Econometrics for Political Analysis

Part of an Introduction to Political Economics

Kevin Lingfeng Li

Table of contents

Part I

Econometrics and Causal Inference

Chapter 1

Causal Inference

1.1 Introduction to Econometrics

Econometrics is the field of applying statistical methods to analyse real-world economic and social science data. While econometrics was initially pioneered by economics, the techniques econometricians developed have been adopted by most of the social sciences, including Political Science. 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?

This book is designed to be both an approachable, but also rigorous, introduction to Econometrics and the use of statistical methods for the analysis of political institutions and actors. This volume specifically covers most topics you would study in a typical econometrics undergraduate course in an economics department, and is likely more in depth than what you would get in a political science department.

- This books is the 2nd in the *An Introduction to Political Economics* series, which focuses on using economic methods, such as microeconomic modelling and econometrics, to study political phenomena. This book is designed to complement the other books in this series.

I assume a solid understanding of basic statistics, including the topics of random variables and distributions, expectation/mean and variance, and correlation. I also assume familiarity with 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 first volume and some topics in the 2nd volume).

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

If you have completed this book, and want more econometrics training, the 4th book in the *An Introduction to Political Economics* series, *Advanced Econometrics for Political Analysis*, builds on this book by adding further statistical and econometric models that help with more advanced topics.

1.2 Potential Outcomes Framework

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}.

1.3 Estimands and Estimators

Causal Estimands

The fundamental problem of causal inference is that we only can observe one of 2 potential outcomes. For example, if you get treatment D, we cannot observe the world where you do not get treatment D. Thus, we cannot estimate individual effects of the causal treatment. However, there are other estimands we can use.

- Note: **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.

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

Estimators, 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, the estimator's average 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.
- Or more intuitively, imagine you are trying to hit a bullseye in archery. Bias is when you might be very accurate, but aiming in the wrong place, thus not hitting the bullseye. A biased estimator is essentially that - we are consistently and systematically making a mistake when estimating the quantity in question.

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.
- Or more intuitively, in the archery example, we are aiming correctly at the bullseye, however, the wind and our muscles are unpredictable, so each shot might be slightly off in different directions. If we average all our shots, we are hitting the middle, but not each individual shot is in the bullseye.

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 and Selection Bias

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

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

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.

However, this naive estimator is a bad idea. Why?

- 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}).

Selection Bias and Confounding Variables

These differences between the treatment and control groups prior to treatment can actually explain some of our results, 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. For example, smoking might be a confounding variable in the example above - people who smoke more often will go to the hospital, and will have worse outcomes than people who did not smoke and did not go to the hospital. Confounders are often the cause of selection bias.

The naive estimator will capture the effect of these confounders, which we do not want - we want to isolate the effect of our treatment D. To make an accurate causal claim, we must get rid of confounding variables and selection bias. How do we do this? - Randomisation, which we will cover next.

Chapter 2

Randomised Controlled Trials

2.1 Random Assignment and Estimation

The **assignment mechanism** is how we decide which observations receive the treatment D. In the last chapter, we discussed how the Naive Estimator is biased, because of selection bias. We can address selection bias by randomly assigning units into either the treatment or control group.

Key Definition: Random Assignment

With random assignment, each observation has an equal likelihood of being assigned to treatment or control, we should expected the treatment and control groups to be similar - thus eliminating selection bias. As a result, the potential outcomes are independent of treatment/control status:

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

If this assumption is met, then we can use the naive estimator to estimate the treatment effect:

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

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. Thus we know that:

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

Now that we have $\mathbb{E}[Y_{1i}]$ and $\mathbb{E}[Y_{0i}]$, we can calculate the average treatment effect:

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

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 of means between the treatment and control group.

Blocking and Stratified Experiments

Blocking, also called stratified experiments, is an extension of the random experiment to deal with some common issues.

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.
- This is a major issue! After all, the core assumption of random experiments is that randomisation makes the treatment and control group similar, eliminating selection bias.

With blocking, you can prevent this from happening.

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. Mathematically:

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

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.

2.2 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 $\widehat{ATE}$.

- 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.

Now, 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. We will get a distribution, just like a random variable distribution. That distribution is the **sampling distribution**

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 this for us.

2.3 Confidence Intervals and Hypothesis Testing

Confidence Intervals

Since there is variability of estimates between samples, we have to create an interval around our sample estimate $\hat{\theta}_j$ to account for this uncertainty. We assume our estimated $\hat{\theta}_j$ is the centre of this distribution, then add some "buffer" to both sides.

Key Definition: Confidence Intervals

A **confidence interval**'s lower and upper bounds are defined as, given a confidence level of 95% (the standard confidence level):

$$\hat{\theta}_j \pm 1.96 \times \hat{se}(\hat{\theta}_j)$$

Where $\hat{se}(\hat{\theta}_j)$ is the standard error of our estimate of how precisely we have estimated the true value of θ_j, introduced in the previous section.

Confidence intervals say that if we repeated the sampling and estimation process many many times (like we did for our sampling distribution), 95% of the confidence intervals we construct from our samples, would correctly contain the true θ_j.

Why 1.96? It is because in a normal distribution, 95% of the data is contained within 1.96 standard deviations, and Central Limit Theorem states that sampling distributions are normally distributed.

Every value in a given confidence interval is a plausible value of the true θ_j in the population. The most important thing is if 0 is included within the confidence interval. $\theta = 0$ means that there is no causal effect.

Hypothesis Testing

In academia, we do not claim we have found a new theory, unless we are quite confident that the old theory was not true. The old theory is called our **null hypothesis**, often notated H_0. This is the old theory that we are trying to disprove.

- Most often, the "old theory" we are trying to disprove is that *there is no causal effect of* D on Y (since it is rare to study something that has already been proven). No relationship means that $\theta_j = 0$

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.

- In general, our new hypothesis is that t*here is a causal effect of* D on Y, or $\underline{\theta_j \neq 0}$

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:

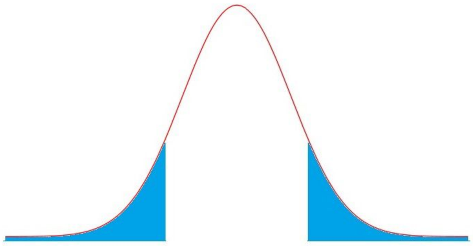
Key Definition: T-test Statistic

The t-test statistic basically tells us how far the parameter we tested is from 0, in terms of standard errors of the parameter:

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

Then, we calculate degrees of freedom:number of observations n, minus the number of variables k, then minus 1: $DF = n-k-1$. With the degrees of freedom, we can find the corresponding t-distribution, labeled t_{n-k-1}. Then, we start from the middle of that t distribution, and go the *number of standard errors* away based on the t-test 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, with its blue highlighted area, 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.

- 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.

2.4 Validity and Limitations of Randomised Experiments

There are two types of validity in Randomised Experiments: Internal and External validity.

Internal validity is about if our experiment accurately captures the average causal effect of our units in our experiment. Essentially - did we estimate the causal effect correctly? Some things that can cause lack of internal validity include:

1. Failure of randomisation: if treatment and control groups are not similar, we violate assumptions of random experiments and that will include selection bias in our estimates.
2. Non-Compliance: Sometimes, our subjects that are assigned to treatment, refuse to comply with the treatment (we cannot force them to). This will mess up the average treatment effect, since some units did not properly undergo the treatment.
3. Attrition: Sometimes, outcomes cannot be measured for some study participants, for example, if they drop out or refuse to answer. This is concerning - because the people who drop out might have some common characteristic (confounding variable), and we will miss this entirely in our estimation.

External Validity is about the generalisation of our conclusions - we know the effect on our experimental subjects, but does this causal effect apply to other units across the world?

- For example, if we do a study in Japan, can we assume that the same effects are applicable in the US? South Africa?

- Generally to obtain this, you want the units included in your observation to be representative of the larger units you want to apply your results to. For example, if you are measuring the causal effect of some treatment on Americans, you want your subjects to be representative of Americans as a whole.

Finally, Randomised Experiments have some **limitations**.

1. Ethical limitations: sometimes, it is unethical to have units take potentially dangerous treatments, or have some units not undergo potential benefits of treatment. We are essentially randomly selecting what happens to people's lives.
2. Practical limitations: often, running experiments is just not possible. For example, let us say you want to see if democracy increases economic growth. To do this, you would need to randomly assign countries to democracy or autocracy (control) groups. But let us be honest, you can't force Canada to be a dictatorship against their will. Often, we will have to use **observational studies** - where we do not control assignment of treatment.

The rest of this course focuses on observational studies. We will introduce regression - one of the most powerful tools in causal inference. Then, we will introduce extensions to regressions that help us address some of the limitations regressions have.

Part II

Multiple Linear Regression

Chapter 3

Regression Models and Estimation

3.1 Linear Regression Model

Introduction

As we discussed in the previous sections, a randomised experiment is the best way of establishing causal relationships. This is because randomise treatments can get rid of the effect of confounding variables. However, 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. By including confounding variables as control variables in our regression model, we can (in theory), isolate the effect of our explanatory variable on our outcome variable.

- In reality, as we will discuss in the later parts of this book, there are often thousands of control variables, many that are not possible to control for. We will introduce further techniques to deal with these.

Before we dive into the linear model, here are some conventional notation that is important:

- The **response variable** (dependent variable) is notated Y. In this book, we will only have one response variable.
- 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 $\overrightarrow{X}$
- Note: our treatment variable (for causal inference) D is considered one of the explanatory variables $\overrightarrow{X}$. We typically define the first of these explanatory variables X_1 as the treatment variable, and all others $X_2, ..., X_k$ as control variables.

Specification of the Linear Model

A 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 of Y given $\overrightarrow{X}$.

- 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 $X = 30$.

The regression model focuses on the expected value of Y given some X value.

💡 Key Definition: Linear Regression Model

Take a set of observed data, with response variable Y, and a number of X variables for n number of observations. Thus, we will have n number of pairs of (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 $\mathbb{E}[Y_i|\overrightarrow{X}_i]$ is the expected value of the conditional distribution $Y_i|\overrightarrow{X}_i$.
- Where the distribution of $Y_i|\overrightarrow{X}_i$ has a variance $Var(Y_i|\overrightarrow{X}_i) = \sigma^2$.
- Where the parameters of the model are the denoted by the vector $\overrightarrow{\beta}$, and contain $\alpha, \beta_1, ..., \beta_k$.

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.
Essentially, ϵ_i is another way to think about the conditional distribution of $Y_i|\overrightarrow{X}_i$, and how not every 30 year old makes the exact same income - there is some variation (and error).

In our model, we have parameters $\alpha, \beta_1, ..., \beta_k$ that need to be estimated (based on our data) in order to create a best-fit line we can actually use. We estimate the parameters and fit the model by using our observed data points $(Y_i, \overrightarrow{X}_i)$, and fitting a best fit line to these points. Our result should take the following form:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki}$$

- Where $\hat{Y}_i$ is our prediction of the value of Y, given any set of $\overrightarrow{X}$ values.
- Notice the error term ϵ_i is not present. This is the expected value of ϵ_i is $\mathbb{E}[\epsilon_i] = 0$.

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

The **sum of squared errors** 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 $Y_i - \hat{Y}$, square that difference, then sum up for all units i in our data.

- Why squared? Well, because we do not care about the direction of our errors (positive or negative), just the size of them. Thus, squaring removes the negative signs so we are only concerned with magnitude.
- Then why not absolute value? There are a few reasons, but generally the primary reason is estimation is much more difficult with absolute value since an absolute value function is not differentiable at its vertex.

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.

3.2 Bivariate Regression Estimation Mechanics

A bivariate regression is a regression model with one explanatory variable X. The **ordinary least squares** (OLS) estimator is concerned with the sum of squared errors. Let us define the sum of squared errors as a function S.

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

The OLS estimator wants to find the parameters that **minimise the sum of squared errors**:

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

Parameter $\hat{\alpha}$

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:

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

To find the value of $\hat{\alpha}$ that creates the minimum value of the SSE, we set the first order derivative equal to 0. We can ignore the -2, since if the sum is equal to 0, then the -2 will have no effect. Now, using properties of summation, isolate $\hat{\alpha}$ as follows:

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

The final step converting to $\bar{Y}$ and $\bar{X}$ is because of the mathematical definition of average. We will plug this $\hat{\alpha}$ equation into our solution for $\hat{\beta}$ to solve that. Once we solve $\hat{\beta}$, we will come back and calculate $\hat{\alpha}$'s value.

Parameter $\hat{\beta}$

Let us find the $\hat{\beta}$ that minimises S by taking the partial derivative of S in respect to $\hat{\beta}$ and setting it equal to 0. This is almost the same as before - use chain rule, then use sum rule to get the deriative:

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

Now, let us plug in our previously solved $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, and we get:

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

Once again, the -2 does not matter as same reason as before. Set equal to 0 to solve for the value of $\hat{\beta}$ that minimises SSE:

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

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:

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

Thus, with this solution, we have estimated $\hat{\beta}$. If we recall, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, so we can quickly solve for that as well. Thus, we now have $\hat{\beta}, \hat{\alpha}$, completing our estimation. We can now put everything into the fitted model:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

The only thing that must be true to estimate bivariate linear regression is that $Var(X) \neq 0$.

3.3 Multiple Regression Estimation Mechanics

Multiple regression, as introduced previously, allows us to add additional control variables:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$$

Similar to our bivariate regression (but with additional variables), our minimisation condition is:

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

However, instead of a best-fit line in a 2-dimensional setting, we now have a best-fit plane in a 3-dimensional space (or higher depending on the number of explanatory variables).

Taking the partial derivatives of each parameter as before, and setting them equal to 0, we get these three first order conditions:

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

Just like before, we can ignore the -2 in the conditions.

The 1st equation, solving for $\hat{\alpha}$, is quite simple, just as in bivariate regression. We get:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 ...$$

The other conditions are difficult to solve by hand, so the computer will do this for us. However, it is still useful to understand what the computer is trying to do - minimise square errors.

The only properties that are required to compute the multivariate regression OLS estimates are:

1. All explanatory variables have sample variability $Var(X_j \in \overrightarrow{X}) \neq 0, \forall X_j \in X$.
2. No two explanatory variable have **perfect collinearity** - which means a correlation coefficient of -1 or 1, which only occurs if the two variables have the same exact values. Mathematically, $X_{ai} \neq X_{ai}, \forall i, \forall X_a, X_b \in \overrightarrow{X}$.

3.4 Multiple Regression Estimation with Linear Algebra

By expressing models in linear algebra, it becomes easier to derive formal results without requiring messy algebra and the summation operator. It is also natural, since our data is stored in spreadsheets (big matrices).

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:

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

- 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.

OLS Estimator with Linear Algebra

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:

$$\begin{aligned} S(b) &= Y'Y - b'X'Y - Y'Xb + b'X'Xb \\ & Y'Y - 2b'X'Y + b'X'Xb \end{aligned}$$

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

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

Differentiating with the vector b yeilds:

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

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

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

Key Definition

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

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

The matrix $X'X$ is invertible for the same criteria as the non-linear algebra solution - that all explanatory variables have non-zero variance, and there is no perfect collinearity.

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

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

Chapter 4

Interpretation and Hypothesis Testing

4.1 Interpretations of Coefficients

When there are multiple explanatory variables $X_1, X_2, ..., X_k$, how do we interpret parameters $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$? I will define $\hat{\beta}_j$ as any one of $\hat{\beta}_1, ..., \hat{\beta}_k$, multiplied to X_j.

First, what does $\hat{\beta}_j$ mean. Consider the regression equation $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki}$.

If we find the partial derivative in respect to X_{ji} of the above equation, we get:

$$\begin{aligned} \frac{\partial \hat{Y}_i}{\partial X_{ji}} &= \frac{\partial \hat{Y}_i}{\partial X_{ji}} \left[\hat{\alpha} + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki} \right] \\ &= 0 + 0 + ... + \hat{\beta}_j + .. + 0 \\ &= \hat{\beta}_j \end{aligned}$$

This shows that $\hat{\beta}_j$ 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 $j = 1, ..., k$.

ℹ 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.

Second, what does intercept $\hat{\alpha}$ mean? Let us take a regression equation, and input $\overrightarrow{X} = 0$:

$$\begin{aligned} \mathbb{E}[\hat{Y}_i | \overrightarrow{X} = 0] &= \hat{\alpha} + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_j X_{ji} + ... + \hat{\beta}_k X_{ki} \\ &= \hat{\alpha} + \hat{\beta}_1(0) + ... + \hat{\beta}_j(0) + ... + \hat{\beta}_k(0) \\ &= \hat{\alpha} \end{aligned}$$

This shows that $\hat{\alpha} = \hat{Y}_i$ when $\overrightarrow{X} = 0$.

ℹ 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.

Binary Y Variable Interpretation

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 Varibale
>
> 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.

4.2 Model Summary Statistics

Estimated Residual Standard Deviation

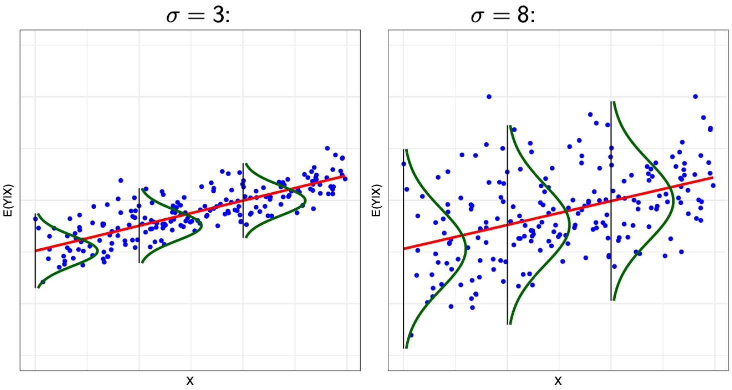
We can derive the estimate of the **residual variance** σ^2 with this formula:

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - k - 1}$$

But what is the residual variance? Recall our regression model: $Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$

We know that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Our estimate of the residual variance $\hat{\sigma}^2$ is our estimate of the variance of the error term ϵ_i's variance. More intuitively, it explains how spread out observed values of Y are from our prediction value $\hat{Y} = E(Y|X)$.

The figure below better showcases this in 2 different models. The red lines are our predicted regression line, and the green lines represent the distribution of our error term ϵ_i:



The residual standard deviation $\hat{\sigma}$ (square root of variance) is consistent throughout a model. This is one of the assumptions of the linear regression model - that errors are consistently distributed, no matter the value of X. This assumption is called **homoscedasticity**.

If $\hat{\sigma}$ varies depending on the value of X, then that is called **heteroscedasticity**. When this occurs, it is often a suggestion that our relationship may not be linear - and we perhaps need to try a few transformations. We will get into transformations in a later chapter.

Total Sum of Squares

The total sum of squares is the total amount of sample variation in Y:

$$\begin{aligned} TSS &= \sum (Y_i - \bar{Y})^2 \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \end{aligned}$$

Where TSS is the total sum of squares, SSM $\sum(\hat{Y}_i - \bar{Y})^2$ is the model sum of squares, and SSE $\sum(Y_i - \hat{Y}_i)^2$ is the sum of squared errors (that we used to fit the model).

SSM (model sum of squares) represents the part of the variation of Y that is explained by the model, while SSE (sum of squared errors) represents the part of the variation of Y that is not explained by the model (hence, why it is called error).

R-Squared Statistic

R-squared is one of the key summary statistics of our model.

Key Definition: R-Squared

R-squared R^2 is a measure of the percentage of variation in Y, that is explained by our model (with our chosen explanatory variables). The percentage of variation in Y explained by our model would be:

$$R^2 = \frac{SSM}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

Since R^2 shows how much of the variation in Y our model explains, it is often used as a metric for how good our model is - however, don't overly focus on R^2, it is just one metric with its benefits and drawbacks.

4.3 Hypothesis Testing

Confidence Intervals

We previously discussed confidence intervals in section 2.4. The mechanics are practically the same, but we replace the sample average treatment effect with $\hat{\beta}_j$ as our sample estimate. To account for sampling variation, we have to create an interval around our estimate $\hat{\beta}_j$ to account for this uncertainty. We assume our estimated $\hat{\beta}_j$ is the centre of this distribution, then add some "buffer" to both sides:

$$\hat{\beta}_j \pm 1.96 \times \hat{se}(\hat{\beta}_j)$$

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.

F-Tests of Nested Models

The **F-test of Nested Models** allows us to compare different regression models. We use a smaller model as our null hypothesis, and a larger model (containing the smaller model) as our alternative hypothesis. More mathematically:

$$\begin{aligned} M_0 :& E[Y] = \alpha + \beta_1 X_1 + ... + \beta_g X_g \\ M_a :& E[Y] = \alpha + \beta_1 X_1 + ... + \beta_g X_g + \beta_{g+1} X_{g+1} + ... + \beta_k X_k \end{aligned}$$

Importantly, all explanatory variables in model M_0 must also be in M_a (hence "nested").

The F-test uses the F-test statistic. This statistic compared the R^2 values of the two models. Let us say the R^2 value of M_0 is notated R^2_0, and the R^2 value of M_a is notated as R^2_a. The F-test statistic essentially measures the difference $R_a^2 - R_0^2$. If the difference is sufficiently large, that means the M_a model has significantly more explanatory power than M_0.

Mathematically, the F-test statistic is as follows, with k_a being the number of explanatory variables in the alternate hypothesis:

$$F = \frac{R^2_{\text{change}}/df_{\text{change}}}{(1 - R^2_a)/[n - (k_a + 1)]}$$

The sampling distribution of the F-statistic is the F distribution with parameters $k - a - k_0$ and $n - (k_a + 1)$ degrees of freedom. We then obtain the p-value from this distribution. The p-values of the F-statistic show the following:

- If the p-value is very small, that means R_a^2 is significantly larger than R_0^2. This is evidence against model M_0, and in favour of the larger model M_a
- If the p-value is large, that means R_a^2 is not much larger than R_0^2. This means there is no evidence against M_0, and M_a is not the statistically significantly better model.

F-tests of nested models can help us determine if we should include certain extra explanatory variables. If our model with more variables is statistically significant, it is an indication that we should include those extra variables.

4.4 Regression in R

We use the *lm()* function to run a regression: The general syntax is as follows:

- Replace *model_name* with your model name, *Y* with the name of your response variable, *X1, X2...* with the name of your explanatory variable, and *mydata* with the name of your dataset.
- Add additional explanatory variables with more + signs, and you can remove them down to a minimum of one *X*

```
model_name <- lm(Y ~ X1 + X2 + X3, data = mydata)
summary(model_name)
```

Confidence Intervals in R

To calculate confidence intervals, we can use the *confint()* command, and simply input the name of our model within:

```
confint(model)
```

F-Test of Nested Models in R

If we want an F-test between two models, we can use the *anova()* function, replacing *model1* with the name of the null hypothesis model M_0, and replacing *model2* with the name of the alternative hypothesis model M_a.

```
anova(model1, model2)
```

Chapter 5

Categorical Explanatory Variables

5.1 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 $X = 0, X = 1$:

$$\begin{aligned} E[Y|X=0] &= \alpha + \beta(0) = \alpha \\ E[Y|X=1] &= \alpha + \beta(1) = \alpha + \beta \\ E[Y|X=1] - E[Y|X=0] &= (\alpha + \beta) - \alpha = \beta \end{aligned}$$

ℹ Interpretation of Binary Explanatory Variables

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 $X = 0$

Thus, 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.

5.2 OLS and Difference-in-Means Estimator

Because of the unique coefficient meanings in a regression with binary explanatory variables, the OLS estimator can take a shortcut when estimated the coefficients α and β:

- Coefficient α is estimated as $\hat{\alpha} = \bar{Y}_0$
- Coefficient β is estimated as $\hat{\beta} = \bar{Y}_1 - \bar{Y}_0$

Where $\bar{Y}_1$ is the sample mean of Y for observations in category $X = 1$, and $\bar{Y}_0$ is the sample mean of Y for observations in category $X = 0$. So essentially, OLS is using a **difference of means** estimator.

Proof:

Let us prove that the bivariate OLS estimator with binary X is equivalent to a difference in means. In Chapter 3, we proved the OLS estimator produces the following $\widehat{\beta}$ solution:

$$\widehat{\beta} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Let us focus on the numerator of the $\widehat{\beta}$ OLS solution. Let us expand out the numerator as follows:

$$\begin{aligned}
\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{N}[X_iY_i - X_i\bar{Y} - X_i\bar{Y} + \bar{X}\bar{Y}] \\
&= \sum_{i=1}^{N} X_iY_i - \sum_{i=1}^{N} X_i\bar{Y} - \sum_{i=1}^{N} \bar{X}Y_i + \sum_{i=1}^{N} \bar{X}\bar{Y}
\end{aligned}$$

We know that X is a binary variable with $X = 1$ being the treatment state t, and $X = 0$ being the control state c.

Note, N indicates the number of observations, and N_t indicates number of observations in treatment group. We can now actually evaluate the 4 summations we have found from expanding, and simplify further:

$$\begin{aligned}
\sum_{i=1}^{N} X_iY_i - \sum_{i=1}^{N} X_i\bar{Y} - \sum_{i=1}^{N} \bar{X}Y_i + \sum_{i=1}^{N} \bar{X}\bar{Y} &= N_t\bar{Y}_t - N_t\bar{Y} - N\left(\frac{N_t}{N}\right)\bar{Y} + N\left(\frac{N_t}{N}\right)\bar{Y} \\
&= N_t\bar{Y}_t - N_t\bar{Y} - N_t\bar{Y} + N_t\bar{Y} \\
&= N_t\bar{Y}_t - N_t\bar{Y} \\
&= N_t(\bar{Y}_t - \bar{Y})
\end{aligned}$$

A weighted average formula means $\bar{Y} = \frac{1}{N}(N_t\bar{Y}_t + N_c\bar{Y}_c)$. We can substitute $\bar{Y}$ in our numerator with that as follows:

$$\begin{aligned}
N_t(\bar{Y}_t - \bar{Y}) &= N_t\left(Y_t - \frac{1}{N}(N_t\bar{Y}_t + N_c\bar{Y}_c)\right) \\
&= N_t\bar{Y}_t - \frac{N_t}{N}(N_t\bar{Y}_t + N_c\bar{Y}_c) \\
&= N_t\bar{Y}_t - \frac{N_t^2\bar{Y}_t}{N} - \frac{N_tN_c\bar{Y}_c}{N} \\
&= \bar{Y}_t\left(N_t - \frac{N_t^2}{N}\right) - \bar{Y}_c\left(\frac{N_tN_c}{N}\right)
\end{aligned}$$

We know that $N_c = N - N_t$, thus $\frac{N_tN_c}{N} = \frac{N_t(N-N_t)}{N} = N_t - \frac{N_t^2}{N}$. Plugging that in to the numerator, and simplifying, we get:

$$\begin{aligned}
\bar{Y}_t\left(N_t - \frac{N_t^2}{N}\right) - \bar{Y}_c\left(\frac{N_tN_c}{N}\right) &= \bar{Y}_t\left(N_t - \frac{N_t^2}{N}\right) - \bar{Y}_c\left(N_t - \frac{N_t^2}{N}\right) \\
&= \left(N_t - \frac{N_t^2}{N}\right)(\bar{Y}_t - \bar{Y}_c)
\end{aligned}$$

Thus, our $\hat{\beta}$ now looks like (putting in the changes we did to the numerator):

$$\hat{\beta} = \frac{\left(N_t - \frac{N_t^2}{N}\right)(\bar{Y}_t - \bar{Y}_c)}{\sum\limits_{i=1}^{N}(X_i - \bar{X})^2}$$

Now, let us expand the denominator:

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N} X_i^2 - \sum_{i=1}^{N} X_i\bar{X} - \sum_{i=1}^{N} X_i\bar{X} + \sum_{i=1}^{N} \bar{X}^2$$

Using the same ideas as previously in the numerator, we can solve the 4 separate sums here, and simplify the denominator further:

$$\begin{aligned}\sum_{i=1}^{N} X_i^2 - \sum_{i=1}^{N} X_i\bar{X} - \sum_{i=1}^{N} X_i\bar{X} + \sum_{i=1}^{N} \bar{X}^2 &= N_t - N_t\frac{N_t}{N} - N_t\frac{N_t}{N} + N\left(\frac{N_t}{N}\right)^2 \\ &= N_t - \frac{N_t^2}{N} - \frac{N_t^2}{N} + \frac{N_t^2}{N} \\ &= N_t - \frac{N_t^2}{N}\end{aligned}$$

Thus, putting our simplified denominator in, our $\hat{\beta}$ now looks like:

$$\hat{\beta} = \frac{\left(N_t - \frac{N_t^2}{N}\right)(\bar{Y}_t - \bar{Y}_c)}{\left(N_t - \frac{N_t^2}{N}\right)}$$

Then simplifying and cancelling out:

$$\hat{\beta} = \bar{Y}_t - \bar{Y}_c$$

Our final estimate of $\hat{\beta}$ with OLS is, as seen above, equivalent to difference of means.

5.3 Polytomous Explanatory Variable

A **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.

- 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 (since these are the only 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.

5.4 Categorical Explanatory Variables in R

Chapter 6

Transformations and Interactions

6.1 Polynomial Transformations

Sometimes, a linear (straight-line) best-fit line is a poor description of a relationship. We can model more flexible relationships that are not straight lines, by including a transformation of the variable X that we are interested in.

Quadratic Transformations

Key Definition: Quadratic Transformation

Quadratic transformations of X_j take the following form:

$$\mathbb{E}[Y_i|\overrightarrow{X}] = \alpha + \beta_1 X_{ji} + \beta_2 X_{ji}^2 + \text{other explanatory variables}$$

If you recall from high-school algebra, an equation that takes the form of $y = ax^2 + bx + c$ creates a *parabola*. A true parabola has a domain of $(-\infty, \infty)$. However, our model often does not need to do this. The best-fit parabola is only used for the range of plausible X values, given the nature of our explanatory variable. For example, if X was age, a negative number would make no sense. Because the parabola's domain often exceeds our plausible range of X values, the vertex of the parabola (where it changes directions) may not be in our data.

We always include lower degree terms in our model. For example, in this quadratic (power 2) model, we also include the X term without the square. To fit a model like this, we simply do the same process of minimising the sum of squared errors. How do we interpret the coefficients β_1 and β_2?

Interpretation of Quadratic Transformations

β_1's value is no longer directly interpretable. This is because we cannot "hold all other coefficients constant", since β_2 also contains the same X variable.

β_2's value also cannot be directly interpreted. If the coefficient of β_2 is statistically significant, we can conclude that there is a non-linear relationship between X and Y. If β_2 is negative, the best-fit parabola will open downwards, and if β_2 is positive, the best-fit parabola will open upwards.

If we want to interpret the magnitude of the model, we are best off using predicted values of Y (obtained using the model equation above).

There is one more thing we can interpret with the quadratic transformation: the **vertex** of the best-fit parabola. The vertex, if we remember our algebra, is either the maximum or minimum point of a parabola. Thus, if we remember from calculus and optimisation, we can find the maximum and minimums through setting the derivative equal to 0. For the quadratic model, this is as follows - we first find the derivative, then set the derivative equal to 0:

$$\begin{aligned}
&\widehat{Y} = \widehat{\alpha} + \widehat{\beta}_1 X + \widehat{\beta}_2 X^2 \\
&\frac{d\widehat{Y}}{dX} = 0 + \widehat{\beta}_1 + 2\widehat{\beta}_2 X \\
&0 = \widehat{\beta}_1 + 2\widehat{\beta}_2 X \\
&-\widehat{\beta}_1 = 2\widehat{\beta}_2 X \\
&X = \frac{-\widehat{\beta}_1}{2\widehat{\beta}_2}
\end{aligned}$$

This point is useful, as it is either the maximum or minimum of our best-fit parabola. This means that at the X value we calculate from this equation, we will either see the highest or lowest expected Y value.

General Polynomial Models

While quadratic models are the most common polynomial transformation, we do not have to stop there. We can continue to add further polynomials (although anything beyond cubic is exceedingly rare):

- Cubic: $E[Y] = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
- Quartic: $E[Y] = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 x^4$

Each higher order coefficient, if statistically significant, indicates that the relationship between X and Y, is not of the previous highest power. For example, if the cubic term β_3 is statistically significant, we can reject a quadratic relationship between X and Y

Remember to always include the lower power monomials within our polynomial model. For example, if you have a quartic transformation, you must also have the linear, quadratic, and cubic terms.

6.2 Logarithmic Transformations

Logarithmic transformations are another form of non-linear transformations. These are commonly used for heavily skewed variables, such as when the explanatory variable is income, wealth, and so on.

In situations with heavily skewed variables, we often replace X in our models with $\log(X)$. Note that in statistics, when we refer to logarithms, we are referring to natural logarithms, such that $\log(X) = \ln(X)$.

Key Definition: Logarithmic Transformation

The logarithmic transformation of explanatory variable X_j takes the following form:

$$\mathbb{E}[Y_i|\overrightarrow{X}] = \alpha + \beta_j \log(X_j) + \text{other explanatory variables}$$

Interpretation of the β coefficient can be a little bit trickier for logarithmic transformations. We could interpret it in the same way we interpret linear regressions: given a one unit increase in the log of X, there is an expected β change in Y. However, this issue is that this does not really say much - I mean, who knows what a *one unit increase in the log of* X even means?

With some properties of logarithms, we can actually create a more useful interpretation. Based on logarithm rules, we know the following to be true:

$$\begin{aligned}\log(X) + A &= \log(X) + \log(e^A) \\ &= \log(e^A \times X)\end{aligned}$$

Now, let us plug this into our original regression model:

$$\begin{aligned}E[Y|X] &= \alpha + \beta \log(X) \\ E[Y|e^A \times X] &= \alpha + \beta \log(e^A \times X) \\ &= \alpha + \beta[\log(X) + A] \\ &= \alpha + \beta A + \beta \log(X)\end{aligned}$$

Now find the difference between $E[Y|e^A \times X]$ and $E[Y|X]$:

$$\begin{aligned}E[Y|e^A \times X] - E[Y|X] &= [\alpha + \beta A + \beta \log(X)] - [\alpha + \beta \log(X)] \\ E[Y|e^A \times X] - E[Y|X] &= \beta A\end{aligned}$$

Thus, we can seen that when we multiply X by e^A, we get an expected βA change in Y. We can make this interpretation more useful by purposely choosing some value A that makes e^A make more sense. For example, if $A = 0.095$, then $e^A = 1.1$, and multiplying by 1.1 is a 10% increase.

ℹ Interpreation: Logarithmic Transformation

When X_j increases by 10%, there is a expected $0.095\beta_j$ unit change in Y

6.3 Interaction Effects

Interactions, also called moderating effects, means that the effect of some X_j on Y is not constant, and depends on some third variable X_k. Essentially, X_k's value changes the relationship between X_j and Y.

- For example, Y could be the chance of a civil war occurring, X_1 is the severity of an economic crash, and X_2 is the development level of a country.
- We could quite reasonably expect that in the effect of a economic crash on a chance of civil war would be significantly higher in developing nations rather than developed.
- Or in other words, the chance that a civil war occurs due to a economic crash is higher in countries like Venezuela, North Korea and Eritrea, compared to the relationship in Norway, Switzerland, and Denmark.
- Essentially, X_1's effect on Y is affected by the value of X_2.

Key Definition: Interaction Effect Regression

Interaction effects are represented by two variables being multiplied together in a regression equation. In the model below, X_1 and X_2 are interacting with each other:

$$\mathbb{E}[Y|\overrightarrow{X}] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \text{other explanatory variables}$$

Note: you can see in the above equation, X_1 and X_2 both are interacted, as well as have their own separate coefficients. We should always include both variables independently along with the interaction effect.

We can mathematically show that the effect of X_1 on Y is not constant - and varies due to the value of X_2. We show this through finding the partial derivative of X_1 on Y, since the derivative is, by definition, the function of the rate of change between X_1 and Y.

$$\begin{aligned}
\hat{Y} &= \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 \\
\frac{\partial \hat{Y}}{\partial X_1} &= 0 + \hat{\beta}_1 + 0 + \hat{\beta}_3 X_2 \\
\frac{\partial \hat{Y}}{\partial X_1} &= \hat{\beta}_1 + \hat{\beta}_3 X_2
\end{aligned}$$

As you can see, the relationship between X_1 and Y here depends on the value of X_2. In more intuitive words, given a one unit increase in X_1, there is an expected $\hat{\beta}_1 + \hat{\beta}_3 X_2$ increase in Y.

We can also find the effect of X_2 on Y:

$$\begin{aligned}
\hat{Y} &= \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 \\
\frac{\partial \hat{Y}}{\partial X_2} &= \hat{\beta}_2 + \hat{\beta}_3 X_1
\end{aligned}$$

With these equations, we can interpret the coefficients of our model more generally.

Interpretation of Interaction Effects

$\hat{\beta}_1$ is the relationship between X_1 and Y, given $X_2 = 0$.
$\hat{\beta}_2$ is the relationship between X_2 and Y, given $X_1 = 0$.

$\hat{\beta}_3$ represents two things. For every one unit increase of X_2, the magnitude of the relationship between X_1 and Y changes by $\hat{\beta}_3$. Similarly, for every one unit increase of X_1, the magnitude of the relationship between X_2 and Y changes by $\hat{\beta}_3$.

α is still the expected value of Y when all explanatory variables equal 0.

The coefficient β_3's significance level tells us if there is a statistically significant interaction.

- If β_3 is not statistically significant, we can often remove the interaction term.
- However, if β_3 is statistically significant, that means we have found two terms that interact.

6.4 Transformations and Interactions in R

Chapter 7

Gauss-Markov Theorem and Causal Inference

7.1 Gauss Markov Theorem

Key Definition: Gauss-Markov Theorem

If these assumptions are true: 1) Linearity of parameters; 2) Random sampling from the population; 3) Non-perfect collinearity; 4) Exogeneity; 5) Homoscedasticity.
The Gauss-Markov Theorem states that the OLS estimator is the best linear unbiased estimator (BLUE) for the coefficients of a linear regression models.

This is important, as if you meet the assumptions required, you can be confident that you are obtaining the best possible coefficient estimates of any linear model. So - if we are interested in accurately finding the causal effect of D on Y, we can be confident we are getting the most accurate estimate.

Linearity of Parameters

The first Gauss-Markov Assumption is that the parameters $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_k$ must be **linear**. This does not mean that the best-fit line has to be linear. Linearity of parameters refers to the coefficients $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_k$ must not be multiplied/divided with each other. Different coefficients must be added together.

For example, a moderating effect regression has explanatory variables multiplied together. However, see that all coefficients $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_k$ are not multiplied together. This is also the case for polynomial transformations, log transformations, and all models we have covered.

Random Sampling from Population

The second Gauss-Markov Assumption is that our observations $(\overrightarrow{X}_i, Y_i)$ must be **randomly sampled from the population**. Remember that in Causal Inference, the population is actually the potential outcomes, and our sampling is done through our assignment mechanism.

Thus, in theory, to get the best linear unbiased estimator for a causal effect, we must have a randomised controlled experiment. We do have some other ways to address not meeting this condition, that we will explore later in the book.

Non-Perfect Collinearity

The third Gauss-Markov Assumption is that of our explanatory variables $X_1, ..., X_k$, no two can have perfect multicollinearity. Perfect multicollinearity means a perfect correlation/relationship between two of our explanatory variables. Or in other words, one explanatory variable is an exact linear function (with no error) of another explanatory variable.

The reason this is required is mechanical - we discussed this briefly when deriving the OLS estimator in chapter 3. If there is perfect multicollinearity, $X'X$ is no longer invertible, thus making the mechanics of OLS impossible.

To avoid this, when choosing explanatory variables, do not choose two variables that measure the same thing. For example, do not include GDP per capita, and GDP per capita in 1000s of dollars, since they are the same variable, just scaled differently.

However, do note that the more correlated two explanatory variables are, the more variance there is in the estimates. This is because OLS tries to "partial" out effects to each variable, but it is hard to tell what effect one X has on Y compared to another X that is very correlated with it.

If we are selecting highly correlated control variables, this issue does not matter at all - after all, we don't care about the effect of control variables, they are just there to control for confounders. However, we probably do not want a highly correlated variable with our main treatment variable - since that can muddy the effect of the treatment variable.

7.2 Exogeniety and Endogeniety

Key Definition

The fourth Gauss Markov Assumption is **Exogeniety**: the error term and regressors are uncorrelated. In other words, the change in X should not affect the expected value of the error. Mathematically:

$$\mathbb{E}[\epsilon_i | \overrightarrow{X_i}] = 0, \forall i$$

The opposite is **Endogeniety**, when an explanatory variable is correlated with the error term.

We actually proved this theorem earlier when deriving the OLS estimator. Let us take the example of bivariate linear regression, since it is easier to show. First, let us recall the first-order conditions to maximise our sum of squared errors, for both $\hat{\alpha}$ and $\hat{\beta}$:

$$\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_i) = 0$$
$$\sum_{i=1}^{n} X_i (Y_i - \hat{\alpha} - \hat{\beta}_1 X_i) = 0$$

We also know that the error/residual is $\epsilon_i = Y_i - \hat{\alpha} - \hat{\beta}$. Let us plug that in to our above equations, we get:

$$\sum_{i=1}^{n} \epsilon_i = 0$$
$$\sum_{i=1}^{n} X_i \epsilon_i = 0$$

These two equations tell us the following:

1. The first equation says the OLS residuals always add up to zero. Thus, the average error/residual is also 0.
2. The second equation shows that the sample covariance and correlation between X and ϵ are uncorrelated.

Thus, these two assumptions result in our Gauss-Markov Assumption of $\mathbb{E}[\epsilon_i|\overrightarrow{X}_i] = 0, \forall i$, where the explanatory variables are uncorrelated, and thus, also have a average error/residual of 0 for all explanatory variable inputs.

When this assumption is met, we say that our explanatory variables are **exogenous**.

However, when this assumption is not met, we say that the specific X_j that are correlated with the error terms are **endogenous regressors**, and that our model has **endogeniety**.

- We will discuss the Instrumental Variable Estimator later, which deals with endogenous regressors.

7.3 Homoscedasticity and Heteroscedasticity

Key Definition: Homoscedasticity

The fifth Gauss-Markov Assumtion is **homoscedasticity** - that the variance of the error term is consistent and constant for all values of the explanatory variables. Mathematically:

$$Var(\epsilon|\overrightarrow{X}) = \sigma^2$$

If this assumption is violated, we have **heteroscedasticity**.

The simplest way to identify heteroscedasticity is to look at the residual plots - a plot with explanatory variables on the x-axis, and residuals on the y-axis.

- If the residuals show no pattern, then there is no heteroscedasticity. If there is a pattern, for example, if the residuals are very small when X is small, and very big when X is big, then we have heteroscedasticity.
- There are also more formal tests for this, but they take more effort than it is worth. Thus, we will not delve into these heteroscedasticity tests.

OLS is still unbiased as long as the first 4 assumptions are met. Heteroscedasticity does not cause bias or inconsistency in the estimates of coefficients. However, if heteroscedasticity is present, it may indicate a better estimator may exist (so OLS is no longer BLUE).

Furthermore, and more useful for us, the usual standard error formula that we use to calculate confidence intervals and hypothesis tests no longer works. This is because the old standard errors are based on the assumption of homoscedasticity.

Thus, we need to modify our standard errors so that they are still accurate. We do this by calculating the **robust standard errors**, and then conducting our tests using these robust standard errors.

- Do not worry about how to calculate these by hand. Our statistical software will include options to calculate robust standard errors.

7.4 Regression Design for Causal Inference

We can use regression for causal inference by including treatment D as an explanatory variable in our model. If we want to use regression, and only regression for causal inference, we have a few options.

If we have random assignment mechanisms in our study, we can just use a single variable regression to determine causal effects, assuming all other Gauss-Markov assumptions are met (homoscedasticity is not required if we use robust standard errors). Many random controlled trials use regression to determine the causal effects.

However, if we are dealing with observational designs without random assignment mechanisms, things become more complicated. In theory, if we control for every possible confounding variable with multiple linear regression, we can accurately calculate the causal effect of a treatment.

- So, we should include every possible confounding variable in our model to control for these confounders.
- However, in social sciences, this is nearly impossible. There are often thousands of confounding variables, some that we might not even know about.
- The rest of this book will focus on ways to slowly account for these confounding variables that still remain after we included what we could in the regression.

7.5 Robust Standard Errors in R

To calculate robust standard errors, we must install and load the **fixest** package.

Then, the syntax is the same as a linear regression with the standard *lm()* function, but we add an additional argument of *se = "hetero"*, which tells R to calculate heteroscedasticity-robust standard errors.

```r
# load fixest package
library(fixest)

# model
model_name <- feols(Y ~ X1 + X2 + X3, data = mydata, se = "hetero")
summary(model_name)
```

Interpretation is the same as the standard linear regression.

Part III

Methods for Causal Inference

Chapter 8

Fixed Effects Models

8.1 Hierarchical Data

Hierarchical data is data that comes in different "clusters" or "levels". For example, if we have data on individuals from multiple different countries, that means our individual observations are clustered at the country-level.

Hierarchical data can also be clustered over time. For example, we might have GDP data for all countries in the world from 1960-2024. Each year (ex. 2024) will have GDP data for all countries, thus, the data is clustered by year. Data clustered over years is often referred to as **panel data** or **longitudinal** data.

Hiearchical data can be clustered over country and year at the same time. The previous example of GDP data can be clustered by year (ex. 2023, 2022, etc.) and clustered by country (ex. USA, UK, etc.).

Why do we care about clusters? Well - this is because one cluster might be very different than another cluster. For example, if we were explaining individual voting turnout between countries, different electoral and cultural factors in each country might explain some of the differences. Another example is the 2008 financial crisis, which may mean 2008 values will be different from 2015 because of circumstances surrounding each particular year.

These differences between clusters affect our regression results. For example, if we want to explain the outcome variable individual voter turnout with the explanatory variable individual education level, some of the effect of different countries and years may be captured in our regression. That means our regression is not accurately measuring the size of effects.

Thus, we need some way to control for these clusters in our data to isolate the effect of our treatment D and accurately assess the causal impact.

8.2 Fixed Effects

Fixed Effects are a way to control for the issue of differences between clusters.

Let us assume that we have m number of clusters in our data. Thus, we have a specific cluster $i \in \{1, ..., m\}$. Each cluster i will have n number of observations, so we will have observation $t \in \{1, ..., n\}$ within cluster i.

Using this framework, every observation can be defined as Y_{it}, which essentially means the Y value of the ith cluster's tth observation. The corresponding explanatory variable values will be notated $\overrightarrow{X}_{it}$.

Key Definition: Fixed Effects Model

A fixed effects model will take the following form:

$$Y_{it} = \alpha_i + \beta_1 X_{1it} + ... + \beta_k X_{kit} + \epsilon_{it}$$

Where α_i is the fixed effect for cluster i, defined as:

$$\alpha_i = \beta_{00} + \beta_{02} D)i2 + ... + \beta_{0m} D_{im}$$

Where $D_{i2}, ..., D_{im}$ are dummy variables for the clusters $2, ..., m$. Cluster 1 is the reference category (like a categorical explanatory variable). β_{00} is the average Y of the reference cluster category (cluster 1), when $\overrightarrow{X} = 0$. β_{0j} is the difference between the average Y of cluster j, and the reference category (cluster 1), when all $\overrightarrow{X} = 0$.

Or in other words, including fixed effects for clusters i means using the clusters as an additional categorical variable in our regression. We can demonstrating this by writing out α_i in our above linear model to get:

$$Y_{it} = \beta_{00} + \beta_1 X_{1it} + ... + \beta_k X_{kit} + \beta_{02} D_{i2} + ... + \beta_{0m} D_{im} + \epsilon_{it}$$

The fixed effect α_i captures the predictors of Y that are shared by all observations within their cluster i. For example, if our fixed effects were by countries, α_i would capture all the predictors of Y that are shared by all observations from that same country. To interpret our coefficients β_j, we would do the same as we previously would, but adding the line - controlling for levels of Y we would expect for that cluster in general.

Two-Way Fixed Effects

Often in Political Economics, we will have 2-way clustered data by both country and year. For example, if you have data on GDP and Democracy level from all countries between 2006-2024, you will have two types of clusters - clusters by country, and clusters by year.

Key Definition: Two-Way Fixed Effects

We can combine these two for two-way fixed effects of both country and year. Two-way fixed effects takes the following form:

$$Y_{it} = \alpha_i + \gamma_t + \beta_1 X_{1it} + ... + \beta_k X_{kit} + \epsilon_{it}$$

α_i represents country fixed effects, exactly as we described in the previous section.
γ_t represents year fixed effects. Why does it have the subscript t? Well, in panel data, for each country, you will have many different years of data (ex. USA will have data between 2006-2024). Thus, within cluster i of the country, each observation t is a different year. Thus, t is the year of the data.

To interpret our coefficients β_j, we would do the same as we previously would, but adding the line - controlling for levels of Y we would expect for that country in that year in general.

This allows us to account for differences in countries and differences in years, and is very very common in Political Economics. You will also sometimes see different variations of this, including State-Year, Country-Decade, District-5Years, or any Geographic-Time clustering.

8.3 Fixed Effects in R

To calculate fixed effects, we must install and load the **plm** package (there is a way to do it in base-r, but plm makes it easier, and there are other packages as well).

```r
library(plm)
```

The syntax is very similar to standard linear regression, however, we add an *index* parameter to indicate what variables we want fixed effects on, and a *model = "within"* parameter to specify we want fixed effects. For two ways, we need an additional parameter *effect = "twoways"*.

```r
# one-way fixed effects
model1 <- plm(Y ~ X1 + X2, data = mydata,
              index = "Cluster Variable Name", model = "within")
summary(model1)

# two-way fixed effects
model2 <- plm(Y ~ X1 + X2, data = mydata,
             index = c("Cluster 1 variable", "Cluster 2 variable"),
             effect = "twoways", model = "within")
```

Important note: Often, the variable *year* is encoded as numeric, but we want it to be a categorical variable for fixed effects, so use the *as.factor()* function to coerce the variable *year*.

Also, you can do this in base-r with the *lm()* function as shown throughout the regression examples, just by including the cluster variable as a categorical explanatory variable, however, the output is not as nice.

Chapter 9

Instrumental Variables Framework

Chapter 10

Other Instrumental Variable Designs

10.1 Continuous Independent Variables

10.2 Examiner Designs

10.3 Shift-Share Bartik Instruments

10.4 Recentered Instruments

Chapter 11

Regression Discontinuity

Chapter 12

Differences-in-Differences

Chapter 13

Survey Experiments