Econometrics for Policy Analysis

Jared Greathouse

5/27/24

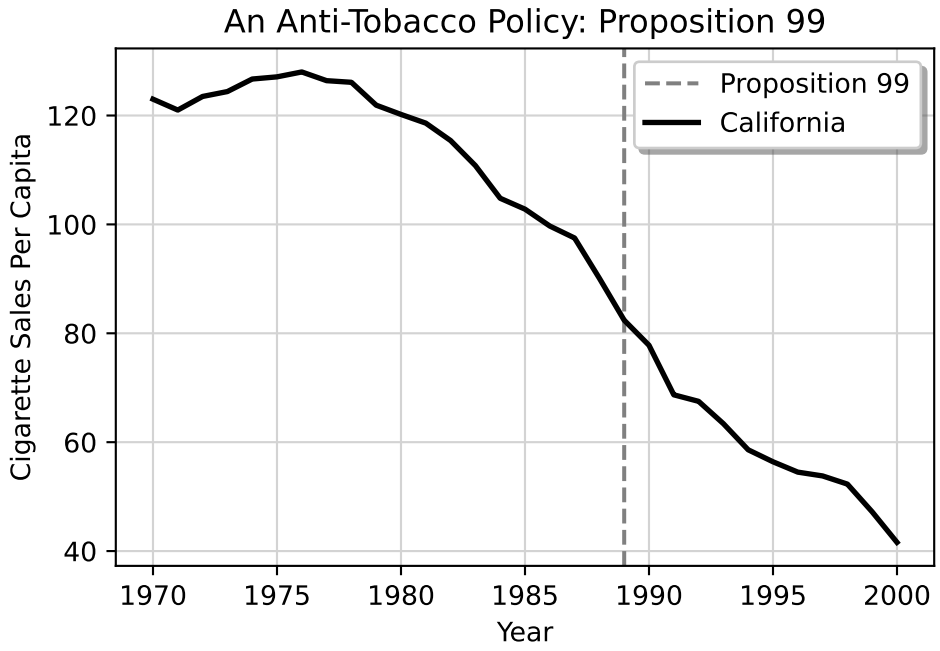
Table of contents

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, para-

doxically), 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 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. I will recommend chapters from each book to read (of course, you need only read one chapter, from either text). 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 R. For Stata users, Statalist is a great resource for Stata. R also is backed by a vast statistician community.

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 a baseline for math, I presume everyone knows the basics of central tendency (that is, mean, median, mode). I also presume knowledge of basic algebra (e.g., solving single variable equations, what a function in math is).

3. 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.

4. 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.

5. 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 linked to their descriptions. In lieu of finding your own dataset, you may use any of these for your class paper, should you wish.

6. Main Takeaways: In addition to statistics, the main goal of the course is to provide the reasoning skills scientists use to understand the world better. These skills will be useful not just in academia or even the professional workforce, but everyday life.

1.4 Class Schedule

Below is the schedule. Unless I write otherwise, all readings for Econometrics for Policy Analysis (EPA) should be done before class, alongside one other chapter of your choosing. The reason for this is because the class is meant to be interactive, i.e., I should not be the only discussant. The lecture notes (i.e., each chapter of EPA) simply sets up the baseline for discussion.

1.4.1 Week 1

- 08-26-2024 (Monday)

Introductions and EPA, C2

- 08-28-2024 (Wednesday)

Basic Probability and Confidence Intervals: EPA C3. IS C2, IS C3 (skim), IDS C2, IDS C3, especially "Discrete Probability" and "Random Variables".

Here, the main things to focus on are probabilities of discrete random variables, as well as t-tests, standard errors, and confidence intervals.

1.4.2 Week 2

- 09-02-2024 (Monday)

University holiday. No clase.

- 09-04-2024 (Wednesday)

Correlation, Coeffcients, and Association (EPA, C3)

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

1.4.3 Week 3

- 09-09-2024 (Monday)

OLS Derived, Watch this: Partial Derivatives, (EPA, C5, skim the derivation) (ROS, C7), IS, C10. Also, Inference for OLS (Gauss-Markov Assumptions). **Today, the research question is due.**

- 09-11-2024 (Wednesday) Data Types and Measurement (EPA, C5), RMSS, C6

1.4.4 Week 4

- 09-16-2024 (Monday) Data Gathering/Cleaning
- 09-18-2024 (Wednesday)

Intro to Data Visualization

1.4.5 Week 5

- 09-23-2024 (Monday)

Histograms and Scatterplots

- 09-25-2024 (Wednesday)

Time Series Graphs

1.4.6 Week 6

- 09-30-2024 (Monday)
- 10-02-2024 (Wednesday)

1.4.7 Week 7

- 10-07-2024 (Monday)
- 10-09-2024 (Wednesday)

1.4.8 Week 8

- 10-14-2024 (Monday)
- 10-16-2024 (Wednesday)

First Draft Due

1.4.9 Week 9

- 10-21-2024 (Monday)
- 10-23-2024 (Wednesday)

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

- 11-25-2024 (Monday)
- 11-27-2024 (Wednesday)

1.4.15 Week 15

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

1.4.16 Week 16

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

1.4.17 Week 17

- 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.

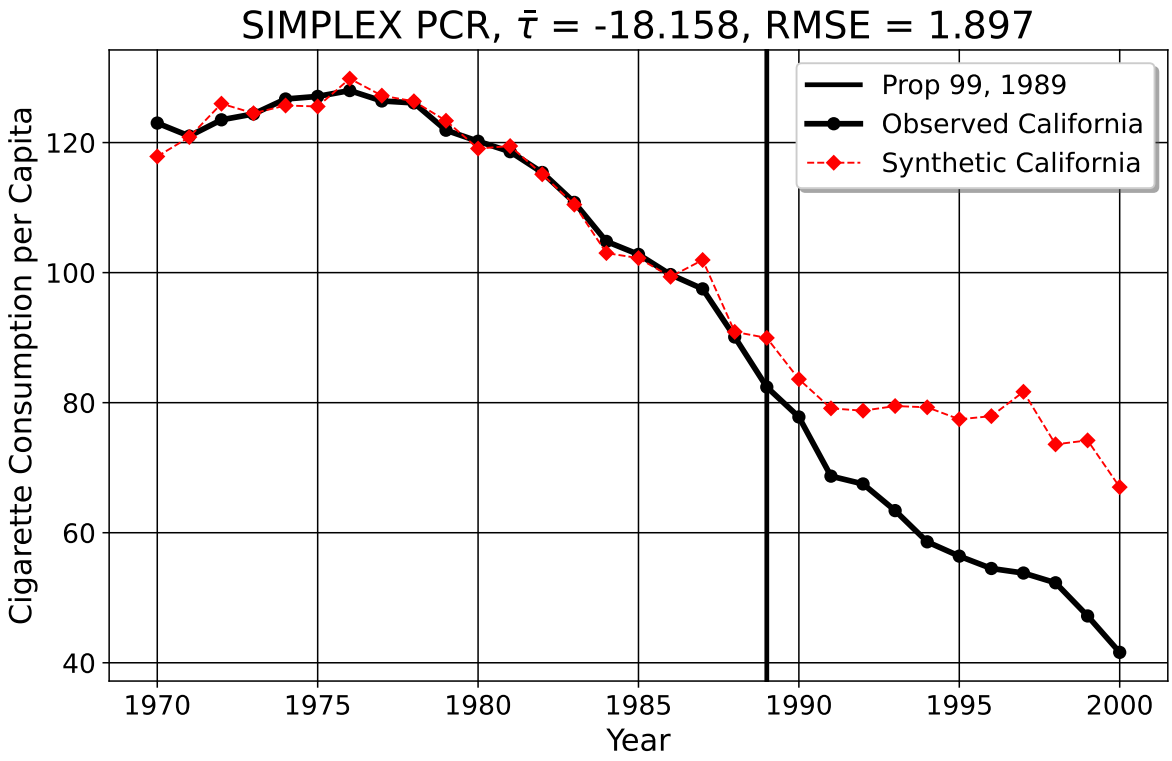
As I've mentioned above, a collection of observations about a set of phenomena is what we call data. Thus, in public policy analysis, data is central to all that we do. One may ask why using data matters at all; the simple reason is that it allows us to resolve disagreements. While people may conduct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

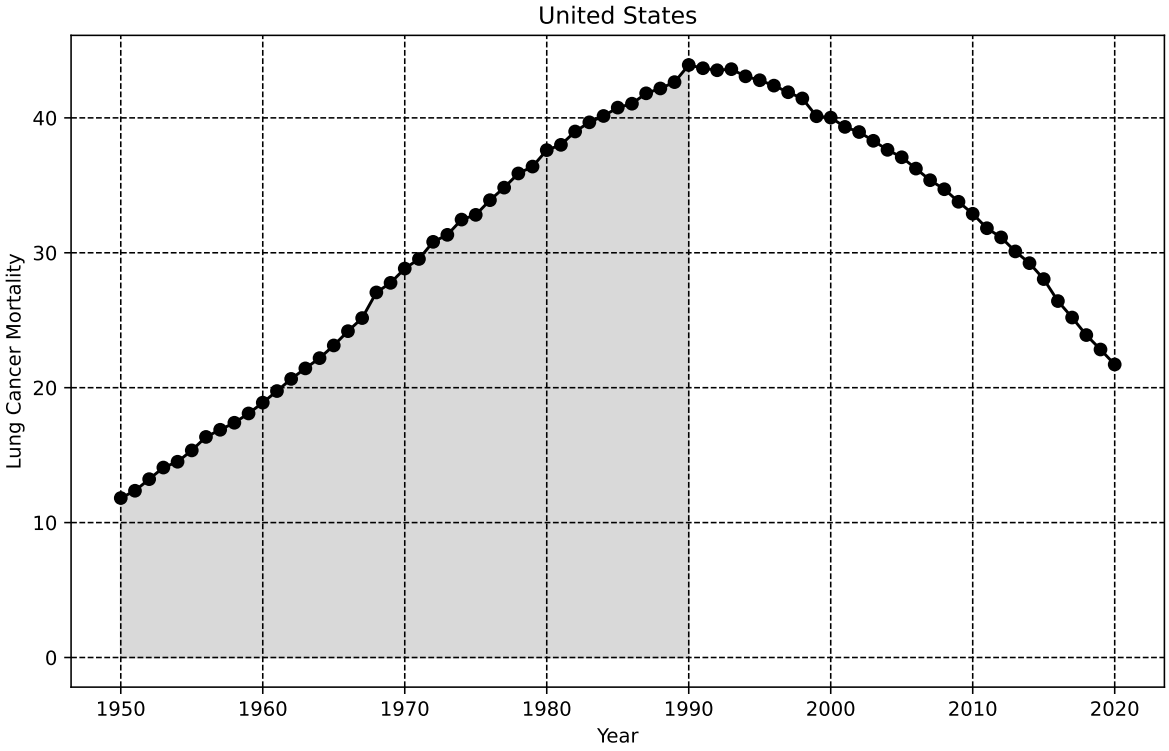
However, there are a few issues with the externality argument. Firstly, being unbuckled turns you into a human projectile. You can hit your passengers or even others outside your vehicle if you're unbuckled. Your market exchange (you buying the car and driving it) is now potentially having second-order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



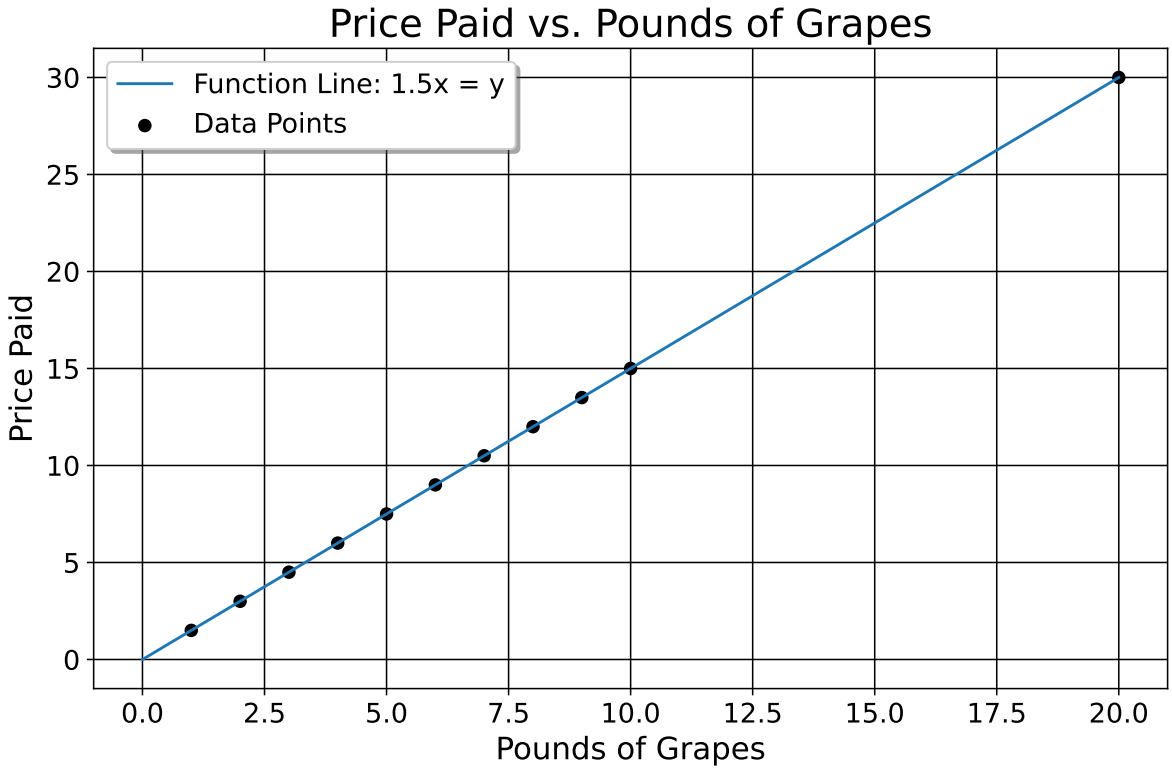
The shaded area represents the period before any state-wide anti-tobacco legislation was passed in the United States. We can see quite clearly the age-standarized lung cancer mortality rates rose in a fairly linear manner in the United States. However, the curve is parabolic; mortality rates were rising 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 OLS Explained

3.1 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.

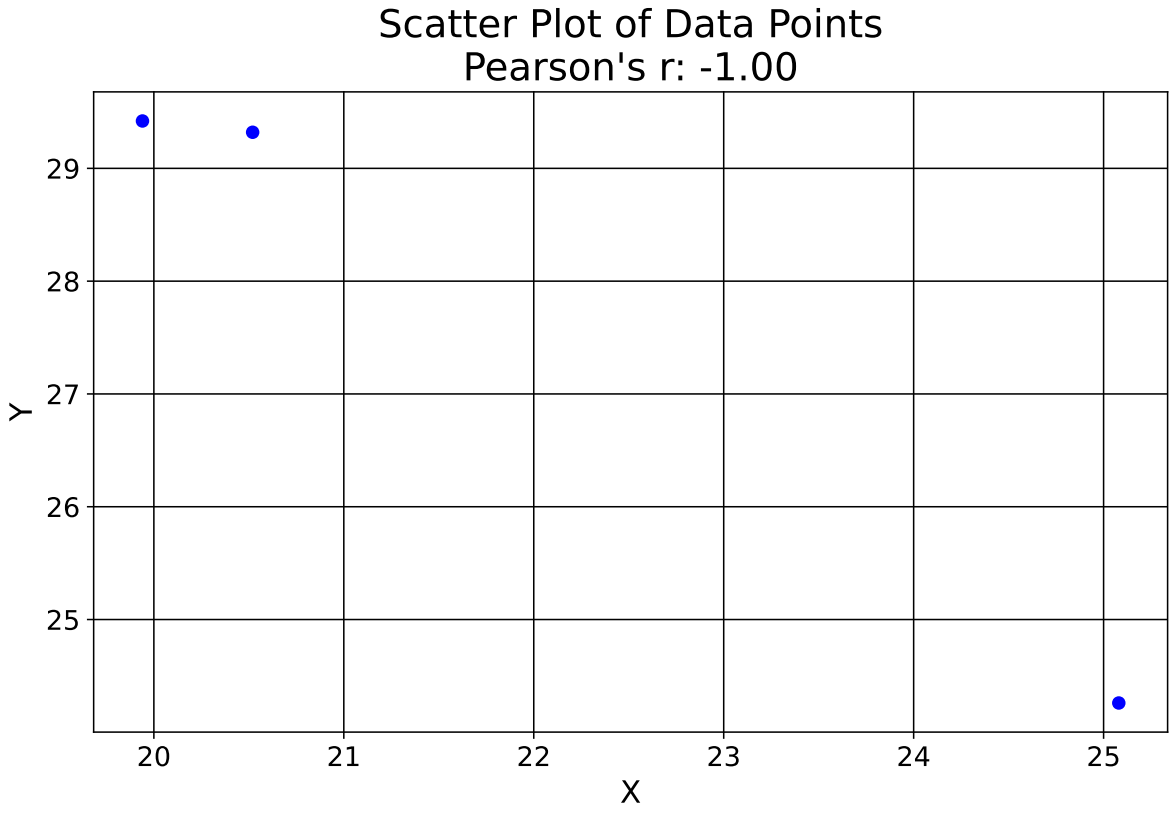


The way we find the m and b for a straight line is the "rise over run" method, in this case

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

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* $u_i = y_i - \hat{y}_i$ are 0, where y_i maps on to the real data and $\hat{y}_i$ is the prediction from the line/function. Here, the letter "u_i" is just a variable for the the distance from the predicted point to the observed point. If the observed values are 10 but we predict 11, then our residual is -1.

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, so we typically do not even need algebra here, we intuitively understand that this is how we calculate expenses. That is, we know from the above that 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, e.g, $(20.52, 29.32), (25.08, 24.26), (19.94, 29.42)$? As we can see here, there is simply no straight line that will fit to all of these points perfectly.



Before, we simply would have to throw our hands up, in a sense, and say that there's no

solution. Normal algebra has failed us in that we cannot find a function which explains the points in this dataset. But, this is what the world is like, right? Take the idea of predicting crime rates in cities. We would presume some *function* exists that generates the crime rate for that city. Some cities are wealthier or poorer than others or of differing racial compositions, or will differ by factors like age composition or alcohol use. Thus… some cities have high crime rates, others have low crime rates, and 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. After all, crime rates are a *random variable* in the sense that the number it takes on is not guaranteed. Sometimes, crime is high, other times its low(er).

So, what can we do? We can't find a function for the line that perfectly explains the crime trends, or the data points in the plot above… But, how about we instead seek the best *possible* line, given our data points? The line which best *fits* (in that it minimizes the residuals) to the data, in that it does our best to approximate the data? To connect these ideas together, 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.

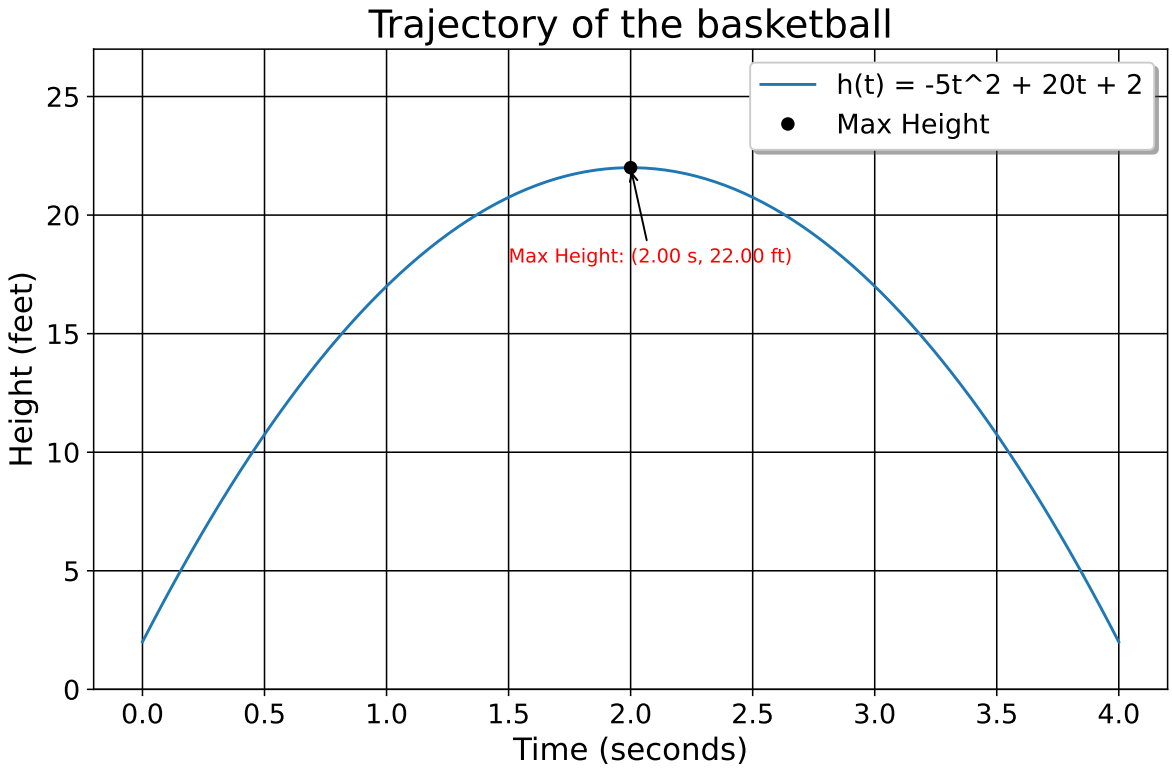
3.2 Arrividerci, Algebra.

To do such a task, 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 baiscs of derivatives, to find the optimal solution. The derivative is the slope of a curve/line at a given point. One interesting property about derivatives is that when we set the derivative to 0 and solve for the variable, we reach a maximum or minimum, usually. For a realistic example, say we shoot a basketball off a 2 foot cliff, which produces a trajectory of $h(t) = -5t^2 + 20t + 2$, where $h(t)$ is our height over time and the 2 represents the fact that we are standing 2 feet above ground, and the line itself represents the ball's trajectory. We can find the *maximum* height of the ball by taking the derivative of the original quadratic function and solving it for 0.

The power rule for dertivatives is where we subtract the exponent value by 1 and palce the original value to be multiplied by the base number. For example, the derivative of $y = 2x^3$ is just $6x^2$. Here is the derivative for the function of the ball's trajectory using the power rule:

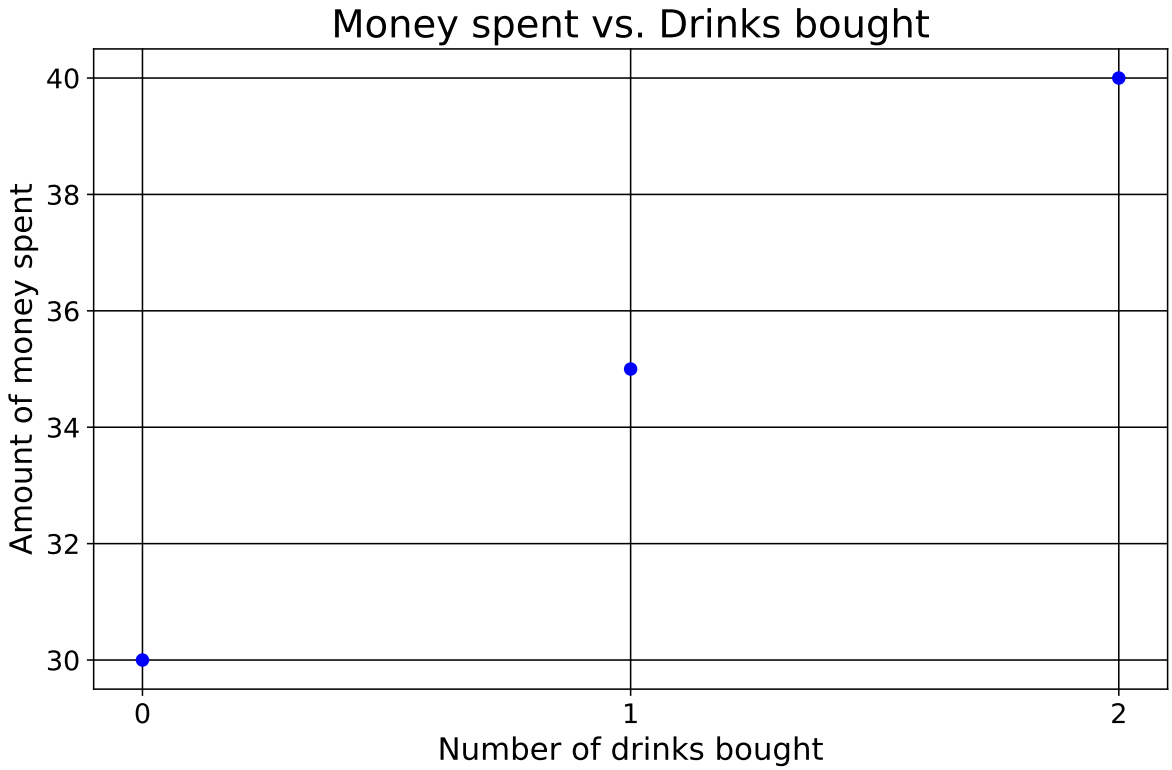
$$h(t) = -5t^2 + 20t + 2 = \frac{d}{dt}(-5t^2) + \frac{d}{dt}(20t) + \frac{d}{dt}(2) = -10t + 20$$

When we solve the derivative for 0, $-10(2) + 20 = 0$, we get the maximum height of 22 feet after 2 seconds. Derivatives 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 squared residuals. This is known as the *ordinary leasy squares* approximation (OLS) for *linear regression*. OLS is the main estimator you'll use for this class.


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

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 known variables, now we must solve for them. Unlike univariable calculus however, here we have multiple variables which each affect the function differently. And in this case, it makes perfect sense: if we wish to derive a function for how much we spend at the store, it's quite possible we may spend money on multiple things and that those things will have different effects on how much you spend in total. With regression, we take the partial derivative with resepct to each variable, holding the other variable constant (that is, not considering the effect of the other variable). If this seems at all abstract to you, I sort of deviate from most econometrics textbooks and I formally derive the OLS betas. precisely, I give a detailed explanation of how the betas are calculated. Note that all of the steps below, Stata, R, or Python does (and optimizes!) for you. I only provide this derivation so you have a source to refer to when you wish to know how adn *why*, exactly, the machine returns the numbers that it returns to you.

Suppose we wish to attend a nightclub. We pay *some* cost of money to enter, and then we pay *some* amount of money per drink. However, unlike the real world where we know the price and entry fee up front and may calculate it, in this case we wish to derive the amount we pay up front, as well as the cost per drink using only reported data.



3.3 An Extended Example

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

3.3.1 Step 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 and y=amount of money we spend that evening. 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.

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

3.3.2 Step 2: List Our Variables We Solve For

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

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

Here, y_i is how much we pay, β_0 is how much we pay to enter, β_1 is how much we pay for each additional drink, and x is the number of drinks we get. Nothing is different, so far, from anything we've discussed above. I've simply substituted m and b with the Greek letter "beta". Precisely, these are the betas we wish to derive.

3.3.3 Step 3: Write Out the Objective Function

OLS minimizes the sum of squared residuals. That is, it seeks the line which best fits the data, given some input value of x. We can represent this with something called an objective function.

$$S = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

The word "argmin" here means "argument of the minimum". The symbols underneath it, β_0, β_1, mean these are the values we seek to estimate which minimize the prediction errors of the line. Everything after "argmin" is the objective function we are minimizing. Precisely, we wish to minimize the sum of the residuals squared. As above, y_i is our observed data points (how much we pay in total given the ith drink bought) 30, 35, and 40. The sigma symbol $\sum_{i=1}^{n}$

means we are adding up the ith squared residual to the nth data point/number of observations (in this case 3).

One may ask *why* we are squaring the residuals. First of all, the squared residuals means that we are dealing strictly with positive numbers (making the solution analytically easier to compute). Using the sum of absolute errors is a non-differentiable function. Also, suppose one observation has a residual of 40 and another of -40. The residuals would cancel out to 0. The squaring also has the property of penalizing worse predictions, realtive to simply summing the residuals. If the observed value is 20 but we predict 25, the error is -5. 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 all errors as created equally, and they are not (or, should not be for desirable statistical properties).

3.3.4 Step 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 (β_0) and how much it costs for each drink (β_1)

$$S = (30 - (\beta_0 + \beta_1 \cdot 0))^2 + (35 - (\beta_0 + \beta_1 \cdot 1))^2 + (40 - (\beta_0 + \beta_1 \cdot 2))^2$$

3.3.5 Step 5: Simplify the Objective Function, cont.

Here we simplify things.

$$S = (30 - \beta_0)^2 + (35 - \beta_0 - \beta_1)^2 + (40 - \beta_0 - 2\beta_1)^2$$

Note that the $\beta_1 \cdot 0$ term goes away because anything multiplied by 0 is just 0. We also simplify the second term because anything multiplied by 1 is just itself in the case of $\beta_1 \cdot 1$.

3.3.6 Step 6: Take Partial Derivatives

To find the values of β_0 and β_1 that minimize S, we take the partial derivatives of S (for "Spending") with respect to β_0 and β_1. Recall that the partial derivative is simply how much the value of the function changes when our other variables that affect our outcome are held constant. In this case, the power rule and chain rule are our friend here. To hear more about combining the power rule and chain rule, see here.

1. **Partial derivative with respect to β_0:**

First, we do:

$$\frac{\partial S}{\partial \beta_0} = -2(30 - \beta_0) - 2(35 - \beta_0 - \beta_1) - 2(40 - \beta_0 - 2\beta_1).$$

Nothing crazy has happened here. I've simply brought the 2 down in front of each set of parentheses and subtracted 1 from the exponent. The -1 comes from the chain rule due to the fact that we consider $(30 - \beta_0)$ as one function, and $(30 - \beta_0)^2$ as another function. The "inner" function $(30 - \beta_0)$ has a derivative of -1 for β_0, since there's no exponent attached to it and 20 is a constant, so its derivative is 0. Therefore, the derivative of the outer function $(30 - \beta_0)^2$ (applying the power rule from above) is $2(30 - \beta_0) \cdot -1$. Now, we simplify. First, distribute the -2 across each term inside the parentheses:

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= -2(30 - \beta_0) - 2(35 - \beta_0 - \beta_1) - 2(40 - \beta_0 - 2\beta_1) \\ &= -2(30) + 2\beta_0 - 2(35) + 2\beta_0 + 2\beta_1 - 2(40) + 2\beta_0 + 4\beta_1.\end{aligned}$$

Then we combine like terms:

$$= -60 + 2\beta_0 - 70 + 2\beta_0 + 2\beta_1 - 80 + 2\beta_0 + 4\beta_1.$$

Now, we group the β_0 and β_1 terms together:

$$= (-60 - 70 - 80) + (2\beta_0 + 2\beta_0 + 2\beta_0) + (2\beta_1 + 4\beta_1).$$

Now, I factor out the 2

$$= (-60 - 70 - 80) + 2(3\beta_0) + 2(\beta_1 + 2\beta_1).$$

So, we get

$$= -210 + 6\beta_0 + 6\beta_1.$$

Now we re-arrange by adding 210 to the RHS

$$6\beta_0 + 6\beta_1 = 210.$$

See how there are two 6s? We may now divide the right and left hand sides by 6 to further simplify

$$\beta_0 + \beta_1 = 35$$

Okay, first parital calculated.

2. **Partial derivative with respect to β_1:**

$$S = (30 - (\beta_0 + \beta_1 \cdot 0))^2 + (35 - (\beta_0 + \beta_1 \cdot 1))^2 + (40 - (\beta_0 + \beta_1 \cdot 2))^2$$

Rewrite this as:

$$S = (30 - \beta_0)^2 + (35 - (\beta_0 + \beta_1))^2 + (40 - (\beta_0 + 2\beta_1))^2$$

We differentiate each term with respect to β_1:

1. The first term:

$$\frac{\partial}{\partial \beta_1}(30 - \beta_0)^2 = 0$$

(since it does not involve β_1)

2. The second term:

$$\frac{\partial}{\partial \beta_1}(35 - (\beta_0 + \beta_1))^2 = \frac{\partial}{\partial \beta_1}(35 - \beta_0 - \beta_1)^2$$

Using the chain rule:

$$= 2(35 - \beta_0 - \beta_1) \cdot (-1)$$

$$= -2(35 - \beta_0 - \beta_1)$$

3. The third term:

$$\frac{\partial}{\partial \beta_1}(40 - (\beta_0 + 2\beta_1))^2 = \frac{\partial}{\partial \beta_1}(40 - \beta_0 - 2\beta_1)^2$$

Using the chain rule:

$$= 2(40 - \beta_0 - 2\beta_1) \cdot (-2)$$

$$= -4(40 - \beta_0 - 2\beta_1)$$

Now, sum the partial derivatives of the individual terms to get the partial derivative of S with respect to β_1

$$\frac{\partial S}{\partial \beta_1} = 0 + (-2(35 - \beta_0 - \beta_1)) + (-4(40 - \beta_0 - 2\beta_1))$$

Simplify this expression:

$$\frac{\partial S}{\partial \beta_1} = -2(35 - \beta_0 - \beta_1) - 4(40 - \beta_0 - 2\beta_1)$$

Expand and combine like terms:

$$\frac{\partial S}{\partial \beta_1} = -2(35 - \beta_0 - \beta_1) - 4(40 - \beta_0 - 2\beta_1)$$

$$= -2 \cdot 35 + 2\beta_0 + 2\beta_1 - 4 \cdot 40 + 4\beta_0 + 8\beta_1$$

$$= -70 + 2\beta_0 + 2\beta_1 - 160 + 4\beta_0 + 8\beta_1$$

$$= -230 + 6\beta_0 + 10\beta_1$$

Just to set it up for the next section, we can again place the constant on the RHS after rearranging

$$6\beta_0 + 10\beta_1 = 230$$

Okay, second partial calculated.

3.3.7 Step 7: Get the Betas

Now we've set up our partial derivatives. Now we can algebraically get our betas. In order to find their values, though, 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 predictions and the observed data. Here are the equations we must solve for (note, we have two because we have two variables, the entry fee and the drinks, but this can easily be extended to more variables)

$$\beta_0 + \beta_1 = 35$$
$$6\beta_0 + 10\beta_1 = 230.$$

Here I use a method called substitution to solve the system, but there are many such ways we can solve this.

3.3.7.1 Step 7.1: Solve for β_0:

The first step of substitution is to solve for one of our variables. I chose the first parital since it is the easiest.

$$\beta_0 = 35 - \beta_1$$

This is pretty simple, we just subtract β_1. We know that the entry fee is some positive number. We also know that β_1 on the RHS has to be less than 35, as this would mean that we get in for free, which does not make sense given the problem at hand. We also know, logically, that β_0 can't be 35, because then this means that drinks are free, which is also inconsistent with the logic of the problem. The reason I'm making such a fuss about this, is because when the regression gives us a number, the model predictions need to either make sense under existing constraints, 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. It also means we must be sensible in interpreting our regression models– if negative values do not make sense for a certain variable, then restrict the interpretation of the model to values that do make sense.

3.3.7.2 Step 7.2: Substitute β_0:

Now, since we know the equation for the constant (the entry fee), we can put this back in the partial derivative for β_1

$$6\beta_0 + 10\beta_1 = 230.$$

Once we do this, we can solve for β_1. Here is the partial for β_1 with this replacement for β_0.

$$6(35 - \beta_1) + 10\beta_1 = 230.$$

To solve, we distribute the 6

$$210 - 6\beta_1 + 10\beta_1 = 230$$

and combine these terms $-6\beta_1 + 10\beta_1$ together

$$210 + 4\beta_1 = 230.$$

Next, we subtract 210 from the RHS

$$4\beta_1 = 20.$$

Finally, we divide by 4

$$\beta_1 = 5.$$

Now, we know our value for β_1!!! We know that for each drink we get, we pay 5 more dollars. Since we now know *this*, we substitute 5 into $\beta_0 + \beta_1 = 35$ where β_1 is to get β_0. We now have one equation to solve

$$\beta_0 + 5 = 35,$$

where we simply subtract 5

$$\beta_0 = 30.$$

Now, we know the entry fee is 30 dollars.

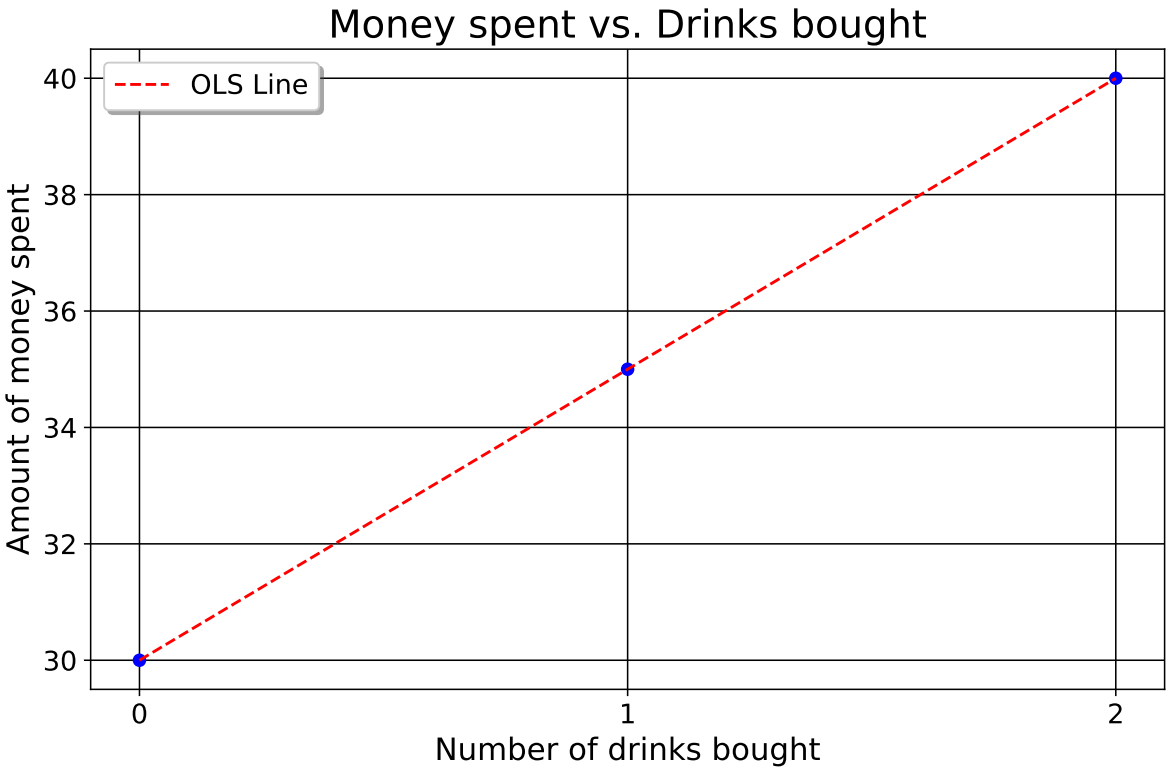
3.3.8 Step 8: Our OLS Line of Best Fit

So, our line of best fit is $\widehat{y} = 30 + 5x$. The way we interpret this is is for every additional drink bought, you pay 5 more dollars. In fact, more precisely, *compared* to those who only wanted to get in the club, 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 because the parital derivative allows for us to (atttempt to) isolate the causal impact of *multiple* variables. In fact, we did so implicitly right here in this derivation! We considered the impact of BOTH entry fees and the number of drinks we bought too. If we wanted to add in another column, say the cost of an additional cigar or some such thing, all we would do is add another beta, β_2, to the original objective function and multiply the corresponding values by whatever the costs were. We would then have three systems of equations to solve for, but the underlying principle of adding in other variables and finding the line that minimizes the distance between the line and points is the same.

3.4 After Running Regressions

Next, we discuss…


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

References