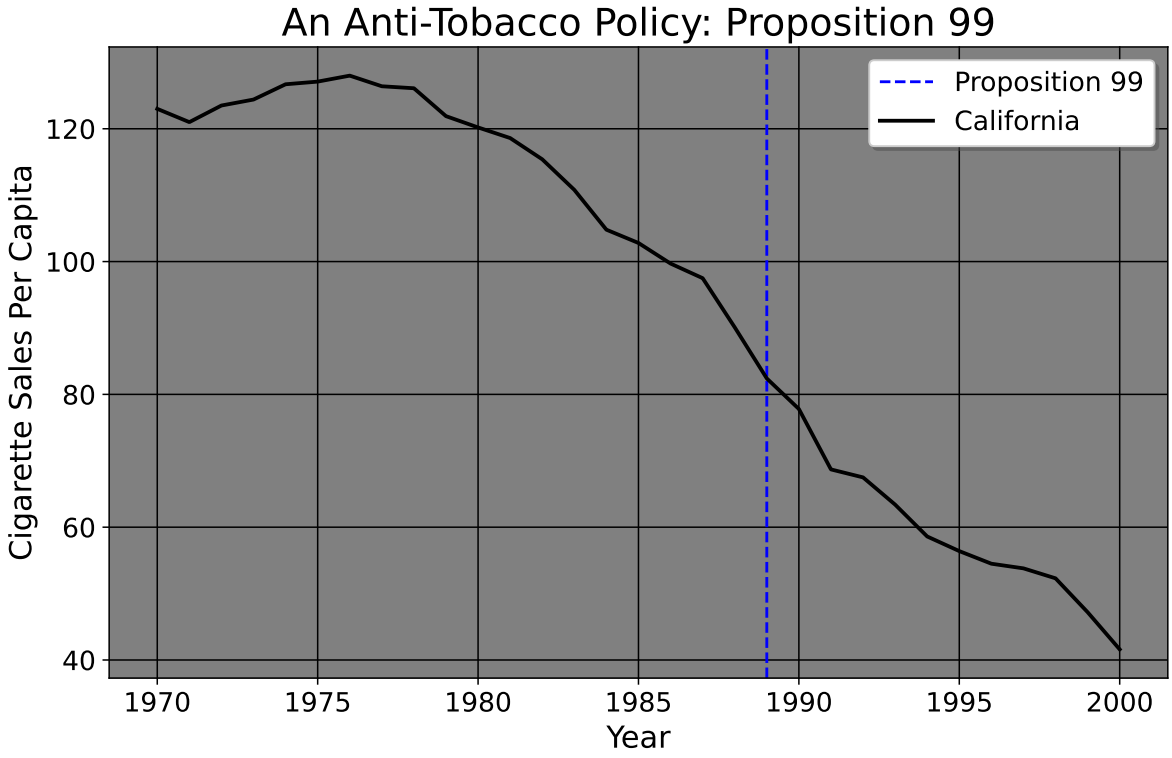
Every day, governments pass laws/public policy to affect some outcome of interest. Policy usually touches thousands if not *millions* of people. From traffic-circles to pop/sugar sweetened beverge taxes, vaccine mandates and universal pre-k programs, cannabis legalization to minimum wages, public policy impacts us all from birth to death.

Policy is never self justifying. It demands evaluation. If California bans tobacco smoking in public, or if New York City implements gun control, presumably we would agree these *likely* impact outcomes like tobacco use or homicide rates, ideally decreasing both of them.

If California's anti-tobacco policy didn't affect smoking rates at all (or worse, if more people began to smoke) or if gun control has 0 impact on homicide rates (or increased them, paradoxically), then surely these could not be justified in the very first place. Before we continue, understand fundamentally these outcomes being affected *are* the point. The only reason that we, as a society, do policy is precisely **because** we think policy affects (or should affect) people somehow. If political science studies "who gets what where", one summation of policy studies might be "what works?" But what policies should we care about? How can we know if they work? This is the starting point for empirical policy analysis. This class discusses the theory and process for how statistical analysis of data may be used to answer policy questions.

1.1 Course Philosophy and Structure

I believe the best way to demonstrate knowledge of policy analysis is through *writing*. As such, there will be no quizzes or in-class exams. Why? It is unrealistic. In real life, rarely do we have an hour and 30 minutes or a ten minute quiz window on the internet to write a full summation of our ideas or think through a question. Typically, we have much more time and resources to help us. In fact, proper use of resources *is what makes a good analyst*: good analysts don't need to remember everything, but they do need to be good at *finding answers* and using them sensibly. In this spirit, you have one assignment. Specifically, you'll write a paper where you derive a research question you find interesting and *apply* the statistical


An Anti-Tobacco Policy: Proposition 99
Proposition 99
California
Cigarette Sales Per Capita
120
100
80
60
40
1970
1975
1980
1985
1990
1995
2000
Year

concepts we cover to answer questions about a real, existing policy. Here is the breakdown of your course grade. The class is broken up into two sections: in the first section, we go over basic probability, correlation, and regression. The remainder of the class covers research for policy analysis.

- 35% of your grade comes from the first draft of the paper, 15% question and 20% draft.
- 60% for the final paper and presentation (respectively, 30 percent each), and
- 5% for attendence.

You will discuss the justification for the policy (including *why* we should care about understanding its effects). You'll gather real data on the policy of interest (including information on the primary variables of interest, relevant predictors/covariates), and outcomes you'll focus on. Finally, after you've defined the research question as well as collected and cleaned your dataset, you'll use the statistical tools we cover (probability theory, descriptive statistics, and regression) to discuss the effects of the policy or intervention. The paper you produce must ask a causal question where there is at least one intervention of interest.

In many senses, public policy is a catch all term covering various disciplines. Public health scholars may care about how banning of abortion in Texas affected fertility rates, or how COVID-19 vaccine/mask mandates affected the COVID-19 case rate per capita compared to other jurisdictions that did not enact these policies. Criminologists may care about how the building of Cop City affected how many people are shot by police, or how a state legalizing cannabis affects crime rates or the consumption of alcohol. Policy historians may care about how Pinochet's 1973 economic policies affected the GDP of Chile or about how Britain's National Health Service of 1948 affected infant mortality. Economists may ask how Hurricane Katrina affected the economy of New Orleans. Environmental scholars may care about how a train derailment affected housing prices. These of course are just some fields; increasingly, advanced empirical methods are used in the business sector and government. Given the array of areas and topics that policy touches, I don't care about what policy or research question you choose to study. To quote Noam Chomsky (who was quoting another MIT professor), the important part isn't what we cover in class; it is about what we discover. The only two stipulations I have is that your rersearch question/outcomes must be 1) quantifiable with **accessible** data that you can use and also must 2) be interesting to you.

1.2 Additional Details

1. If I feel the concept is important, it'll be in the lecture notes or we will discuss it. I will also assign external readings to be done before class.
2. There is no required textbook (aside from this one!) for this course. Various free textbooks exist such as Introductory Econometrics with R, Introductory Statistics, Intro to Modern Statistics, Regression and Other Stories, Intro to Econometrics, Intro to Political Science Research Methods, and many others. The Policy Department at Georgia

State also recommends Introduction to Research Methods or Research Methods for the Social Sciences. The corresponding lecture will focus on the content that each respective chapter covers. Note that these books cover different aspects of the course in different levels of depth (Gelman's book *Regression and Other Stories* is obviously mainly about regression, one of the last math topics we cover, whereas the others are more rudimentary).

3. The same is true for software– I don't care which of these you use, but the only ones I know well are Stata, Python, and (to a lesser degree) R. For Stata users, Statalist is a great resource for Stata. R also is backed by a vast statistician community. I will sometimes include code blocks for Stata and Python in the text.

1.3 Helpful Notes from Me

1. Sun Tzu said every battle is won before it is fought. To reverse the perspective, as Ben Franklin said, if you fail to prepare, prepare to fail. The fact that the paper is the only assignment you have, in effect, means that I expect quality questions, idea, and analyses written in a professional manner. I do not expect perfection, or material at a level beyond the main content, but preparation is your best friend in this course.

2. As corollary to the preceding points, please *do* contact me if you have questions. Policy data analysis is what I do in my research every day. I love what I do, and I love discussing this topic with others. If you have any questions about the ideas we cover in class or have any difficulties, you may always meet with me or contact me otherwise. Thinking of your research question early, asking me for feedback, and so on helps more the earlier you talk to me.

3. Do not simply communicate with me. In addition, feel free to communicate with your classmates. This is something I only really learned the value of as a PHD student, so I figured I would advise the same to you. As an extension of this, I will consider allowing for collaboration on the final paper in groups of two, *with my permission.* For such papers to be considered, I must hear the research question well in advance, as well as the exact ideas on the data, analysis, and relevance of the question overall.

4. As you'll see by skimming the sections of EPA, I frequently use graphics that I construct from real datasets which I link to. On my GitHub page, you'll find these datsets, and more, linked to their descriptions. In lieu of finding your own dataset, you may use any of these for your class paper, should you wish.

1.4 Class Schedule

Below is the schedule. All readings for Econometrics for Policy Analysis (EPA) should be done before class. The other book chapters (unless I write otherwise) are optional.

1.4.1 Week 1

- 08-26-2024 (Monday)

Introductions and EPA, C2

- 08-28-2024 (Wednesday)

Required: EPA C3.

Optional: IS C2, IS C3 (skim), IDS C2, IDS C3, especially "Discrete Probability" and "Random Variables".

A refresher on averages. Also covers t-tests, standard errors, and confidence intervals

1.4.2 Week 2

- 09-02-2024 (Monday)

University holiday. No clase.

- 09-04-2024 (Wednesday)

Basic Asymptotic Theory (the Law of Large Numbers, Law of Iterated Expectations, and the Central Limit Theorem)

1.4.3 Week 3

- 09-09-2024 (Monday)
- 09-11-2024 (Wednesday)

Correlation, Coeffcients, and Association (EPA, C3)

Here we cover basic correlation in 2 Dimensions, mainly using scatterplots and contingency tables.

1.4.4 Week 4

- 09-16-2024 (Monday)

Required: Watch this: Partial Derivatives OLS Explained

Optional: (ROS, C7), IS, C10. Also, Inference for OLS (Gauss-Markov Assumptions). **Today, the research question is due.**

- 09-18-2024 (Wednesday)

Gauss-Markov Assumptions (from the previous chapter)

1.4.5 Week 5

- 09-23-2024 (Monday)

Panel Data

- 09-25-2024 (Wednesday)

Intro to Treatment Effects

1.4.6 Week 6

- 09-30-2024 (Monday) Required: Data Types and Measurement (EPA, C5) Optional: RMSS, C6

Data Gathering/Cleaning (Sampling, Measurement)

- 10-02-2024 (Wednesday)

Data Visualization

1.4.7 Week 7

- 10-07-2024 (Monday)

Writing for Policy Analysis: The Introduction and Literature Review

- 10-09-2024 (Wednesday)

Writing for Policy Analysis: The Background

1.4.8 Week 8

- 10-14-2024 (Monday)

Writing for Policy Analysis: Data

- 10-16-2024 (Wednesday)

Writing for Policy Analysis: Methods

1.4.9 Week 9

- 10-21-2024 (Monday)

Writing for Policy Analysis: Results and Conclusions

- 10-23-2024 (Wednesday)

First Draft Due, Presentations begin.

1.4.10 Week 10

- 10-28-2024 (Monday)
- 10-30-2024 (Wednesday)

1.4.11 Week 11

- 11-04-2024 (Monday)
- 11-06-2024 (Wednesday)

1.4.12 Week 12

- 11-11-2024 (Monday)
- 11-13-2024 (Wednesday)

1.4.13 Week 13

- 11-18-2024 (Monday)
- 11-20-2024 (Wednesday)

1.4.14 Week 14

- 12-02-2024 (Monday)
- 12-04-2024 (Wednesday)

1.4.15 Week 15

- 12-09-2024 (Monday)
- 12-11-2024 (Wednesday)

1.4.16 Week 16

- 12-16-2024 (Monday)

2 Data and Policy Studies

2.1 What is This Thing Called Science?

Science at its core is a process we use to understand observable phenonmena. It is based on using logic and observations of the senses to form coherent and simple understandings about the world. Data, or a collection of observations, is fundamental to being able to conduct scientific research. We use data in our daily lives to make conclusions; we don't call it as such, but we do. Note here that data is not a living, breating concept: it requires interpretation by us. We use principles of science to interpret data and the analyses we conduct upon data. As we learn in middle and high school, science typically begins with asking questions or defining a problem.

Suppose our current problem involves commute time to school or work, and we don't wish to walk. In this case, that's our question: "What's the ideal way to get to school/work?" We then gather information. Chances are we may use Google Maps or Waze to guide us. In this context, these tools provide us with the information we need, namely, *estimates* of how long our commute will be. And, assuming we wish to get to our destination as fast as possible, we make *inferences* or conclusions about the ideal way to take based on the GPS' options. If GPS says the highway takes 15 minutes but the backstreets which avoid highways take 35 minutes, we will typically elect to use the highway since that takes us to our destination the quickest.

There's still two more steps to do, though: test our hypothesis and draw conclusions about the actual observed facts. This means that we must, in real life, leave home and take the way we decide to take. When we get to our destination, we form conclusions about how actually taking the highway went. Of course, we repeat this idea multiple times; eventually, we "typically" take a certain direction to work or school precisely because we have the expectation the highway way will, on average, be preferable to *alternative* ways. This is a simple example, yet it illustrates the central point: in scientific inquiry, we ask questions, draw on available information, form ideas, take actions based on that information, and draw conclusions or plan accordingly based on testing the validity of that observed information. We don't call this science in daily life, but that's exactly what it is. The steps I've outlined so far are present in every field from public policy to physics, albeit with a little more sophisitcation.

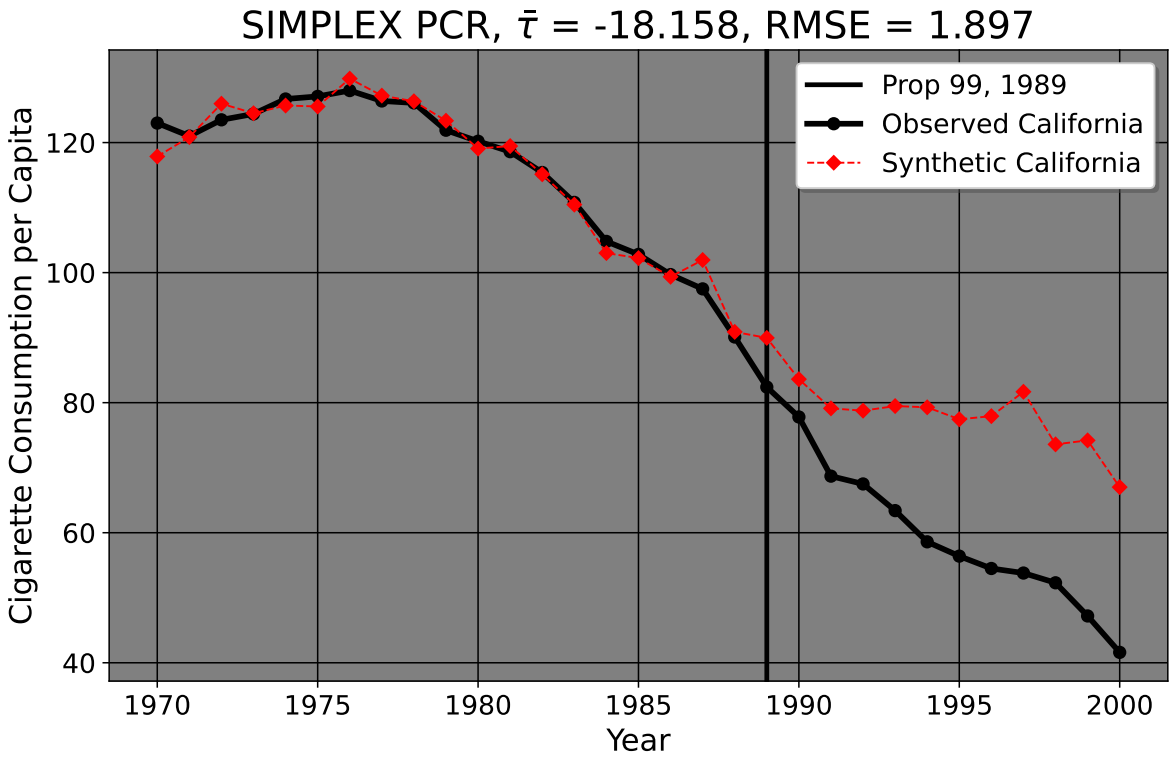
As I've mentioned above, a collection of observations about a set of phenomena is what we call data. Thus, in public policy analysis, data is central to all that we do. One may ask why using data matters at all; the simple reason is that it allows us to resolve disagreements. While people may conduct different data analyses and obtain different results and even reach

different conclusions, the main idea is that we can look into the real world and obtain concepts that map on to metrics that we think are important and test them against our expectations. After all, everyone can have opinions or views on things, but the useful part is *testing out* our expectations against reality. That way, we can have a better sense of what's more likely to be true if a certain policy happens/is passed.

Traditionally, data analysis in the policy space has three goals in mind. The first is descriptive analysis of a phenomenon or topic. In this setting, we simply use raw or lightly transformed data to visualize understanding or relationships between variables (braodly, this is called analysis of variance). For example, we may ask (the classic political science question of) why some countries are wealthy or more developed and others poor/underdeveloped. We could classify *units of analysis* (schools, cities, states, or any other entity) by some criteria (Global South, Southern United States, New England, Metro Atlanta) and compare different metrics of income between them. We may take the average income of each unit and make graphics which show disparities between them. At a deeper level, we may wish to explain the factors which lead to these disparities. So say for cities, we may wish to understand how urbanicity, distance to the capital of the state, age composition, racial composition, and political status of the mayor explains variation in income levels for that city or a set of cities. These sorts of studies can help us point out disparities (for example, maybe cities of the United States that are mostly black or Native American in racial composition have 10,000 less dollars compared to mostly white areas) or identify broader trends. A second goal of policy analysis is prediction. A common problem in macroeconomics is the forecasting of GDP trends. Of course, the only way we may do this is by collecting data on GDP or some other measure we can observe across time and applying statistical techniques to try and predict how GDP/unemployment trends would look under a certain set of assumptions.

A third need for data in policy analysis is for the purposes of estimating the impact of some policy or intervention on some outcomes. Recall the example from the syllabus of Proposition 99, where California wished to reduce tobacco smoking. This intervention raises an immediate question for policy analysis: namely, "what was the *effect* of this intervention on the actual smoking rates we see?" This is a question we may collect tobacco sales data on, for at least California. After data collection (or even prior, in this case), we can form hypotheses. A hyopothesis is a declarative/interrogative, testable statement about the world. It is like a hypothetical in the sense that we try to imagine the effect of a policy on an outcome so that we can answer questions about it. Here, we can hypothesize that Proposition 99 has a *negative* impact on tobacco smoking. Negative here is not intended in the normative sense; presumably most people reading this do not smoke (tobacco, anyways) or think that smoking is wonrg or immoral. Instead, here "negative" means that the policy might decrease the tobacco sales per capita compared to what they would have been otherwise. To test this, we can use statistical analysis to compare California to other states that didn't do the policy.

The plot shows the cigarette pack sales per 100,000 for California from the years 1970 to 2000 (our dependent variable). The thick black line denotes the observed values for California,


SIMPLEX PCR, $\bar{\tau}$ = -18.158, RMSE = 1.897
Prop 99, 1989
Observed California
Synthetic California
Cigarette Consumption per Capita
120
100
80
60
40
1970
1975
1980
1985
1990
1995
2000
Year

and the vertical black reference line shows the year that Proposition 99 (the independent variable/treatment) was passed. As I mentioned above, we typically wish to produce an estimate of California's cigarette consumption in the years following 1989, had Proposition 99 never been passed. This line is denoted by the red dashed line. After we do our analyses/estimations, we can discuss what the implications are. In other words, was the policy effective by some appreciable margin? Are there other outcomes concerns to consider?

2.2 4 Steps of Data Analysis

Broadly speaking, we can think of data analysis being broken into 5 distinct concepts. I summarize them below.

2.2.1 Identifying Policy Problems

As we've discussed above, the first step in this process is simply asking questions. What kind of questions? Policy questions. Knowing what specific questions to ask though can be tricky. Policy is a giant field. Of the thousands of questions we could ask, how do we know which ones will be the most pressing or timely? In other words, how do we know that this is a problem that policy *needs* to be enacted for? How can we identify programs whose analysis benefits the citizenry or other interested parties? Put simpler, who cares? Why do we want to do this study or answer this question? Who stands to benefit?

2.2.2 Gathering Data

Even once we've identified the problem, how do we go about gathering real data to answer questions? If we can't get data that speaks to the issues that we're concerned about, we can't obtain answers that are useful.

2.2.3 Cleansing Data

In real life, datasets do not come to us wrapped in a pretty bow ready for use. Cleaning data (or organizing it) can be a very messy affair in the best of times. In order for us to answer our questions, the data we obtain must be organized in a coherent way such that we can answer questions at all. If you wish to plot the trend lines of maternal mortality in Romania compaed to 15 other nations and your data are not sorted by nation and time, **trust me**, the plot you'll get will not just look terrible, but you can't glean any trends or patterns from it. What's worse, you may not even know improper sorting is the casue of the problem until you bother to look at your dataset again. So, it is best to have good habits developed early.

2.2.4 Analyzing Data

For analysis, we apply statistical analysis in order to answer the questions we're asking, using the dataset we've now cleaned. Such techniques can range from simply descriptive statistical analysis to complex regression models. From such models, we sometimes wish to make inferences to a bigger population, but sometimes more specific statistics (e.g., the average treatment effect on the treated units) are of interest.

2.2.5 Presenting the Results

Now that we've done analysis, we can finally interpret what the findings mean. We attempt to draw conclusions based on our results and come up with avenues for future research or other relevant aspects of interest. In this section, we typically try and say why our findings are relevant.

2.3 Identifying Policy Problems

2.3.1 Justifications For Policy

Before we can do any analysis though, we have to take a step back. We have to ask ourselves how we know a problem exists in the first place. There are two broad justifications that policy is based on: negative externalities and social good, but the main point of both justifucations is "*harm*".

2.3.2 Externalities

The idea of externalities comes from microeconomic theory, which says that efficient markets will affect only those parties who willingly participate in transactions. Particularly in the case of negative extrnalities, or externalities which harm others, we could use public policy to rectify this.

Consider a very simple example: seatbelts. In physics, any force that is not stopped by an equal, oppostite force will keep going. So, if you're in a car crash while driving at 60 miles per hour while unbuckled, the car stops. You, however, don't stop: you keep going, 60 miles per hour through the windshield. No public policy is needed just yet. So far, any cost that comes from a transaction has been borne by you, the driver. By the way, I'm not kidding: one of the arguments against seatbelts was literally that using seatbelts should be a personal decision *if* it does not put others at risk. Additionally, industry also argued against mandatory seatbelt laws on the grounds that it was the government interfering between the transactions of a consumer and the seller.

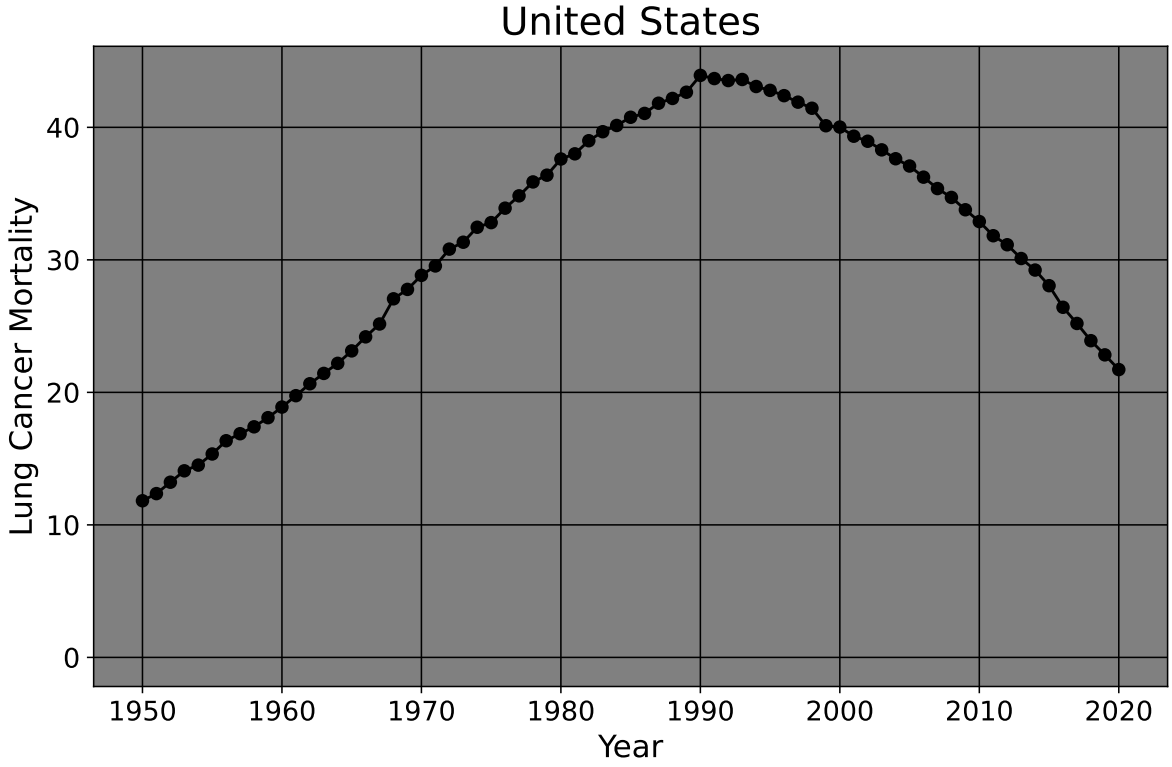
However, there are a few issues with the externality argument. Firstly, being unbuckled turns you into a human projectile. You can hit your passengers or even others outside your vehicle if you're unbuckled. Your market exchange (you buying the car and driving it) is now potentially having second-order effects on others by you not using a seatbelt. So, the government may wish to mandate seatbelts while driving in order to prevent these negative externalities which come in the form of medical bills or death. To address the argument of indsutry above, that seatbelt laws would raise costs of production, this raises an important moral dilemma: does the harm caused to the business of having to install seatbelts matter more than the human harm caused by a society where seatbelts are optional? Also, we are human beings. We have imperfect knowledge. We know for fact that we don't have all the answers, to paraphrase Socrates. We also don't know if the actions we do will ultimately hurt someone else. We live in a probabalistic world (which we will return to later). Indeed, we could argue against laws banning DUI in precisely this manner, saying that we don't know if the intoxicated driver will harm someone until they do. But, as with seatbelts, we never know if there will be another passenger on the road or a child playing in the street. So, we rarely know if we're *actually* putting peoples' lives in danger by driving drunk or unbuckled. We can't know if an externality will occur until it does, usually. Thus, the next view (social good) adopts a different form of reasoning.

2.3.3 Social Good

Moreover, the externality justification isn't typically the way we think about things from a public policy perspective. Usually, we have social welfare goals in mind. This can come in the form of harm reduction or prevention measures. When we argue for public education, for example, we typically don't do so because we think that the private schools won't educate citizens enough (even though they won't), and that public school will be to decrease inefficient education markets. In fact, we typically don't think of education (in our formative years anyways) as a market at all. We usually argue for public education because we think that education has *inherent* benefits, and that being denied a certain level of education necessitates an inherent harm. Imagine for a moment how the literacy rate of the United States would look if school was completely optional. We likely would not complain about GDP loss, we'd likely complain about a society where lots of people can't read the cereal box or function within society in a decent manner. In other words, society has a vested interest in keeping people safe, educated, and healthy to some degree. So we mandate seatbelt laws, basic schooling, and other laws/regulations in service of these ends. Importantly, "these ends" *does not* have a right or wrong answer. The goals of policy are ultimately decided by people within the society. However, knowing the goals of a policy and reasons for its existence helps us ask meaningful questions about it. Following the above discussion, a natural research question that follows is "How did seatbelt laws affect the rate of car accident injuries and deaths?"

2.3.4 Why Is Tobacco a Problem?

As we've discussed above, harm or necessity is typically a standard we look to in order to determine if policy is needed. As I've mentioned, California passed Proposition 99 in 1989 to reduce smoking rates. But, how did we know there was a problem to begin with? To do this, we can grab data on lung cancer morality rates from 1950 until today. Presumably, of course, we view lung cancer as harmful and somehthing we wish to prevent.



The shaded area represents the period before any state-wide anti-tobacco legislation was passed in the United States. We can see quite clearly the age-standarized lung cancer mortality rates rose in a fairly linear manner in the United States. However, the curve is parabolic; mortality rates were rising every single year until the zenith in 1990. Mortality began to fall when the first large scale anti-tobacco laws were passed. Of course, the *degree* to which these laws were the cause of this decrease is an empirical question (especially since lung cancer develops over time, the decrease after 1990 suggests other thing may have also contributed to the decline in behaviors that led to the decreease in mortality). However, given the clear increase in lung cancer rates and other obvious harms of tobacco smoking in the preceding decades, policymakers in California and the voters, in fact, became increasingly hostile to tobacco smoking in public and in other crowded areas. So, California passed legislation in 1988 (as did at least a dozen other states from 1988 to 2000) to decrease smoking rates.

Had I not plotted this trend line, people (from the tobacco industry in 1970, for example) could simply say "Well, nobody *knows* if lung cancer mortality is a problem. How do we know if there's a problem here? I don't think one exists." This plot makes a powerful case that lung cancer is indeed a problem which must be addressed due to the persistent rise in mortality. Data in other words provides intellectual self-defense; if you posit that a problem exists, then this should be demonstrable using datasets that speak to the issue at hand. As a consequence of this, if a problem does exist (be it tobacco smoking or the impact of racial incarceration/arrest disparities), we can then look for policies that attempt to mitigate or solve the problem. That way, we can go about doing analysis to see which policies are the most effective.

3 Summary

At this point, it's clear that data and data analysis are critical to public policy. It allows us to visualize trends, identify the effects of interventions, and reach conclusions on the basis of this evidence. However, the "how" we reach conclusions part matters, since the methodology we use to reach conclusions fundamentally affects what we can conclude in the very first place. The next two lectures cover probability and asymptotic theory; these form the foundations of quantitative public policy analysis.

Part I

Mathematics and Econometric Theory

4 Basic Probability Theory

Human beings are awestruck at uncertainty in everyday life. In the elder days, the Greeks consulted Oracles at Delphi, the Vikings Seers, the samurai onmyōji, and, more recently, horoscopes/birth charts to make sense of happenings. Of these, however, only one has taken the throne of mathematical statistics: probability. Probability is a formalized system which allows us, under differing philosophies (Frequentism and Bayesianism), to rigorously make sense of events that occur.

More concretely, probability is a measure of the likelihood that an event will occur. It ranges from 0 to 1, where 0 indicates impossibility and 1 indicates certainty. Generally, there are two kinds of probabilities econometricians and policy analysts are concerned with: discrete random variables and continuous random variables. Why *random*? Well, because the event *may occur or not*. If a coin had only heads, only tails, or a die only had the number 1 on it, there's no uncertainty anymore and probability wouldn't be needed. But in real life, outcomes are essentially never guranteed. *Discrete* random variables have finite values they can take on. (heads or tails for the coin). By extension, a continuous random variable can take on infinitely many values. Suppose we have data on the width of a coke can or the amount of time in minutes spent studying. These are uncountable in the sense that they can have so many different values that they can't be easily counted.

To fix ideas, let's begin with the idea of a sample space, which we denote by the uppercase Greek letter "Omega", Ω. This represents the set of all possible outcomes for some *instance* or experiment. For example, suppose we have a die of 3 sides numbered 1, 2, and 3, which we cast to see what number faces up. In our case, $\Omega = \{1, 2, 3\}$. Any collection of these outcomes we call events. How then do we assign probability to this event? Say we ask for the probability of getting an odd number for a singular die cast, or $A = \{1, 3\}$. What are the odd numbers in $1, 2, 3$? 1 and 3. Since there are two of these, over three possible outcomes, the probability is just two-thirds. Formally, the way we'd write this is

$$\mathbb{E}\left[\mathbf{1}\left\{A\right\}\right] = 1 \times P(A) + 0 \times P(A') = (2)\,\frac{1}{3} + 0 = \frac{2}{3}$$

Here, A is our event of interest (getting a 1 or 3), and A' (*not* A) is the probability of A not occuring.

4.1 Descriptive Statistics

Probability is rarely used in a vacuum, though. We typically, in the policy sciences, wish to take a given outcome from a set of outcomes and draw conclusions from it. To do this, we use descriptive statistics (also called *moments*).

4.1.1 Means: Arithmetic and Median

The first moment is called the **average/arithmetic mean**. The formula for the mean, also called the *expected value* (denoted by $\mathbb{E}$) is

$$\bar{x} = \mathbb{E}[X] = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where $\bar{x}$ is the mean, N is the number of values, and x_i represents the i-th value in the sequence. The uppercase Greek letter "sigma" means summation, or $\sum_{i=1}^{N} x_i$. It adds the values from $i = 1$ to N. For a discrete random variable, the expected value is

$$\mathbb{E}[X] = \sum_{i=1}^{N} x_i \cdot P(x_i).$$

So for our die, the expected value is

$$\mathbb{E}[X] = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} = \frac{1+2+3}{3} = 2$$

For a coin, the expected value (assuming it is fair) is 0.5. Suppose we have a room of 10 men and 40 women, where women take the value of 1 and men the value of 0. The average number of women in the room is $\frac{40}{50}$. This means the expected value of women in the room is .8. In other words, if we randomly selected a person from the room 10 times, we'd expect about 8 of them to be women. We can take averages with things aside from die and coins too. Naturally, if the amount of water in one giant jug was 5 liters and in another jug there is 7 liters, the average liters of water in the sample is $\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{2}\sum_{i=1}^{2} 5 + 7 = 6$ liters. Now we should distinguish between the population and sample statistics: the sample is that subset of the population that we can get. We can rarely sample every single American in the country (the population), but a random (or representative) sample of 3000 Americans, say, is just fine. This difference is important: outside of simulations, we never can get every single datapoint for all our interventions of interest. So, we collect a sample which approximiates that population we are truly interested in.

The median, or the middle number, is also a type of average. It is less influenced by outliers than the average is. Suppose we have a dataset of years of education across a group of peopl in a neighborhood, $A = \{5, 6, 7, 9, 18\}$. The middle number here is 7 (since two numbers lie to

the let and right of 7). But let's consider the issue deeper: suppose we were to use the average years of education at the average. For us, we have

$$\frac{1}{5} \times \sum_{i=1}^{5} 5 + 6 + 7 + 9 + 18 = \frac{45}{5} = 9$$

The mean and median produce differing values. If we were to use the mean, we'd conclude the average person in this sample is in high school. When in fact, as a raw number, the modal respondent is a middle schooler with one elementary schooler. Thus, we can see that the average is influenced by outliers (in this case, somebody in graduate school). The classic joke is that when Bill Gates walks into a bar, everyone, *on average*, is a billionaire.

4.1.2 Variance

The **variance** is the second moment. For a random variable X, the sample variance is denoted as $s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$. For an intuitive example, suppose we have two middle schoolers in a room, one who reads at 6th grade level and the other at 8th grade level, $A = \{6, 8\}$. The sum of squared differences of each of these datapoints from the mean is (7) is 2, since 6 is 1 less than 7, and 8 is 1 more than 7. So, our sample variance is 2. The variance simply reflects the average distance of each data point from the center/mean of our observations. In practice however, we must correct for uncertainty. So we subtract by 1 in the denominator. This is called Bessel's correction. Subtracting 1 factors in uncertainty, since practically we are unsure about the "true" average in a population.

The square root of our sample variance is what's called the standard deviation from the mean. Why standard? The raw variance is *not* in the same units as our original data. When we take the square root, we may interpret this as the "standard" distance from the mean. For this simple example, we can round the standard deviation down to 1. When we think about it, it makes sense. If you're at a middle school where the average reading level is 7th grade, people who read at 6th grade level are simply 1 year below the average, and those at 8th grade 1 year more than the average.

Some might wonder why we're squaring these differences from the mean. The squared differences gives more weight to outliers, or datapoints that are very far from the mean. Suppose $A = \{6, 8, 18\}$. The average of this is 10.6. Person 1 and 2 are only 4.6 and 2.6 years less than the mean. But someone in middle school with a graduate in college reading level at 18 years is very, very, very far from the mean (practically speaking). The squared differences themselves are 21.79, 7.13, and 53.77, and the sample standard deviation is roughly 6.43. Had we not squared the differences, we'd get 4.89. So, we square the larger differences to assign more weight to large outliers, since not doing so would basically treat the middle schooler with a college graduate reading level as roughly equal to those who are much closer to the average of 10.

4.2 Hypothesis Testing

In public policy we oftentimes wish to test hypotheses. A hypothesis is a statement about the world that we wish to determine the validity of. For example, we could hypothesize that the average math score for a scool is 86, or we can hypothesize that black people use welfare less than white people. We are always testing our hypothesis (which we call the research hypothesis, H_R) against a scenario where this hypothesis is wrong (the null hypothesis, H_0). That is, we start off by assuming the math score is not different from 86 or that blacks and whites use welfare at the same rates. We only change our minds in light of compelling evidence. If this confuses you, imagine we had a courtroom where the burden of proof is now shifted on the defense to prove their client innocent. We would never be okay with presuming guilt. No, we'd say that the people making the positive claim are the ones who must supply enough evidence to convince us otherwise. In research, one way doing this is by using something called a t-test.

4.2.1 One Group T-Test

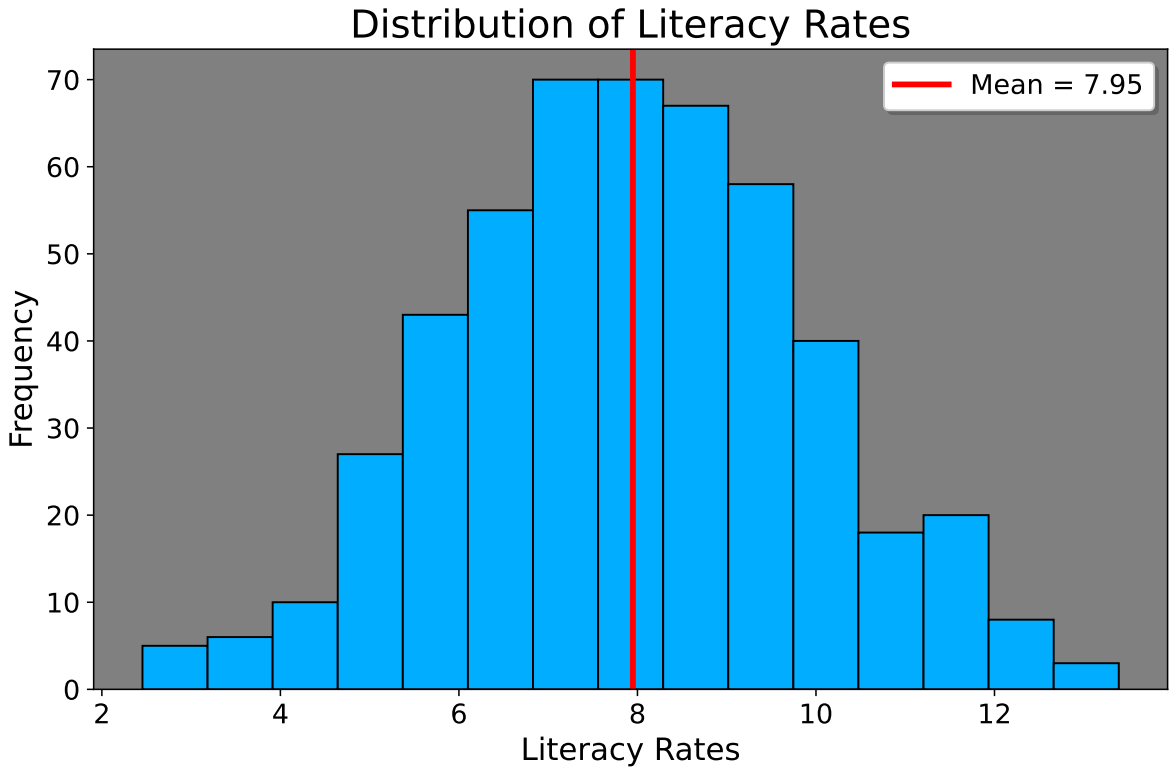
First we cover the one-sample t-test, where we compare our research hypothesis against some known/predefined statistic. If I ask you what you think the average literacy rate is in the population of Americans, you may give different answers like "I think it's at the 9th grade level", "I think the average literacy level is less than the 6th grade level" or "I think the average literacy rate is different from 0". Each of these forms sets of testable hypotheses. In the first case, H_R is "The literacy rate is equal to 9th grade." In the second case, H_R is "The literacy rate is less than the 6th grade level." Finally, we'd say H_R for case 3 is "I think the literacy rate in America is not 0" (or, some significant portion of the population can read). Formally, we denote these hypotheses as

$$H_R : L = 9$$
$$H_R : L < 6$$
$$H_R : L \neq 0.$$

Our corresponding null hypotheses (the ones we start by assuming) are

$$H_0 : L \neq 9$$
$$H_0 : L > 6$$
$$H_0 : L = 0.$$

we simply take a of sample the population somehow (which is usually taken care of for us in census data/compiled statistics). We then calculate the sample average of the grade level


Distribution of Literacy Rates
Mean = 7.95
Frequency
Literacy Rates
0
10
20
30
40
50
60
70
2
4
6
8
10
12

of our sample (ranging from 0 being illiterate and 16+ which means postgraduate). Let's visualize this.

This is a histogram. It shows a distribution of data. To better conceptualize it, imagine the height of the histogram (the y axis) represents the number of people in the data who take on the values on the x-axis. So, roughly 70 people have a literacy level of 8. In this case I generated a sample of 500 people with a small amount of variance. The average literacy rate in this population is 8 (for 8th grade). We now wish to see if our population mean (of 8th grade) is different from the researcher mean (9, 6, and 0). To do this, we need what's called a t-statistic, which is a measure of deviation from the mean *when taking into account* the standard deviation from the mean and sample size. The formula for this simple t-statistic is

$$t = \frac{\bar{x} - \mu_R}{\frac{s}{\sqrt{n}}}.$$

Let's parse these terms. In the numerator we take the difference of our sample mean ($\bar{x}$, the mean we in fact observe in our dataset) versus our hypothetical mean that we are testing our sample mean against, denoted as μ_R ("myoo-sub R"). The denominator is the standard error, which is the standard deviation divided by the square root of our sample size. For example, here's how we'd do this with the null hypothesis for 0 (that is, our sample mean is different from 0).

$$t = \frac{\bar{x} - \mu_R}{\frac{s}{\sqrt{n}}} = \frac{7.949 - 0}{\frac{1.9983}{\sqrt{500}}} = 88.95.$$

In other words, our average literacy grade level is 88 times that of what we would expect given our standard error.

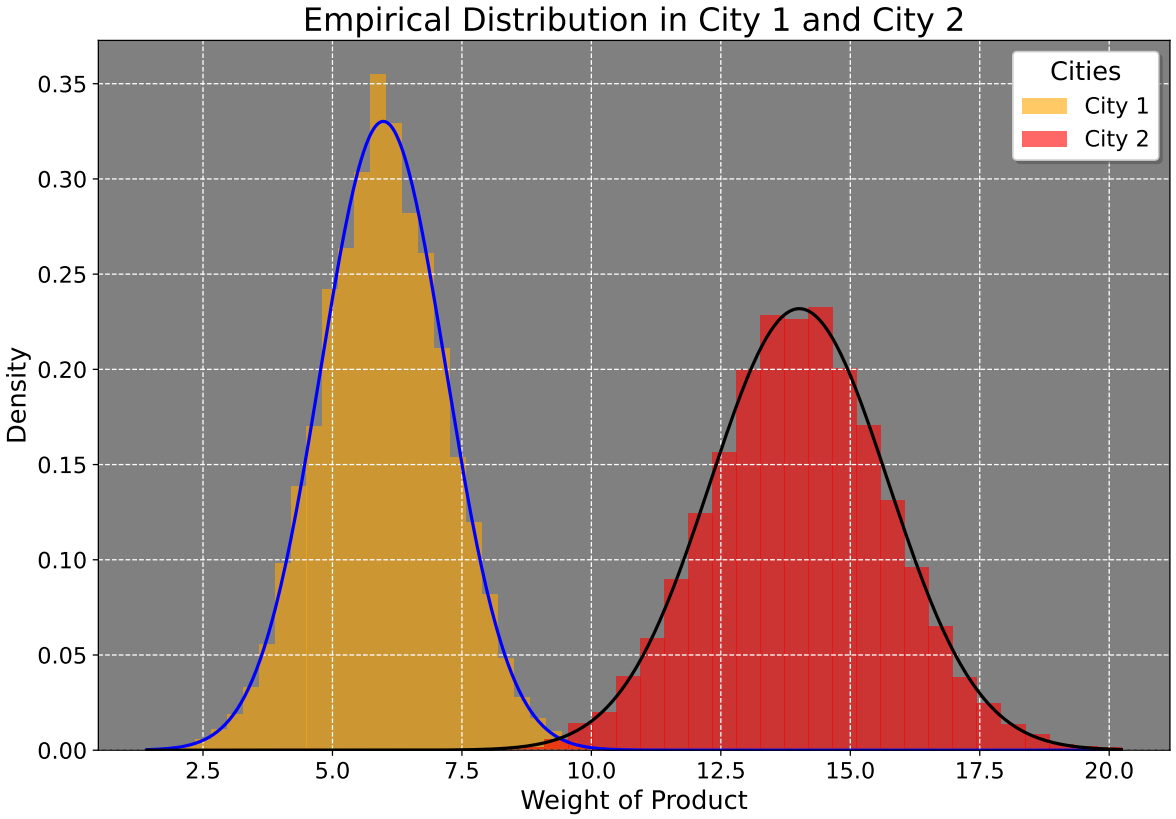
4.2.2 Two-Group T-Test

We can also do a 2-group t-tests, where we wish to compare average group differnces. We can compare men and women, one city to another city, one city to many cities, and so on. For our purposes though, we'll just compare one city to another one. I generate a sample of 20000 where one city has a mean of 6 and another of 14 with respective standard deviations of 1.5 and 3.

Here is how we'd calculate the t-statistic in this case.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}$ is the averge of a city and s denotes the variance for each city. We plug in the values. For the denominator:


Empirical Distribution in City 1 and City 2
Cities
City 1
City 2
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Weight of Product

$$\sqrt{\frac{1.5}{10000} + \frac{3}{10000}} = \sqrt{\frac{1.5+3}{10000}} = \sqrt{\frac{4.5}{10000}} = \sqrt{0.00045}$$

which yields

$$\sqrt{0.00045} \approx 0.0212.$$

Now, calculate the t-statistic:

$$t = \frac{6 - 14}{0.0212} = \frac{-8}{0.0212} \approx -377.36$$

4.3 Uncertainty Around the Mean

Typically we are concerend with the uncertainty of our estimates. Uncertainty around the mean is typically expressed through **confidence intervals**. A confidence interval provides a range of values that, under certain conditions, contains the true population mean.

4.3.1 Confidence Intervals and the Normal Distribution

To understand confidence intervals, it's essential to first grasp the role of a normal/Gaussian distribution. The normal distribution is a continuous probability distribution characterized by its bell-shaped curve. We call it a continuous distribution because unlike coin flips, other data points can take on many values such as homicide rates, COVID-19 rates, and other metrics that can't be broken into simple, countable groups. Most real-world phenomena, when measured, tend to follow a normal distribution (we will return to this in the lecture on asymptoyic theory). A normal distribution is defined by its mean (μ) and standard deviation (σ). The great thing about a normal distribuition is that we can prove that 68 and 95% of the data lie within 1 and 2 standard deviations of the mean. We will exploit this fact to construct a confidence interval.

4.3.2 Constructing a Confidence Interval

The most common confidence interval is the 95% confidence interval. This means that if we were to take many samples and construct confidence intervals for each of them, approximately 95% of these intervals would contain the true population mean. Typically, we have only one sample to work with (and we rely on asymptotics to argue for the validity of our confidence intervals), but methods such as bootstrapping (where we simulate many such samples) may be employed to do this too. For our current purposes though, we construct a confidence interval for the population mean μ, we use the sample mean $\bar{x}$ and the standard error of the mean as we've defined them above.

For a 95% confidence interval, we use the critical value from the standard normal distribution, typically denoted as t^*. For a 95% confidence level, $t^* \approx 1.96$ (since approximately 95% of the values lie within 1.96 standard deviations from the mean in a standard normal distribution). The 1.96 number is a good approximation for all sample sizes greater than 30; otherwise, a different t-statistic would be used. In the old days, t-tables were used to do this, but now software handles this for us. We usually interpret confidence intervals that contain 0 (say, [-1,1]) as being insignificant. By extension, if the CI does not conatain 0 (say, [3,5]), we say it is significantly different from 0 (or, that the means are much different from one another). Note, 95% CIs do not mean 95% of the sample data lies within this interval. Confidence intervals are statements about the mean. Instead, it means that if we were to take many samples and construct intervals in the same way for a mean of interest, 95% of those intervals would contain the true population mean.

4.3.2.1 One Group T-Test CI

I generated a 10,000 person sample of incomes (in 1000s). The true average is 50. We think the average is 60. The standard deviation is the square root of 2. To estimate the confidence interval, we compute

$$0.0277 = 1.96 \times \frac{\sqrt{2}}{\sqrt{10000}}$$

to get our standard error of the mean. Now, in order to characterize the range that the mean falls within, we simply do

$$\text{CI} = \mu \pm \text{Margin of Error} = 50 \pm 0.0277 = (49.986, 50.014).$$

We interpret this as "Our sample mean is 50 thousand dollars. We are 95% confident that given the data, the real mean lies between 49.986 and 50.014 thousand dollars." Since both these numers are less than 60, our research hypothesis ($H_R = 60$) is likely incorrect, as 60 does not fall within these estimates.

4.3.2.2 Two-Group T-Test CI

Now, we can revisit the city example using product weight from the above and see if the means significantly differ. We typically use this kind of t-test in situations where we wish to compare one group to another. The formula for the CI for the difference between two means is given by:

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We know the values from the above, so we plug them in. We also use the critical value of 1.96, since this is the value we use for a 95% CI. First we compute the standard error using the variances and sample sizes for both groups:

$$SE = \sqrt{\frac{1.5}{10000} + \frac{3}{10000}} = \sqrt{\frac{4.5}{10000}} = \sqrt{0.00045} \approx 0.0212.$$

We now have a margin of error (using the critical value 1.96 as above) of:

$$\text{Margin of Error} = t \times SE = 1.96 \times 0.0212 \approx 0.0413$$

We already know the mean difference is -8, so now we just plug that in and solve:

$$CI = (-8) \pm 0.0413.$$

So, the 95% confidence interval for the difference in means is:

$$(-8 - 0.0413, -8 + 0.0413) = (-8.0413, -7.9587).$$

This interval suggests that City 1 consumes significantly less, on average, than City 2. By the way, for those curious, if we just reversed the order of the numerator, we'd get the same result but it would be $(7.9587, 8.0413)$, where we'd say that City 2 consumes significantly more than City 1.

4.4 A Brief Word on Practical Significance

To conclude, a word of caution: as we can see, the magnitude of the t-statistic and tightness of the CIs tend to scale with sample size. Consider the two group case:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

We can see that an increase in the sample size leads to a decrease in the overall denominator. Supoose for the denominator $s^2 = 4$ for both groups. If both groups have the size of 40, then we just have $\sqrt{.1 + .1}$. But if both groups have a sample size of 400, the then we have $\sqrt{.01 + .01}$. The reason this matters is because researchers oftentimes interpret the t-statistic and whether it's greater than 1.96 as a measure of practical importance. But this is wrong! Since our t-statistic is guaranteed to increase with sample size, per the formulae above, at certain sample sizes it would be hard for our confience intervals to NOT contain 0/be significantly different.

What this means as a matter of practicality is to always keep in mind your sample size and what would matter practically to people in real life. If you estimate that the price of one

brand of bottled water, for example, costs 0.05 dollars more than another brand across all 50 states, and your t-statistic is 70 and your CI is [0.01,0.06], then do not claim (in isolation anyways) that this difference is very meaningful or earth-shattering, since they make either a penny more or 6 cents more. In other words, the findings are statistically significant, but practically they are meaningless. The only way for this to really matter to anyone is by having some metric about how much each brand sold (in terms of individual water bottles) for us to reach any firm conclusion about how much this average difference matters. I say this because I do not want for you, in real life or in your papers, to apply these ideas mechanically. I want you to always keep in mind how statistics maps on to the real world.

5 Summary

Probability is the stepping stone into using statistical methods. It is the foundation of decisionmaking in business, economics, and policy analysis. For many, the concepts covered here will be new material– indeed, the term "statistics" or "data analysis" can be intimidating to people at first glance.

I believe the best way to introduce these topics is to keep a balanced perspective between mathematics and application. However, this course only scratches the very surface; the world of quantitative methods in policy analysis is a big one. For those of you interested in graduate school or who wish to use statistics for your future job, your mastery of this essential material will not be in vain.

6 Asymptotic Theory

Across the earth, there are around 7.5 sextillion sand grains across all beaches and deserts. However, mathematics isn't bound by Earthly constraints. As I mentioned in the previous chapter, probability and statistics is about quantifying uncertainty in order to draw conclusions. However, it is now time to understand the very basics under which we *can* draw conclusions to start with. To do this, we investigate basic asymptotic theory. It is the very underpinning of statistics, in particular explaining the circumstances under which we can be confident about our estimates. We defined confidence intervals rather quickly in the previous lecture. Here, we will add to your toolkit with which to understand when they are valid.

6.1 Law of Iterated Expectations

Suppose we wish to commute to and from Georgia State University from Marietta, Georgia. As we've discussed in the previous chapter, we can think of some variable c (commute time) as a random variable, as its value is not guaranteed until we actually leave home and arrive. Suppose we are now interested in the average time it takes us to arrive using I-75 South, conditional on us taking the South or North depending on our starting point. We think I-75 South will take 15 minutes, compared to I-75 North which we think will take 20. We can formalize this also as $x = \{1, 0\}$, where we take North being coded as 1, else as 0 if we take North. So, we leave home (or school) using either highway, and record how long it took to get to the destination. Say, we record the value of 30 minutes in the morning and 40 in the afternoon. Do we conclude that this is how long it takes to get to school on averge, and that we should use some other interstate? No. Why not? This is only one estimate from one day. There are all kinds of things that could have been going on in the morning or afternoon that might influence your travel time, most notably traffic, construction, or other random events for any given path we choose. So, what are we left to do? The only thing we can do, is collect more data.

So suppose we take this same highway for one month, using only 75 South and North (unlike say I-285). We record the amount of time it took us to get to school and home that day using either way (that is, we record our commute time to and from school each day for both ways, taking the average separately for each week). Mathematically, we express this as the expected commute time conditional on the chosen route x: $\mathbb{E}[c|x = 1]$ (expected commute time given we take North) and $\mathbb{E}[c|x = 0]$ (expected commute time given we take South). The *Law of Iterated Expectations* (LIE) states that the overall expected commute time for I-75 (to Marietta,

anyways) is the average of these conditional expectations, weighted by how often each route was taken. Formally: $\mathbb{E}[c] = \mathbb{E}[\mathbb{E}[c|x]]$. This means that the overall average commute time $\mathbb{E}[c]$ is the expected value of the conditional expectations $\mathbb{E}[c|x]$

So, if after a month of data collection, we find that the average commute time is 17 minutes, this is the unconditional expectation $\mathbb{E}[c]$, calculated by taking the average of the commute times across both routes across all days. This approach, where we repeatedly take the average of an outcome/variable given some condition (here, the route we take), reflects the Law of Iterated Expectations.

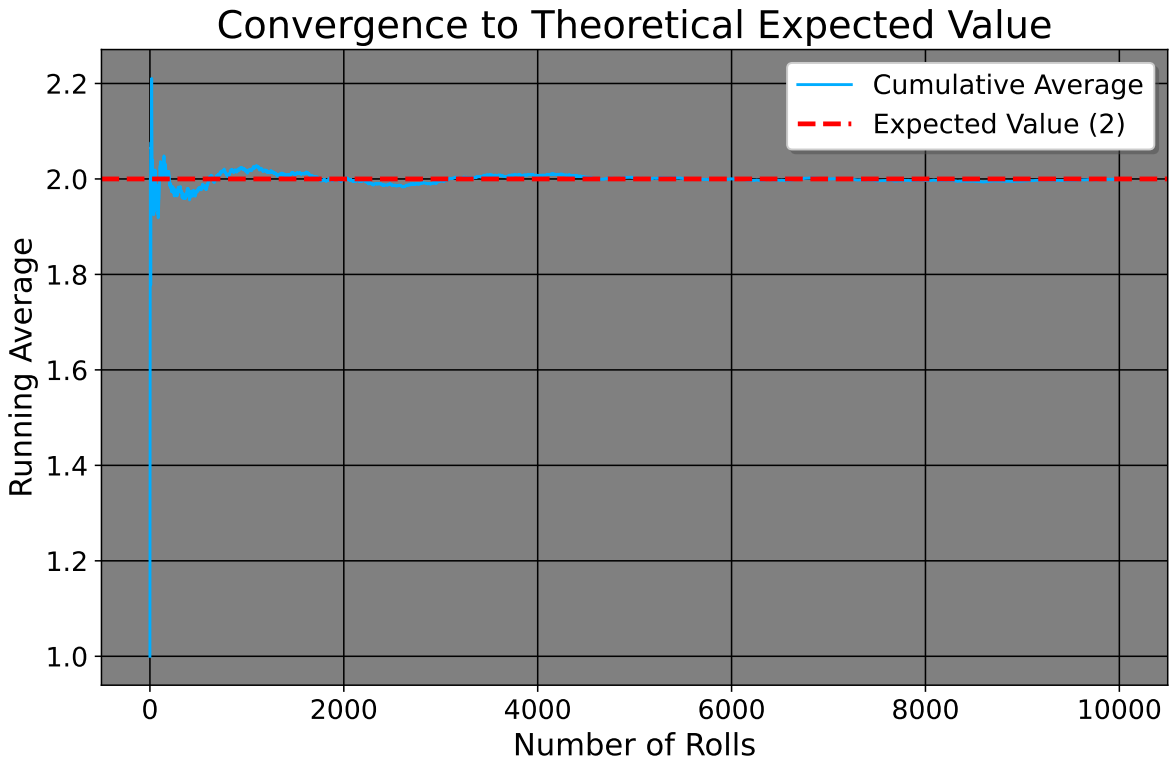
6.2 Law of Large Numbers

Why might this be true, though? Why, after taking all of the averages across a whole month do we arrive at 17, which is a lot closer to 15 than we first thought? Surely, this number is much quicker than the 30 minutes we took the first day? The reason for us getting this value is because of what we call the Law of Large Numbers, or

$$\lim_{N \to \infty} P\left(\left|\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right| < \epsilon\right) = 1.$$

Here is what the math says: as our sample size N increases without bound (that is, as we take the highway more and more and more and more and more… to an infinite amount of times), the probability that the average of our individual empirical daily commute times x_i approaches the *true* population value is 1. In other words, the more estimates we take, the closer, in probability, we come to our population estimate. You see, the first day of 30 minutes was simply a sample of 1. We had nothing else to base our ideas off of aside from whatever GPS tells us. Maybe there was traffic or some other unforeseen thing. However, as we take the highway more times, we tend to get a better sense of how long it'll take to get places, what lanes to use, and so on and so forth. To further prove this, the expected value of a three sided die numbered 1 to 3 is 2. If we cast the die three times and then take the average of what we get, we should see the cumulative empirical average converge to the theoretical average.

As it turns out, that's exactly what we do see. As researchers, what this means is that drawing from a large sample tends to be better than a small one for empirical. For example, if someone has a sample size of 20, we likely would not be okay with generalizing one particular aspect of this sample (say, weight or political affiliation) to everyone in the same city, as we'd need more data points to average over. Ideally, if we're trying to sample American public opinion, we don't want to use only 2 American states, ideally we'd have all the state data available to use.


Convergence to Theoretical Expected Value
Cumulative Average
Expected Value (2)
Running Average
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0
2000
4000
6000
8000
10000
Number of Rolls

6.3 Central Limit Theorem

By understanding the Law of Iterated Expectations (LIE) and the Law of Large Numbers (LLN), we can now delve into the Central Limit Theorem (CLT), which helps us characterize the overall distribution of commute times. That way, we can do things like calculate the confidence interval of our commute times, hoping it will approximate the true one. The CLT says:

$$\frac{\bar{x_i} - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} N(0,1)$$

or that as the sample size increases, the probability of our empirical distribution approaches[a normal distribution. Aside from the distribution of our measurements, our sample mean, by LLN, approaches the population mean. If this seems at all abstract to you, we may simulate this. Here, I define the average commute time to be 15 with a standard deviation of about 3 minutes (between 12 and 18 minutes). I then simulate the commute time across 20000 commutes (since I didn't wish to drive down Interstate 75 20,000 times!). We can clearly see from the GIF that the empirical distribution (that is, the commute times we experience) quickly approaches the *true* population average time as we commute more and more.

This convergence supports the LLN, which tells us that our sample mean will approach the true mean $\mu = 15$ as we collect more data. Additionally, notice how our confidence intervals get tighter and tighter given an increase in sample size. Note that this result is intuitive: as we collect more data, it would make sense that we have a better picture about the underlying data (such as, the sample mean and variance). So, with this in mind, it makes sense that we have a better sense of our uncertainty of our estimates, as the confidence interval shows. So, with the commuting example, the first day took us a half hour. We suspected that the commute would be 15 and 20 minutes respectively (the expected value of which is 17.5 minutes). So, in terms of minutes, we could hypothesize that the true travel time takes between 15 minutes or 45 minutes. In other words, we're quite uncertain. But, as we collect more data, the uncertainty decreases, tightening to be around the true value.

Additionally, the CLT allows us to visualize how even if the original distribution of the variable is not normal (say, a coin toss where we only have two outcomes, heads or tails), the distribution, given enough samples, *converges* to a normal distribution.The plot above flips a coin 1000 times, plotting the empirircal distribution after every 100 flips. We can see that the distribution, despite having only two outcomes, converges to a normal one. Thus, CLT allows us to construct confidence intervals and make inferences about a population, given a large enough sample.

6.4 Sampling

Before we continue however, we need to say a few words about sampling, and the idea of collecting a random, probibalistic sample versus a non-random sample. The reason this matters

is because no matter how we calculate our statistics, we need to be sure that our statistics can actually map on to the actual constructs we want them to. A random sample, in principle, is the idea that everyone from a given population has the same chance to be included in the sample for some study. For example, if we wanted to survey the public opinion of Georgia State students, one potential way of doing this would be to get every single active GSU email from the University and put it into a spreadsheet. We then could generate a variable in the spreadsheet called a "Bernoulli" variable, or a variable that takes on the values 0 or 1. We can then, after setting the number of observations (defined as the number of emails we have), define the probability with which a given variable takes on 0 or 1 to be 0.5. In other words, everyone has an equal chance of being included in the survey, 1 means youre included, else 0. Note, there are other, much more sophisticated ways to do this, but this is one way of taking a random sample.

What would a non-random sample look like, then? Well, in principle it's where everyone in the population does not have an equal chance of being included in the sample. But, this is a simple definition. It does not explain why it's bad or why it's harmful to researchers. So, let's consider a few kinds of non-random sampling. The most obvious one is something called convenience sampling. As the name suggests, we take a sample based on whoever is around us. This can be in our class, on the street, or in our friend group. For example, I have two coworkers, one goes to MIT and the other went to Columbia, both for econ degrees. If I wanted to know how good the average person in the United States is at calculus or math, then I cannot simply assess their skillset and reach a conclusion. Why not? Well, MIT and Columbia are incredibly selective schools that *select for* math skills. Furthermore, the fact that I know them or work with them is likely correlated to my research interests which demand a good background in statistics/applied mathematics. Similarly, even if we personally randomly asked people on the street, this is still not a random sample. Why? Well, there are likely underlying reasons to people being in certain locations. If you're in the biology department, it's likely that most people you meet will have an uncanny knowledge about germ theory or anatomy. And we'd expect this, we'd say something's wrong if these were not true.

At this point, you're likely asking what any of this has to do with the previous discussion of asymptotic theory or policy analysis for that matter. Well, the sample we collect is directly related to the results that we obtain. If our goal is to approximate math knowledge in America, it's inapprorpiate to only sample students from the engineering department at Georgia Tech or the economics students at Chicago. To say it differently: even if we did collect data from every single student at every single engineering department in the United States, this would map on to the population of engineering students, at best, *not* the United States as a whole. In statistics language, our estimate of the "true" mean would be biased, in this case significantly upwards. Even though our confidence intervals would be more precise as our sample size increases (that is, the more engineering students we ask), we would still never converge to the true parameter because our sample is so dissimilar to the population of people we care about, effectively giving us the right answer to the wrong question. When we discuss regression, the importance of sampling will become even more apparent, but for now suffice to say that it's important that the sample be as representative of the underlying population as possible.

7 Summary

Asymptotic theory simplifies statistical analysis by encouraging us to think about the "true" population of interest. It provides us with tools to derive more accurate and precise estimates from. But, as a more general rule for policy analysis (and life), it demands us to think with an infinte mind instead of a limited one. In other words, we should always ask ourselves as researchers "if we could sample everyone, would this statistic I just calculated be close to reliable?" One practical implication of this is having a sufficiently large sample size in order to increase the probability of being closer to the true mean. However, as the previous section discusses, these asymptotics are only as justified as the quality of the sample. We need, in other words, our sample to be as representative as possible of the broader population of interest. In the next lecture on correlation, aside from sampling, we will discuss the basics of statistical design in order to have valid results from statistical analysis.

8 Correlation and Association

As Gregory House once said, "The cave man who heard a rustle in the bushes checked out to see what it was lived longer than the guy who assumed it was just a breeze". House's point here is that the human ability for hypothetical reasoning and the ability to form associations in our minds was critical to our survival and development. In this case of course, we learned to check the bush since it could be a sabretooth or another human who wanted to eat us or steal our food, and we also learned to reason that checking things out, spear in hand, was more likely to keep us safe than rolling the dice by not checking. We will discuss the aspect of statistical reasoning that covers hypothetical reasoning later, but for now we cover simple correlation and measures of association. In statistical analysis, we typically begin by looking at bivariate correlations. We calculate a statistic called "Pearson's r". The formula for Pearson's r, the Pearson correlation coefficient for continuous variables, is:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}.$$

Let's parse these terms, shall we? The numerator is the formula for what we call the covariance between two random variables. It is the sum of the differences between the individual datapoints for each variable and its average, divided by the product of the standard deviations of each variable. More practically, r represents how variables tend to move together, given how much their values move internally (hence the standard deviation term). As a quick example, let's compute the Pearson correlation coefficient r for the following data points $x = \{1, 2, 3, 4\}$ and for y we have $y = \{2, 4, 5, 7\}$.

Here is the computation…

First we calculate the means of x and y:

$$\bar{x} = \frac{1+2+3+4}{4} = 2.5 \quad \bar{y} = \frac{2+4+5+7}{4} = 4.5$$

Then compute the differences from the mean for each data point:

$$x_i - \bar{x} = \{-1.5, -0.5, 0.5, 1.5\} \quad y_i - \bar{y} = \{-2.5, -0.5, 0.5, 2.5\}.$$

Then calculate the sum of the products of these differences:

$$\begin{aligned}
&\sum(x_i - \bar{x})(y_i - \bar{y}) = \\
&(-1.5 \cdot -2.5) + (-0.5 \cdot -0.5) + (0.5 \cdot 0.5) + (1.5 \cdot 2.5) = \\
&3.75 + 0.25 + 0.25 + 3.75 = 8.
\end{aligned}$$

Next, we compute the sum of the squared differences for x and y. For x we have

$$\begin{aligned}&\sum(x_i - \bar{x})^2 = \\ &(-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 = \\ &2.25 + 0.25 + 0.25 + 2.25 = 5.\end{aligned}$$

For y we have

$$\begin{aligned}&\sum(y_i - \bar{y})^2 = \\ &(-2.5)^2 + (-0.5)^2 + 0.5^2 + 2.5^2 = \\ &6.25 + 0.25 + 0.25 + 6.25 = 13.\end{aligned}$$

Now, we plug these in:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{8}{\sqrt{5} \cdot \sqrt{13}} = \frac{8}{\sqrt{65}} = \frac{8}{8.062} \approx \boxed{0.993}.$$

Okay, so our correlation coefficient is 0.993. We'd say there's a strong, positive linear relationship between our variables here (whatever they happen to be). Similarly, if r were negative, we'd say there's a strong negative relation. Of course, if the coefficient were, say, 0.01 or -0.01, we'd say there's pretty much no relationship between the variables.

Here's how we'd visualize these.

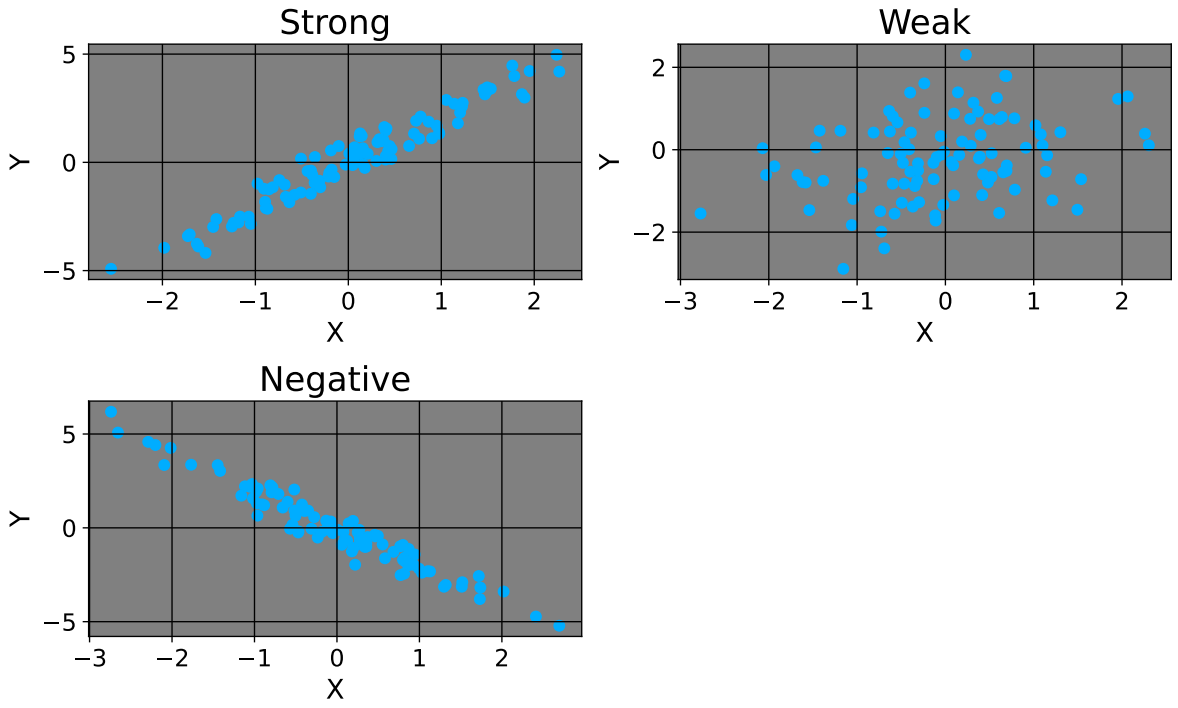
We can think of r as the degree to which one variable moves with another variable. If the relationship were just $r = 1$ or $r = -1$, we'd say for every one increase in x, there's a guaranteed increase/decrease in y respectively because they move identically. For a simple example of a perfect linear relation, consider a dataset with the number of games an NBA team won in one column versus how many they lost in another column. For simplicitly, let's consider the first ten games. If a team won 5, they also lost 5. If a team won 6, they must've lost 4. There's a perfect, inverse linear relaionship between these; to win the first game necessarily means you've lost 0 games yet, and to lose the first means you've won none just yet.

By the way, we can do the exact same thing with categorical vairables (e.g., race and promotions). We could consider a setting of two variables where we have gender as our predictor and an outcome for "promotion" (being promoted or not).

We can posit that gender may affect the likelihood of someone being promoted, perhaps men are more likely to be promoted than women. In this situation, we can graph the proportion of men who were promoted versus not. The plot above suggests that men were much more likely to be promoted than women are.

For those who care about the "statistical significance" of r …

As for mean differences, we can also use the t-statistic for the correlation coefficient to determine statistical significance. So let's use the example above. First, here is the formula for the t-statistic:



$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the Pearson correlation coefficient and n is the number of data points. Given:

$$r \approx 0.993$$
$$n = 4,$$

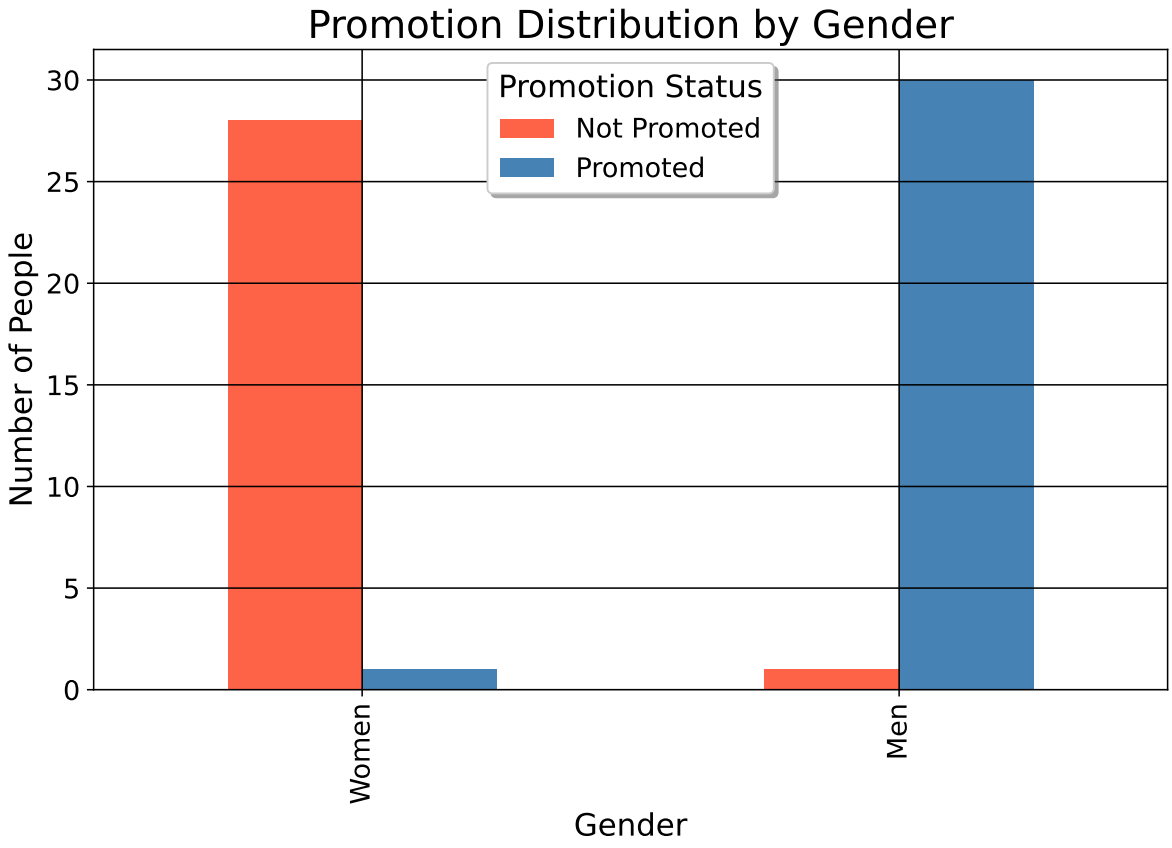
we can begin by plugging in our values. First, we compute the variance in the denominator by squaring our r:

$$r^2 = 0.993^2 = 0.986049$$

Then, we compute how much our one variable x does NOT explain the variance of the other variable:

$$1 - r^2 = 1 - 0.986049 = 0.013951$$

For our numerator, we compute:


Promotion Distribution by Gender
Promotion Status
Not Promoted
Promoted
Number of People
0
5
10
15
20
25
30
Women
Men
Gender

$$\sqrt{n-2} = \sqrt{4-2} = \sqrt{2} \approx 1.414$$

Next, in the denominator, this simplifies to:

$$\sqrt{1-r^2} = \sqrt{0.013951} \approx 0.118$$

Finally, we compute:

$$t = \frac{0.993 \cdot 1.414}{0.118} \approx \frac{1.404102}{0.118} \approx 11.899$$

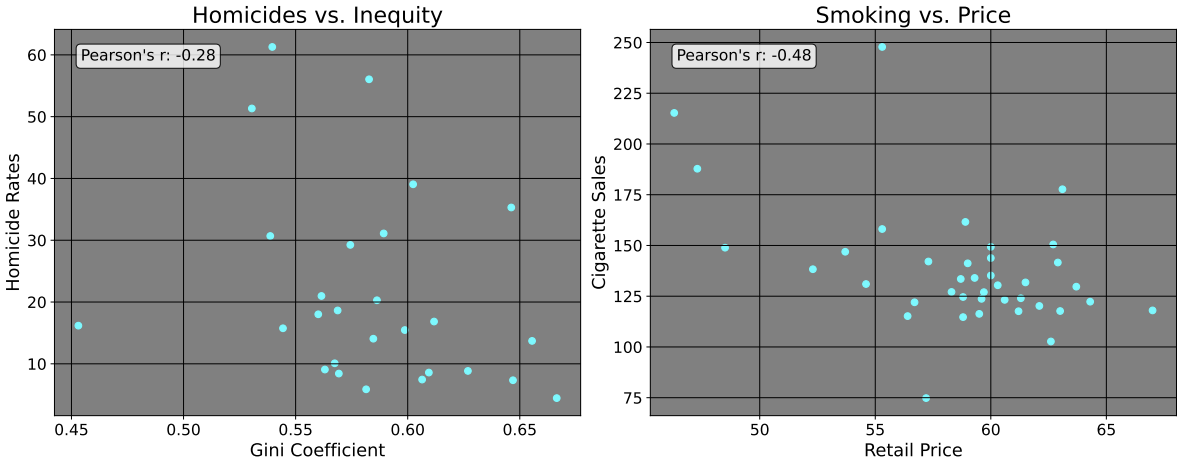
Thus, the t-statistic for the given Pearson's r is approximately $\boxed{11.899}$. Since $11.9 >> 1.96$, we'd say the linear association is statistically significant.

> **!** Important
>
> You will never do this by hand (at least I never have!), I include this section only to show it's possible.

8.1 A Prelude To Regression

Typically, when we discuss correlation, we use scatterplots to visualize the association between variables. A scatterplot is simply a chart which depicts the co-movement of variables. On the x-axis we plot our independent variables/predictors (or, the things we think explain a given outcome) and the y axis plots the variable we think is being affected by the one on the x-axis. Below, I plot in the left panel the homicide rate per 100,000 versus the state-specific Gini coefficient for 27 Brazilian states in the year 1990. The right panel plots the average retail price of cigarettes versus cigarette consumption per capita for 39 American states in the year 1980.

Well now, what do we see here? We see the plotted datapoints along with the Pearson's r. We can see a negative correlation coefficient reported, where a one unit increase in the Gini coefficient leads to a decrease in the homicide rate in Brazil for this year. We also observe a negative relationship between the retail price of cigarettes and the consumption of cigarettes, where an increase in price leads to a decrease in the amount of cigarettes consumed. This result especially should be pretty intuitive: all else equal, as the price of a good increases, the demand for said good generally decreases. However, what about the leftmost plot? The Gini coefficient is a measure for inequality, where 1 denotes one person has all the money and everyone else nada/nothing. A Gini coefficient of 0 means everyone has the same amount of money, and a measure of anything in between is some intermittent level of inequality. Well,



this result is puzzling in a bivariate lens: typically, we associate income inequality and poverty with an increase in things like homicide and other kinds of crime. This is where we begin to think atypically. While correlation is a useful measure sometimes, it alone is inadequate for serious policy analysis. Below, I explain why.

8.2 The First Exercise of the Statistical Mind

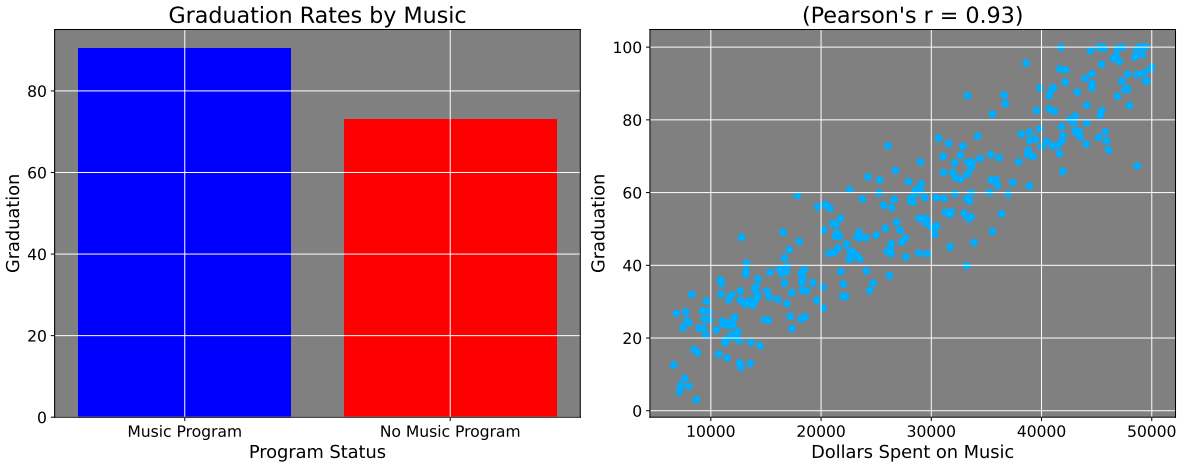
Earlier this morning, I was on Facebook and one of my friends posted a picture that cited a story of a public school music teacher named Annie Ray winning a Grammy. In the article, they say [caps theirs]

> THE FACTS ABOUT THE IMPORTANCE OF MUSIC EDUCATION DON'T LIE… Schools that have music programs have significantly higher graduation rates than do those without music programs (90.2 percent as compared to 72.9 percent).

For visualization, let's do some graphing shall we? The plot on the left plots the rates from the quote, the right plot is simulated data.

This quote caused fire alarms to sound in my head. Why? Because the article (and reporting on music education more broadly) misleadingly discusses these statistics. This does NOT mean these statistics are wrong in terms of their computation. Presumably whoever did this used statistical software to get these numbers, and I trust that the numbers are accurate.

My criticism is about practical implication. The heavy suggestion from the block quotation is the music programs are *causing* this 17.3 percentage point difference in graduation rates. When we read things like "The facts don't lie", there's this air of certainty that these statistics are being reported with.



And in fairness, this is not a *completely* crazy idea: music does indeed help people learn languages. It likely helped me learn Spanish and made me a better overall reader. It's also correlated with spatial skills. In fact, in another life, I was a music theorist who could do harmonic analysis of chord progressions and tell you what I thought the artist was trying to communicate. And as a matter of fact, music studies likely made me an even better econometrician, because one thing music analysis teaches you about is *context*. And in larger context, these statistics seem misleading.

The suggestion is that if more schools *just* had music programs, we'd see higher graduation rates by 17.3 percentage points, on average. And yes, to some degree, if a school now has a music program and did not have one before, it is true students now have the opportunity to learn music. They may even go on to study and succeed in music now that they have the opportunity to do so. But how could we estimate this? How many people would this even affect, exactly?

Most people do not want to be musicians. Learning music, like any other art form or professional skillset, is a non-trivial investment of time and money. Everyone doesn't have the means or honestly dedication to pursue it, esepcially in light of other potential desires or opportunities. Those who do decide to become musicians (not just in classically trained in school, but generally) may have personal qualities that differ from other students in ways that affect whether they graduate or other aggregate metrics of success. So if 20 more students in a school of 5000 and a graduating class of 300 reap the benfits of music, is this really enough to move the proverbial needle on the graduation rate for an entire school or county? Not likely.

More to the point, not every school has the *means* to have a music program, nevermind a well funded music program. According to statistics reported by Yamaha, 8% of all public schools in the United States don't have **any** arts programs at all (music, theatre, or dance).

In the U.S., schools are funded by property taxes in public schools and by extremely wealthy donors at private schools. This means that the wealthier public schools will be located in

wealthier districts which naturally has a bigger tax base. What do those districts have more of? Money! Status, class, opportunity. Instruments do not grow on trees, they cost money; not every school has 80,000 dollars for a Steinway piano. By the way (no, I did not look this up before writing it), the same study, according to Yamaha's reporting, said

> The study also noted that a disproportionate number of these students without access to arts education are concentrated in major urban or very rural areas where the majority of students are Black, Hispanic or Native American, and at schools with the highest percentage of students eligible for free/reduced-price meals.

I have to emphasize again, I didn't suspect this because I looked it up before, I suspected it because this is what it means to think like a statistician. It means you have to think in a multivariate way, where more than one thing can affect something else. Here, this idea is pretty obvious: how do we know that these graduation rates are not explained in part (if not entirely) by baseline differences in socioeconomic status of communities (poverty, low employment rates, larger contextual factors), opportunities of individual families (say, personal connections individual families may have that others don't), as well as the effects those factors have on individual students? When we think of it like this, simply suggesting that music programs would be a great solution is not so convincing.

These aside, there may also be what we call "selection bias" here, where some students are able to *select* into/decide to go to a given school. For example, some schools have magnet programs. The one I was in was for the International Baccalaureate Program; other high schools have STEM magnet programs or music/arts programs where some schools literally recruit middle school students who want to pursue these things. And when they select for these qualities, it becomes hard to isolate the impact of a music program on a graduattion rate, becasue they're already selecting for highly motivated students (who mostly but not exclusively are from wealthier districts or better-to-do families). These are the kinds of students who would perform well anyways, even if music wasn't there.

This again does not discount the real benefits of music education or the arts more generally! But when we read statistics, as with harmonic analysis of music chords, we also need to understand the context they exist within, and when we do this we begin to see that the relationship is not as clear cut as simple descriptive statistics might seem. As we see from the simulated plot, some schools have a graduation rate of almost nobody, which also happen to have low levels of music funding. And while I couldn't find specific examples of this when I looked up data on it, numbers this low do happen anecdotally. And when we see numbers like this, we must ask ourselves "How are these sets of schools different from these sets of schools?"

As we will see when we cover regression analysis, the world rarely works off of pretty, linear functions. Measures of simple association do not mean that something is causing another thing, the world is much too complex for that. As researchers and as human beings, we must constantly be skeptical of simplistic claims and investigate them when they seem too good to be true.

8.3 Tying This in With Asymptotic Theory

As we learned before, statistics is justified oftentimes on large scale asymptotics, where we consider an infinte population of units to sample from. In this framework, there's the idea of the existence of a population coefficient/average which exists only in theory and a sample (a subset of the population) which we use to calculate a mean which approximates that "true" value.

Suppose we have a spreadsheet at hand for the year 2018. The first column of said dataset/spreadsheet is the name of an American high school. The second column of the spreadsheet is some variable $m \in \{0, 1\}$. This, in English, means the column takes the value 0 or 1, where 0 means a school does not have a music program and 1 means they do have one. The third column of the dataset is the respective graduation rate of that school for 2018.

High School	Music	Graduation
Ben Franklin	1	92
Paul Revere	0	75
Alexander Hamilton	1	90
Thomas Jefferson	0	73
George Washington	1	88

Suppose now we're interested in the effect of music programs on graduation rates. Formally, we may denote our observed outcomes as

$$y_i = \begin{cases} y_i^0 & \forall\ i \in \mathcal{N}_0 \\ y_i^1 & \forall\ i \in \mathcal{N}_1 \end{cases}$$

where $\mathcal{N}_1$ is the set of schools that has a music program versus $\mathcal{N}_0$, or the set of schools that do not, where y^1 is the graduation rate we observe when $m = 1$ and y^0 is the graduation rate we observe when $m = 0$.

If we simply believed the article I cited above, we'd say that there is some effect of music education on the graduation rate (or the average of the difference between the graduation rates of schools with and without music programs, $\frac{1}{N}\sum_{i=1}^{N}(y^1 - y^0)$). Following the article, we'd say that this difference around 17.3 percentage points (in the table above, it's around 16).

But this cannot be! As we've discussed, many more factors will affect not just whether a school has a music program and their graduation rates. So, if we were to simply compute $\frac{1}{N}\sum_{i=1}^{N}(y^1 - y^0)$, the number 16 (or 17) would be wrong, **even if we had access to every single high school and college on the planet**. Because of the baseline diffenreces between these schools, our average mean difference (think a two-group t-test) would never converge to the population mean for the "effect" of music education. We would say that this is a *biased*

estimate becasue our estimate would be very far from the true effect, since the true effect is likely much smaller than 17.

8.4 Implications

The reason this matters for policy analysts is because when taken to its conclusion, mistaken correlation for causation or not thinking about things in a systemic, multivariate way could result in pumping more money into music programs in order to fix failing schools, a much wider and more sophisticated problem.

It leads to things like pumping more money into police deparmtents to decrease crime rates, even though crime has been falling in general for decades in the U.S. and there's no real link between militant policing measures and crime reduction.

It also leads to things like mass like many modern governemnts doing things like making dating apps and sponsoring mass dating events (yes, really!) in order to increase birth and marriage rates.

The idea of course is that people aren't having kids or getting married *because* they're not meeting one another. And if more people met, the more kids they'd have. And to a degree this is true due to things like social media, so the underlying logic makes plenty of sense. But then, once we agree to this probelm, we have to ask *why* are people not meeting one another and having kids or getting married? The real problem is *a lot* more systemic. Especially in South Korea, Japan, and China, the reasons are mostly having to do with labor issues, changes in gender norms, and broader social factors which lead to people not wanting to have kids.

9 Summary

Correlation and causation is a very delicate topic in statistics. Whenever we're doing research, we must always be careful to structure our studies in such as way that our findings are not corrupted by other factors, even if our software tells us we've found a "statistically significant" correlation. I'm sure if we took the t-statistic of music programs and graudation (or funding of music programs, as my scatterplot does), we'd find a very high t-statistic for the correlation coefficient. Equally, we'd find a high t-statistic for countries reporting increases in loneliness and low fertility rates. But I don't care, and neither should you, since there are other factors which contribute to graduation aside from music, and a lot more factors driving fertility rates than simple loneliness or lack of opportunity to meet people. Thinking causally can be a challenging thing. After multiple years of torment, you will learn to think like this as if by muscle memory since it'll be so routine to you. We will cover this in more detail in our chapter on treatment effect estimation.

10 OLS Explained

This is the chapter on regression. We begin by covering data types. Then, we review the idea of a function and how it relates to a line. After a review of derivatives, we finally cover the computation of regression coefficients, inference, fit, and OLS assumptions.

10.1 Math Preliminaries

> **!** Important
>
> This is a statistics course, not a math course. However, math is the language which underlies statistics. This is the *only* part of the course where we use any calculus to explain ideas. Furthermore, you'll never be expected (from me) to use calculus for any of your assignments.
> *With this said*, I do use the calculus to explain what is going on in regression. The calculus explains the *why*, which I think is always important to know if you're implementing or interpreting regression. I presume however that you are like myself when I was in undergrad, in that you've either never taken calculus (me) or had little exposure to it. So, I link to tutorials on the calculus concepts that we employ here. Consult these links, if you'd like (they sure helped me in preparation of this chapter). However, they are completely optional. I try to explain everything as best I can, so they are provided as a supplement to the curious who want a better understanding.

10.1.1 A Primer on Data Types

For any dataset you ever work with, you'll likely have different variables (columns). Regression is no exception. The predictors for regression must be numeric, naturally. These take a few different types. Note that if it is not numeric, we call it a "*string*".

The most common kind is a ratio variable (a value we may express as a fraction/continuous variable), such as the employment rate.

State	Year	Employment Rate (%)
Alabama	1990	55.3

State	Year	Employment Rate (%)
Alabama	1991	56.1
California	1990	62.1
California	1991	61.5
Georgia	1990	58.4
Georgia	1991	59.2

A dummy variable is a binary variable that indicates the presence or absence of a characteristic. A dummy variable (also called an *indicator* or *categorical* variable) is a variable that takes on the values 0 or 1. For example, a simple dummy indicates whether a respondent in a survey is a male or a female. In this case, the number 1 "maps" on to the value for male, and 0 for female. Note that in this case, the simple average of these respective columns returns the proportion of our observations that are male or female.

Respondent ID	Gender (Male=1, Female=0)
1	1
2	0
3	1
4	0

Dummies can also be used to capture unobserved variation across groups. For instance, when predicting homicide rates across states like Alabama, California, and Georgia for 1990 and 1991, we can include dummy variables for each state. These dummies help account for unique, stable characteristics of each state, such as culture, that are hard to measure directly. In other words, if we think something "makes" Alabama, Alabama, compared to California or Georgia, we can include these kinds of variables to capture that unobserved variation. When including dummies in regression, we must always omit one category from the regression (for reasons we will explain below). So, for example, we could include Alabama/Georgia dummies or Georgia/California dummies, where California and Alabama would be what we call the *reference group*.

State	Year	Alabama (1/0)	California (1/0)	Georgia (1/0)
Alabama	1990	1	0	0
Alabama	1991	1	0	0
California	1990	0	1	0
California	1991	0	1	0
Georgia	1990	0	0	1
Georgia	1991	0	0	1

There is also a notion of an *ordinal* variable, where the data at hand must obey a specific order. Suppose we ask people in a survey how religious they are on a scale from 1 to 10, where 1=Atheist and 10=Extremely Religious. Here, order matters, because 1 has a very different meaning from 10 in this instance. An ordinal variable has a clear, ordered ranking between its values.

Respondent ID	Religiosity (1-10)
1	3
2	7
3	5
4	10

These are the data types you will generally work with. When we move on to real datasets, their meaning will become much more apparent.

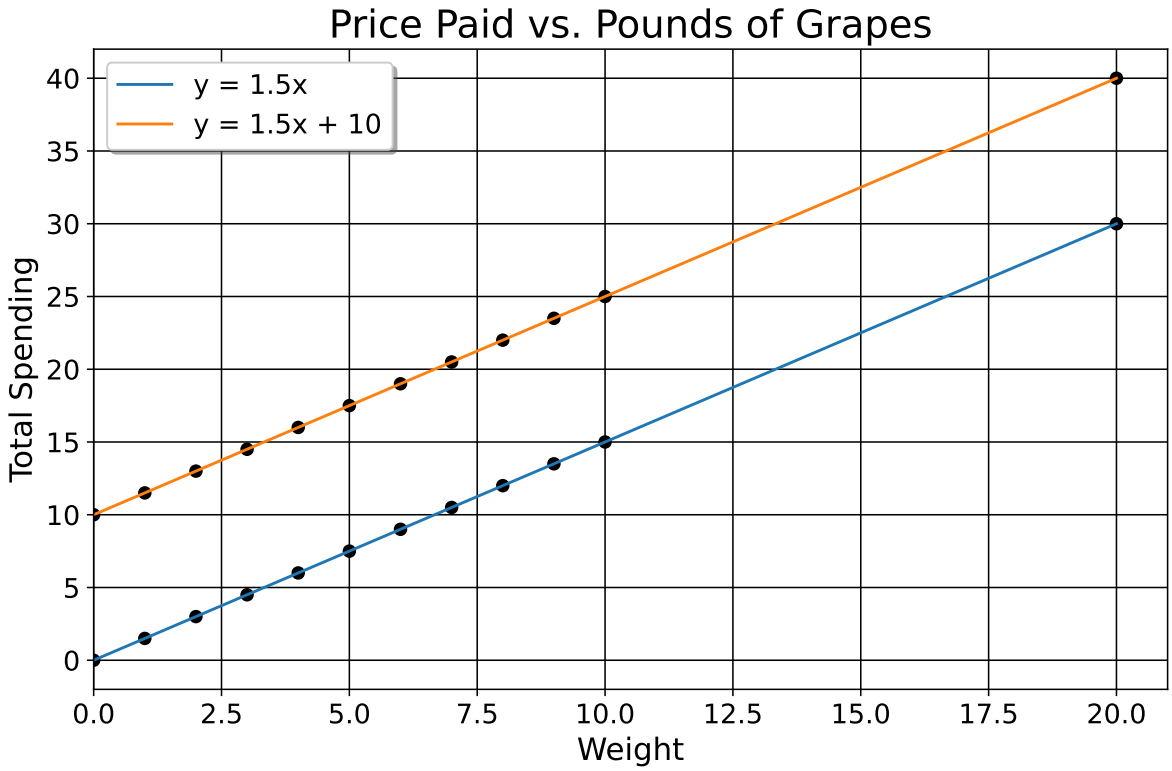
10.1.2 Review of Lines and Functions

In middle school, we learn about the basics of functions in that when we plug in a number, we get another number in return. If you're at the grocery store and grapes are 1 dollar and 50 cents per pound, we just weigh the grapes and multiply that number by 1.5. This could take the form of $(0, 0), (1, 1.5), (2, 3)$, and so on. In fact, we can represent these data points in a table like this

x	y
0	0
1	1.5
2	3

These points form a line, the equation for which being $y = mx + b$. We can also think of this line as a function. It returns a value of y given some values for previous expenditures and pounds of grapes. Here, y is how much we pay in total, m is the rate of change in how much we pay for every 1 pound of grapes bought, and b is our value we pay if we get no grapes.

For this case, the function for the line is $y = 1.5x$. For here, $b = 0$ because in this case, how much we pay is a function of pounds of grapes only. We could add a constant/b, though. Suppose we'd already spent 10 dollars, and now how much we spend is a function of both some previous constant level of spending, and new amount of grapes bought. Now, our function is $y = 1.5x + 10$. The way we find the m and b for a straight line is the "rise over run" method, in this case


Price Paid vs. Pounds of Grapes
y = 1.5x
y = 1.5x + 10
Total Spending
Weight
0
5
10
15
20
25
30
35
40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{3-0}{2-0} = \frac{3}{2} = 1.5$$

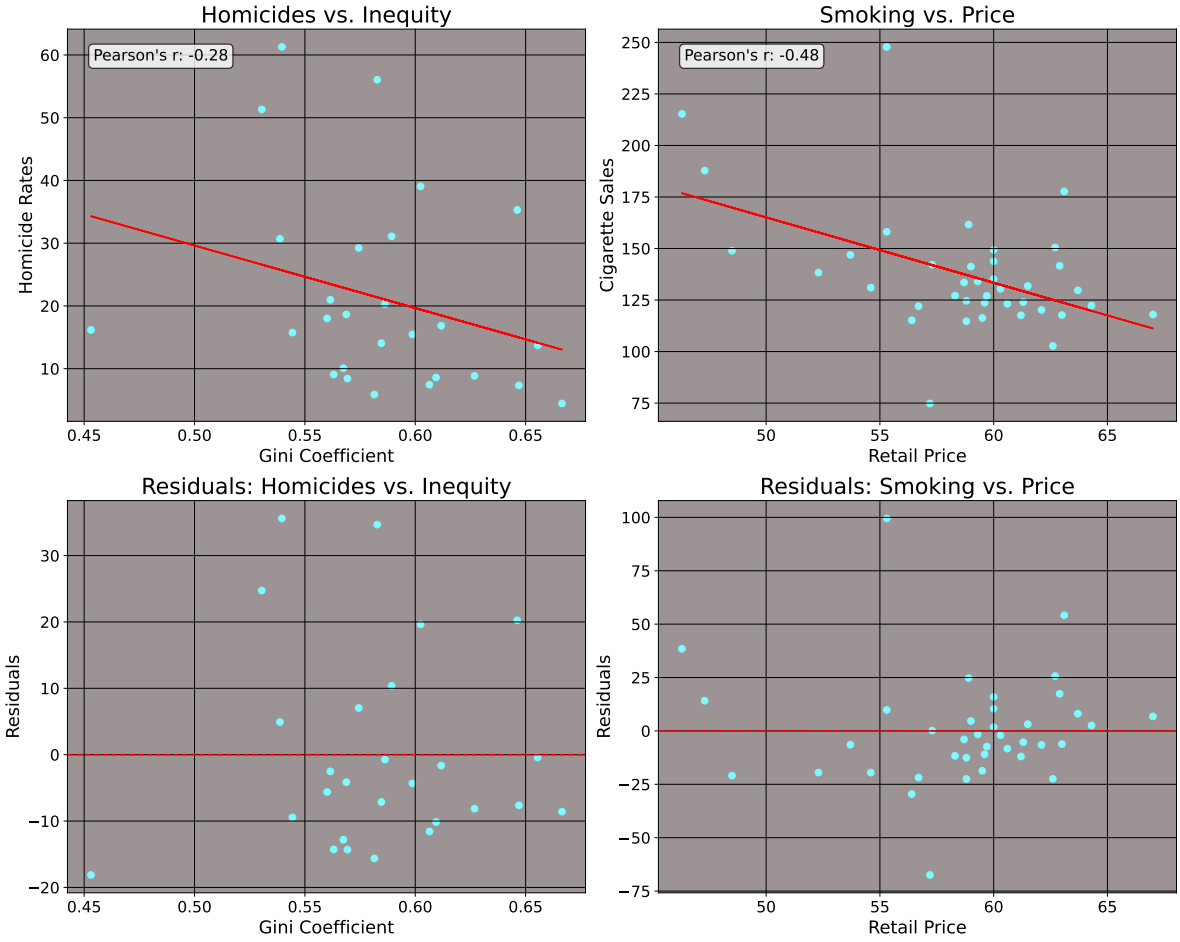
Lines fit to points sometimes have discrepancies between the line and the data we see. We call this the residual, or $\hat{\epsilon} = y - (mx + b)$. Also notice how the rate of change, or $1.5x$, is the same at every point on the line: in this instance, you pay a dollar and fifty cents for *any* amount of grapes we get.

Regression, fundamentally, is about fitting a line (or a plane) to a set of data given some inputs. Going forward, I will use the words "outcome variable" or "dependent variable" to refer to the thing that we are studying the change of, and "predictors", "covariates", or "independent variables" to refer to the variables we think affect our outcome. In the grape example, our outcome is the total amount of money we spend, and the singular predictor we use is how many pounds of grapes we get. With all this being said though, this example is very simplistic. After all, the necessary information is known to us up front (price and weight). But… what if the data we have at hand are not nice and neat in terms of a function? Suppose we consider a more challenging example.

Take the idea of predicting crime rates in Brazilian states in the year 1990 using the inequality level as a predictor, or data on the consumption of cigarettes in American states in the year 1980 using price as a predictor. We would presume some *function* exists that generates the crime rate for that Brazilian state in that year, or that consumption level for that American state in that year. We would also imagine, naturally, that the covairates inequality and tobacco price would affect these outcomes. But, does it make sense to expect for some deterministic function to predict these values, given some input? No.

The homicide rate or cigarette consumption rate in any state anywhere is not guaranteed. In some states, homicides or tobacco consumption is high, other times its low. Why? Well for homicide, some states are wealthier and some are poorer. Some states vary by racial compositions, or will differ by factors like age composition, alcohol use, and gun ownership. Thus… some cities have high homicide rates, others have low homicide rates. We can reason accordingly for cigarette consumption of American states. Naturally, one reason for this would be the price of cigarettes, as one might expect, since people tend to not want to buy more of a good as the price increases (well… usually.) The number of young people in that state may mean that younger people are risk takers and may be more likely to smoke than adults (or alternatively, young people may perceive smoking as something for older adults and smoke less). Levels of alcohol taxation may matter as well, since alcohol may be a substitute for tobacco, so states with higher taxation may smoke more, on average. Also, plain and simple measures like culture (and other unobservable things) may play a role. We can plot these data for illustration

Here, I draw a line between these input variables and the observed outcomes in question. The x axis represents, respectively, inequality and price and the y axes represent homicide rates and cigarette consumption. It is quite obvious though that no *deterministic* function exists here


Homicides vs. Inequity
Pearson's r: -0.28
Homicide Rates
Gini Coefficient
Smoking vs. Price
Pearson's r: -0.48
Cigarette Sales
Retail Price
Residuals: Homicides vs. Inequity
Residuals
Gini Coefficient
Residuals: Smoking vs. Price
Residuals
Retail Price

for either of these cases, as we have residuals. The line *imperfectly* describes the relationship between 1) inequality and homicide and 2) retail price of cigarettes and cigarette consumption per capita. So, we can't find one line that fits perfectly to all of these data points. But, even if we cannot find a perfect, deterministic function that fits to all of these data points, how about we find the *optimal* line in the sense that it best *estimates* m and b by having the lowest possible values for $\hat{\epsilon}$ at every point on the line? We can see how this relates to the grape analogy above: the line passes through all of the observed data points, meaning it is optimal in the sense that the line has the lowest possible residual values.

10.2 Arrivederci, Algebra, Ciao Derivatives.

To do this though, we've now reached a point in the course where simple algebra is no longer our best guide. We now *must* use calculus, specifically the basics of derivatives and optimization. I mentioned the word *optimal* above to refer to the fit of the line, but now we're going to get a much better sense of what is meant by this.

> **!** Important
>
> Okay, so here I'm *kind of* lying. You actually **don't** *need* to say farewell to algebra (completely) to derive regression estimates, but that process "requires a ton of algebraic manipulation". For those brave of heart who know algebra well, you can *probably* just watch the series of videos I just linked to and skip to this section of the notes, but I **do not** recommend this at all. I find the calculus way via optimization a lot more intuitive.

Firstly though, a primer on derivatives. The derivative is the slope of a curve/line given a very small change in the value of the function. It is the instantaneous rate of change for a function. We do not need derivatives for linear functions, since we know what the rate of change is from real life ("a dollar fifty *per* pound", "two *for* five", etc.). For example, the derivative of $50x = y$ is just 50, since that is the value y changes by for any increase in x. For a constant (say, 5), the derivative is always 0, since no matter what value x takes, its value does not change at all.

When we set the first derivative of a function to 0 and solve, we reach an optimal point on *the original function*, usually. An optimal point (or "critical point") is a place on the function where the value of the function is at the lowest or highest possible value the function can reach over a given set of inputs. We can find the critical points by solving an optimization problem. An optimization problem takes the form of $f(x)$

$$\underset{\theta \in \Theta}{\operatorname{argmin}} f(\theta) \text{ s.t. } g(\theta) = 0, h(\theta) \leq 0,$$

where there's some function $f(\theta)$ (called the *objective function*) that is minimized over a set of $g(\theta)$ equality constraints and $h(\theta)$ inequality constraints. The word "argmin" here means

"argument of the minimum". It means that we are seeking the values, or *arguments*, which minimize the output of that function. The symbols underneath this, $\theta \in \Theta$ represents the coefficients we are solving for. These are the values which will minimize the function. But this is all abstract. Let's do some examples.

10.2.1 Power Rule

Suppose we shoot a basketball while we stand on a 2 foot plateau, which produces a trajectory function of $h(t) = -5t^2 + 20t + 2$. Here $h(t)$ is a function representing the ball's height over time in seconds. The $-5t^2$ reflects the downward pull of gravity. The $20t$ means that you threw the ball at 20 miles per hour originally. And the 2 means that we're standing 2 feet above solid ground. We can find the *maximum* height of the ball by taking the derivative of the original quadratic function and solving it for 0. In this case, we use the power rule for derivatives. The power rule for dertivatives, expressed formally as $\frac{d}{dx}(x^n) = nx^{n-1}$, is where we subtract the exponent value of a function by 1 and place the original value to be multiplied by the base number. For example, the derivative of $y = 2x^3$ is just $6x^2$, since $3 - 1 = 2$ and $2 \times 3 = 6$. The derivative of x^2 is $2x$. With this in mind, we set up our objective function

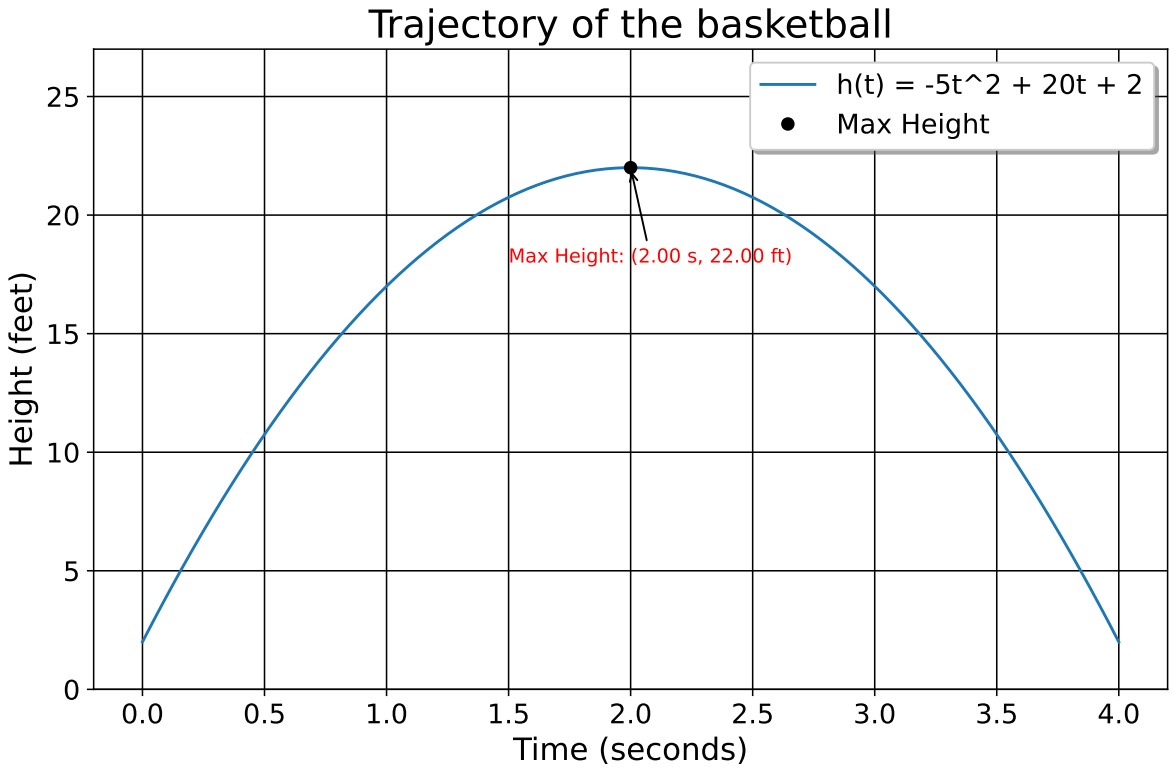
$$H = \underset{t \in \mathbb{R}_{>0}}{\operatorname{argmax}} \left(-5t^2 + 20t + 2\right).$$

where t (a postitive real number) is time in seconds, expressed as $\mathbb{R}_{>0} = \{t \in \mathbb{R} \mid x > 0\}$. We seek the value of t where the ball is at the maximum height possible.

We then follow the rules I've explained above, differentiating with respect to ($w.r.t$) each term.

$$\begin{aligned} &h(t) = -5t^2 + 20t + 2 = \\ &\frac{\mathrm{d}}{\mathrm{d}t}(-5t^2) + \frac{\mathrm{d}}{\mathrm{d}t}(20t) + \frac{\mathrm{d}}{\mathrm{d}t}(2) = \\ &t(5 \times 2) + 1(20) = \overbrace{-10t + 20}^{\text{Derivative}} \end{aligned}$$

Let's break this down. The derivative of $-5x^2$ must be $-10x$ by the power rule. All this means is that for every additional second, the ball falls by ten more feet due to gravity. The 20 reflects the initial velocity that we threw the ball at, or the 20 feet per second I mentioned above. And the derivative for 2 is 0 because time has no influence on the height of ground we threw the ball from, we started from where we started from, gravity is what it is, and thus this should not affect the rate of change of the height of the ball. We can then solve the derivative for 0 to find the optimal point.


Trajectory of the basketball
h(t) = -5t^2 + 20t + 2
Max Height
Max Height: (2.00 s, 22.00 ft)
Height (feet)
Time (seconds)
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

$$
\begin{aligned}
&-10t + 20 = 0 \Rightarrow \\
&-10t = -20 \Rightarrow \\
&\frac{-10t}{-10} = \frac{-20}{-10} \Rightarrow \\
&\boxed{t = 2}
\end{aligned}
$$

Okay, so the ball is at its zenith after 2 seconds. We may now plug in the value of 2 into the original function to get the maximum height of the ball:

$$
\begin{aligned}
&-5t^2 + 20t + 2 \Rightarrow \\
&-5(2)^2 + 20(2) + 2 \Rightarrow \\
&-20 + 40 + 2 = 22
\end{aligned}
$$

From here, we can see why we set the derivative to 0. Since the derivative is the rate of change of the function at a specific point, and we know the ball is rising vertically since we shot it upwards, we also know the ball must be slowing down over time as it rises. The maximum point is simply the height where the speed of the ball is 0 miles per hour, and therefore not changing anymore so that it may be pulled to earth.

t	$h(t)$	$h'(t)$
0	2	20
1	17	10
2	22	0
2.5	21.25	-5

We can know that we are at a maximum by taking the second derivative of our function, which would just be -10. When the second derivative, expressed as $\frac{\mathrm{d}^2}{\mathrm{d}t^2} < 0$, we know that we are at a maximum point. And since $-10 < 0$, we know that this is a maximum point.

10.2.2 Chain Rule

The next rule of differentiation to know is something called the chain rule. The chain rule is called that because of the *chain* reaction we can think of when we think of the way the value of one function affects the value of another function. In mathematics this is called a *composite function*.

A common example in economics is where we wish to maximize profits. Suppose we are the manager of shipping for a company. Our job is to ship kilos of product to a city so that a wholeseller, or a *distributor*, may sell to vendors who will in turn sell to customers. However,

we are not doing business in a competitive market. The local distributor has a monopoly over entry ports, and is able to set prices. The distributor will give us 75,000 dollars per x kilos of product we give them. As the sellers, we must come up with the right amount of kilos to sell such that we maxmize our profits, given the price we face. Ideally, we could snap our fingers and sell them the product in an unlimited manner. In other words, the profit we'd make per ki would simply be $r(x) = 75x$.

But unfortunately, we do have costs. We have a set of *fixed costs* comprising baseline expenditures of doing business with the distributor. For example, we may have transportation costs to move our product from home to the city, a certain amount of money for fuel, and other costs that we simply cannot avoid paying in order to do business at all (in the ball analogy, this would be how far we are off the grouund originally). We have *variable costs* which are things that we as the producers, in the very short run, have direct control over, say, how many drivers we have or how much plastic we use, the number of workers on our farms, and so on (in the ball analogy above, this would be how hard we throw the ball). And finally, we have marginal costs, or the costs that we bear as the business for each new ki produced (the downward pull of gravity in the above example).

For this example, our cost function is $c(x) = 0.25(x+6)^2 + 191$ (note that all of these numbers are in 1000s, but we shall solve the profit exactly as I've written it). As we've discussed above, the constant will at some point go away, and we will be left with only the factors of production that actually change (our marginal and variable costs). But I'm getting ahead of myself. First, we have to think about what profit means.

In microeconomic theory, our profit function is $\pi(x) = r(x) - c(x)$, or the difference between how much we make in dollars versus how much it costed us in dollars to produce all that we've sold. If we produced nothing, we'd be losing money since we still have to pay fixed costs (200 in this case). However, as we produce one *more* ki, we slowly increase the amount of money we make until our revenue equals our total costs. This is known as the break even point, when we're not profiting or operating at a loss, we're making just enough to remain in business. After this point, as we produce and sell more, our revenue begins to grow relative to our costs (say, as we get more efficient in distributing workers and tools to make the product). However at some point, the rate of change of profit will be equal to 0. Why? Well, we can't keep producing forever because we do not have infinite resources. This means that while we may detect a change in profits from 1 kilo to 100, at some point it will not make sense to produce another kilo because the cost it takes to make another kilo is greater than the revenue we would make from selling it. While we'd still be making profit at that level, or profits would not be maximized.

Like the ball example, we're looking to ascend the profit function by producing more until the rate of change of profit is at 0. So, to maximize profit, we differentiate each component of the profit function, $\dfrac{\mathrm{d}\pi(x)}{\mathrm{d}x} = \dfrac{\mathrm{d}r(x)}{\mathrm{d}x} - \dfrac{\mathrm{d}c(x)}{\mathrm{d}x}$. When we take the derivative of both revenue and cost functions, combine them together, and solve for 0, we can find the point at which our overall profits are maximized. As before, we may express this as an objective function, where

our *objective* is to find the level of production value that maximizes our profit. Our objective function in this case looks like

$$\underset{x\in\mathbb{R}}{\operatorname{argmax}} \left(\underbrace{75x}_{r(x)} - \underbrace{[0.25(x+6)^2 + 191]}_{c(x)} \right).$$

We consider these functions separately

$$\begin{aligned} r(x) &= 75x \\ c(x) &= 0.25(x+6)^2 + 191 \end{aligned}$$

and then take their derivatives. Beginning with revenue,

$$\begin{aligned} r(x) &= 75x \\ \boxed{\frac{\mathrm{d}r}{\mathrm{d}x} = 75}. \end{aligned}$$

Simple enough. Profit increases by 75,000 per ki sold. Now, for $c(x) = 0.25(x+6)^2 + 191$, we apply the chain rule. The outer function (or everything outside the parentheses here) is $u(y) = 0.25y^2 + 191$. The inner function is $y(x) = x + 6$.

First, we differentiate the outer function $u(y)$ by applying the power rule:

$$\frac{\mathrm{d}u}{\mathrm{d}y} = 0.25 \cdot 2y = 0.5y.$$

The constant vanishes, and the 2 from the quadratic comes down and we multiply it by 0.25. Now we differentiate the inner function:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}(x+6) = 1.$$

The chain rule states: $\frac{\mathrm{d}}{\mathrm{d}x}u(y(x)) = \frac{\mathrm{d}u}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x}$. Applying the chain rule:

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{\mathrm{d}u}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x} = 0.5y \cdot 1 = 0.5(x+6).$$

So, we are left with:

$$\frac{\mathrm{d}u}{\mathrm{d}x} = 0.5(x+6).$$

After distributing the one-half term:

$$0.5(x+6) = 0.5x + 0.5 \cdot 6 = 0.5x + 3.$$

Therefore, the final result is:

$$\boxed{0.5x + 3}.$$

Next, we combine the derivatives together in the original equation

$$\begin{aligned}
&\frac{\mathrm{d}\pi(x)}{\mathrm{d}x} = \frac{\mathrm{d}r(x)}{\mathrm{d}x} - \frac{\mathrm{d}c(x)}{\mathrm{d}x} = \\
&75 - (0.5x + 3) = \\
&75 - 3 - 0.5x = \\
&72 - 0.5x.
\end{aligned}$$

Now we solve for 0:

$$\begin{aligned}
&72 - 0.5x = 0 \\
&72 = 0.5x \\
&\frac{72}{0.5} = \frac{0.5}{0.5}x \\
&\boxed{\text{Optimal Production Level: } 144 = x}.
\end{aligned}$$

To verify that we are maximizing, we take the second derivative of the profit function, $\frac{\mathrm{d}^2\pi}{\mathrm{d}x^2} = \frac{\mathrm{d}^2r}{\mathrm{d}x^2} - \frac{\mathrm{d}^2c}{\mathrm{d}x^2}$. To do this, we do:

$$\frac{\mathrm{d}}{\mathrm{d}x} = [75 - 0.5x - 3] = -0.5.$$

The derivative of the two constants are both 0, so those are deleted. All we're left with is the linear term. Since $\frac{\mathrm{d}^2\pi}{\mathrm{d}x^2} < 0$, we are at a maximum point. Therefore, the number of kilos we should sell our distributor is 144. What is profit? $(75x) - (0.25(x+6)^2 + 191)$ where $x = 144$. First, calculate $(x+6)^2$:

$$(x+6)^2 = (144+6)^2 = 150^2$$

$$150^2 = 22500$$

Next, calculate $0.25(x+6)^2$:

$$0.25(x+6)^2 = 0.25 \times 22500 = 5625$$

This is the sum of our marginal costs and the variable costs. Next we calculate 75×144:

$$75 \times 144 = 10800.$$

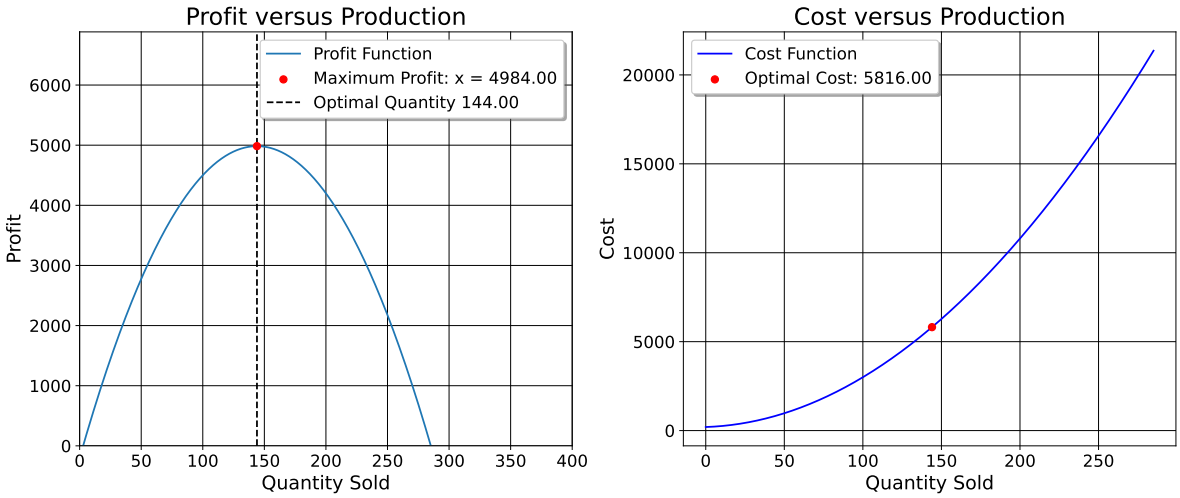
This is our total revenue. Now we add the marginal, variable, and fixed costs:

$$5625 + 191 = 5816.$$

Subtract total revenue from total costs

$$10800 - 5816 = 4984$$

Therefore, since I said all these numbers were in 1000s, our profit from 144 kis sold is $\boxed{4,984,000}$. To verify that none of what we just did is voodoo, we can check this by plotting the profit function.



Ok, all done for now. We will use derivatives to solve for the value which minimizes the sum of residuals squared. This is known as *ordinary least squares* regression (OLS), also called *linear regression*, or simply just "regression". OLS is the main estimator you'll use for this class, and it is the main foundation of econometric analysis for public policy research. It will be much more involved than what we just did, but this provides the mathematical foundation for regression as an estimator.

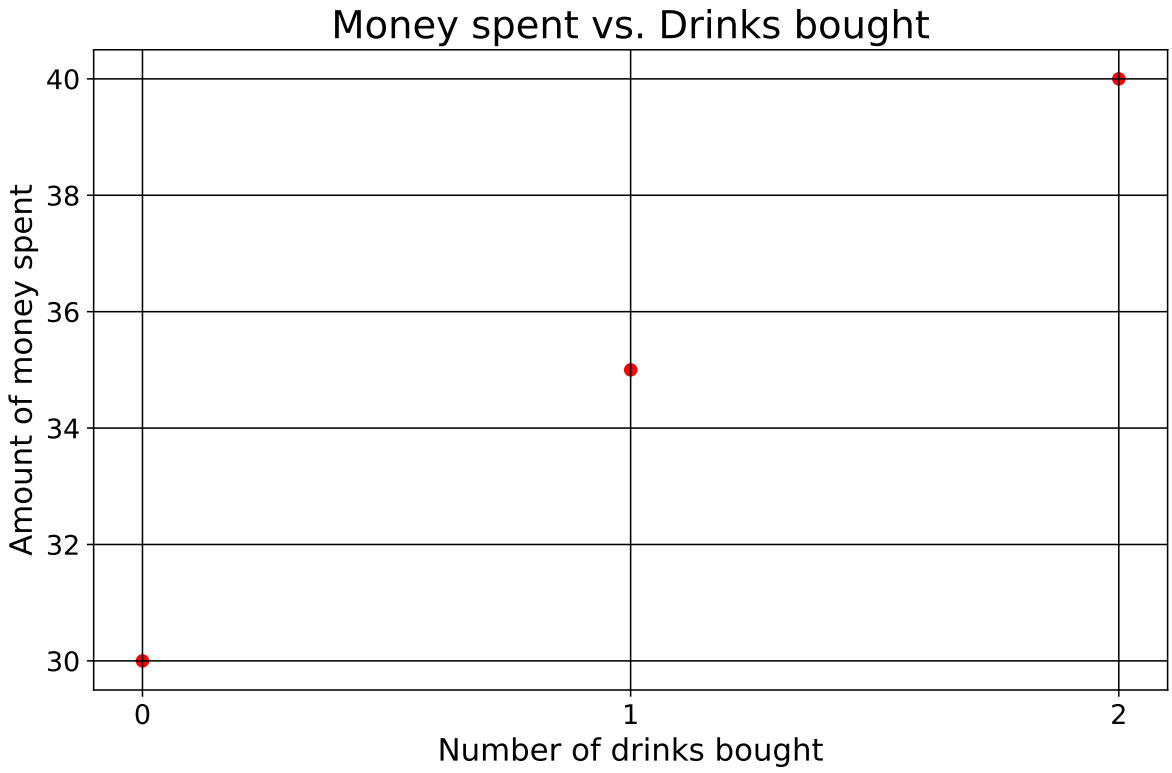
10.3 An Extended Example

To introduce OLS, we can return to the equation of a line ($y = mx + b$) where m and b are variables. Unlike the above examples where m and b were once known variables, now they

are unknown quantities we must solve for. Below, m and b will take on the values of β_1 and β_0 respectively. With OLS, we have multiple predictors (typically), each of which affect the output of the function differently.

In the multivariable case, we take the *partial derivative w.r.t.* each variable (that is, assuming that the other variables do not change). If this seems at all abstract to you, I will provide a detailed, clear derivation of the betas. Note that for all of the steps below, Stata, R, or Python does (and optimizes!) this process for you. I only provide this derivation so you have a source to refer to when you wish to know how and *why*, exactly, the machine returns the numbers that it returns to you. I also believe a clear explanation of the math will help you understand how to interpret the results that we get.

Before we continue, let's fix ideas. Suppose we wish to attend a nightclub and we wish to express how much we pay for that evening as a function (our outcome variable, a ratio level variable). At this nightclub, our outcome is a function of two things. We pay *some* one-time cost of money to enter, and then we pay *some* amount of money per new drink we buy (where number of drinks is also a ratio level variable). I say "*some*" because unlike the real world where we would know the price and entry fees by reading the sign, in this case we wish to estimate these values with only two known variables: how many drinks we bought and how much we paid.



10.3.1 List the Data

Say that we have data that looks like $(0, 30), (1, 35), (2, 40)$, where x= number of drinks we buy (0,1,2) and y=amount of money we spend that evening (30,35,40). In spreadsheet format, this looks like:

x	y
0	30
1	35
2	40

If you want to, calculate the rise-over-run of these data points to derive m and see what the answer might be in the end. Below though, we proceed by deriving what m *must be*.

$$m = \frac{35 - 30}{1 - 0} = ...$$

10.3.2 Define Our Econometric Model

We begin by defining our model. That is, we specify our outcome and the variables which affect our outcome (the values we're solving for, entry price and drink cost). Our model of how much we pay given some entry fees and additional drink costs looks like:

$$y_i = \beta_0 + \beta_1 x_i$$

Here, y_i is the total amount of money we spend that evening in dollars given the i-th drink, $\widehat{\beta}_0$ is how much we pay (also in dollars) to enter, $\widehat{\beta}_1$ is how much we pay for the i-th drink, and x is the total number of drinks we get. Nothing **at all** about this is different, so far, from anything we've discussed above. I've simply substituted m and b with the Greek letter β ("beta") into the already familiar equation for a line.

10.3.3 Write Out the Objective Function

Now that we have our model, next let's think about what the objective function would be. We already know that we wish to minimize the residuals of the OLS line. So, we can represent

the objective function for OLS as

$$S = \text{argmin} \sum_{i=1}^{n} \hat{\epsilon}^2$$

$$S = \text{argmin} \sum_{i=1}^{n} (y_i - \hat{y})^2$$

where $\hat{y}$ is defined as

$$S = \underset{\hat{\beta}_0, \hat{\beta}_1}{\text{argmin}} \sum_{i=1}^{n} (y_i - (\overbrace{\hat{\beta}_0 + \hat{\beta}_1 x}^{\text{Predictions}}))^2.$$

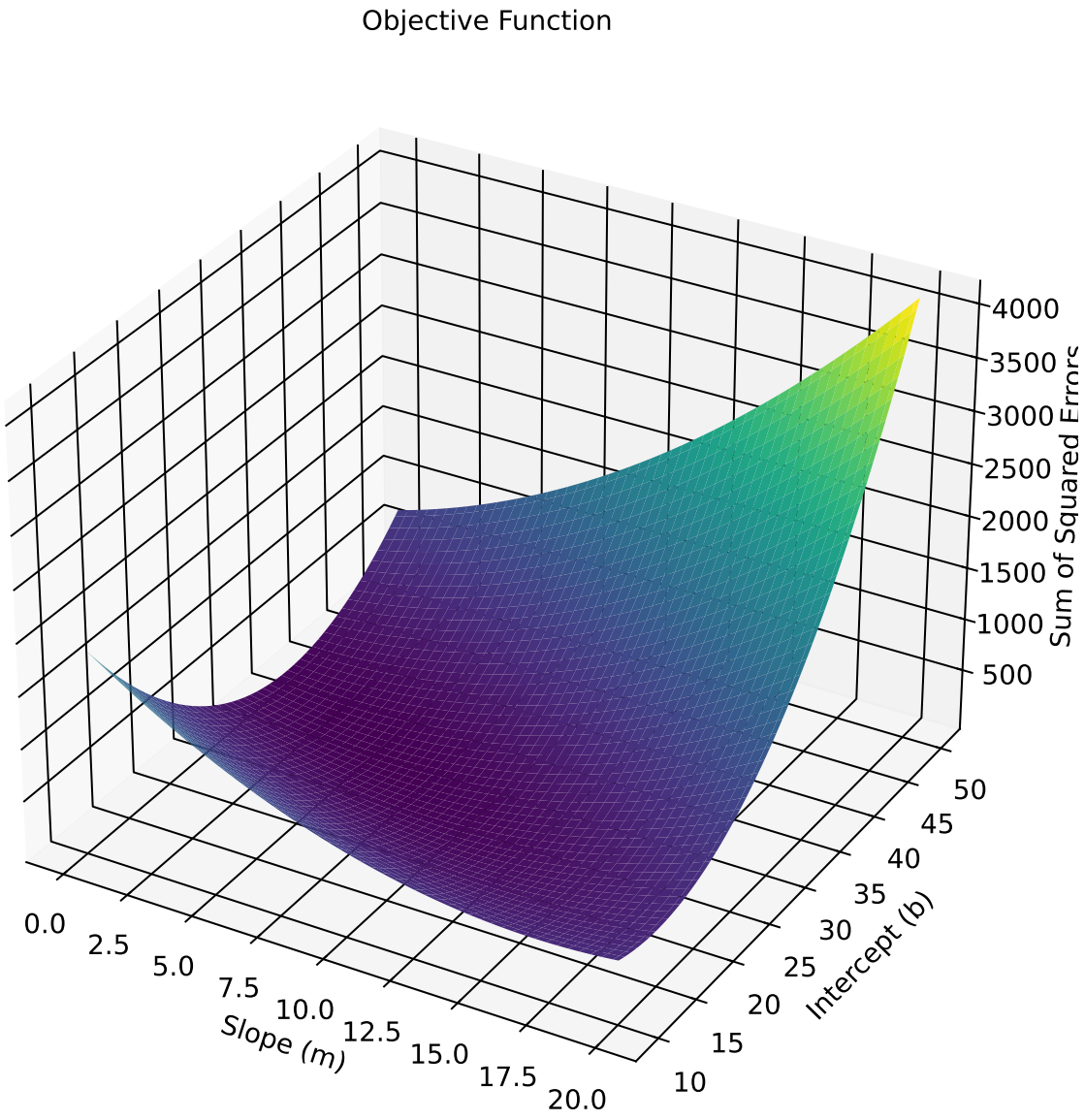
As above, we call the solutions $\hat{\beta}_0, \hat{\beta}_1$ *optimal* if they return the lowest possible values the function S can produce. What values can S produce? The sum of the squared residuals. The sigma symbol $\sum_{i=1}^{n}$ means we are adding up the i-th squared residual to the n-th data point/number of observations (in this case 3). This means that the line we compute will be as close as it can be to the observed data at every single data point. By the way, just to show the objective function is not an optical illusion or arcane expression, we can literally plot the objective function in three dimensions, where we have the slope and intercept values plotted against our total squared residuals.

I plot the values returned by the objective function given the datapoints and some value for our coefficients. Of course, we wish to get the values for the betas which produce the lowest possible values for this plane. Clearly, the intercept can't be 50 and the slope 20 (as this maximizes the residuals!).

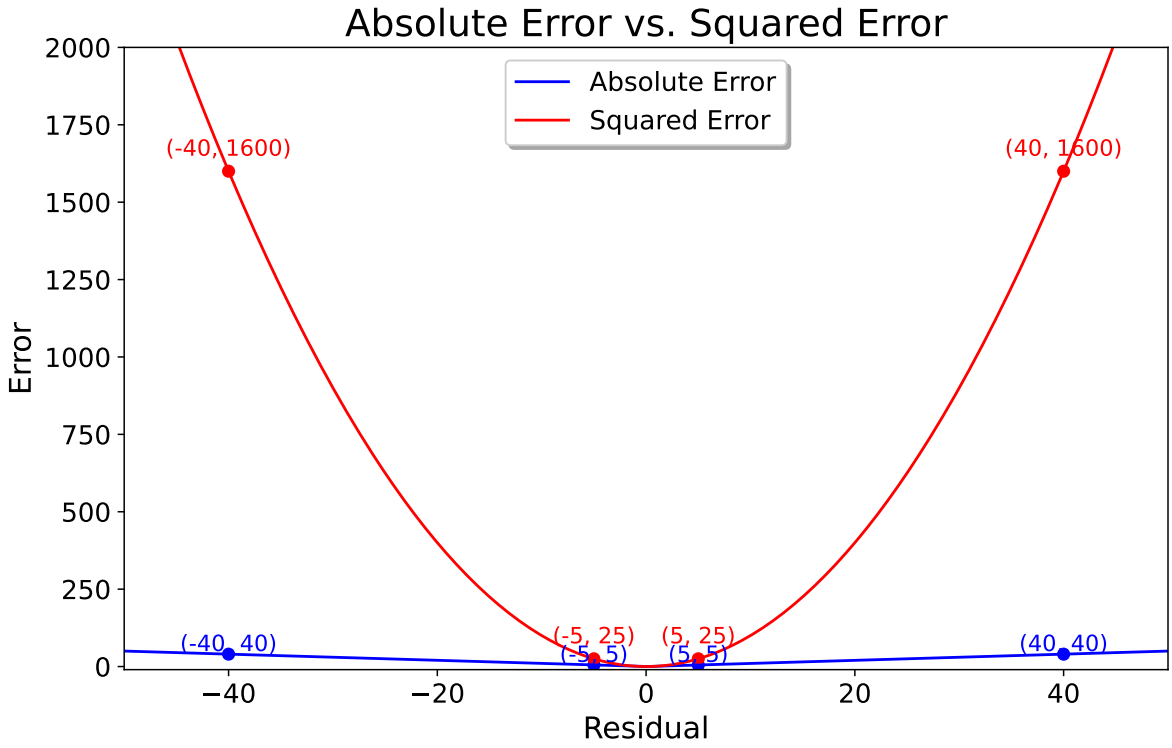
The middle formula, $(y_i - \hat{y})^2$, re-expresses the residuals. This expression should look familiar. Recall the formula for the variance of a variable $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$? Here, we are taking our real observations y and subtracting some *expected value* $\hat{y}$ (y-hat) from it. Because we are minimizing the residuals (or, the model prediction error), another way of thinking about OLS is that is the line that maximizes the explained variance given our independent variables. The lines that has the least-wrong predictions is also the line that explains its variation patterns best.

As with the variance, one may ask *why* we are squaring the summed $\hat{\epsilon}_i$, instead of say the absolute error. First of all, minimizing the raw sum of the predicted residuals (that is, without squaring them) is a non-differentiable function. We could use the raw sum of the errors as an objective function (called the *mean absolute error* instead of the *mean squared error*), but have fun doing that, as due to the non-differentiable nature of the absolute value function, we would need to use numerical methods, such as gradient descent, to compute its solution. By no means impossible… just computationally less tractable.

Using the squared residuals means that we are dealing a quadratic function which, as we did above, is easily differentiable. The squaring of residuals also penalizes worse predictions.


Objective Function
Slope (m)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Intercept (b)
10
15
20
25
30
35
40
45
50
Sum of Squared Errors
500
1000
1500
2000
2500
3000
3500
4000

Indeed, just as with the variance, all residuals *should* not be created equally. If the observed value is 20 but we predict 25, the residual is -5. Its squared residual is 25. But if the observed value is 40, and we predict 80, the "absolute" error is -40 and the squared error of is 1600. If we did not square them, we would be treating a residual of 5 as the same weight as a residual of 40. For proof, we can plot these



10.3.4 Substitute Into the Objective Function

First, we can substitute the real values as well as our model for prediction into the objective function. We already know the values x-takes. You either buy no drinks, 1 drink, or 2. So with this information, we can now find the amount of money we pay up front ($\hat{\beta}_0$) and how much it costs for each drink ($\hat{\beta}_1$)

$$S = \underbrace{(30 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0))^2}_{\text{Term 1}} + \underbrace{(35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1))^2}_{\text{Term 2}} + \underbrace{(40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2))^2}_{\text{Term 3}}$$

And when we look at the notation carefully, we see all of this makes sense: we are adding up the differences between our outcome, and the predictions of our model.

10.3.5 Take Partial Derivatives

To find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, we take the partial derivatives of S with respect to (w.r.t.) $\hat{\beta}_0$ and $\hat{\beta}_1$. Here is a short sketch of how we do this: For simplicity, I break this into two sections, one section per coefficient. In this case, the chain rule and power rules for differentiation are our friends here. To hear more about combining the power rule and chain rule, see here. First, we differentiate *w.r.t.* $\hat{\beta}_0$ (entry fees), then we do the same for $\hat{\beta}_1$ (drink fees).

1. **Partial derivative w.r.t. $\hat{\beta}_0$:**

Here is our full objective function:

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial}{\partial \hat{\beta}_0}\left[(30 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0))^2 + (35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1))^2 + (40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2))^2\right].$$

By the chain rule, we can take the partial derivative by applying the power rule to the outer functions and the linear differentiation rule to the inner functions:

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial}{\partial \hat{\beta}_0}(30 - \hat{\beta}_0 + \hat{\beta}_1 \cdot 0)^2 + \frac{\partial}{\partial \hat{\beta}_0}(35 - (\hat{\beta}_0 + \hat{\beta}_1))^2 + \frac{\partial}{\partial \hat{\beta}_0}(40 - (\hat{\beta}_0 + 2\hat{\beta}_1))^2.$$

In the first term, we have $\hat{\beta}_1 \cdot 0$, so we keep the other part of the function but the $\hat{\beta}_1$ goes away since that's what anything multiplied by 0 means. So, the quadratic power goes outside the parentheses, and the derivative of $-\hat{\beta}_1$ is just -1. So by application of the chain rule, we get this result:

$$\frac{\partial S}{\partial \hat{\beta}_0} = 2(30 - (\hat{\beta}_0)) \cdot (-1) + 2(35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1)) \cdot (-1) + 2(40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2)) \cdot (-1).$$

See how all these three terms are multiplied by -1 and 2? Well, by distributive property, we know we can factor out the 2 and negative 1. That returns this result:

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\left[(30 - \hat{\beta}_0) + (35 - \hat{\beta}_0 - \hat{\beta}_1) + (40 - \hat{\beta}_0 - 2\hat{\beta}_1)\right].$$

Next I rearrange everything inside of the brackets:

$$(30 + 35 + 40) - (\hat{\beta}_0 - \hat{\beta}_0 + \hat{\beta}_0) + (\hat{\beta}_1 - 2\hat{\beta}_1) =$$
$$(105) + (-3\hat{\beta}_0) + (-3\hat{\beta}_1).$$

Finally, we just distribute the 2:

$$\begin{aligned} &-2\left[105-3\hat{\beta}_0-3\hat{\beta}_1\right] \\ =&\boxed{-210+6\hat{\beta}_0+6\hat{\beta}_1}. \end{aligned}$$

This is our partial derivative for the first coefficient, or for the entry fee.

2. **Partial derivative w.r.t. $\hat{\beta}_1$:**

We can follow a similar process for this partial derivative:

$$\frac{\partial S}{\partial \hat{\beta}_1}=\frac{\partial}{\partial \hat{\beta}_1}\left[(30-(\hat{\beta}_0+\hat{\beta}_1\cdot 0))^2+(35-(\hat{\beta}_0+\hat{\beta}_1\cdot 1))^2+(40-(\hat{\beta}_0+\hat{\beta}_1\cdot 2))^2\right]$$

Using the chain rule, this looks like:

$$\frac{\partial S}{\partial \hat{\beta}_1}=\frac{\partial}{\partial \hat{\beta}_1}(30-\hat{\beta}_0)^2+\frac{\partial}{\partial \hat{\beta}_1}(35-(\hat{\beta}_0+\hat{\beta}_1))^2+\frac{\partial}{\partial \hat{\beta}_1}(40-(\hat{\beta}_0+2\hat{\beta}_1))^2.$$

We apply the power rule to the outer terms and the linear differentiation rules to each inner term. As before, the 2 simply goes in from of the parentheses and the derivative of $\hat{\beta}_1$ is taken. Note that for the first term, $\hat{\beta}_1$ is multiplied by 0, so since this is multiplied by the outer function, the first term vanishes completely.

$$\frac{\partial S}{\partial \hat{\beta}_1}=2(2(35-(\hat{\beta}_0+\hat{\beta}_1))\cdot(-1)+2(40-(\hat{\beta}_0+2\hat{\beta}_1))\cdot(-2).$$

As we can see, the 2 again is a common term, which we again put outside in brackets:

$$\frac{\partial S}{\partial \hat{\beta}_1}=2\left[-1\cdot(35-(\hat{\beta}_0+\hat{\beta}_1\cdot 1))-2\cdot(40-(\hat{\beta}_0+\hat{\beta}_1\cdot 2))\right].$$

When we simplify the inner terms, we get:

$$\frac{\partial S}{\partial \hat{\beta}_1}=2\left[-(35-(\hat{\beta}_0+\hat{\beta}_1))-2\cdot(40-(\hat{\beta}_0+2\hat{\beta}_1))\right].$$

We apply the distributive property again for the 1 and 2:

$$2\left[-35+\hat{\beta}_0+\hat{\beta}_1-80+2\hat{\beta}_0+4\hat{\beta}_1\right].$$

Now we rearrange by putting the same terms next to each other:

$$2\left[\hat{\beta}_0 + 2\hat{\beta}_0 + \hat{\beta}_1 + 4\hat{\beta}_1 - 35 - 80\right].$$

and simplify by combining them together:

$$2\left[3\hat{\beta}_0 + 5\hat{\beta}_1 - 115\right].$$

Thus after distributing the 2 inside of the brackets, the partial derivative of S with respect to $\hat{\beta}_1$ is:

$$\boxed{6\hat{\beta}_0 + 10\hat{\beta}_1 - 230}.$$

Now we've taken the partial derivatives of both our variables, entry fees (which we presume are constant) and the number of drinks we buy. This can be represented like:

$$\nabla S = \begin{bmatrix} \frac{\partial S}{\partial \hat{\beta}_0} \\ \frac{\partial S}{\partial \hat{\beta}_1} \end{bmatrix} = \begin{bmatrix} -210 + 6\hat{\beta}_0 + 6\hat{\beta}_1 \\ -230 + 6\hat{\beta}_0 + 10\hat{\beta}_1 \end{bmatrix}.$$

Technically, in mathematics, we'd call this the gradient. Before we continue though, do not lose sight of our goal: all these two equations represent are the instantaneous rates of change in our sum of squared residuals given some change in the variable in question. Our goal is still to find the the values for these betas that minimize our objective function.

10.3.6 Get the Betas

Okay, no more calculus. We can now return to *algebraland* to get our betas, with a slight modification.

Remember how above after we calculated the normal derivative of the profit function or the ball trajectory we just solved the derivative for 0? In that case, we had only one variable, t or x. Well, now we don't just have one variable! We have 2 $\hat{\beta}_1$ and $\hat{\beta}_0$.

As before, we still want to set these both equal to 0 because at the point both equal 0, our sum of squared residuals is no longer rising or falling (or, it is the the critical point on the surface I plotted above). So, let's write these equations again (I divided the first equation by 6 since all of its terms were divisible by 6). To solve both equations simultaneously, first we add the constants to both RHS of both partial derivatives:

$$\hat{\beta}_0 + \hat{\beta}_1 = 35$$
$$6\hat{\beta}_0 + 10\hat{\beta}_1 = 230.$$

Here I use a method called substitution to solve the system, but there are many such ways we can solve this. In substitution, we solve one equation first and substitute the solution for a variable into the other equation. I solve the first parital for $\hat{\beta}_0$ since it is the easiest. So, we subtract $\hat{\beta}_1$:

$$\hat{\beta}_0 + \hat{\beta}_1 = 35 \Rightarrow$$
$$\hat{\beta}_0 = 35 - \hat{\beta}_1.$$

Okay, so this is our expression for β_0. Since we now know the expression for the constant (the entry fee), we can plug this into the partial for $\hat{\beta}_1$ where the β_0 currently is and solve for $\hat{\beta}_1$. We do:

$$6\hat{\beta}_0 + 10\hat{\beta}_1 = 230$$
$$6(35 - \hat{\beta}_1) + 10\hat{\beta}_1 = 230.$$

Next, we distribute the 6

$$210 - 6\hat{\beta}_1 + 10\hat{\beta}_1 = 230$$

and combine the terms $-6\hat{\beta}_1 + 10\hat{\beta}_1$ together. That returns this result:

$$210 + 4\hat{\beta}_1 = 230.$$

Next, we subtract 210

$$4\hat{\beta}_1 = 20.$$

Finally, we divide by 4

$$\boxed{\hat{\beta}_1 = 5}.$$

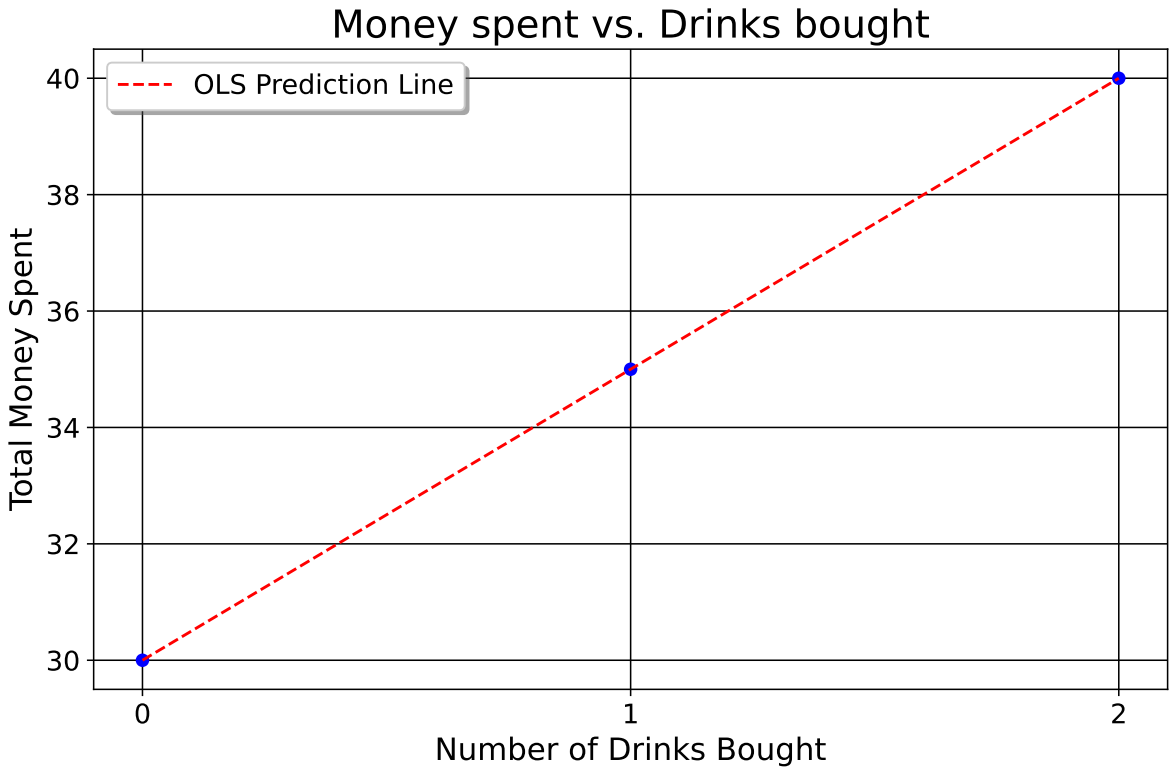
Now, we know our value for $\hat{\beta}_1$!!! We know that for each drink we get, we pay 5 more dollars. Since we now know *this*, we substitute 5 into $\hat{\beta}_0 + \hat{\beta}_1 = 35$ where $\hat{\beta}_1$ is. Then, we have one equation to solve for, with our goal being to get the value of $\hat{\beta}_0$. We can put this value into the partial derivative for $\hat{\beta}_0$:

$$\hat{\beta}_0 + 5 = 35.$$

Now, we simply subtract 5 from the RHS

$$\boxed{\hat{\beta}_0 = 30}.$$

The entry fee is 30 dollars.


Money spent vs. Drinks bought
OLS Prediction Line
Total Money Spent
Number of Drinks Bought
40
38
36
34
32
30
0
1
2

10.3.7 Our OLS Line of Best Fit

So, our line of best fit is $\hat{y} = 30 + 5x$. In social science, you'll hear people throw around terms like "when we controlled for *this*" or "*adjusted for*" another variable, or "when we *hold constant* these set of variables". This is what they mean by it! They, in the plainest language possible, mean that the dependent variable changes by *that amount* for every unitary increase in an independent variable, assuming the other values of the function do not change change. That's exactly what the partial derivative is, the change in a function given a change in one variable for that function. So here, assuming the club has a flat entry fee that does not change on a person to person basis, the amount the function changes by for every new drink is an increase of 5 dollars. Or, *compared* to the scenario where you only wanted to get in the club (and not drink at all, where $x = 0$), you spend 5 more dollars per each new drink you get. One may ask why we did this at all. Why *bother* with the partial derivative approach and the messy system of equations, why not simply display a regression table and go through the practical interpretation? After all, assuming we just did the following in Stata: `reg y x`, we would've gotten the exact same set of results that I just did quicker.

The primary reason is pedagogical. OLS was never derived for me in quite this manner in undergrad. So I believe you should see it done with a simple, tractable, familiar example, even though you'll never do this for any work you ever do. This way, OLS is not a computerized black box you mindlessly use for a dataset- you actually can *see* where the numbers come from in a simplified way.

10.4 Inference For OLS

Now that we've conducted estimation, we can now conduct inference with these statistics we've generated. Indeed, this is the primary point of this at all, in a sense. We *want* to know if these estimates are different from some null hypothesis. To begin, recall the notation of $\hat{\epsilon}_i$ which denotes our residuals for the regression predictions. We can use this to generate the standard error/the uncertainty statistic associated with the respective regression estimate. We can begin with the residual sum of squares, calcualted like $\text{RSS} = \sum(\hat{\epsilon}_i)^2$. Put another way, it is all the variation *not* explained by our model. If $RSS = 0$, as was the case in the above example, then we have no need for inference since there's nothing our model does *not* explain. We then can estimate the variance of the error like $\hat{\sigma}^2 = \frac{\text{RSS}}{n-p}$, where n is our number of observations and p is our number of predictors (including the constant). We divide by $n - p$ because this takes into account our model's residual degrees of freedom, or our model's freedom to vary. Note as well that when $n = p$, the error variance is not defined, meaning for OLS to be valid we need less predictors than observations. For a more detailed explanation of degrees of freedom, see Pandey and Bright (2008).

For an example of how a regression table is presented, consider the above example that estimates the impact of tobacco prices on consumption across states

	Estimate	Std. Error	T-statistic	R^2
Intercept	323	56.3691	-3.31	
Coefficient (retprice)	-3.17	.958191	-3.028	0.22

As we know from above, this suggests that a one dollar increase in the price of cigarettes implies a reduction of 3 in the rate of tobacco sales per capita. As we've discussed with normal descriptive statistics/classic hypothesis testing, we can also compute confidence intervals for these regression estimates. To do this, we need a standard error for our regression estimates. We compute:

$$\frac{\frac{\text{RSS}}{n-(k+1)}}{\sum(x_i - \bar{x})^2}$$

We already know RSS from above. Then, we add the differences of each point for x and its mean. This is:

$$\frac{\frac{26015.2655}{37}}{765.81} = 0.958191.$$

Now that we have this, we can calculate the t-statistic for the beta, which is simply the coefficient divided by the standard error. We can also calculate confidence intervals for coefficients too. The formula for this should appear quite familiar

$$\beta_j \pm t_{\alpha/2,\text{df}} \cdot \text{SE}(\beta_j)$$

Here, β_j is the coefficient of our model, t is our test statistic (1.96 ususally), α is our acceptance region (0.05 in most cases if we want a 95% confidence interval), SE is our standard error as we've computed it above. For the price example, we do $-3.171357 + (1.96 \times .958191)$ and $-3.171357 - (1.96 \times .958191)$, returning a 95% CI of $[-5.112836, -1.229877]$. There's a little rounding error, but that's what we get. The way to interpret the CI is as follows: given our sample size and input data, the true parameter of the effect of price on tobacco smoking rates lies within the range of -5.1 and -1.2. In other words, if some assumptions hold (which we will discuss below), a dollar increase in price may decrease the tobacco smoking rate by as much as 5 or as little as 1.

Just as we discussed in the preceding chapters, lots of statistics is justified asymptotically, based on the law of large numbers and CLT. In other words, as n tends to infinity, $\lim_{n\to\infty}$, our betas will converege to the true population value and the standard errors will shrink. Ergo, as these shrink, the confidence intervals will tighten, meaning our estimates will be more precise. A practical consequence of this is that as a very general rule, having more observations in your dataset makes your OLS estimates more precise and less biased. For the above for example, we would not trust these estimates as much because they come from one year only. Ideally,

to get a better sense of how price increases affect consumption, we'd need to collect these observations over time and adjust for other things that may affect cigarette consumption.

10.4.1 Goodness of Fit Measures for OLS

We typically thhink of two goodness of fit statistics when using OLS, the R-squared statistics and the Root Mean Squared Error

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Here, for R-squared, we have two terms: first, $\text{SS}_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared residuals. SS_{res} quantifies the amount of variance unexplained by the independent variables in the model, and SS_{tot} is just the total amount of variance of y. Think of R-squared as a ratio/percentage of how well the model explains our outcomes. An R-squared of 1 reflects the simple example we derived above, where the model perfectly explains the variation of our outcomes. R-squared usually scales with the amount of predictors in the model (that is, as we add more variables, the R-squared will increase). However, I think the best way to think about R-squared is a measure of how well our model does, compared to the average of our outcomes. In the nightclub example, if we just took the average of our outcomes, we'd guess for that evening, you'd spend 28.3 dollars. But, in this case, the model significantly outperforms this since it explains the variation perfectly. Note, that it is possible to have a negative R-squared. It is very rare, and it basically means that your model does a worse job than the simple average of the outcomes. I've seen this in my work, but it is very rare. I've only encountered it in the wild maybe twice. In the above example, including just price as a predictor explains about 22 percent of the variation, which is not bad considering it's only one variable!

The Root Mean Squared error, or RMSE, is exactly as it sounds: it is the square root of our average of our SS_{res}. For the tobacco, example, our RMSE is $\frac{26015.2655}{3726.516} = 26.516321$. In English, this simply means that when we use price to explain consumption for the year 1980, the model is off, on average, by about 27 packs. Again, considering that the average cigarette consumption in 1980 was 137,being off by 26 packs isn't so bad. It suggests, as one would expect, that we've explained our dataset fairly well using price as an explanatory variable. As RMSE approaches 0, we explain our variation better, with an RMSE of 0 being perfection. Note that other goodness of fit metrics do exist; however, these are the most common ones you'll encounter.

10.5 Assumptions of OLS

Keep in mind, despite all the detail we've discussed so far, do not lose sight of the larger picture: OLS is simply a mathematical estimation method. Its *validity* for explaining the

external world (aside from having quality data) relies on a few assumptions (collectively called the Gauss-Markov assumptions) being defensible. I say defensible instead of true because practically they are never true. After all, this is statistics: almost all of statistics is true. All statistical research (pretty much, outside of simulations) is at least partly wrong because we live in a probalistic world where we don't have all the answers. In other words, the assumptions are only as valid as we can defend for them. Below, I detail them.

10.5.1 Assumption 1: Linear in Parameters

The very first assumption is that the parameters for our model are linear. The classic regression model for OLS is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \epsilon_i.$$

We call this a linear model. Why? How do we know if it's a linear relationship, and what might violations of this look like? Let's say we're buying weed. Say the price per quarter ounce is 80 dollars. The impact of β_1 is the same everywhere in the function, $y = 80x$. But step back and ask ourselves, from the seller's persepctive, if this makes sense: does it make sense for weed to cost the same for every weight amount? No! Why not? Well, for one, let's say you're selling a full gram or pound of weed. That's ***so much*** weed that weed(wo)men/people will charge much much more for lone individuals who wish to buy this much. So while it may be 80 for a quarter ounce, it'll now be, say, 900 per pound. In fact, we could express this as a piecewise function

$$\beta_j = \begin{cases} 80 \text{ if } x < 1 \\ 900 \text{ if } x > 1. \end{cases}$$

Why might this be done? Firstly, that's so much more product than the average person could smoke or use. So, anyone interested in this would need to pay premium prices for such an outlandish amount. Also, it allows the dealer to get the pound of weed off their hands– relative to ounces, pounds of weed are much more likely to be noticed by police and therefore punished by the law harsher. So, the quicker they sell, the quicker they may re-up. So, for the normal crowd of people who do not but pounds, they pay one price. For those who are abnormal in how much they demand (say, the distributor for the connect for cocaine markets), they pay another price altogether. We see price discrimination in legal markets too, such as Sams Club. We can see that a regression model here IS NOT linear in parameters, since the slope of the line will change at different values of the function.

People often confuse this assumption with non-linear values of our independent variables as they relate to our outcome. They conflate nonlinear regression

$$y_i = \beta_0^2 + \beta_1^2 x_{i1} + \cdots + \beta_K^2 x_{iK} + \epsilon_i$$

with

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \cdots + \beta_K x_{iK}^2 + \epsilon_i,$$

or an OLS model with non-linear relations between the inputs and the outcomes. Let me explain why this is wrong, because as it turns out, we can indeed model curvature. I've already given an example of when we'd have a nonlinear realtionship in terms of our betas. Now I discuss non-linearities in terms of our predictors. Let's say we wish to model how much someone can dead lift given some input of age. Let's say the true population parameter for the OLS model is 6 (we ignore the constant for exposition purposes)

$$y_i = 6x_i$$

What is our value for 0? 0, since you're not yet born. For age 10? 60. For age 30? 180. For age 80? 480. I think we can already see why this relationship being modeled would be silly: it presumes that the older you get, the stronger one is as a hard and fast rule. Which, generally speaking, is true… but we also know that at some point, as with all things, glory fades. Someone that was once strong and in shape will not (in general) always be that way because the body declines with the passage of time. How do we take this into account for our regression model, though?

$$y_i = \beta_1(x_{i1} \times x_i) + x_i \equiv y_i = \beta_1 x_i^2$$

We simply square the original value of age, keeping its linear form in the regression model. That way, when age 4 is input in the model, the number our regression model reads in the computer is 16. When age 10 is put into the model, it reads 100. Of course, as one would expect, there's likely some critical point for this function, where people begin to be able to lift less given some values of age. We never know this of course, but OLS can be used to estimate it in the manner that we've done.

Another example of being able to account for non-linearities from economics is the idea of modeling how much produce one may produce given a set of labor inputs. Suppose we're cooking cocaine. With just two people, you can get work done, but it won't be a lot. With three people, you can do more, and more with each additional person. However, there's an idea in economics called diminishing marginal returns for the factors of production (in this case labor). You may be able to cook a lot with 10 or 20 people, but when you have 40 or 50 people, at some point we end up producing less because there's too many proverbial cooks in the kitchen. So, if we wished to model output of cocaine as a function of labor, we'd likely wish to square the "number of workers" part of our model since it does not make sense to expect production to increase perfectly linearly with every new human working with you. So you see, the linear in parameters assumption deals with our betas impact on our predictor variables, not the input values of our predictor variables.

Note that when we include such terms in our model, called *interaction terms* in econometrics (I may cover this more later), we must include the linear term in the model as well. In Stata, this would look something like `reg y c.labor##c.labor`. Under the hood, this includes in the model `reg y labor labor2`.

10.5.2 Assumption 2: Random Sample

We next presume that we've collected a random sample of our population. The name *random* sample is something of an antiquated, romantic name to denote the idea that the sample we've collected is representative of the population we wish to map on to. Suppose we wish to investigate the relationship between introducing all virtual menus at a restaurant (the kind you scan on your phone) to see if it increases how much money they make for [*random marketing reasons*]. We take all the restaurants in Buckhead and Sandy Springs in Atlanta as a sample, comparing the ones that did this intervention to the ones that didn't do the intervention. We get a coefficient of 10,000, suggesting that this intervention increased money made, on average, by 10,000 dollars compared to the places that didn't do this. The issue with this idea is that our sample is not a random sample of the population. Sandy Springs and Buckhead, in particular, are among the wealthiest areas in Atlanta. We can't generalize the effet of this intervention to the population (restuarants in Atlanta, say) becuase our units of intertested are decidedly *not* representative of Atlanta's entire culinary scence. They have very specific customers that make a generalization to the bigger population a bad idea.

Another example can come from sports. Say we posit a relationship between height and skill at basketball. We take a sample of NBA players, run a few regressions for relevant metrics and have our software spit out coefficients at us. Can we generalize to the population? No!! The NBA is one of the most selective sports leagues on the planet. The NBA selects for height and skill, among other things. The worst player on the NBA bench is like a god from Olympus compared to the average human, physically and in terms of skill. They are **not** representative of even a human 2 standard deviations above the mean.

So, we cannot use NBA players generalize to the population, *unless* of course we are concerned only with NBA players. The same would apply to the first example: if we care only about the high-income restaurants in the city, then that's great, but assuming we wish to generalize more broadly, we will need more data from other, more diverse units that have more information encoded in their outcome about the sample.

10.5.3 Assumption 3: No Perfect Collinearity

The simple way to think about this one is we cannot include redundant variables in our regressions. Suppose we wish to predict the ticket sales of NBA teams. In our regression, we include the number of games won as well as the number of games lost (2 variables). Well, these are mirror images of each other. The number of games you won is a direct function of the total games minus the number you lost, and the number you lost is a direct and perfect function of the total minus the number you won.

By extension, suppose we wish to compare women to men (say we wish to test that men earn more/less than women on average). We take data on 500 respondents who we've sampled randomly across a company. We have one column that denotes the respondent as male and

the other as female. We cannot include both male and female columns in our models, these are perfect linear functions of one another. A female is necessarily not coded as male, and male is necessarily not coded as female. Practically, this means we must choose when we use a categorical variable in our models. Say our regression includes age ang gender as a predictor. If category 1 of gender is female and category 0 is male, then if the beta for "gender" is -30, we would interpret the beta for gender as "holding constant age/compared to men of the same age group, female respondents earn about 30 dollars less than men." By extension, the coefficient for male (if we decided to include this group as the group of interest) would just be 30, with a similar interpretation in the other direction.

10.5.4 Assumption 4: Strict Exogeneity: $\mathbb{E}[\epsilon_i | x_{i1}, \ldots, x_{iK}] = 0$

Next we presume strict exogeneity. Formally, this means the average of our errors, given the set of covariates we've controlled for, is 0. It means our predictor variables may not be correlated with the error term. Note that the error term is different from the residuals: the error term includes unobserved characteristics that also may affect the outcome. In other words, we cannot omit important variables from our model.

For example, say we wish to see how the number of firefighters sent to a call affects the amount of damage from that fire. We wish to measure the association in other words. We conclude that there's a positive relationship between number of firefighters sent and damage. Say, we use the number of trucks sent to a call to predict the damage in dollars for a sample size of 10,000,000 calls. We find from the bivariate model that every truck you send increases damage by 30,000 dollars. So, we elect to send *less* people to future calls. Is this a good idea? No!!!! People will die like that.

Presumably, the firefighters are not pouring gasoline on the fire, so perhaps we've *omitted* things from our model that might influence *both* how many people we send as well as fire damage. What else should we control for? Maybe, building size, building type, neighborhood income status, local temperature, and other relevant predictors to ensure that we are not blaming the outcomes on a spurious relationship. Indeed, on some level we would expect for the size of the fire to be correlated with the number of people sent to fight it. Thus, when we do not control for other relevant factors, our coefficients, no matter how precise, suffer from *omitted variable bias*. Strict exogeneity is pretty much **never** met in real life, but it basically posits that there's no other critical variable missing from our regression model that may explain our outcome. This is also why it matters to critically think about the variables one will use in their regression model *before* they run regressions.

11 Summary

This undoubtably is the most weighty chapter, both in terms of mathematics and in terms of practical understanding. Regression is one of the building blocks for policy analysis, in addition to solid theoretical background and contextual knowledge of the policy being studied. The reason I chose to cover this first, in the first few weeks of the class instead of waiting until the end, is because I believe that the only way to *truly* understand regression is by use in applied examples. This is what you'll wrestle with in your papers.

12 Causal Inference

Statistics teachers oftentimes proudly declare to their students that correlation is not causation when emphasizing the idea that just because two things move together that doesn't mean that one thing is causing the other thing. We've discussed examples of this before. So, the question that (at least for me) itched in the back of my mind was "Okay. Well, what *is* causality then? What does it mean for a thing to cause another thing?"

Our final chapter covers treatment effects/causal inference for policy analysis. Strictly speaking, we could have multiple courses on this. So, as an introduction, this chapter seeks to provide you with the basic philosophy of causal inference, specifically, what it is as a concept, how it's used in the policy sciences, and how we may use regression to implement basic causal designs for research.

> **!** Important
>
> In this chapter, I (in addition to the basic philosophy of causality) introduce one of the basic causal inference methods in econometrics and public policy: the difference-in-differences design. It is designed specifically for impact analysis (that is, how one policy affected some specific outcome). Even though it is the last thing we cover, it is *not* a requirement for your final papers. You may use normal regression to study mere associations if you so choose.

12.1 What Is Causality?

As I mentioned in the chapter on correlation, humans have *evolved* to think hypothetically. It is how we have survived for as long as we have. Causal inference demands that we imagine another world that we believe could exist, but doesn't exist. In history, we'd call this a "counterfactual", so termed because we are talking about a fictional scenario that happened in contrast to observed facts. We as human beings do this all the time.

- How would the American economy have evolved post 1860 if the Civil War never happened?
- What if a school did a new math curriculum? Would math scores improve?

- How would gun homicide statistics look, 6 months from now, if a state *didn't* pass gun control policies?
- How would a grocery store's in store sales have evolved if it didn't implement all self checkout scanners?
- Did Nebraska's repeal of the tampon tax affect tampon use? How would tampon sales have evolved if Nebraska didn't get rid of the tax?
- How would New Orleans' outward migration have looked if Hurricane Katrina didn't happen?

A counterfactual, at its heart, is the way a metric, outcome, or construct *would have* looked in a world where what did happen (some treatment, policy, or intervention), did not happen. However, we get but one copy of reality. We can't literally look at the United States where the Civil War happened (the reality we have) and one where it didn't happen (not that we'd really want to, by the way). We can't have a school with one grade level has two math curriculums (the current one and new one) at once, and even if we could, how could we know the new curriculum is the driver of grades instead of something else? We can't see the same city that has banned guns and not done so, or a state that both taxes and doesn't tax tampons. Thus, counterfactuals are things we can estimate, guess about, and speculate on, but never see in real life. Before we get into how we'd estimate counterfactuals statistically, though, let's use a more relatable example.

Suppose I'm going to school today. I think the way I take to school (Way A) is quicker than Way B. This gives us a set of two ways to take, \(d \in \{0,1\}\) (read as "d in 0 1"), where \(d=0\) means we've taken Way B and \(d=1\) means we've taken Way A. The outcome of interest \(y\) is the commute tone associated with each way we take, Each way we take, expressed formally as \(d \mapsto y\left(d\right)\), or our commute time being a function of the road we choose. We may represent the outcomes of each way as \(y^A\) and \(y^B\), where naturally \(y^A\) is how long it takes if we take my way and \(y^B\) is how long it takes if we go the other way. The "treatment effect" of Way A is \(\tau = y^A - y^B\). Here, \(\tau\) (the Greek letter "t-ow") is the difference in minutes between the way it took me by taking my way, and the time it *would've* taken me if I'd taken Way B. In fact, I did this as I wrote this. I used Google Maps to tell me how long the drive from my apartment to Georgia Tech would be. Using the highway it takes 14 minutes. But, one of the options when I avoid highways takes 23 minutes.

Way Taken	Indicator d	Commute Time	Outcome y
Way A	$d = 1$	$y^A = 14$ min	$y = y^A = 14$ min
Way B	$d = 0$	$y^B = 23$ min	$y = y^B = 23$ min
Treatment Effect	τ	$y^A - y^B = -9$ min	N/A

Suppose I do indeed take Way A, as I would, and that it in fact takes 14 minutes. Does this mean the effect of Way A is -9, or, my way being quicker by 9 minutes? No, not exactly. Maybe I do take \(y^A\), but traffic builds and it doesn't on the other way. Or, maybe \(y^A\) still would take 14 minutes, but the other way, \(y^B\), happens to take 20 minutes instead of 23, meaning our treatment effect is now \(14-20=-6\). The problem inherent here is I cannot take both ways at once. I have a choice to make, and once I choose I must commmit to it. I can *either* take my way or the other way, I can't do both on the same day at the same time. Thus, because of this choice, I can only guess as to what \(y^B\)'s travel time actually would have been for me on that day. Only one outcome exists in reality. Mathematically, we may represent this as $y = dy^A + (1 - d)\, y^B$. If we take Way A, we get \(y=y^A \times 1 +\left(1-1\right)y^B\), or just \(y^A\) since anything multiplied by 0 is just 0 and \(y^A \times 1\) is just \(y^A\). If we take Way B, we get \(y=y^A \times 0 +\left(1-0\right)y^B\), or just \(y^B\) because now \(y^A \times 0=0\) and \(\left(1-0\right)y^B\) is just \(1 \times y^B\). This means that the counterfactual is *inherently* unobservable. Short of time machines where we can peer into alternate universes, the counterfactual is something we have to estimate.

12.2 Randomized Controlled Trials

Establishing causality and generating counterfactuals are all about comparisons. Typically, we compare a group of one or more units that did an intervention or policy to units that did not do the same policy. We use regression as a vehicle to facilitate this comparison. Before we do this for a real policy example though, let's think about how this is done in a (close to) ideal setting.

In medicine, we must test drugs in order to see if they work before we allow them to be used on humans in a broader sense. We use randomized controlled trials to try to establish the efficacy of drugs. A randomzied controlled trial is a form of study design where we, as the reasearchers, assign a treatment at random to a certain number of people or units or entities. Those who get the treatment we call the *treatment group*, those who do not get it are called the *control group* (or, sometimes we call the untreated group the *donor pool*). When we say "random assignment", we mean that we assign the treatment such that each person has an equal probability of getting the treatment. There are many such way to do this in reality, but at its heart we essentially use a computer to flip a coin across \(N\) individuals/units to determine if it gets treatment. If treatment assignment is truly random, this now means that any *other* covariates that may influence the outcome do not predict treatment status or outcome information.

Say we wish to study the impact of a vaccine on recovery time. We cannot just give the vaccine to some people and not others in a non-random way because maybe other variables are influencing recovery rates. Perhaps those who took the vaccine are younger on average than those who didn't. Or, maybe they had bettter baseline health characteristics. This means, on average, those who took the vaccine would recover from COVID (say) quicker than the

control group, not completely because of the vaccine but because they were already healthier or younger on average compared to the control group. We'd say something is wrong if they *didn't* recover quicker. Alternatively, maybe there are just unobserved factors we can't see which explain why the treatment group did better, to a degree.

When the coin flip decides who gets the vaccine, then in a large enough representative sample, our treatment and control groups are *balanced* across all confounders, on average. We say "balanced" because when all study participants of all ages, races, and so on are **equally likely** to be given the vaccine or not, the average difference in recovery time can be better attributed to the vaccine instead of other factors such as age. Thus, it is very important to ensure that our control group is balanced across all relevant areas which may affect the outcome. If our treatment and control groups are balanced, we may compute the average treatment effect of the treatment as $\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} y_i^1 - y_i^0$. This is just the average of the raw differences in the outcomes of the treatment group and control group, where \(y_i^1\) is the observed outcomes for all of our treated units (recovery time, in this case) and \(y_i^0\) represents the average recovery time for all those in the control group.

To illustrate the idea of balance in a public policy setting, I generate synthetic data on 100 individuals who, at their job, enter some program which may increase income. Individuals are aged from 18 to 50. However, age may correlate with income. Older people tend to have more work/professional experience than younger people, on average. So, simply comparing the outcomes of adults versus the outcomes of a younger group may be ill-advised, as maybe those who make more money and participated in the program would have made more money anyways, without the program, due to different baseline levels of experience due to age. To test this, I iteratively assign some probability of treatment to all of them from \(0.05 \leq p \leq 0.5\) in increments of 0.01. We can see, from the GIF below, that when the probability of being treated is \(0.5\), the differences across both age and pre-existing incomes vanishes. The control group is on average a year older than the treatment group, and the income difference between them vanishes to an absolute difference of 54 dollars, where the treatment group makes more than the people who didn't do the jobs program. So now that we've randomized the treatment, we can have people take the program and see how it affects their incomes.

12.3 Problems With Randomization

The central issue with randomization is that there are some interventions (in fact, most of them) that researchers simply cannot randomize. After all, many treatments of interest have explicit assignment mechanisims (i.e., this neighborhood has high crime rates *therefore* we elect to send more police as a response to crime). Even if the rationale for doing the treatment is not given, sometimes our available set of control units may differ in important ways from the unit that's treated.

In February of 2023, Turkey had an earthquake. Suppose we're interested in the effect of this earthquake on the local economic outcomes for the affected cities or the entire country.

Well, researchers cannot randomize earthquakes to strike certain cities versus others, and even if we could this would be morally unacceptable. So assuming we were comparing cities in Turkey that were affected to those that weren't, the affected areas may differ in their baseline characteristics from unaffected areas. For example, maybe poorer areas were more vulnerable than richer ones. For a cross country comparison, maybe bulding codes would explain the differences in the effect of the earthquake which, in turn, affect the economic implications for Turkey versus another unexposed nation.

Another example is cannabis legalization. We cannot flip coins to have some states legalize cannabis and others not. Canabis' legality is decided by the preferences of the legislature. Thus, we run into the problem of selection bias (as in, maybe some states are more likely to legealize cannabis than others). We also run into counfounding biases. If we wish to see how legal cannabis affected alcohol sales for Oregon, then we need to consider what other factors may affect alcohol consumption aside from the policy of interest. That is, Oregon may differ from other states (say, Alabama or Mississippi) on key characteristics that makes the causal comparison unreasonable. Maybe the price of alcohol between Oregon and a set of others states was not similar enough. Maybe Oregon simply had different economic conditions that made alcohol consumption more or less likely. Perhaps cultural factors would lead to higher level of alcohol consumption anyways, absent cannabis legalization. The fact that we cannot randomize means that researchers cannot make plausible the unconfoundedness assumption (or, lack of omitted variable bias) which underlies OLS regression models.

12.4 Difference-in-Differences

Even though we cannot randomize all treatments/policies, does this mean that we cannot do policy analysis at all? No. Modern econometrics has developed a slew of methods for doing policy analysis when the intervention of interest simply *cannot* be subject to a controlled expriment. I now introduce the difference-in-differences method (DD), using Proposition 99 as an example case. DD is a method used for panel data. Up until this point, we have presumed that we are working with only one collection of units at one time point. This is called a cross-sectional dataset, where we have \(N>1\) units and \(T=1\) time periods.

Country	Year	Units	Value
Mexico	2000	1	50
Guatemala	2000	1	45

The opposite of this is called a time-series dataset, where we have $N = 1$ units and $T > 1$ time periods.

Country	Year	Units	Value
Mexico	2000	1	50
Mexico	2001	1	55

A panel dataset is where $N > 1$ and $T > 1$.

Country	Year	Units	Value
Mexico	2000	1	50
Mexico	2001	1	55
Guatemala	2000	1	45
Guatemala	2001	1	50

To implement DD, we need a few key ingredients: First, we need a treatment of interest with a clear before and after point. In our case, Proposition 99 was passed in 1988, and enacted in 1989. So, we have a clearly defined treatment point. We also need at least one unit that experienced the treatment, and at least one that doesn't experience the treatment. In our case, it's Caifornia versus 38 controls.

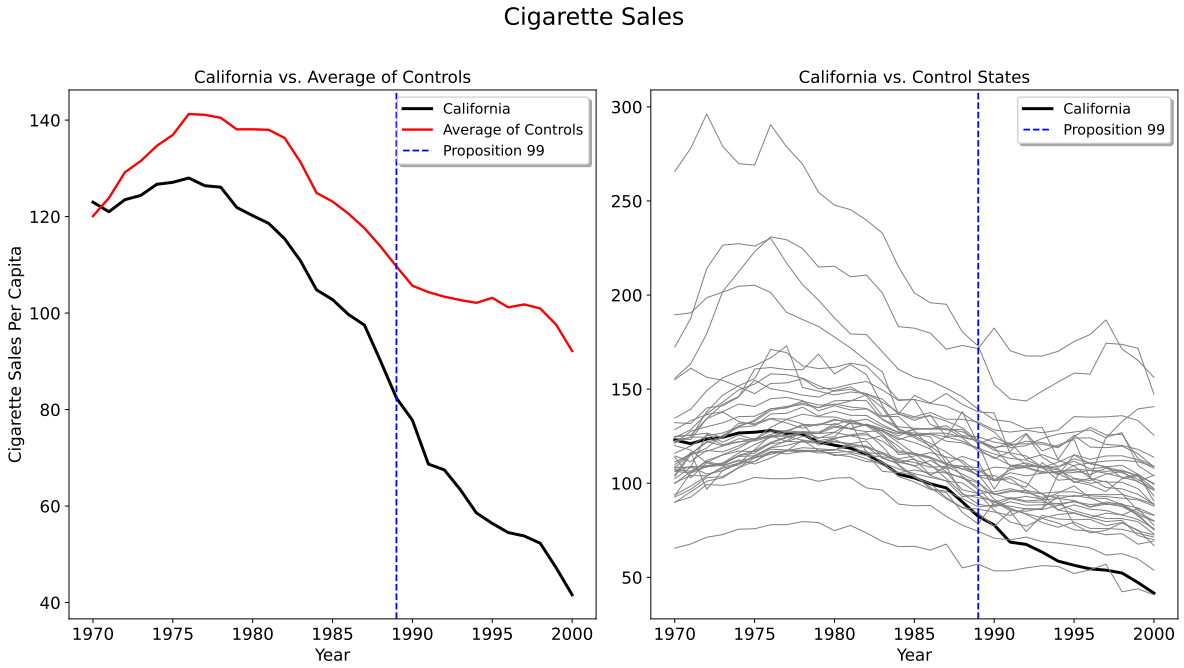
Unlike in a setting where we have a randomized trial, our control group will not be balanced on covariates/outcomes with the treatment group in the pretreatment period. That is, if we take the average of our treatment and control group outcomes in the pre period, the numbers will not have the same or very close values as they did in the GIF above when I demonstrated randomization. But what if we don't need the means to be balanced exactly? What if we just need for the *trends* of the groups to be balanced?

After all, ideally, the only thing separating the control group and treatment group is that one group was treated and the other wasn't. In other words, the difference in values is attributable, in part, to the timing of the treatment. So, if the two groups have similar trends in the pre-intervention period, all we now need us a time trend to account for the time specific differences from the beginning of our study to the intervention point. Thus, DD asks us, as analysts, to accept a singular condition as plausible: namely, that the intercept adjusted average of our controls is a good enough proxy for how the treated unit's outcomes would look absent treatment. This is called the "*parallel trends assumption*". It means that if the intervention never happened, the trend of our control group would move in the same way as the average of the treated unit. In this setting, the average of our control group explains unit specific variation in the outcome (that is, unit-specific unobservable things such as culture).

For parallel trends for DD to hold, the control group therefore must be as similar as possible to the treatment group before the treatment was done. Why? because the quality of our controls is what we use to build our counterfactual. For example, let's say Honolulu implements an anti-crime policy. Can we use New Orleans or St. Louis as comparison cities? Likely

not. The latter two area high crime areas, being regularly distinguished in national crime statistics. They are heavily urbanized places, with vastly different cultural makeups, climates, and settings. Therefore, we wouldn't expect for these two cities to be sufficiently comparable enough to Honolulu to warrant a good causal comparison.

In our case, we have \(N=39\) and \(T=31\). We have \(T_0=19\) preintervention periods (from 1970 to 1988) and \(T_1=12\) post intervention periods from 1989 to 2000. Here, each unit is indexed to the letter \(i\). For our purposes, \(i=1\) is California, and the other \(i=\{2 \ldots 39 \}\) units are the control states. In panel data, each of the units corresponds to only *one* time period. Each of the 39 units, in this case, has 31 rows.



Here, we plot the cigarette consumption of California versus the average of controls, as well as the individual outcomes for the control states. We can see that in 1970, California is fairly similar to the average of the other 38 untreated states. However as the years progress, the average trend of control units slopes upward by a lot, whereas California's smoking trends grew relatively little and begin to precipitously fall in and after 1975. When we look at the plot on the right, this makes a little more sense: there are a few states (Kentucky and New Hampshire) at the top of the plot whose tobacco sales drastically increased between 1970 and the mid 1970s.

12.4.1 Estimating DD

As one might expect, the main workhorse of DD is simply OLS regression, the topic of our previous chapter. I will give Stata and R code below so we can streamline DD estimation, but I think it helps *a lot* to understand what's happening from the perspective of regression at first.

DD is simply a form of OLS regression, where we seek the line that minimizes the prediction error between our independent variables and our outcome. However in this case, our outcome and predictors are special. Our dependent variable in this regression model, \(y_{1t}\), is the pre-intervention outcomes of our treated unit, California. Our independent variables are an intercept and the year-wise average of the 38 control states, \(\bar{y}_{\text{co},t}\). The simple DD model looks like

$$y_{1t} = \beta_0 + \beta_1 \bar{y}_{\text{co},t} \quad \text{s.t. } \beta_1 = 1.$$

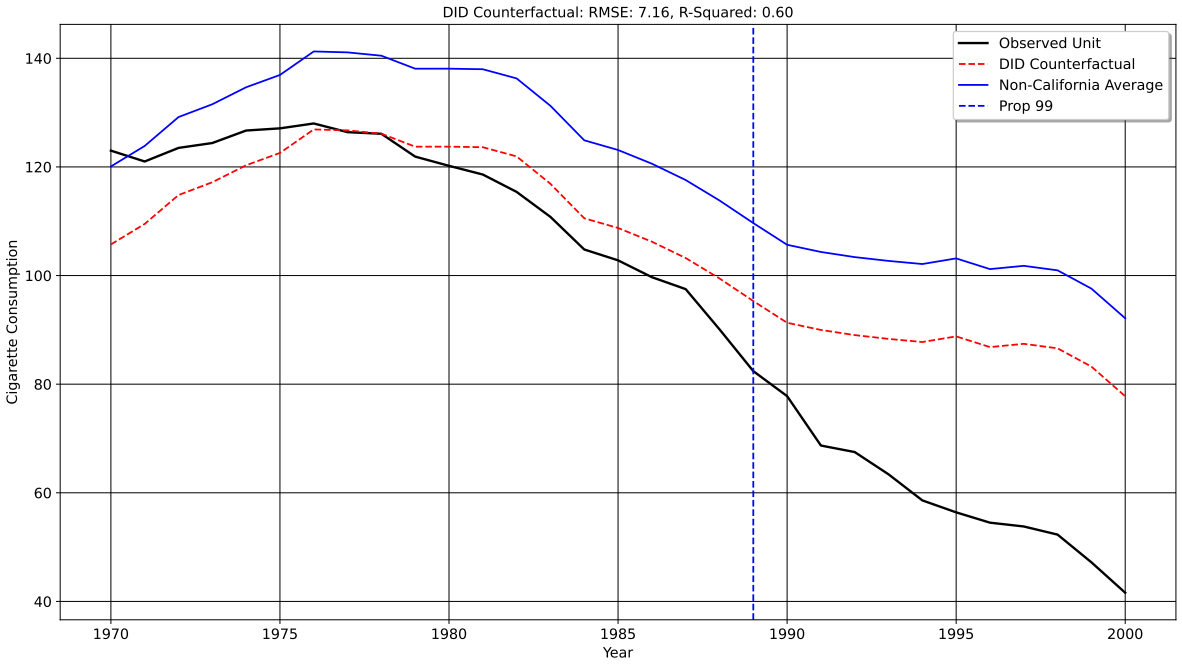
Notice here how we force the coefficient for the average of controls to be 1. Why? Well, DD presumes that a pure average of our controls, adjusted only by a time-based intercept, is a good enough proxy for our counterfactual. As a consequence of this, DD presumes that all of our control units are good control units for the treated unit(s). All control units are treated equally. The constant here for Proposition 99 is roughly \(-14\). After we estimate this model (again just for the preintervention period), we predict the remaining values for the post-intervention period. How do we do the prediction, by the way? Well, it's easy! We add or subtract whatever value we get for β_0 to/from the mean of controls! The dataset itself which we use to conduct the regression looks like:

Table 12.5: DD Dataset

California	Control Group Mean
123	120.08
121	123.86
123.5	129.18
124.4	131.54
126.7	134.67
127.1	136.93
128	141.26
126.4	141.09
126.1	140.47
121.9	138.09
120.2	138.09
118.6	137.99
115.4	136.29

California	Control Group Mean
110.8	131.25
104.8	124.9
102.8	123.12
99.7	120.59
97.5	117.59
90.1	113.82

I plot the results of this regression model below.

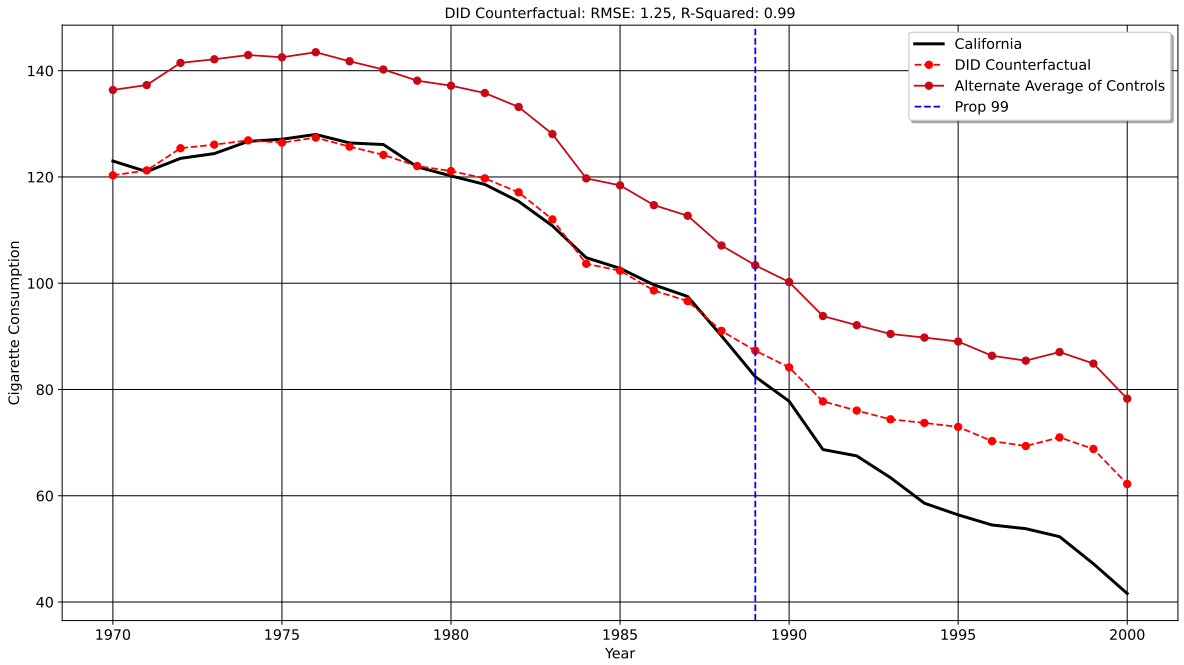


Well, what do we see here? We see the counterfactual produced by DID, using all 38 controls. As promised, the counterfactual predicted by DID is simply the original average of the control units minus 14.359! Seriously, that's all it is. Now that we've predicted our counterfactial, we now calculate the treatment effect for California. We calculate the average treatment effect on the treated unit as \(ATT = \frac{1}{T_1 - T_0} \sum_{T_0 +1}^{T} (y_{1t} - \hat{y}_{1t})\). Or, the average of the differences between what we in fact observed (California under treatment) and what we did not observe (California's intercept adjusted untreated outcomes that represent the cigarette consumption trends absent Prop 99). The ATT suggests that Prop 99 decreased tobacco consumption by about 27.349 packs per capita, with a 95 percent confidence interval of \(\left[-32.522,-22.177\right]\).

Okay, now we have this counterfactual. But is it a good one? How can we tell if the counterfactual is plausible here? Can't we do better? We can begin by recalling a metric of fit, the RMSE.

In this case, think of the RMSE as a metric of parallel-ness. The smaller the RMSE is, the better our pre-intervention predictions are. The bigger it is, the worse our pre-intervention predictions are. This matters because if we have good pre-intervention model predictions, we're likely to have better post-intervention counterfactual predictions. If we have worse pre-intervention model predictions, we are less likely to take seriously the predicted counterfactual. The RMSE here is about 7. Which isn't so bad, but look at what this means practically: The DID counterfactual underpredicts the true values for California in the pre-intervention period from around 1970 to 1975. Beyond this, it also misses the true values for the years of 1980 through 1988. This is particularly bad because if your predictions diverge significantly from the treated unit's observed values in the years *right before* the intervention takes place, why would we think that the post-intervnetion predictions are valid? This imbalance comes from a violation of the parallel trends assumption.In other words, the reason for such bad predictions are the fact that we included all 38 control units, all of which may not be comparable to California's tobacco consumption trends. This means that we must be smart about which units we are comparing California to, since if we include poor control units, we will have poor predictions.

What if we used a different control group though? Suppose we altered the control group to be a smaller subset of controls. Say I use Montana, Colorado, Nevada, Connecticut. After all, why not? Montana, Colorado, and Nevada are all geographically quite close to California, and Connecticut has a similar preintervention trend of tobacco smoking to California.



Well now! This certainly looks **a lot** more parallel than when we used all of our control units!! When we use the limited set of controls, our treatment effect becomes -13.647, with a

confidence interval of $[-14.549, -12.745]$. Think of how big of a reduction this is: when we used all control units (some of which are clearly different from California's pre-1989 tobacco smoking trends), we come up with a decrease of 27 packs per capita, but when we use a more limited pool of controls, we get an ATT which implies a reduction of 13.6 packs per capita. That's pretty much a reduction of **100%** in terms of the average treatment effect! We have cut our treatment effect *in half* by virtue of having a better control group. We also get tighter confidence intervals after using a more similar control group, meaning we are a lot more confident about the effectiveness of the treatment. In addition to the pre-intervention average trend of controls "looking" more parallel, we can confirm that the regression model with the alternate control group is superior to the original model in that the new control group reduces the bias from parallel trends imbalance. The RMSE for the alternate control group shrinks by **140%** compared to the original model.

> **Note**
>
> As with OLS, rarely will you do DD in the manner I've described it above. However, I feel that Stata and R, through their excellent computing capabilities, can largely obscure what's going on when we use `reg` or `didregress`. Also, we can extend DD to instances where many units are treated at different time points. However, for the introductory level, I think DD with a single treated unit is more than adequate as a starting point.

Learning Goals

- Know how to make your own callouts.
- Be able to mess with some SCSS/CSS styling.

12.4.2 Estimating DD in Stata

Part II

Applied Research Methods

Pandey, Shanta, and Charlotte Lyn Bright. 2008. "What Are Degrees of Freedom?" *Social Work Research* 32 (2): 119–28. https://doi.org/10.1093/swr/32.2.119.