Econometrics for Social Scientists

Volume I: Causal Inference and Estimation

Kevin Lingfeng Li

Table of contents

Preface

This book is the 1st book in a sequence on **Econometric Methods for Political Analysis**.

1. Volume I: Causal Inference and Estimation (this book) introduces causal inference, randomised experiments, the ordinary least squares estimator, the instrumental variables estimator, and quasi-experimental designs.
2. Volume II: Further Causal Estimators expands on the first book by discusses other estimators, such as the maximum likelihood estimator, selection on observables estimators, and generalised methods of moments estimator.
3. Volume III: Prediction and Forecasting introduces prediction methods such as non-linear prediction, classification, cross-validation, as well as time-series models.
4. Volume IV: Multivariate Measurement discusses topics in latent variable measurement, dimensional reduction, and measurement of textual data for statistical analysis.

This series is designed to be both an approachable, but also rigorous, introduction to Econometrics and the use of statistical methods for the analysis of social and political actors and phenomena.

I assume a solid understanding of basic probability and statistics, including the topics of conditional probabilities, random variables and distributions, expectation/mean and variance, and correlation.

I also assume a solid understanding of mathematics, including algebra, single variable calculus, and some linear algebra. While you will be able to still learn from this book without a solid mathematical background, you will gain much more from understanding the mathematics behind the methods.

To see what mathematics is specifically required, or to refresh on the mathematics needed, consult the **Quantitative Methods** sequence (particularly the 1st volume and some topics in the 2nd volume).

I will also add some reference code for the R-programming language and Stata in case you are interested in implementing these methods on your own.

Part I

Introduction to Causal Inference

Chapter 1

Causal Inference

1.1 Correlation is not Causation

Econometrics is the field of applying statistical methods to analyse real-world social science data. Econometrics has two primary goals:

1. **Causal Inference**: Establishing how one feature directly causes another feature. This is essential to understanding the world around us and designing better policies. Key point: correlation causation.
2. **Predictive Inference/Forecasting**: Given data we have, how can we predict the values of data we do not have? For example, what will sales be next year? GDP? Who will win the next election? What are the likely costs/effects of a policy?

We will focus on causal inference in this book, as it is the most important aspect of econometrics.

Variables are **correlated** when one variable changing is associated with the other variable changing on average. For example, an increase in X causes an increase/decrease in Y would mean X and Y are correlated.

There are three reasons why variables are correlated with each other.

1. There could be a causal effect of X on Y, causing the correlation.
2. There could be a third variable, W, causing both X and Y to change.
3. There is a causal effect of Y on X.

The most important thing about econometrics and causal inference is that **correlation does not equal causation**. This is because of the 2nd reason listed above - our correlation between X and Y could be caused by a third variable W, called a **confounder**.

For example, *Ice Cream Sales* and *Number of Fatal Shark Attacks* are two highly correlated variables in the United States. Does this mean that selling ice cream causes fatal shark attacks? No!

The reason this relationship exists is because of another variable - the *weather*. The weather, when it is sunny, causes both ice cream sales and more people to go to the beach, which causes more fatal shark attacks. However, there is no direct link between ice cream sales and the number of fatal shark attacks.

The goal of econometrics is to distinguish between correlation and causation. We want to "partial out" the effect of confounding variables and isolate the causal effects.

This book will introduce a series of different **estimators** that will help us isolate and estimate these causal effects.

1.2 Potential Outcomes and Causal Effects

A **causal effect** is a change in some feature of the world Y, that would directly result from a change in some other feature D. Essentially, change in D causes change in Y.

> Key Definition: Potential Outcomes
>
> Causal effect implies that there are **potential outcomes**. Imagine that there are 2 worlds, that are exactly the same until treatment D occurs. In one world, you get the treatment D, and the other world, you do not get this treatment. Since these 2 worlds are identical besides the treatment D, the difference between the world's Y outcomes are the effect of our treatment D.
>
> In the real world, we only observe one of these realities - either a unit i gets, or does not get, the treatment. The other world that we do not observe is called a **counterfactual**.

Thus, there are two states of the world in the potential outcomes framework:

- The control state $D = 0$ is the world where a unit does not receive the treatment D. Y_{1i} is the potential outcome for unit i, given it is in the treatment state $D_i = 1$.
- The treatment state $D = 1$ is identical to the control state, with the only exception that a unit receives the treatment D. Y_{0i} is the potential outcome for unit i, given it is in the control state $D_i = 0$.

The **individual causal effect** τ of the treatment D for any unit i is $\tau_i = Y_{1i} - Y_{0i}$. Since the two states are identical except for treatment D, the resulting difference must be as a result of treatment D. However, in the real world, we do not have parallel worlds (unfortunately) - we only observe one outcome: either unit i gets the treatment $D_i = 1$, or does not get the treatment $D_i = 0$.

> Key Definition: Observed Outcomes
>
> The **observed Y outcome** (in the real world) of any unit i is given by the equation:
>
> $$Y_i = D_i \times Y_{1i} + (1 - D_i) \times Y_{0i}$$

This equation might be a little abstract, however, it is easy to understand by plugging D_i in:

$$[Y_i|D_i = 0] = 0 \times Y_{1i} + (1 - 0) \times Y_{0i} = Y_{0i}$$
$$[Y_i|D_i = 1] = 1 \times Y_{1i} + (1 - 1) \times Y_{0i} = Y_{1i}$$

Intuitively, if the observation is in the control state $D_i = 0$, we observe potential outcome Y_{0i}. When an observation is in the treatment state $D_i = 1$, we observe potential outcome Y_{1i}.

Stable Unit Treatment Value Assumption

Take two units i and j. The Stable Unit Treatment Value Assumption (SUTVA) is the assumption that unit j getting the treatment D, does not affect the outcomes of unit i.

This is important - because if this assumption were to be violated, we would have more than two potential outcomes. If unit i were affected by j's treatment status, we would not only have the potential outcomes of unit i being in treatment or control, but also would have to consider unit j being in treatment or control.

1.3 Estimands and Estimator Properties

Causal Estimands

An **estimand** is the quantity we are trying to estimate (i.e. what we are interested in). One of the estimands is the Average Treatment Effect:

> Key Definition: Average Treatment Effect
>
> **Average treatment effect (ATE)** is the average of all individual treatment effects:
>
> $$\tau_{ATE} = \mathbb{E}[\tau_i] = \mathbb{E}[Y_{1i} - Y_{0i}] = \frac{1}{n}\left(\sum Y_{1i} - \sum Y_{0i}\right)$$

There are also other treatment effects we can use to estimate. The **average treatment effect on the treated (ATT)** is the treatment effect of only units who recieved the treatment $D_i = 1$

$$\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]$$

The **average treatment effect on the controls (ATC)** is the treatment effect of units who only did not recieve the treatment $D_i = 0$

$$\tau_{ATC} = \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 0]$$

The **conditional average treatment effect (CATE)** is the treatment effect of units, given they have some other variable X value. For example, if X is gender, the CATE could be the treatment effect on only females.

$$\tau_{CATE} = \mathbb{E}[Y_{1i} - Y_{0i}|X = x]$$

Estimator Bias and Variance

The above causal estimands are not directly calculable, and we to estimate them with an **estimator**.

Bias is when an estimator consistently and systematically poorly estimates the estimand. Or in other words, an estimator is biased, if on average, the estimate of our estimand (over many tries of estimation), is not actually the true value of the estimand. That means something is consistently off with our estimator - we might be consistently overestimating by 5%, or underestimating, etc. Mathematically:

$$\mathbb{E}[\hat{\theta}_i] = \mathbb{E}[\theta], \text{ where } \hat{\theta} \text{ is the estimate}$$

Variance is the difference between our estimations derived from our estimator - i.e. the consistency.

- For example, you might have an unbiased estimator, where our average estimate is the actual causal estimand. However, while the average is correct, the variance of our estimates is very wide.

Ideally, we want an unbiased estimator that has low variance. We will explore many different types of estimators for causal effects throughout this book, each with its own bias and variance.

1.4 Naive Estimator of Observed Outcomes

The **naive estimator** is an estimator that only compares our observed outcomes, without any comparison to the counterfactual potential outcomes:

$$\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$$

Or more intuitively, the average observed outcome Y of those in the treatment group, minus the average observed outcome Y of those in the control group.

This is often what many people initially do when trying to find a causal effect. Essentially, we are comparing units that are assigned to treatment, and the units that are not assigned to treatment, and their observed outcomes.

- In other words, the naive estimator is looking at the correlation between the treatment D and the observed outcomes of Y, without considering the counterfactual.

The naive estimator is a bad idea. Remember, our treatment effects are supposed to be comparing to two potential outcomes of the same unit. We are supposed to compare Y_{1i} to Y_{0i}.

- However, in this scenario, we are not comparing the potential outcomes of the same individual. We are comparing the outcome of some observation A in treatment Y_{1A} and the outcome of some other observation B in control Y_{0B}.
- But what if observation A and B are different? There outcomes may not be due to the treatment D, but because of the differences between A and B. This is why counterfactual comparison is important - when we compare the potential outcomes of the same unit in control and treatment groups, we can be confident of the affect of the treatment, since it is the same unit of observation for both groups.

We can prove this mathematically. We start with the naive estimator:

$$\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$$

Then, we do a little algebra trick - we add a new term to this equation, and then subtract the same term. The two new terms thus cancel each other out to 0.

$$= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i - 1]$$

Then, we rearrange the terms, then simplify, getting the result:

$$\begin{aligned} &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \end{aligned}$$

If we look at the final result, we can divide it into 2 parts:

1. The first part, $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]$, is the average treatment effect of the treated τ_{ATT} that we introduced previously.
2. The second part $\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$ is what we call the **selection bias**. Intuitively, it is the difference between the treatment and control groups prior to the treatment taking place (hence the potential outcome being Y_{0i}).

1.5 Selection Bias and Confounding Variables

The differences between the treatment and control groups prior to treatment is captured in our naive estimator, which is why our results with the naive estimator are **biased**.

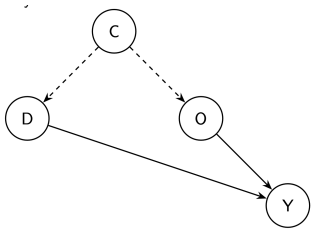
For example, if we are measuring the question *does going to the hospital make you more healthy*, and we simply measured the outcomes of people who went to the hospital and did not go to the hospital, we might see that in general, people who did not go to the hospital are healthier!

- Does this mean that going to the hospital makes you unhealthier? No! It is because more unhealthy people choose to go to the hospital in the first place. Thus, the hospital has generally more unhealthy individuals in it. The hospital might perform miracles on these people, but they are still not as healthy as the healthy people who did not need to go to the hospital.
- The differences between the people who chose to go to the hospital versus the people who did not go to the hospital explains the differences in our outcome, not the actual treatment that the hospital provided.

This is selection bias - when our treatment and control groups are fundamentally different and unequal even prior to treatment.

> Key Definition: Confounders
>
> A **confounder** is a variable that is explaining the differences in the treatment and control groups.
>
> Confounding variables result in selection bias, and are why correlation does not equal causation. In order to accurately calculate causal effects, we need to find some way to eliminate the effect of confounding variables.

For example, look at the figure below:



In the figure above, D is the treatment group, and O is the observed group. C is some confounding variable correlated with D, that affects whether an observation i gets the treatment D or control C.

When we calculate the naive estimate (or correlation), our causal estimate captures both the effect of $D \rightarrow Y$, but also the effect of $D \leftrightarrow C \rightarrow O \rightarrow Y$, since D and C are correlated.

- This second effect through the correlation of D and C is called the **backdoor path**.
- Both the direct $D \rightarrow Y$ and the backdoor path are included in the naive estimator (and correlation).

However, the actual causal effect of treatment D on Y is only the section of $D \rightarrow Y$, which does not include the backdoor path. So, we need to find some way to only look at $D \rightarrow Y$, and eliminate/partial out the effect of the backdoor path.

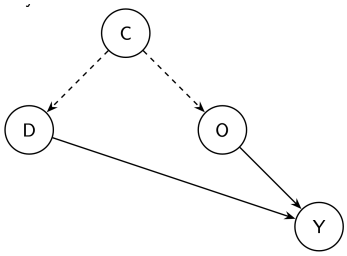
To make an accurate causal effect estimate, we must get rid of confounding variables, selection bias, and the backdoor path. How do we do this? The best method is randomisation, which we will cover next.

Chapter 2

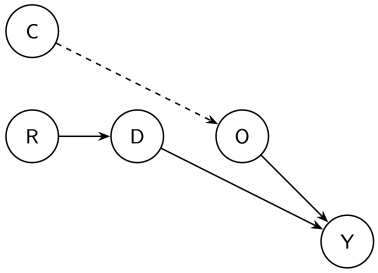
Randomised Controlled Trials

2.1 Random Assignment and Confounders

The **assignment mechanism** is how we decide which observations receive treatment D. In the last chapter, we discussed how confounder C affects which observations get the treatment, introducing selection bias.



We can address this confounder C and eliminate the backdoor path $D \leftrightarrow C \rightarrow O \rightarrow Y$ by randomly assigning units into either the treatment or control group:



With random assignment mechanism R, now units are assigned to control randomly, not based on the confounder variable C. Thus, C and D should no longer be correlated, thus removing the backdoor effect, selection bias, and the influence of confounder C.

Key Definition: Random Assignment

With random assignment, selection bias and the influence of confounder C is eliminated, thus the control group and treatment group should be very similar. Thus, the potential outcomes are independent of treatment/control status:

$$\mathbb{E}[Y_{1i}|D_i = 1] \approx \mathbb{E}[Y_{1i}|D_i = 0] \quad \text{and} \quad \mathbb{E}[Y_{0i}|D_i = 1] \approx \mathbb{E}[Y_{0i}|D_i = 0]$$

2.2 Difference-in-Means Estimator

If this assumption of random assignment is met, then we can use the naive estimator to estimate the treatment effect. This is because we have eliminated the confounders that cause selection bias.

Mathematically, we can prove this. Recall the naive estimator:

$$\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$$

The selection bias term $\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] = 0$ under the above assumptions of randomisation. Thus, there is no longer selection bias, and confounding variables will have been accounted for.

Our observed potential outcomes in our randomised experiments are Y_{1i} and Y_{0i}. We know that the treatment group D_i does not affect our potential outcomes, since units are the same in both treatment and control. Thus we know that:

$$\mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}]$$
$$\mathbb{E}[Y_{0i}|D_i = 0] = \mathbb{E}[Y_{0i}]$$

$\mathbb{E}[Y_{1i}]$ is the expected value of Y in the treatment group. Basically, we just find the mean Y value of all observations of the treatment group.

$\mathbb{E}[Y_{0i}]$ is the expected value of Y in the control group. Basically, we just find the mean Y value of all observations in the control group.

Thus with these, we can find the average treatment effect:

Key Definition: ATE Estimate of a Randomised Experiment

The estimate of the average treatment effect of a randomised controlled trial is:

$$\hat{\tau}_{ATE} = \mathbb{E}[\tau_i] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] = \bar{Y}_t - \bar{Y}_c$$

Where $\bar{Y}_t$ is the average Y value of the treatment group, and $\bar{Y}_c$ is the average Y value of the control group. Thus, the causal effect is simply a **difference-in-means** of the two groups.

Or in other words, to calculate the average treatment effect of a randomised experiment, we:

1. Calculate $\bar{Y}_t$, the average Y value of all observations in the treatment group.
2. Calculate $\bar{Y}_c$, the average Y value of all observations in the control group.
3. Find the difference $\bar{Y}_t - \bar{Y}_c$. That is our average treatment effect.

For this estimate of the ATE to be true, there must be random assignment of treatment, and the treatment and control groups must be similar to each other on all confounding variables.

If this assumption is violated (either due to lack of random assignment, or failure of randomisation), we can no longer estimate the causal effects with the difference-in-means estimator. This is because if the control and treatment groups are not similar to each other on all confounders, there is the possibility of selection bias in our difference-in-means estimator.

Finally, note that we can also use a different estimator, the OLS estimator (with simple linear regression), to get the same final estimate. We will discuss this in chapter 3.

2.3 Uncertainty and Standard Errors

Intuition of Uncertainty

Remember how we randomly assigned units to treatment or control? What if we ran the experiment again? The treatment and control groups would very likely not be exactly the same, and thus, we would get a slightly different causal effect.

Thus, we have some uncertainty with our causal estimate - re-running the experiment might result in a different answer. The ATE we have calculated is only our specific sample average treatment effect (SATE), often notated $\hat{\tau}_{ATE}$ or $\hat{ATE}$.

- Why sample? Well, through random assignment, you are basically "randomly sampling" potential outcomes - since randomly choosing one unit to be in treatment/control means not seeing the other counterfactual potential outcome.

Thus, we need some mechanism to account for sampling variability and how rerunning the experiment might result in slightly different results. We do this with sampling distributions and standard errors.

Sampling Distributions and Standard Error

Imagine that we take a sample from a population (or some random assignment mechanism). Then, we find the average treatment effect of the sample $\hat{\tau}_{ATE}$.

That is a **sample estimate**, which is often notated $\hat{\theta}$. (I use θ, since this idea of uncertainty can be applied to any estimate, not just average treatment effect).

Then, let us take another sample from the same population (or do another random assignment), and find the sample estimate. This will be slightly different than the first sample, since we are randomly sampling. That is another sample estimate.

We keep taking samples from the same population (more random assignments), and getting more and more sample estimates.

Let us plot all our sample estimates $\hat{\theta}$ (different $\hat{\tau}_{ATE}$ values) into a "histogram" or density plot. The x axis labels the possible $\hat{\tau}_{ATE}$ values, and the y axis is how frequently a specific sample estimate occurs.

The result is a distribution, just like a random variable distribution. That distribution is the **sampling distribution**.

According to **central limit theorem**, the sampling distribution approximates that of a **normal distribution** (or t-distribution if our sample size is small). We know that a normal distribution is defined by two parameters - mean and variance.

Standard deviation is the square root of variance. The standard deviation of our sampling distribution is what is called the **standard error** of our estimate.

> Key Definition: Standard Error
>
> A **sampling distribution** is the imaginary distribution of estimates, if we repeated the sampling and estimation process many, many times.
> The **standard error** is the standard deviation of the sampling distribution. It is often notated $SE(\hat{\theta})$.
> The computer/software we use will calculate an estimate for us.

2.4 Hypothesis Testing and Causal Inference

We know there is some uncertainty with our sample estimate, as defined by the standard error. So, how do we know if we actually have a causal effect with this uncertainty?

What we do is hypothesis testing: a way to test, given a certain level of uncertainty, whether or not we believe there is a causal effect.

We start off with the status-quo "old theory", and try to disprove it. This status quo theory, called the **null hypothesis** and notated H_0, is typically that *there is no causal effect of* D on Y. The new theory we have come up with, and are trying to prove, is called the **alternate hypothesis**, often notated H_1 or H_a.

Mathematically, our hypothesis are:

$$H_0 : \tau_{ATE} = 0 \quad \text{and} \quad H_1 : \tau_{ATE} \neq 0$$

We assume that the null hypothesis is true, unless we are 95% confident that we can reject the null hypothesis, and only then, can we accept the alternative hypothesis.

How do we actually test these hypotheses? First, we have to calculate a t-test statistic:

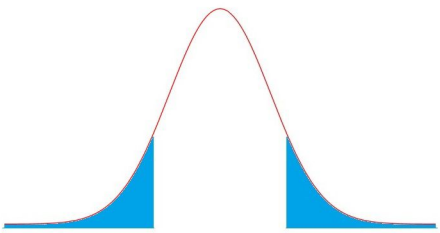
Key Definition: T-test Statistic

The t-test statistic tells us how far our estimate is from the null θ_0, in terms of standard errors:

$$t = \frac{\hat{\theta}_j - \theta_0}{\widehat{SE}(\hat{\theta}_j)}$$

Then, we find the corresponding t-distribution based on the degrees of freedom ($n - 2$). Then, we start from the middle of that t distribution, and go the *number of standard errors* away based on the t-statistic. We do this on both sides from the middle of the t-distribution.

Once we have found that point, we find the probability (area under the distribution) of a t-test statistic of ours, or more extreme, could occur. The figure below shows this probability (called the p-value):



Key Definition: P-value

Essentially, a p-value is how likely we are to get a test statistic at or more extreme than the one we got for our estimated θ_j, given the null hypothesis is true. If this is less than 0.05 (5%), then we reject the null hypothesis as no longer true.

- So if the p-value is above 0.05, there is a high chance that the null hypothesis is true. Thus, we cannot reject that there is no causal effect of D on Y.
- If the p-value is lower than 0.05, then there is a low chance that the null hypothesis is true. Thus, we can reject the null hypothesis, and conclude a causal effect of D on Y.

2.5 Balance Table and Stratified Experiments

Balance Tables

If the treatment and control groups are not similar (in regard to key confounding variable values), our difference-in-means estimator will be biased with selection bias.

To confirm this assumption is met, researchers will often show a **balance** table before or their estimation process. A balance table is essentially a table that shows the average difference in values of confounders in both treatment and control groups. For example, below is a balance table:

	Confounder 1	Confounder 2	Confounder 3	Confounder 4
Control Group Values	11.503***	1.474***	0.405***	81.983***
	(0.196)	(0.060)	(0.015)	(2.980)
Treatment - Control	−0.069	−0.062	0.009	−1.760
	(0.241)	(0.074)	(0.019)	(4.179)
Num. obs.	562	565	560	565

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The key row to look at is the *treatment-control* row. The numbers not in parentheses the difference between treatment and control for the corresponding confounder. The numbers in parentheses are the standard errors of the estimated difference. Stars (see legend below the table) show significance of a t-test (we will discuss this in the next section).

If no *treatment-control* for key confounders is significantly different (no stars), then randomisation has suceeded, and indeed, our control and treatment group are similar.

Blocking and Stratified Experiments

Blocking, also called stratified experiments, is an extension of randomised experiments to ensure that randomisation does not fail. Imagine that you have four units in your experiment that you have to assign to treatment/control. Their pre-treatment outcomes are $Y_{0i} = \{2, 2, 8, 8\}$. This means that you have a 1/3 chance to end up with the random assignment of $\{2, 2\}$ in one group and $\{8, 8\}$ in the other group, which would result in selection bias.

With blocking, you can prevent this from happening. The procedure is as following:

1. Before randomisation, you separate your sample of N units into J subgroups.
2. Within each group, randomly assign units to treatment and control group (essentially, smaller randomised experiments within a bigger experiment).

For example, we could divide our prior example into 2 subgroups: $\{2, 2\}$ and $\{8, 8\}$. Then, within each group, randomly assign one observation to treatment, and one to control. Thus, we are guaranteed to get units from both subgroups in both our treatment and control groups.

To estimate our effects for blocking experiments, we will have to take the weighted average of each subgroup's average treatment effect (ATE), with the weights being the proportion of units each group accounts for:

$$\tau_{ATE} = \sum_{j=1}^{J} \frac{N_j}{N} \tau_j$$

Where N is the total number of observations, J is the total number of subgroups, j is one of the subgroups, N_j is the number of units within subgroup j, and τ_j is the ATE of the subgroup j.

2.6 Validity and Issues with Experiments

Randomisation, when it works, is magical - it can help us obtain the best estimates of causal effects that are possible.

But, randomised controlled trials do have several issues that can affect our validity of our estimates and conclusions:

1. **Failure of randomisation**: if for some reason, randomisation does not result in control and treatment groups being similar, we cannot accurately estimate the average treatment effect. Blocking should help us deal with this, but it is still possible for randomisation to fail.
2. **Non-Compliance**: Sometimes, despite assigning certain units to treatment or control, the units do not comply and do the opposite (i.e. people assigned to control still take the treatment). This is a huge issue since researchers are not gods - we cannot force people to take treatment. The biggest threat non-compliance creates is that perhaps, a confounder is causing certain units to be more likely to not comply. We will deal with this issue in the later parts of this book with Instrumental Variables.
3. Attrition: sometimes, it is not possible to measure the outcomes of some people in a study, either due to people moving away, passing away, or refusing to answer surveys or have their measurements be taken. This once again is an issue - since perhaps a confounding variable is causing this issue.

However, the biggest issue with randomised controlled trials is the impracticality and infeasability of them in many situations, especially in social science research.

- For example, let us say you want to run an experiment on how democracy affects economic growth. It is nearly impossible to randomly assign countries to be a democracy. First, if you assign, for example, Russia to be a democracy, you have no power to make them actually follow through with your study. Second, there are ethical concerns about randomly allocating millions to democracy or autocracy.
- Another major issue is that even when experiments are theoretically feasible, they can be too expensive to implement. Randomised Controlled Trials are extremely expensive even when they are possible to run.

Thus, we will need to introduce ways to address for confounders and estimate causal effects for the vast majority of cases where we will not be able to run a randomised controlled trial.

The rest of the book focuses on these techniques. A brief overview of them:

- Simple Linear Regression will not be of much use for this, as it typically only is feasible in ideal random situations, however, it is the core that many further methods build on, and can help us explore continuous treatment variables.
- Multiple Linear Regression, in theory, can control for confounding variables. However, the limitation of this is that for an accurate estimation, all possible confounding variables (including unobservable or unmeasureable) confounders have to be included to prevent bias, which is very infeasible most of the time.
- Instrumental Variables Estimator, in theory, can deal with both non-compliance and confounders, however, there are also drawbacks to this method that we will cover later.
- Finally, we will discuss quasi-experimental methods, like regression discontinuity and differences-in-differences, which are the most popular methods of causal inference today.

Part II

Ordinary Least Squares Estimator

Chapter 3

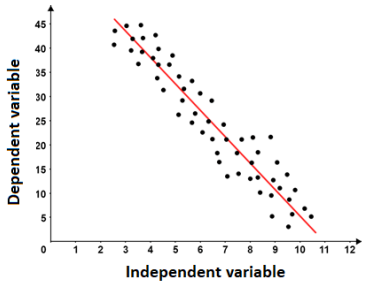
Simple Linear Regression and Causal Inference

3.1 Introduction to Regression

So far, we have looked at only binary treatment variables D, that can either be $D = 0$ control group or $D = 1$ treatment group. However, a lot of the relationships we are interested in are not binary. For example, you might be interested in how someone's income affects their likelihood to vote. Or how someone's age affects their political beliefs. Or how a country's GDP affects the chance a country becomes a democracy. Income, age, and GDP are all treatment variables that are not binary.

How can we explore the causal relationship between a continuous treatment variable D (often instead notated as X in statistics), and a continuous Y variable (we will focus on non-continuous Y later in chapter 5)?

One way is that we can explore this relationship with a linear best-fit line. A best-fit line is useful, since the **slope** of the line represents how much Y changes for every unit change of D.



The graph here introduces some common terminology:

- The **independent variable**, also called the **explanatory variable** or **treatment variable**, often notated X (or D in casual inference settings) is the variable we believe to be doing the causing.
- The **independent variable**, also called the **response variable** or **outcome variable** is notated Y. This is the variable that is being caused.

With regression, we are interested in how X causes Y. Given a set of data with n number of observations, each observation i being a data point (X_i, Y_i), the goal of regression is to fit a best-fit line through this data. With this best-fit line, the slope will tell use the average affect of a one unit change in X on Y.

3.2 Simple Linear Regression Model

A linear regression model is the specification of the conditional distribution of Y, given X (or D in causal settings).

- Why a distribution? For example, if X was age and Y was income, at age $X = 30$, not every single 30 year old makes the same amount of money. There is some distribution of incomes Y at age $X = 30$.
- Why conditional? Well, if we believe X causes Y, then we should expect the distribution of Y values to change based on X values. If Y values do not change when X changes, then there is no causal effect.

The linear regression model focuses on the **expected value** of the conditional distribution, notated $\mathbb{E}[Y_i|X_i]$. Essentially, as X changes, the expected value of Y changes.

Key Definition: Simple Linear Regression

Take a set of observed data with n number of pairs of (D_i, Y_i) observations. The linear model takes the following form:

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- Where $\mathbb{E}[Y_i|X_i]$ is the expected value of Y, given some value of X for observation i.
- Where X_i is the observed value of X for observation i.
- Where the coefficients of the model are β_0 (the intercept) and β_1 (the slope). These are unknown quantities - we need to estimate these by finding what intercept and slope make the best best-fit line.

We can also write the linear model for the value of any point Y_i in our data, which is more common:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Where Y_i is the observed value of Y for observation i.
- Where X_i is the observed value of X for observation i.
- Where coefficients β_0 and β_1 need to be estimated.
- Where ϵ_i is the error term function - that determines the error for each unit i. This error term essentially reflects the "distribution" that we discussed earlier.

As I noted before, we need to estimate intercept β_0 and slope β_1 before we can use our model. This will be covered in the next section.

Once we have the estimates of β_0 and β_1, what can we do with this model?

1. Causal inference and estimation: given a few (important) conditions that have to be met, β_1 can be interpreted as the causal effect of X on Y (or the average treatment effect). We can also use the regression model to estimate the average treatment effect of binary D treatment variables, and randomised controlled trials.
2. Prediction: Once we have the intercept and slope, we can plug in X values and get a prediction of what the Y value is likely to be. This book does not focus too much on prediction, because most real-world things are not linear-straight-line relationships, so the linear regression model may not make the most accurate predictions.

Simple Linear Regression is also a very important building block for all advanced causal inference and predictive techniques, many that we will introduce later in this book and series.

3.3 Simple Linear Regression Estimation

How do we estimate our parameters $\alpha, \beta_1, ..., \beta_k$? Obviously, we want our model to be accurate - so we can estimate it through reducing the errors of models (more specifically, the sum of squared errors).

Key Definition: Sum of Squared Errors

The **sum of squared errors** (SSE) is as follows:

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_{1i} - ... - \hat{\beta}_k X_{ki})$$

Or more intuitively, it is exactly as it sounds - find the error of our prediction $Y_i - \hat{Y}$ for every observation i, square that difference, then sum up for all units i in our data.

The **Ordinary Least Squares (OLS) Estimator** estimates our parameters $\alpha, \beta_1, ..., \beta_k$ through finding the values of $\alpha, \beta_1, ..., \beta_k$ that minimizes the sum of squared errors for our predictions.

A bivariate regression is a regression model with one explanatory variable X, and a fitted simple linear regression takes the form $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$.

Let us define the SSE as function S. The OLS estimator wants to find the parameters that **minimise SSE**:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 = \arg\min_{\hat{\alpha}, \hat{\beta}} S(\hat{\alpha}, \hat{\beta})$$

First Order Conditions

Let us first look at the parameter $\hat{\alpha}$. How do we find what value $\hat{\alpha}$ minimises the sum of squared errors? We know through calculus, that deriving the function to find the first order condition can accomplish this:

$$\frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = \frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} \left[\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \right]$$

First, ignore the summation. The partial derivative of the internal section, using chain rule, is the following:

$$\frac{\partial}{\partial \hat{\alpha}} \left[(Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \right] = -2(Y_i - \hat{\alpha} - \hat{\beta} X_i)$$

But how do we deal with the summation? We know that there is the sum rule of derivatives $[f(x) + g(x)]' = f'(x) + g'(x)$. Thus, we know we just sum up the derivatives to get the derivative:

$$\frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = \sum_{-i=1}^{n} \left[-2(Y_i - \hat{\alpha} - \hat{\beta} X_i) \right] = -2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)$$

Let us do the same for $\hat{\beta}$: find the derivative:

$$\frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = \sum_{i=1}^{n} \left[-2X_i(Y_i - \hat{\alpha} - \hat{\beta} X_i) \right] = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i)$$

We set the two derivatives equal to 0 (first order conditions) We can ignore the -2 out front.

Key Definition: First Order Conditions of OLS

Thus, our first order conditions are a system of equations:

$$\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

Now we have to solve this system of 2 equations. Let us first solve for $\hat{\alpha}$ in terms of $\hat{\beta}$:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = \sum_{i=1}^{n} Y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} X_i = 0 \\ -n\hat{\alpha} = -\sum_{i=1}^{n} Y_i + \hat{\beta}\sum_{i=1}^{n} X_i \\ \hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\hat{\beta}\sum_{i=1}^{n} X_i = \bar{Y} - \hat{\beta}\bar{X}\end{aligned}$$

Now, let us substitute our calculated $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ into the $\hat{\beta}$ condition and solve: for $\hat{\beta}$:

$$\begin{aligned}0 = \sum_{i=1}^{n}\left[X_i(Y_i - [\bar{Y} - \hat{\beta}\bar{X}] - \hat{\beta}X_i)\right] = \sum_{i=1}^{n}\left[X_i(Y_i - \bar{Y} - \hat{\beta}(X_i - \bar{X}))\right] \\ = \sum_{i=1}^{n}\left[X_i(Y_i - \bar{Y}) - X_i\hat{\beta}(X_i - \bar{X})\right] = \sum_{i=1}^{n} X_i(Y_i - \bar{Y}) - \hat{\beta}_1\sum_{i=1}^{n} X_i(X_i - \bar{X})\end{aligned}$$

Here are a few properties on summation that will help us solve this equation:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \\ \sum_{i=1}^{n} X_i(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ \sum_{i=1}^{n} X_i(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

With these rules, we can transform what we had before into:

$$0 = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) - \hat{\beta}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Key Definition

Then solve for $\hat{\beta}$ to get the **OLS estimator** for bivariate regression:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_{XY}}{\sigma_x^2}$$

When $\hat{\beta}$ is the coefficient of our treatment variable D, then the OLS estimator produces our estimate of the average treatment effect (we will discuss if this is a good estimate later)

3.4 Interpretations of Coefficients

We have now estimated our coefficients with the OLS estimator. How do we interpret parameters $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$? What do these estimates mean?

I will define $\hat{\beta}_j$ as any one of $\hat{\beta}_1, ..., \hat{\beta}_k$, multiplied to explanatory variable X_j. $\hat{\beta}_j$ is the rate of change between X_{ij} and $\hat{Y}_i$, holding other explanatory variables $X_{1i}...X_{ki}$ constant. This is the case with any $\hat{\beta}_j$ parameter $j = 1, ..., k$.

Interpretation of $\hat{\beta}_j$

When X_j increases by one unit, there is an expected $\hat{\beta}_j$ unit change in Y, holding all other explanatory variables constant.

If coefficient $\hat{\beta}_j$ is multiplied to our treatment variable of interest D, then that means $\hat{\beta}_j$ is the estimate of the average treatment effect of increasing D by one.

- Assuming D is continuous. Iff not the case, i.e. only control and treatment categories, see the binary explanatory variable section.
- NOTE: we have not discussed yet if this estimate is good or not. See the following chapter for more information on this.

What does intercept $\hat{\alpha}$ mean? Let us take a regression equation, and input $\overrightarrow{X} = 0$, which turns all $\hat{\beta}_j$ and X_j into 0s. Thus, we are left with $\hat{\alpha} = \hat{Y}_i$ when $\overrightarrow{X} = 0$.

Interpretation of $\hat{\alpha}$

When all explanatory variables equal 0, the expected value of Y is $\hat{\alpha}$

Interpreting in Terms of Standard Deviation

Sometimes, it is hard to understand what changes in Y and X mean in terms of units.

- For example, if we are measuring "democracy", what does a 5 unit change in democracy mean? Is that a lot?

We can add more relevant detail by expressing the change of Y and X in standard deviations.

How do we calculate this? Well, let us solve for the change in $\hat{Y}$ given $X = x$ and $X = x + \sigma_X$. This will tell us how much $\hat{Y}$ changes by given a increase of one standard deviation in X.

$$\begin{aligned}\mathbb{E}[\hat{Y}_i|X = x + \sigma_X] - \mathbb{E}[\hat{Y}_i|X = x] &= [\hat{\alpha} + \hat{\beta}(x + \sigma_X)] - [\hat{\alpha} + \hat{\beta}(x)] \\ &= \hat{\alpha} + \hat{\beta}x + \hat{\beta}\sigma_X - \hat{\alpha} - \hat{\beta}x \\ &= \hat{\beta}\sigma_X\end{aligned}$$

To get the change in $\hat{Y}$ in terms of standard deviations of Y, we just divide $\hat{\beta}\sigma_X$ by σ_Y.

Interpretation in Terms of Standard Deviation

For a one-std. deviation increase in X_j, there is an expected $\hat{\beta}\sigma_X/\sigma_Y$-std. deviation change in Y.

3.5 Categorical Explanatory Variable Interpretation

Binary Explanatory Variables

Binary explanatory variables are variables with 2 values, 0 and 1.

- These are extremely common in econometrics - as our treatment variable D often takes two states: $D = 1$ is the treatment group, and $D = 0$ is the control group. We know that in a regression, D is included as an explanatory variable X.

Binary explanatory variables will change the interpretations of our coefficients. We can "solve" for these interpretations given the standard linear model $E[Y] = \alpha + \beta X$, given X has two categories $X = 0, X = 1$:

$$E[Y|X = 0] = \alpha + \beta(0) = \alpha$$
$$E[Y|X = 1] = \alpha + \beta(1) = \alpha + \beta$$
$$E[Y|X = 1] - E[Y|X = 0] = (\alpha + \beta) - \alpha = \beta$$

Thus, we can interpret the coefficients α and β as follows:

Intepretation of Binary Explanatory Variables

When X is a binary explanatory variable:

- α is the expected value of Y given an observation in category $X = 0$
- $\alpha + \beta$ is the expected value of Y given an observation in category $X = 1$
- β is the expected difference in Y between the categories $X = 1$ and $X = 0$

We can see that β is measuring the difference between the two categories.

- In fact, β actually becomes a **difference-in-means** test, meaning that if β is statistically significant, we can conclude a significant difference in the mean Y between the two categories.

Remember that our ATE in our randomised experiment was estimated with a difference in means. Thus, if we include all possible confounding variables, the β coefficient of D will estimate the ATE.

Categorical Explanatory Variable

A **categorical polytomous** variable is one with 3 or more categories that are unranked. A classic example is the variable *country*, which is a categorical variable with all the different countries included in a dataset such as Argentina, France, Mexico, etc.

How do we run a regression with polytomous explanatory variables? What happens is that we divide the variables into a set of dummy binary variables (see how dummy variables are interpreted in section 3.2).

- Dummy binary variables are created for all except one of the categories in our variable. Each dummy variable has two values - 1 meaning the observation is in the category, and 0 meaning the observation is not in that category.
- The category without a dummy variable is the **reference/baseline** category. Essentially, when all other dummy variables are equal to 0, that is referring to the reference/baseline category (the intercept)

Key Definition

Thus, a **polytomous explanatory variable** with n number of categories in X, we would create $n-1$ dummy variables, and input it into a regression equation as follows:

$$E[Y] = \alpha + \beta_{x=1}X_{x=1} + ... + \beta_{x=n-1}X_{x=n-1}$$

Where α is the mean of the reference category n, and the other categories $1, ..., n-1$ get their own dummy variable.

For example, take the following polytomous variable: *company*, which contains the categories *microsoft, google,* and *apple*. Let us create dummy variables for 2 of the 3 categories:

- *Google* will become the first dummy variable X_g. When $X_g = 1$, that observation is part of the *google* category. When $X_g = 0$, that observation is NOT a part of the *google* category.
- *Apple* will become the second dummy variable X_a. When $X_a = 1$, that observation is part of the *apple* category. When $X_a = 0$, that observation is NOT a part of the *apple* category.
- *Microsoft* will not get its own dummy variable. This is because when both *apple* and *microsoft* $X_g = X_a = 0$ that is referring to the *microsoft* category (observations not a part of either previous category).

Mathematically, this is how it would be represented in a regression equation:

$$E[Y] = \alpha + \beta_g X_g + \beta_a X_a$$

To find the expected value of each category, we would do the following:

$$\begin{aligned} E[Y|X = \text{Google}] &= E[Y|X_g = 1, X_a = 0] = \alpha + \beta_g(1) + \beta_a(0) = \alpha + \beta_g \\ E[Y|X = \text{Apple}] &= E[Y|X_g = 0, X_a = 1] = \alpha + \beta_g(0) + \beta_a(1) = \alpha + \beta_a \\ E[Y|X = \text{Microsoft}] &= E[Y|X_g = 0, X_a = 0] = \alpha + \beta_g(0) + \beta_a(0) = \alpha \end{aligned}$$

Thus, from these above equations, we can see the interpretation of the coefficients:

- α is the expected value of the reference category, in this case, *microsoft*.
- β_g is the expected Y difference between the *google* category and the reference category *microsoft*. The statistical significance of this coefficient would be a difference of means test between the two categories.
- β_a is the expected Y difference between the *apple* category and the reference category *microsoft*. The statistical significance of this coefficient would be a difference of means test between the two categories.

Interpretation of Polytomous Explanatory Variables

β_j is the expected difference in Y values between category j and the baseline category.
α is the expected value of Y of the baseline category.

The coefficient p-values of β_j are a difference-of-means test between two categories, and not a statistical significance test of the entire categorical variable.

3.6 Hypothesis Testing and Causal Inference

Robust Standard Errors

So far, we have focused on using standard errors of parameter estimates that were introduced in section 2.2.

However, these "default" standard errors rely on the assumption of **homoscedasticity** (see section 3.4). However, homoscedasticity is frequently violated.

When homoscedasticity is violated, we have to use an alternative estimation of the standard error: **heteroscedasticity-robust standard errors**.

- These robust standard errors are typically larger than the homoscedastic "default" standard errors.
- Because homoscedasticity is frequently violated, and that robust standard errors are more "conservative", we typically use robust standard errors as the "default" in econometrics.

The calculation of robust-standard errors will be done with calculators, and it is not important to know how the mathematics work.

Hypothesis Testing of Parameters

We previously discussed hypothesis testing in section 2.5. The mechanics are practically the same, but we replace the sample average treatment effect with $\hat{\beta}_j$ as our sample estimate. Generally, for regressions, our hypotheses that we test are:

- $H_0 : \beta_j = 0$ - i.e. there is no relationship between X_j and Y
- $H_1 : \beta_j \neq 0$ - i.e. there is a relationship between X_j and Y

Just like previously discussed in section 2.5, we calculate a t-statistic:

$$t = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$$

The t-test statistic tells us how far the estimate is from 0, in terms of standard errors of the estimate.

Then, we have to consult a t-distribution (see section 2.5). We find the probability (area under the distribution) of a t-test statistic of ours, or more extreme, could occur. This is the p-value: how likely we are to get a test statistic at or more extreme than the one we got for our estimated β_j, given the null hypothesis is true.

- So if the p-value is very high, there is a high chance that the null hypothesis is true.
- If the p-value is very low, then there is a low chance that the null hypothesis is true

Generally, in the social sciences, if the p-value is less that 0.05 (5%), we can **reject the null hypothesis**, and conclude the alternate hypothesis.

Chapter 4

Multiple Linear Regression and Causal Inference

4.1 The Multiple Linear Regression Model

In the social sciences, randomisation is often not possible. Linear regression allows us to estimate a model with both our treatment and outcome variables, as well as a series of **control** variables.

In theory, by including every single confounding variable as a control variable in our regression model, we can partial out the effect of confounders and find the average treatment effect.

The **response variable** (outcome variable) is notated Y. The **explanatory variable** (independent variable) is notated X. There is often more than one explanatory variable, so we denote them with subscripts $X_1, X_2, ..., X_k$. We sometimes also denote all explanatory variables as the vector $\overrightarrow{X}$.

- Our treatment variable D is considered one of the explanatory variables $\overrightarrow{X}$.

A linear regression model is the specification of the conditional distribution of Y, given $\overrightarrow{X}$. The linear regression model focuses on the **expected value** of the conditional distribution, notated $\mathbb{E}[Y_i|\overrightarrow{X_i}]$. Essentially, as $\overrightarrow{X}$ changes, the expected value of Y changes.

- I say **distribution** because there are often a range of Y outcomes, each with their own probabilities, for any given X. For example, if X was age and Y was income, at age $X = 30$, not every single 30 year old makes the same amount of money. There is some distribution of incomes Y at age $X = 30$.

Key Definition: Linear Regression Model

Take a set of observed data with n number of pairs of $(\overrightarrow{X_i}, Y_i)$ observations. The linear model takes the following form:

$$\mathbb{E}[Y_i|\overrightarrow{X_i}] = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

- Where the distribution of $Y_i|\overrightarrow{X_i}$ has a variance $Var(Y_i|\overrightarrow{X_i}) = \sigma^2$.
- Where the coefficients of the model (that need to be estimated) are $\alpha, \beta_1, ..., \beta_k$.

We can also write the linear model for the value of any point Y_i in our data:

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$$

- Where ϵ_i is the error term function - that determines the error for each unit i.

4.2 Multiple Regression Estimation

Multiple regression, as introduced previously, allows us to add additional control variables. Similar to our bivariate regression (but with additional variables), our minimisation condition is:

$$(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ...) = \arg \min_{(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ...)} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} ...)^2 = \arg \min_{(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ...)} S(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ...)$$

Taking the partial derivatives of each parameter as before, we get these first order conditions::

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} ...) = 0$$
$$-2 \sum_{i=1}^{n} X_{1i} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} ...) = 0$$
$$-2 \sum_{i=1}^{n} X_{2i} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} ...) = 0$$
$$\text{and so on for } X_{3i}, ..., X_{ki}$$

This system of equations is way too difficult to solve. Instead, we can use linear algebra. We start with the linear model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

The ith observation can be written in vector form as following:

$$Y = X_i' \beta + \epsilon_i, \text{ where } \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } X_i = \begin{bmatrix} 1 \\ X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix}$$

- The X_i' in the equation is the transpose of X_i, to make matrix multiplication possible.
- The first element of the X_i matrix is 1, since $1 \times \alpha$ gives us the first parameter in the linear model.

Since our model has n different observations of i, we can express this into vector form, with the X_i' and β being vectors within a vector.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1'\beta + \epsilon_1 \\ X_2'\beta + \epsilon_2 \\ \vdots \\ X_n'\beta + \epsilon_n \end{pmatrix} = \begin{pmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ X_n'\beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Since β vector appears as a common factor for all observations $i = 1, ..., n$, we can factor it out and have an equation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

We can expand the $X_1', ..., X_n'$ vector into a matrix. Remember that each $X_1', ..., X_n'$ is already a vector of different explanatory variables. Thus, we have a model in the form:

$$Y = X\beta + \epsilon, \text{ where } X = \begin{bmatrix} 1 & x_{21} & ... & x_{k1} \\ 1 & x_{22} & ... & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & ... & x_{kn} \end{bmatrix}$$

- Where the notation for elements of X is x_{ki}, with i being the unit of observation $i = 1, ... n$, and k being the explanatory variables index.
- Where Y and ϵ are $n \times 1$ vectors (as seen above), and β is a $k \times 1$ vector.
- The first row of X is a vector of 1, which exists because these 1's are multiplied with α in our model.

Now, let us estimate our coefficients. Let us define our estimation vector $\hat{\beta}$ as:

$$\hat{\beta} = \arg\min_b (Y - Xb)'(Y - Xb) = \arg\min_b S(b)$$

We can expand $S(b)$ as follows:

$$S(b) = Y'Y - b'X'Y - Y'Xb + b'X'Xb \;=\; Y'Y - 2b'X'Y + b'X'Xb$$

Taking the partial derivative in respect to b, then setting equal to 0, we get:

$$\frac{\partial S(b)}{\partial b}\bigg|_{\hat{\beta}} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix}\bigg|_{\hat{\beta}} = 0$$

Differentiating with the vector b yields:

$$\frac{\partial S(b)}{\partial b} = -2X'Y + 2X'Xb$$

Evaluted at $\hat{\beta}$, the derivatives should equal zero (since first order condition of finding minimums):

$$\frac{\partial S(b)}{\partial b}\bigg|_{\hat{\beta}} = -2X'Y + DX'X\hat{\beta} = 0$$

Key Definition

When assuming $X'X$ is invertable, our OLS estimator solution for $\hat{\beta}$ is:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Once we have estimates of $\hat{\beta}$, we can plug them into our linear model to obtain fitted values:

$$\hat{Y} = X\beta = X(X'X)^{-1}X'Y$$

4.3 Linear Probability Model

If our response variable is binary, (i.e. Y only has two values, 0 and 1), then our interpretation differs slightly.

We treat Y as a variable of two categories, $Y = 0$ and $Y = 1$. The output $\hat{Y}_i$ indicates the probability of a specific observation i of being in the $Y = 1$ category - we can also interpret probability in percentages by multiplying by 100.

Interpretation with Binary Y Variable

For a one-unit increase in X_j, there is an expected $\hat{\beta}_j \times 100$ percentage point change in the probability of being in category $Y = 1$.

$\hat{\alpha}$ is the expected probability of being in category $Y = 1$ when all explanatory variables equal 0.

This model with binary Y is also called the **linear probability model**.

Chapter 5

Causal Estimation with Regression and OLS

5.1 Gauss Markov Theorem

5.2 Endogeniety

5.3 Heteroscedasticity

5.4 Omitted Variable Bias

5.5 Model Selection for Causal Inference