Econometrics for Policy Analysis

Jared Greathouse

Invalid Date

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.

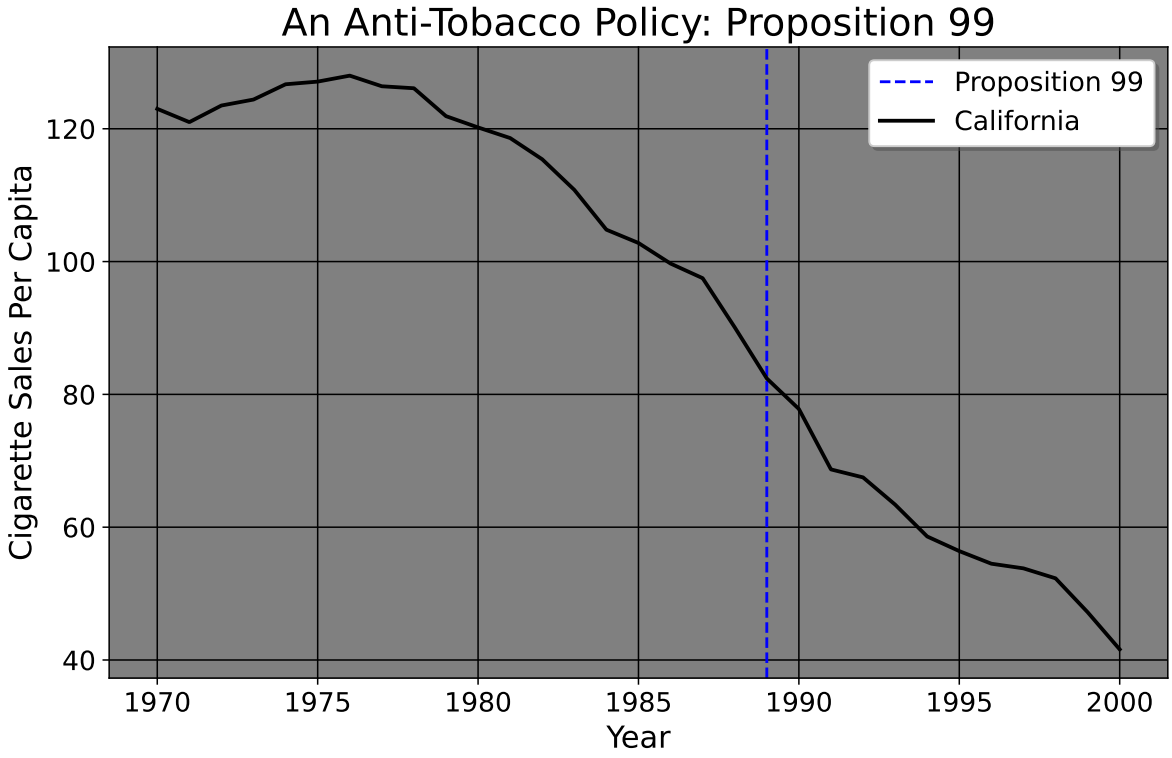
Every day, governments pass laws/public policy to affect some outcome of interest. Policy usually touches thousands if not *millions* of people. From traffic-circles to pop/sugar sweetened beverge taxes, vaccine mandates and universal pre-k programs, cannabis legalization to minimum wages, public policy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

1.4.2 Week 2

- 09-02-2024 (Monday)

University holiday. No clase.

- 09-04-2024 (Wednesday)

Construction of Confidence Intervals.

1.4.3 Week 3

- 09-09-2024 (Monday)

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

- 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.

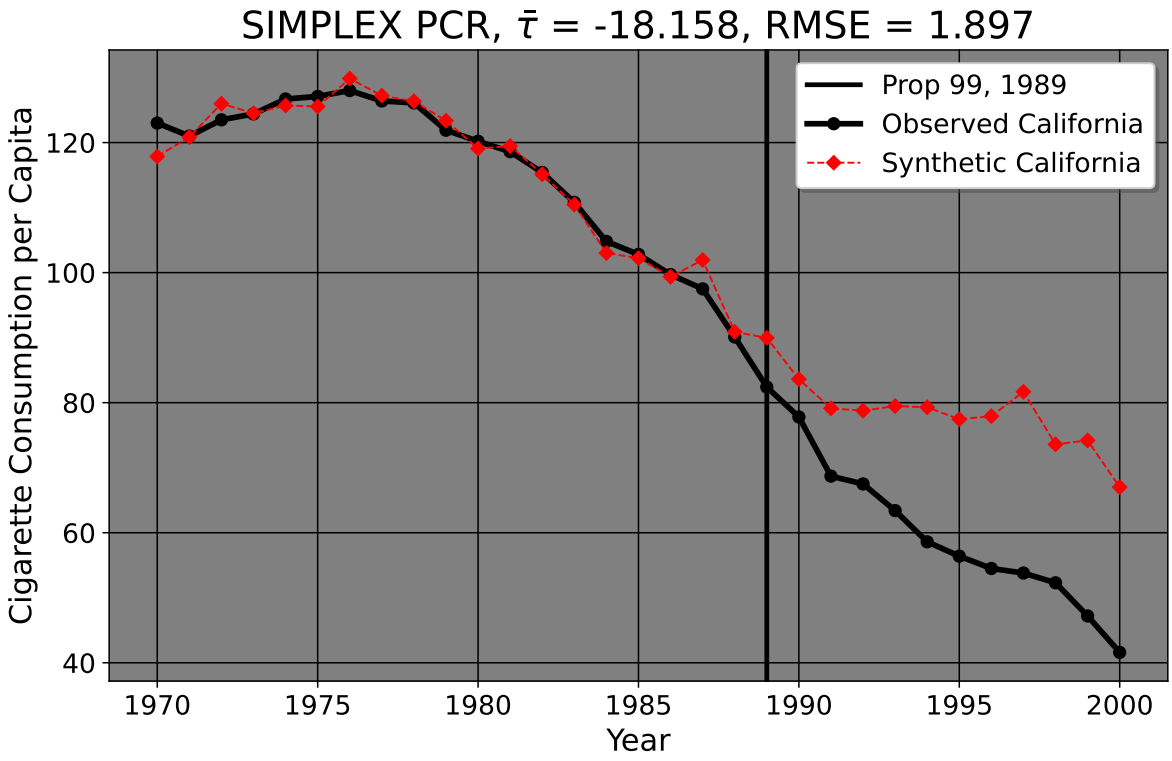
As I've mentioned above, a collection of observations about a set of phenomena is what we call data. Thus, in public policy analysis, data is central to all that we do. One may ask why using data matters at all; the simple reason is that it allows us to resolve disagreements. While people may conduct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

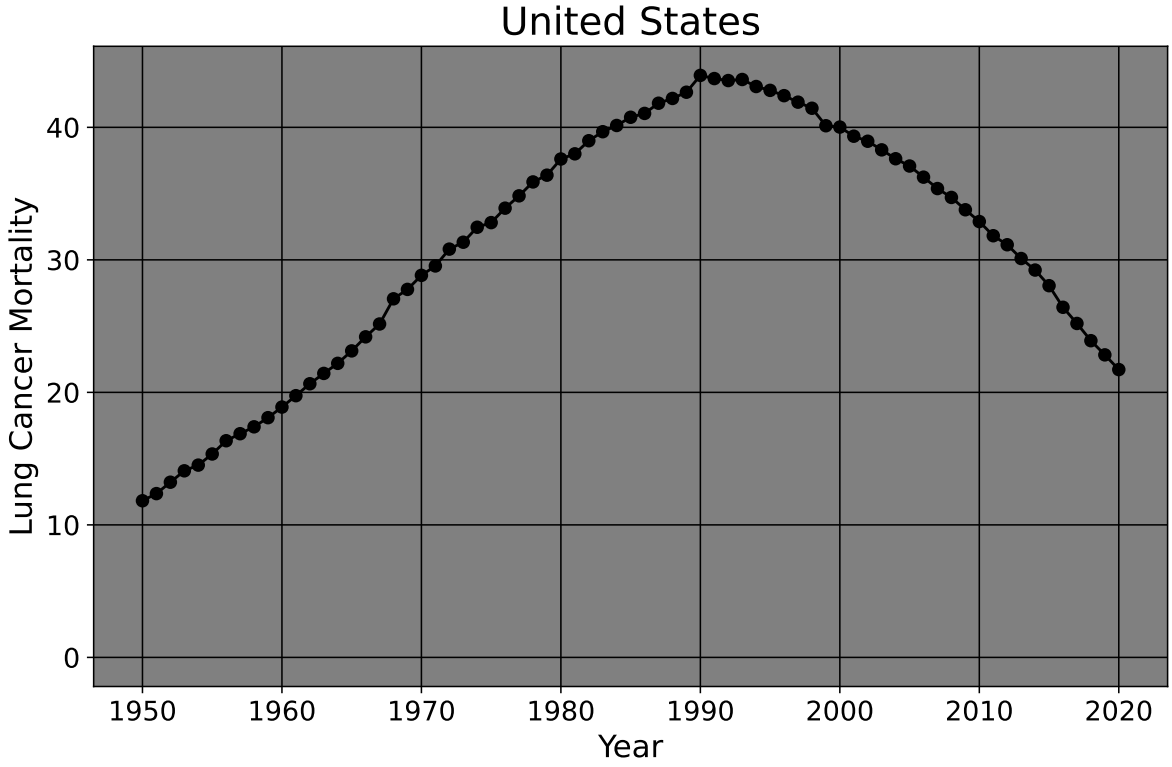
However, there are a few issues with the externality argument. Firstly, being unbuckled turns you into a human projectile. You can hit your passengers or even others outside your vehicle if you're unbuckled. Your market exchange (you buying the car and driving it) is now potentially having second-order 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.

Part I

Mathematics and Econometric Theory

3 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.

3.1 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*).

3.1.1 Means: Arithmetic and Median

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

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

In our formula for the average is $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, 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

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

We can take averages with things aside from die too. 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.

The median, or the middle number. It is a type of average. Suppose we have a dataset of years of education $A = \{5, 6, 7, 9, 18\}$. 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, most of our respondents are in middle school with one elementary schooler. Thus, we can see that the average is influenced by outliers (in this case, somebody in graduate school).

3.1.2 Variance

The **variance** is the second moment. For a random variable X, the varaince is denoted $\text{Var}(X)$, measures the spread of its values around the mean. For a discrete random variable, it is calculated as: $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ For our die, the variance is:

$$\text{Var}(X) = \frac{1}{3}[(1-2)^2 + (2-2)^2 + (3-2)^2] = \frac{1}{3}[1+0+1] = \frac{2}{3}.$$

Note that so far, we've assumed that we have the population (or, every observation we are interested in). In practice, we typically have only a *sample*, or a smaller subset of the population. We will return to this later, but somtimes this affects the way we calculate statistics. For example, while the formula for the mean stays the same, the *sample* variance's formula becomes

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

3.2 Applications

We can use these two moments to answer questions about a group of individuals/units (say, people, cities, etc). For example, say we have the incomes (in 1000s) for 5 people in one year across two groups. For Group A, we have $A = \{50, 50, 50, 50, 50\}$, and for Group B we have $B = \{30, 50, 50, 50, 70\}$. The mean and variance for Group A are just 50 and 0 respectively. 50 is the only number we have, so its average is just that, and since it's the only number we have, it cannot deviate at all from this mean. For Group B, our mean is

$$\mathbb{E}[X] = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{5}(30+50+50+50+70) = \frac{1}{5} \times 250 = 50$$

and the variance is

$$s^2 = \frac{1}{5-1}\left[(30-50)^2 + (50-50)^2 + (50-50)^2 + (50-50)^2 + (70-50)^2\right]$$

which gives us

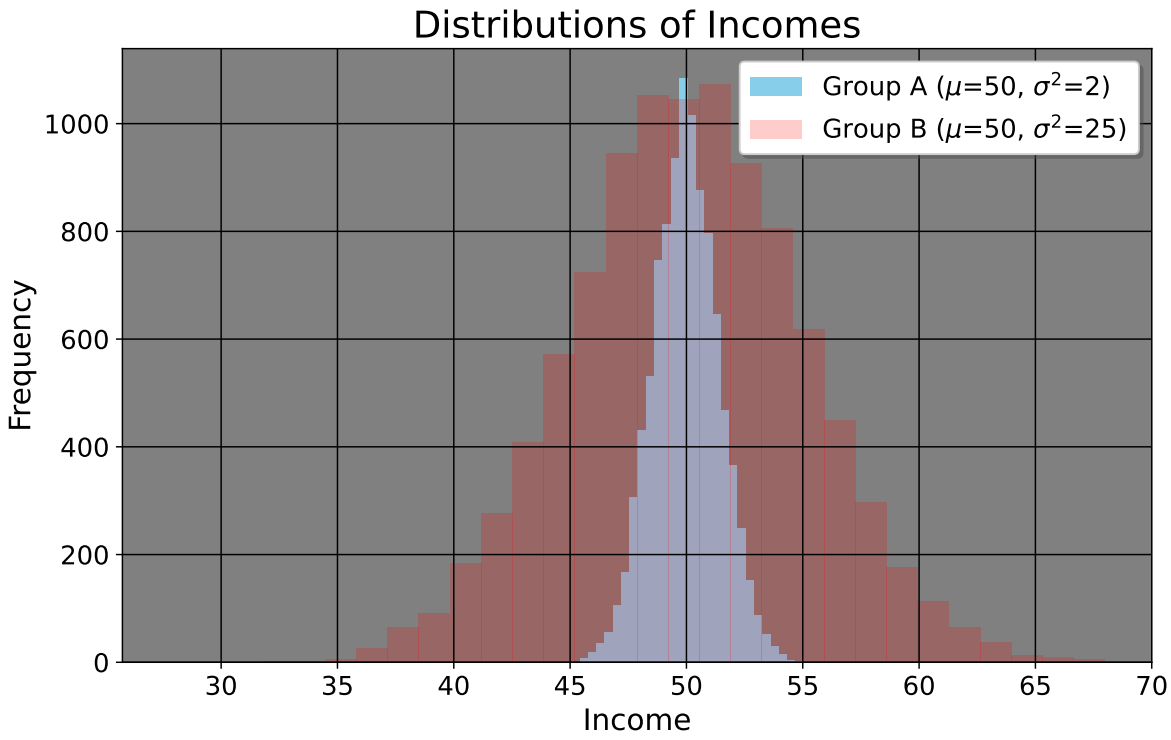
$$= \frac{1}{4}\left[400+0+0+0+400\right]$$

and finally

$$= \frac{800}{4} = 200.$$

So, what do we have here? On average, these groups are the same… but their variance is different. What practically might this mean, why is this useful? Because it allows us to make

better inferences on things such as income distribution. To see why, consider the following plot of 10000 hypothetical incomes across two groups of people.



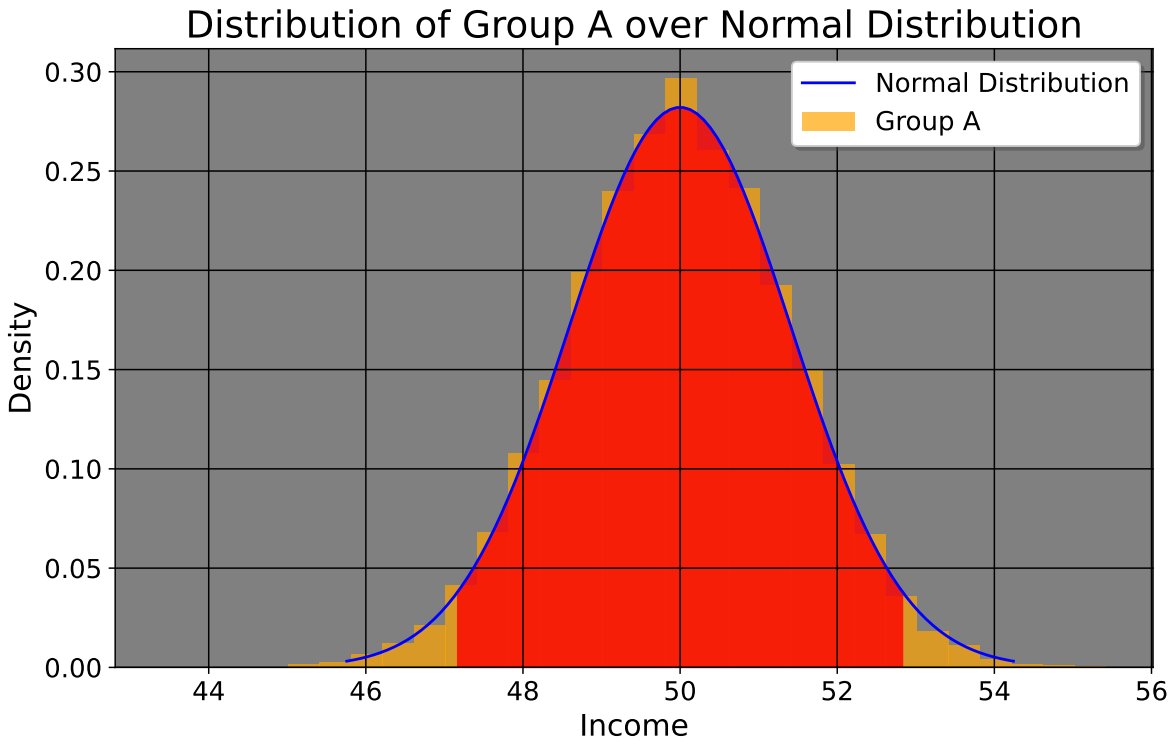
Now, the difference is more apparent. If we only consider the average income of people in both groups, we'd walk away saying that the income distributions are the exact same. But this is wrong, Group A is much tighter distribution of income than Group B because its variance/spread is smaller.

3.3 Uncertainty Around the Mean

Typically, we wish to see how uncertain our estimates are. The way we do this is by calculating the margin of error for the mean. We take such uncertainty into account by computing the standard error of the mean, SEM $= \frac{\sigma}{\sqrt{N}}$. Here, lowercase sigma, σ, is the sample standard deviation and N is our sample size. The sample standard deviation is how far our entire dataset is from the mean. it is the square root of the variance. In the case of Group A from above, our sample size is 10000. We have 10000 incomes to pick from. The expectation/average of these in Group A, as we've said, is 50. The standard deviation of this group is 2.

Before we can discuss the margin of error, though, we must say some words about probability distributions. The distributions of both income groups are what we'd call *normally* distributed

in the sense that most of the values we see are within 1 standard deviation of the mean. In fact, we can show that 68% of our observations lie within one standard deviation of the mean. What this means here, is that if the mean if 50, 68% of our respondents incomes are between 48.6 and 51.4. Let's visualize this for group A.



Typically though, scientists are interested in our uncertainty within two standard deviations of the mean, where 95% of our data lie. To do this, we need the number for the normal distribution that reflects how many standard deivations we are interested in. In practice, the number we use is 1.96. That is, we are using all values from our sample that are within 1.96 standard deviations from our observed mean. Putting this all together, our margin of error for the average income for Group A is

$$0.0277 = 1.96 \times \frac{1.4142135623730951}{\sqrt{10000}}.$$

Now, in order to characterize the range that the mean falls within, we simply do

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

What we're looking at here, is our 95% Confidence Interval. Why 95%? Because that is the range we are intersted within, the area under the normal curve where 95% of our data lie. 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 I simulated this dataset, we know for fact that the true mean does lie within here! But in real life, the confidence intervals will usually never be this tight.

4 Summary

Probability is the stepping stone into using statistical methods. It is the foundation of decisionmaking in business, economics, and policy analysis. This course lays the groundwork for undergraduates to think and reason about public policy using statistics nad probability. 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.

So, 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.

5 A Primer on 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.

5.1 Law of Iterated Expectations

Suppose we wish to commute to Georgia State University in the morning 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, conditional on us taking the highway. We think the highway will take 15 minutes, compared to another way which we think will take 20. We can formalize this also as $x = \{1, 0\}$, where we take the highway being coded as 1, else coded as 0 if we take the other way. So, we leave home and we take the highway, and record how long it took to get there. Say, we record the value of 30 minutes. Do we conclude that this is how long it takes to get to school on averge, and that the other wau is 10 minutes faster? No. Why not? This is only one estimate. There are all kinds of things that could have been going on 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, and we record the amount of time it took us to get to school that day. After doing this, we take the expectation of c conditional on us having taken the highway, or $\mathbb{E}[c] = \mathbb{E}[\mathbb{E}[c|x = 1]]$. After we take this average/expectation aross all of these days, we get the number 17. This is what we call the Law of Iterated Expectations, or LIE. It's the idea that by repeatedly taking the average of a variable given some input (highway or not), we approach the actual value of the average without conditioning on which way we take.

5.2 Law of Large Numbers

Why might this be true, though? Why, after taking all of the averages across a whole month so we arrive at 17, which is a lot closer to 15 than we first thought, and is even quicker than the highway on the first day? The reason is 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, and even then GPS is not always accurate. 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. As researchers, what this means is that drawing from a large sample tends to be better than a small one. 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. Note that in this course, we are always working with a sample we collect, never ever the population.

5.3 Central Limit Theorem

Putting these together, we can now consider things like how we'd 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. Putting LIE and LLN together, the central limit theorem says:

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

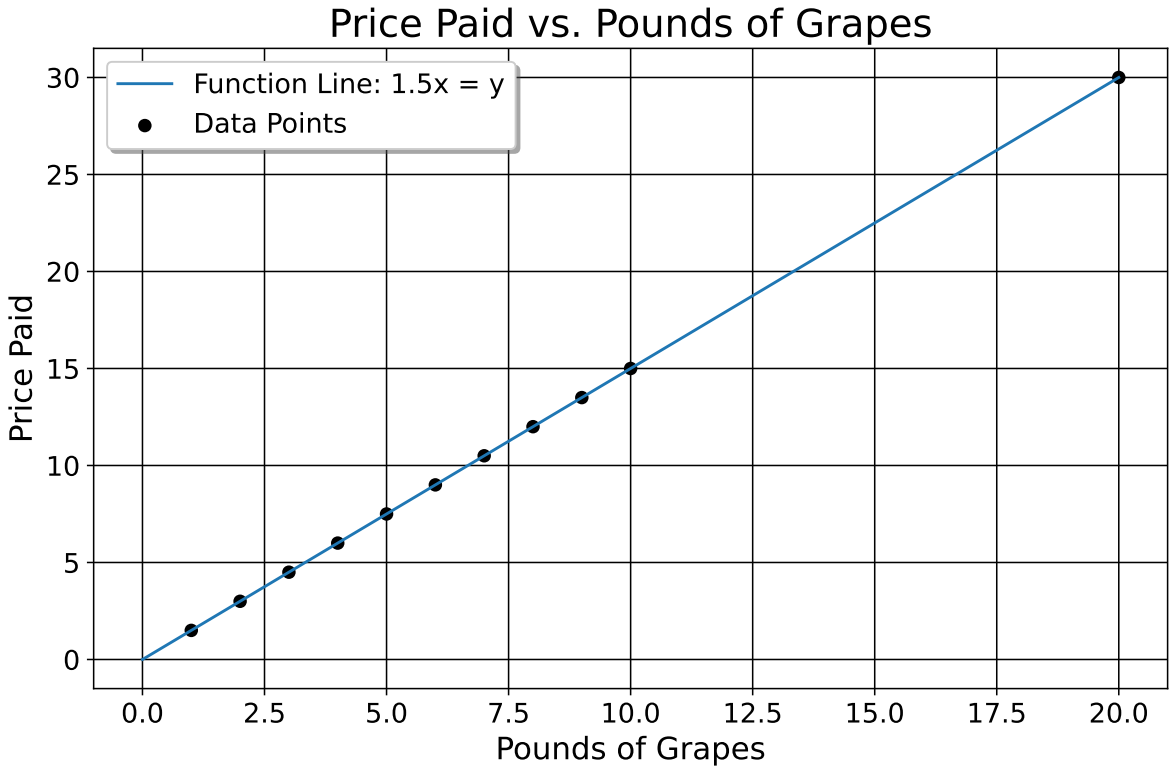
or that as the sample size increases, the probability of our empirical distribution approaches a normal distribution and our sample mean, by LLN, approaches the population mean. If this seems at all abstract to you, we may simulate this. Here, I define the average to be 15 with a standard deviation of about 3 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.

As policy reserachers, these laws have applicability across a variety of the situations we're concerned with.

6 OLS Explained

6.1 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. For $2x = y$ for example, if we plug in 2, we get 4. If we plug in 5, we get 10. 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. These points form a line, the equation for which being $y = mx + b$. Here, y is our outcome, m is the change in the price of grapes for every new pound of grapes bought, and b is our value we pay if we get no grapes.

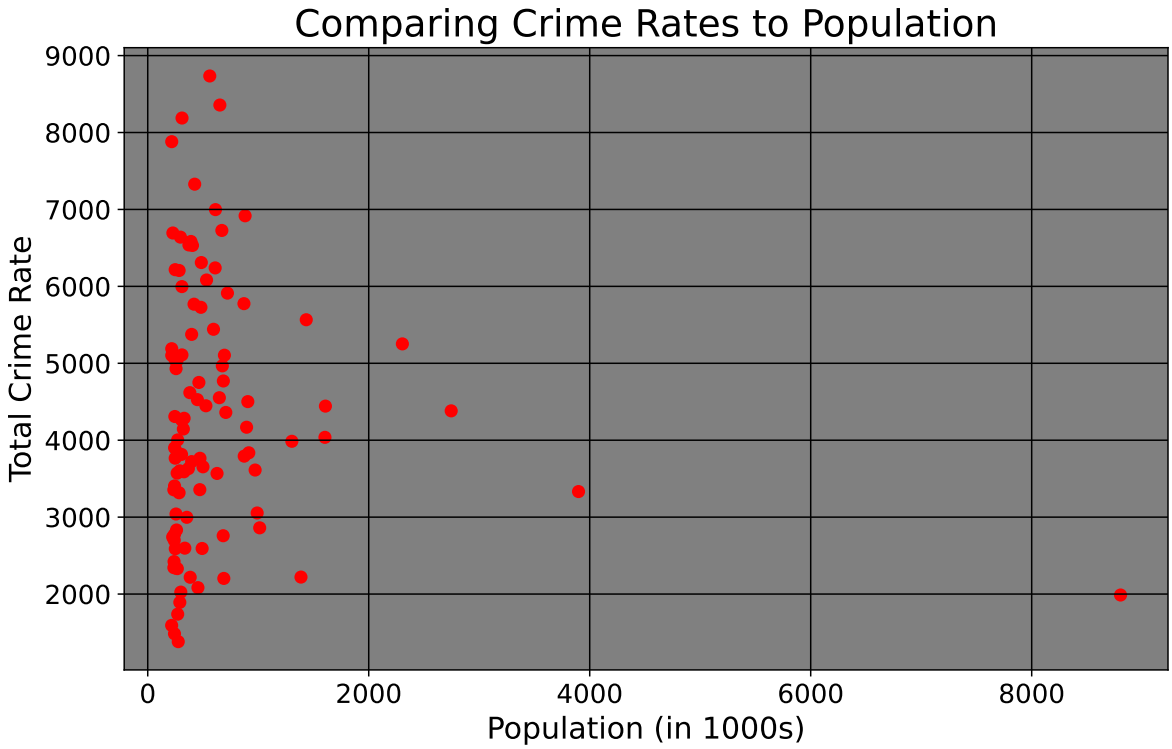


The way we find the m and b for a straight line is the "rise over run" method, in this case

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

For this case, the function for the line of price paid for grapes is $y = 1.5x$. Notice here how the line explains how much we pay perfectly (in other words, the line exactly matches the data points). This means our *residuals* $\hat{\epsilon}_i = y_i - \hat{y}_i$ are 0, where y_i maps on to the real data and $\hat{y}_i$ (which we call y-hat) is the prediction from the line/function. Here, the letter "$\hat{\epsilon}_i$" is just a variable for the the distance from the predicted point to the observed point. If the observed value for our first data point is 10 but $\hat{y}_1 = 11$, then $\hat{\epsilon}_1$ is -1. Note that I add the "hat" to the the greek symbol "epsilon" to denote that they are *estimates* which come from the model, not "error" (which includes residuals *not in the model*, as I explain below)

However, this is because we have a case where all the necessary information is known (price and weight). We know the price of grapes or gas or movie tickets. The algebra is so simple we that we intuitively understand that this is how we calculate expenses. With just these simple known facts, we could calculate the amount of money we pay for grapes at 1 pound or a trillion pounds, assuming the price does not change. But…. what if the data we see are not nice and neat in terms of a function?



Take the idea of predicting crime rates in cities using the population of the city as a predictor. Now, how do we calculate the rise over run? We would presume some *function* exists that

generates the crime rate for that city. However, is quite obvious that no deterministic function exists here. Before, we simply would throw our hands up, in a sense, and say that there's no solution. Algebra has failed us in that we cannot find a function which perfectly explains these data points. But, this is not a fool's errand. After all, this is what the real world is like, right? Crime rates are a *random variable* in the sense that the rate it takes on for a given place and time is not guaranteed. Sometimes, crime is high, other times its low(er). Why? Well, some cities are wealthier and some are poorer. Some cities vary by racial compositions, or will differ by factors like age composition, alcohol use, and gun ownership. Thus… some cities have high crime rates, others have low crime rates. Indeed, it would be very unreasonable to expect to find a singular function that perfectly explains the variation in crime rates across one or more cities.

So, what can we do? We can't find a function for the line that perfectly explains the crime trends… But, how about we instead seek the best *possible* line? The line which minimizes the residuals between what we actually see and what we would predict, given our datapoints? Can we, in a sense, derive a certain function that represents how crime trends change given some population input? To connect this idea to the first example, suppose I asked you to find the function for the line that tells me how much we pay for grapes, *even when I do not* give you the price. I only give you the weight of grapes you buy, and how much you paid for them.

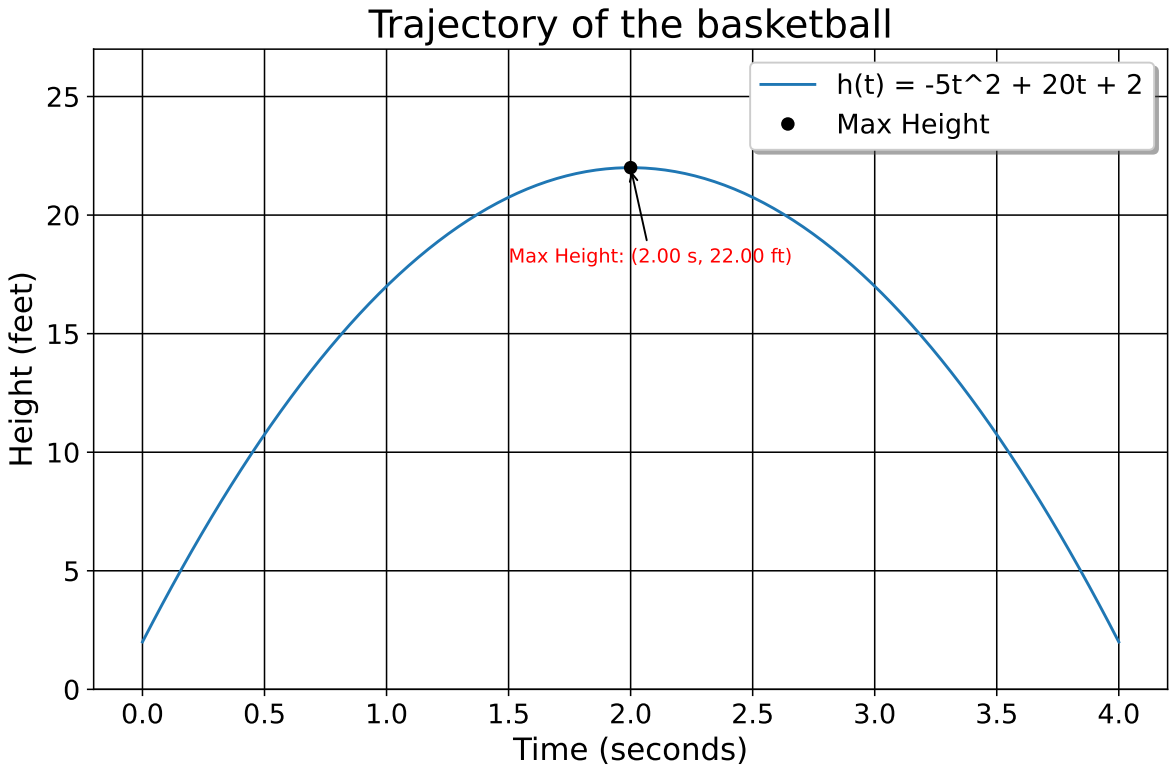
6.2 Arrividerci, Algebra.

To do this, 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. The derivative is the slope of a curve/line at a given point. One very useful property about derivatives is that when we set the derivative of a function to 0 and solve for the variable, we reach a maximum or minimum point on the original function, usually. Optimization problems, in general, take the form of

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

where there's some function $f(\theta)$ (called the *objective function*) that is minimized (or sometimes maximized) over a set of $g(\theta)$ equality constraints and $h(\theta)$ inequality constraints. In our case for regression, we have no constraints in our objective function since we desire the best fitting line possible.

For example, say we shoot a basketball off a 2 foot cliff, which produces a trajectory function of $h(t) = -5t^2 + 20t + 2$, where $h(t)$ is a function representing the ball's height over time in seconds and the 2 represents the fact that we are standing 2 feet above ground. We can find the *maximum* height of the ball by taking the derivative of the original quadratic function and solving it for 0.


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

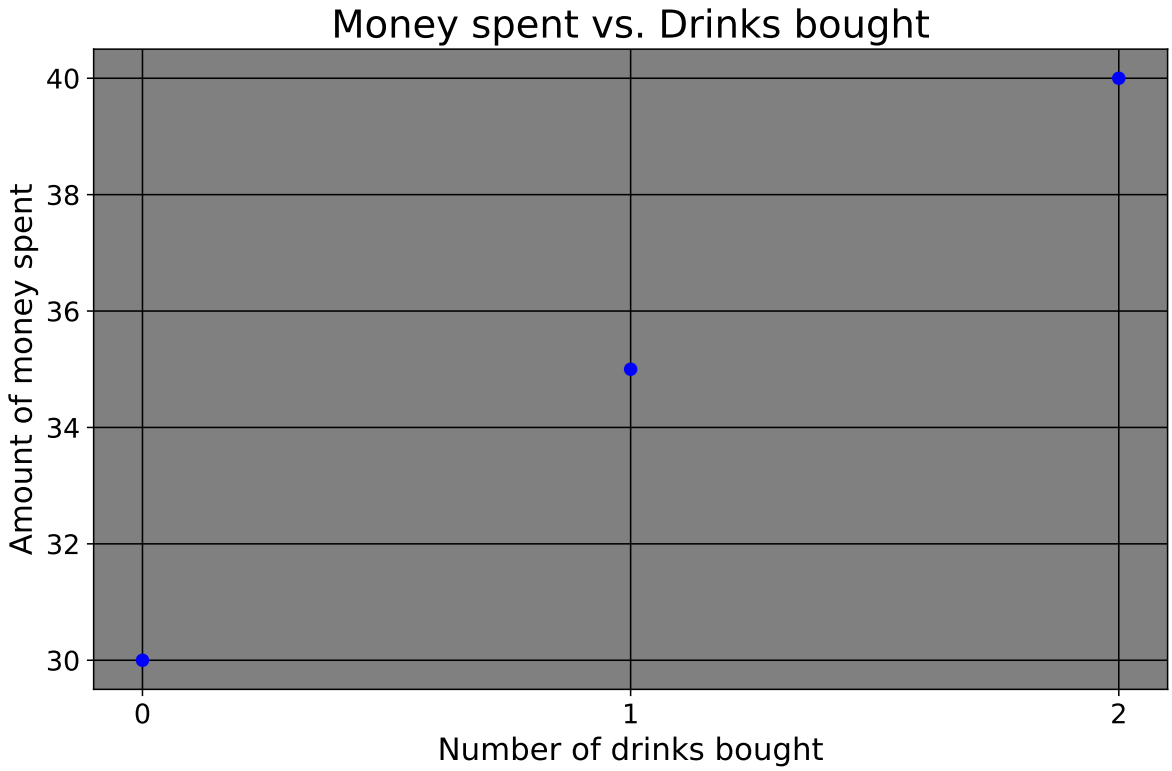
In this case, we use the power rule for derivatives. The power rule for dertivatives is where we subtract the exponent value of a function by 1 and place the original value to be multiplied by the base number. For example, the derivative of $y = 2x^3$ is just $6x^2$, since 3 minus 1 is 2 and 2 multiplied by 3 is 6. Here, we can apply this exact principle to obtain the derivative for the function of the ball's trajectory:

$$h(t) = -5t^2 + 20t + 2 = \frac{d}{dt}(-5t^2) + \frac{d}{dt}(20t) + \frac{d}{dt}(2) = -10t + 20$$

We can then solve the derivative for 0. When we solve the derivative for 0 by multiplying by 2, $-10(2) + 20 = 0$, we get the maximum height of 22 feet after 2 seconds (strictly speaking, if we wanted to be sure that this was a maximum, we could take the second derivative). Derivatives, for this reason, play a key role in minimizing or maximizing functions such as the ones we just did. We will use them to minimize the sum of $\hat{\epsilon}_i^2$. This is known as *ordinary least squares* regression (OLS), also called *linear 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.

To introduce OLS, we can think of 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. Unlike the earlier examples though, with regression we have multiple variables (typically), each of which affect the output of the function differently. And in this case, it makes perfect sense: if we wish to derive a function for crime rates, it's quite possible many things (such as the factors I mentioned above) may have differing impacts on the realized crime rate for a city. To deal with this with regression, we take the *partial derivative* with resepct to each variable, holding the other variable constant (that is, assuming that the other variables do not change, only the current one does). If this seems at all abstract to you, I will provide a detailed, clear derivation. Below, I deviate from most econometrics textbooks. I formally derive the OLS betas using derivatives. 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.

To fix ideas, suppose we wish to attend a nightclub and we wish to express how much we pay for that evening as a function. In this case, how much we pay 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 drink. 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 bough and how much we paid.


Money spent vs. Drinks bought
Amount of money spent
40
38
36
34
32
30
0
1
2
Number of drinks bought

6.3 An Extended Example

> **!** Important
>
> This is where you may start skimming. It is okay to not understand everything here.

6.3.1 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). 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.

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

6.3.2 Define Our Econometric Model

Our population model of how much we pay given some entry fees and additional drink costs looks like:

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

I remove the hats from these because there's some model out there that exists to quantify this relationship between how much we spend given some entry fees, and drink fees, but here it is something we must estimate. Here, y_i is the total amount of money we spend that evening in dollars, $\hat{\beta}_0$ is how much we pay to enter, $\hat{\beta}_1$ is how much we pay for the i-th drink, and x is the 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 "beta" into the already familiar $y = mx + b$.

6.3.3 Write Out the Objective Function

We can represent the objective function to solve OLS as

$$S = \text{argmin} \sum_{i=1}^{n} \underbrace{(y_i - \hat{y})^2}_{\text{objective function}} \quad .$$

I call the objective function "S" for "spending", but we can call it any letter we'd like. $\hat{\epsilon}_i$ are simply calculated like $y_i - \hat{y}$, or the difference between the observed values and our model's prediction. From the above, our models predictions are simply the two variables we've discussed, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. We call this variable y-hat. In fact, our objective function may be rewritten completely by decomposing $\hat{y}$ as the model's coefficients

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

The word "*argmin*" here means "argument of the minimum". The symbols underneath it, $\hat{\beta}_0, \hat{\beta}_1$, are the coefficients we are solving for. We call the solutions $\hat{\beta}_0, \hat{\beta}_1$ *optimal* if they return the lowest possible values the function can produce. What values are these, what values does this objective function produce? The sum of $\hat{\epsilon}_i^2$. 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), meaning that the line will be as close as it can be to the observed data at every single data point.

One may ask *why* we are squaring the summed $\hat{\epsilon}_i$. First of all, minimizing the raw sum of the predicted $\hat{\epsilon}_i$ (that is, without squaring them) is a non-differentiable function. We would need to use numerical methods to compute its solution. Using $\hat{\epsilon}_i^2$ means that we are dealing a quadratic function (making the solution differentiable). Also, suppose one observation has a residual of 40 and another observation has a residual of -40. Adding these residuals would cancel out to 0. The squaring also has the property of penalizing worse predictions. After all, all $\hat{\epsilon}_i$ 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 as a residual of 40, and this does not make sense.

6.3.4 Simplify 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 ($\hat{\beta}_1$)

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

6.3.5 Take Partial Derivatives

> **Note**
>
> If you do not wish to follow this section, then at least skip to this section following my calculation of the partial derivatives.

To find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, we take the partial derivatives of S 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. I define one set of inner and outer functions, beginning with the power rule for the outer function, and then taking the derivative for the inner function. The partial derivative w.r.t each coefficient is simply the sum of these chain rules. 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. $\hat{\beta}_0$:**

First, we denote our inner functions.

$$
\begin{aligned}
f_1 &= 30 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0) \\
f_2 &= 35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1) \\
f_3 &= 40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2)
\end{aligned}
$$

Here are our outer functions:

$$S = f_1^2 + f_2^2 + f_3^2.$$

Now, following the chain rule, we have:

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial S}{\partial f_1} \cdot \frac{\partial f_1}{\partial \hat{\beta}_0} + \frac{\partial S}{\partial f_2} \cdot \frac{\partial f_2}{\partial \hat{\beta}_0} + \frac{\partial S}{\partial f_3} \cdot \frac{\partial f_3}{\partial \hat{\beta}_0}.$$

The first step of the chain rule is using the power rule for the outer functions:

$$
\begin{aligned}
\frac{\partial S}{\partial f_1} &= 2f_1 \\
\frac{\partial S}{\partial f_2} &= 2f_2 \\
\frac{\partial S}{\partial f_3} &= 2f_3.
\end{aligned}
$$

All we've done here is used the power rule, taking the 2 from the exponent and putting it on the outside of each term's outer function. For the next step of the chain rule, we take the derivatives of each term's inner function w.r.t. $\hat{\beta}_0$:

$$\frac{\partial f_1}{\partial \hat{\beta}_0} = -1$$
$$\frac{\partial f_2}{\partial \hat{\beta}_0} = -1$$
$$\frac{\partial f_3}{\partial \hat{\beta}_0} = -1.$$

Since the coefficient for $-\hat{\beta}_0$ is negative 1, each inner function's derivative for $\hat{\beta}_0$ is just -1. Now, we substitute these inner derivatives back into the f_i functions:

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}_0} &= 2f_1 \cdot (-1) + 2f_2 \cdot (-1) + 2f_3 \cdot (-1) \\
&= -2f_1 - 2f_2 - 2f_3 \\
&= -2(30 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0)) - 2(35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1)) - 2(40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2))
\end{aligned}$$

Now, we simplify this expression. Let's first expand each term by distributing the 2:

$$= -60 + 2\hat{\beta}_0 - 70 + 2\hat{\beta}_0 + 2\hat{\beta}_1 - 80 + 2\hat{\beta}_0 + 4\hat{\beta}_1.$$

Next I rearrange. I add parentheses for readability:

$$= (-60 - 70 - 80) + (2\hat{\beta}_0 + 2\hat{\beta}_0 + 2\hat{\beta}_0) + (2\hat{\beta}_1 + 4\hat{\beta}_1).$$

Finally, when we combine like terms, we end with

$$= \boxed{-210 + 6\hat{\beta}_0 + 6\hat{\beta}_1}.$$

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

Following from the above, here are our inner functions

$$f_1 = 30 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0)$$
$$f_2 = 35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1)$$
$$f_3 = 40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2).$$

Then the outer functions:

$$S = f_1^2 + f_2^2 + f_3^2.$$

Just as before, here's the chain rule to get the partial derivative

$$\frac{\partial S}{\partial \hat{\beta}_1} = \frac{\partial S}{\partial f_1} \cdot \frac{\partial f_1}{\partial \hat{\beta}_1} + \frac{\partial S}{\partial f_2} \cdot \frac{\partial f_2}{\partial \hat{\beta}_1} + \frac{\partial S}{\partial f_3} \cdot \frac{\partial f_3}{\partial \hat{\beta}_1}.$$

First I apply the power rule to the the outer functions:

$$\begin{aligned}
\frac{\partial S}{\partial f_1} &= 2f_1 \\
\frac{\partial S}{\partial f_2} &= 2f_2 \\
\frac{\partial S}{\partial f_3} &= 2f_3.
\end{aligned}$$

Next, we'll do the inner functions:

$$\begin{aligned}
\frac{\partial f_1}{\partial \hat{\beta}_1} &= 0 \\
\frac{\partial f_2}{\partial \hat{\beta}_1} &= -1 \\
\frac{\partial f_3}{\partial \hat{\beta}_1} &= -2
\end{aligned}$$

Note how $\hat{\beta}_1$ **isn't** in term 1 since it is multiplied by 0

$$\underbrace{(30 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0))^2}_{\text{Term 1}},$$

its derivative must be 0. This means the first outer function goes away because $f_1 \times 0 = 0$. So, I omit this from my derivation. Next, we'll substitute these remaining two derivatives back into the expression for the outer function:

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}_1} &= 2f_2 \cdot (-1) + 2f_3 \cdot (-2) \\
&= -2f_2 - 4f_3 \\
&= 2(35 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1)) - 4(40 - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 2)).
\end{aligned}$$

Now, we combine like terms:

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2(35 - \hat{\beta}_0 - \hat{\beta}_1) - 4(40 - \hat{\beta}_0 - 2\hat{\beta}_1).$$

To simplify this expression, first we'll expand each term by distributing the -2 and -4. Here is term 1:

$$-2(35 - \hat{\beta}_0 - \hat{\beta}_1) = -70 + 2\hat{\beta}_0 + 2\hat{\beta}_1.$$

Now we expand term 2:

$$-4(40 - \hat{\beta}_0 - 2\hat{\beta}_1) = -160 + 4\hat{\beta}_0 + 8\hat{\beta}_1.$$

Now, combine like terms:

$$= (-70 + 2\hat{\beta}_0 + 2\hat{\beta}_1) + (-160 + 4\hat{\beta}_0 + 8\hat{\beta}_1).$$

I move the constants outside of the parentheses. I add adding parentheses for readability:

$$= (-70 - 160) + (2\hat{\beta}_0 + 4\hat{\beta}_0) + (2\hat{\beta}_1 + 8\hat{\beta}_1).$$

After subtracting the constants and combining like terms once more, we're left with

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

6.3.6 Get the Betas

Now we've set up our partial derivatives, so we can now return to normal scalar algebra to get our betas (in practice this is done with *matrix* algebra). Either way, we must solve a system of equations. The reason we must solve them as a system of equations is because we need for BOTH values to be the ones which minimize the distance between the model's predictions and the observed data. Here are the equations we must solve for (note, we have two equations because we have two variables, but this can readily be extended to more variables):

$$\hat{\beta}_0 + \hat{\beta}_1 = 35$$

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

Here I use a method called substitution to solve the system, but there are many such ways we can solve this. I choose to solve the first parital first since it is by far the easiest.

6.3.6.1 Solve for $\hat{\beta}_0$:

The first step of substitution is to solve for one of our variables.

$$\hat{\beta}_0 = 35 - \hat{\beta}_1$$

This is pretty simple, we just subtract $\hat{\beta}_1$. Before I continue, notice a few interesting things about this: we know, by construction of the problem, that the entry fee is some positive number. We also know $\hat{\beta}_1$ has to be less than 35. Imagine a $\hat{\beta}_1$ that would equal or be greater than 35: this would mean that we get in for free or, enen sillier, are *paid* to enter, which does not make sense given the problem at hand (even though it's theoretically possible, I suppose). We also know that $\hat{\beta}_0$ can't be 35, because then this means that drinks are free. I should note that these are not impossible: maybe girls get in free before 12 or you're a celebrity who's paid to be there.

> **!** Important
>
> The reason I'm making such a fuss about this and carrying on with the example in this manner is because I want you to think of regression as a way of explaining the world, not just a mathematical set of equations. When the regression gives us a number, the model predictions need to either make sense under existing constraints/theory, or be interpreted sensibly. If the model predicts, for example, negative age or that you have negative cups of coffee per day, this suggests we need a different modeling strategy or that something's been coded wrong. If negative values do not make sense for a x-values for a certain variable, we should not include negative values in our interpretation of our results.

Now, we continue to solve.

6.3.6.2 Substitute $\hat{\beta}_0$:

Since we know the expression for the constant (the entry fee), we can plug it into the partial for $\hat{\beta}_1$,

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

Once we do this, we can solve for $\hat{\beta}_1$. Here is the partial for $\hat{\beta}_1$ plugging in the expression for $\hat{\beta}_0$

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

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

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

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 $\hat{\beta}_0$.

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

Now, we simply subtract 5 from the RHS

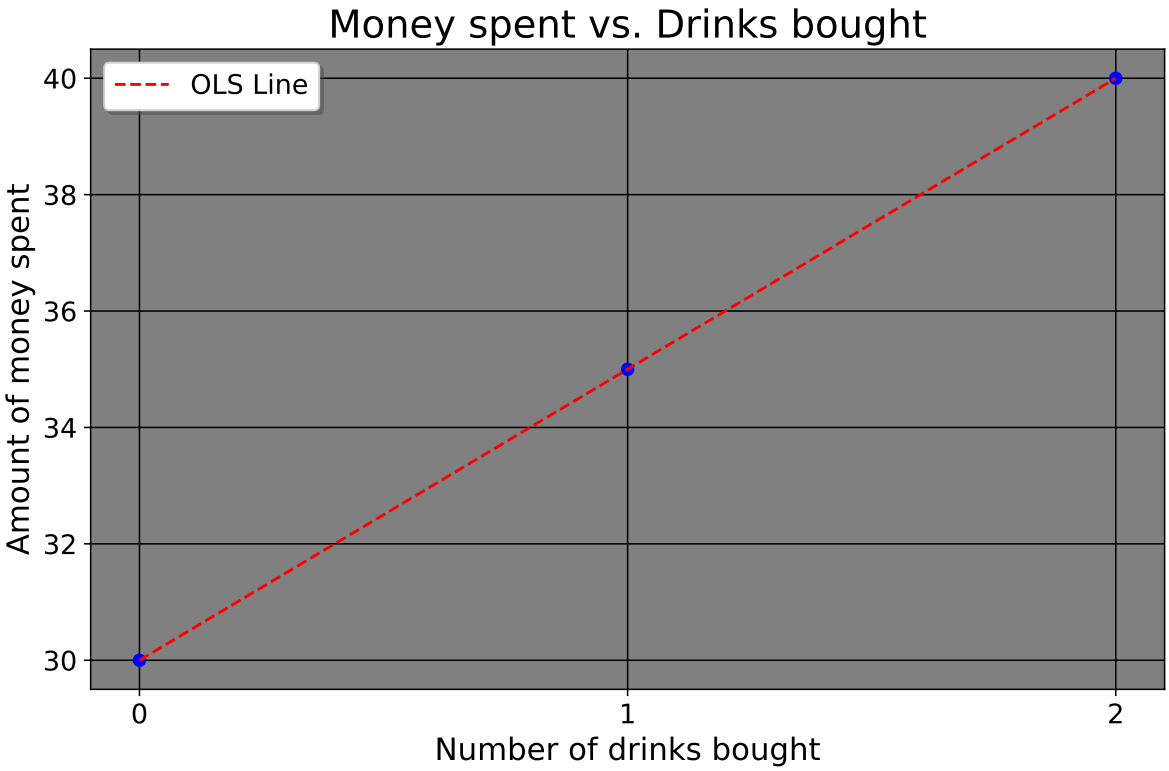
$$\boxed{\hat{\beta}_0 = 30}.$$

The entry fee is 30 dollars.

6.3.7 Our OLS Line of Best Fit

So, our line of best fit is $\hat{y} = 30 + 5x$. The way we interpret this is for every additional drink bought, you pay 5 more dollars. In fact, *compared* to those who only wanted to get in the club (and not drink), you spend 5 more dollars than they do, per each new drink you get. We can now predict how much money we would spend if we bought 5 drinks or 500 drinks.

At the outset, one may ask why we did this at all. Why bother with the partial derivative approach and the messy system of equations? The reason is firstly 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. Additionally, when we understand how the partial derivative allows us to look at the change of a function with the change of *one* variable in that function, we can quickly see how the parital derivative allows for us to (atttempt to) isolate the causal impact of *multiple* variables and go beyond the two-variable case. Strictly speaking, we did just that right here in this derivation! We considered the impact of BOTH entry fees and the number of drinks we bought too (after all, suppose entry was free, then we have no constant). If we wanted to add in another variable, say the cost of an additional cigar, all we would do is add another beta, $\hat{\beta}_2$, to the original objective function and multiply the corresponding values by whatever the costs for that were. We would then have three systems of equations to solve for. However, the underlying principle is the same. Here is our completed regression line.


Money spent vs. Drinks bought
OLS Line
Amount of money spent
Number of drinks bought
40
38
36
34
32
30
0
1
2

6.4 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).

With this, we can estimate the variance-covariance matrix as

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2(X^T X)^{-1}$$

This allows us to compute our standard error: $\text{SE}(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta})_{jj}}$.

> **!** Important
>
> Again, I will never ever ask you to calculate the standard error or t-statistic without the necessary coefficients to calculate it algebraically. The reason I'm showing you this is just so you have a very clear sense of where these numbers are coming from,

As we've discussed with normal descriptive statistics/classic hypothesis testing, we can also compute confidence intervals for these regression estimates. 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 and df is our degrees of freedom. Then, we can interpret our coefficients accordingly. Suppose a confidence interval for a coefficient 6.7 lies within 6.06 and 16.44. We interpret this as follows: given the dataset and input values (and assumiing the OLS assumptions are valid, which we will detail below), we are 95% confident that the true value for that coefficient lies within the range of 6.06 and 16.44. In other words, there's a "width" of this confidence interval by about 10.

Just as we discussed in the preceding chapters, OLS's estimates (conditional on the assumptions being met, which we will cover after this) is only 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.

6.5 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 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.

6.5.1 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. 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.

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.

6.5.2 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 (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. 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.

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. They select 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. So, we cannot use them to 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.

6.5.3 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. 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.

6.5.4 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. For example, say we wish to evaluate how the number of firefighters sent to a call impacts the amount of damage from that fire. We conclude that there's a positive relationship between number of firefighters sent and damage, 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, 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.

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. It means our predictor variables may not be correlated with the error term. In economics we would say that we need for our variables to be exogenous in that the predictors influence the outcome, not the outcome influencing the predictors. This is also why it matters to critically think about the variables one will use in their regression model *before* they run regressions.

7 Summary

Part II

Applied Research Methods

8 Writing for Policy Analysis: THe Introduction

The intro should…

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.