Introductory Econometrics for Political Economy

An introduction to statistical inference and causal inference for the Social Sciences

Kevin Lingfeng Li

Table of contents

Preface

This book focuses on how we can use econometrics to analyse empirical data to prove or reject our hypotheses. This book starts out with the fundamentals of probability and inference. Then, it moves to regression anlaysis. Finally, it moves to methods regarding causal inference.

This book expands on my other book, "An Introduction to Political Economy", as this book provides empirical techniques in which we can evaluate the models disucussed in that book. However, this book can also be used as an independent resource for the study of econometric and statistical techniques for Political Science and other Social Sciences.

This book is meant to be a relatively approachable introduction to the field. However, as Political Economy depends on many economic tools, an rough understanding of Algebra, Single Variable Calculus, and simple Linear Algebra is required. You do not need to be a math wizard, or even good at solving mathematical problems - you simply need an understanding of the intuition behind some key techniques. This book comes with a companion manual - Essential Mathematics for Political Economy. It is recommended that anyone interested in Political Economy glance at the topics covered in the manual, to ensure that they have the mathematical background necessary to succeed.

I created this book as a way to revise for my exams, as well as provide a handy booklet where I could reference all the things I learned throughout my undergraduate and postgraduate degrees. I hope that this guide to Political Economy can be useful to not just me, but others also interested in the field.

This book will use the R language for some examples, as well as providing code for some other languages (Stata, Python). This is not a coding course, so I will not introduce the basics of R.

Part I

Probability and Statistical Theory

Chapter 1

Concepts in Probability

1.1 Basics of Sets

A **set** is the collection of objects, while the **elements** of the set are the specific objects within a set. A capital letter is used to represent a set, for example, set A. A lowercase letter represents an element within the set. For example, element a is a part of set A.

There are a few common types of sets we use often:

- N is the set of natural numbers - the numbers we use to count from 1: $\{1, 2, 3...\}$
- Z is the set of integers - non-decimal numbers both negative and positive: $\{..., -2, -1, 0, 1, 2...\}$
- R is the set of all real numbers - any numbers that are on the number line, including decimals

We can define sets in multiple ways.

- We can list out each element of a set. For example $A = \{1, 2, 4, 6\}$
- We can define them with an interval. For example, $A = [0, 1]$ means all values within the range of 0 to 1, including 0 and 1. Brackets [] mean inclusive of the end points, and parentheses () mean excluding the end points. We can also mix and match brackets and parentheses.
- We can define them in formal notation. For example, $A = \{x : 0 \leq x \leq 1, x \in R\}$. This literally means: x such that x is between 0 and 1, including 0 and 1, and x is in the set of all real numbers R.

Useful notation tips:

- The semicolon : means "such that", and the sign $\in$ means "in" or "belongs to".
- When we put absolute value bars around a set, like $|A|$, that refers to the number of elements within that set.

1.2 Set Operators

There are a few different set operators that are important to understand.

An **intersection** of sets A and B, formally notated $A \cap B$, indicates the elements that are both within A and B at the same time.

- For example, if $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$, then $A \cap B = \{2, 3\}$, since those are the elements that are contained in both A and B at the same time.

A **union** of sets A and B, formally notated as $A \cup B$, indicates elements that are in either A, B, or both A and B.

- For example, if $A = \{1, 2, 3\}$ and $B = \{2, 3, 4\}$, then $A \cap B = \{1, 2, 3, 4\}$.
- $A \cup B = A + B - A \cap B$. We subtract $A \cap B$ since that part is counted twice in both A and B, so we need to get rid of it once to avoid over-counting.

The **complement** of set A is everything that is not in A, but still within the universal set. The complement is denoted as A' or A^c.

- For example, if the universal set contains $\{1, 2, 3, 4, 5\}$, and $A = \{1, 2\}$, then $A' = \{3, 4, 5\}$

A **subset** A has all its elements belonging to another set B. This is notated $A \subset B$.

- For example, if $A = \{1, 2\}$, and $B = \{1, 2, 3\}$, then A is a subset of B since all of A's elements belong to set B as well.
- There are two types of subsets. **Proper subsets** are subsets when the number of elements in A is less than the number of elements in B : $|A| < |B|$. **Imrpoper subsets** are when the number of elements in A is less than or equal to the number of elements in B : $|A| \leq |B|$

1.3 Basic Properties of Probability

Kolmogrov's Axioms are the key properties of probability:

1. For any event A, the probability of A occurring is between 0 and 1. Mathematically: $0 \leq Pr(A) \leq 1$
2. The probability of all events in the sample space S is 1. Mathematically: $Pr(S) = 1$. The sample space is the set of all possible events.
3. If we have a group of mutually exclusive events $A_1, A_2, ..., A_k$, then the probability of those events all occurring is the sum of their probabilities. Mathematically, $Pr\left(\bigcup A_i\right) = \sum Pr(A_i)$
    - Note: mutually exclusive events are events that cannot occur at the same time together.

Other important properties to note include:

- $Pr(A') = 1 - Pr(A)$ - the probability of the complement of A, is equal to 1 minus the probability of A
- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$ - this is because of a property of unions, as shown in section 2.2

1.4 Joint and Conditional Probability

Joint Probability is the probability of two or more events occurring simultaneously. The joint probability of events A and B is notated $Pr(A \cap B)$.

For example, in a deck of cards, A could be the event of drawing an ace, and B could be the event of drawing a spade. Thus, $Pr(A \cap B)$ would be the probability of drawing a card that was both an ace and a spade.

Conditional Probability is the probability of one event occurring, given another has already occurred. Probability of event A, given event B has occurred, is notated as $Pr(A|B)$

To calculate the conditional probability, we use the following formula:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

1.5 Bayes' Theorem

Bayes' theorem is arguable the most important theorem in all of probability and statistics. Thus, instead of just telling you Bayes' theorem, we will actually derive it.

We start with the definition of conditional probability, as seen in the last lesson

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

Let us rearrange the equation by solving for $Pr(A \cap B)$. We will get:

$$Pr(A \cap B) = Pr(A|B) \times P(B)$$

Now, let us consider the conditional probability of $B|A$ (the opposite way around). We use the same conditional probability formula, but switch the B and A. We get:

$$Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)}$$

Let us rearrange the equation by solving for $Pr(B \cap A)$. We will get:

$$Pr(B \cap A) = P(B|A) \times P(A)$$

Now we have two different equations, one for $Pr(A \cap B)$, and one for $Pr(B \cap A)$. Based on the commutative property of sets, we know that $Pr(A \cap B) = Pr(B \cap A)$. Thus, the other parts of the equation must also be equal to each other:

$$Pr(A|B) \times Pr(B) = Pr(B|A) \times Pr(A)$$

Now, let us solve for $Pr(A|B)$. After we do this, we will get the final form of **Bayes' Theorem**:

$$Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$$

Each part of Bayes' Theorem has a name. They are commonly referenced, so it is useful to know their names:

- $Pr(A|B)$ is the conditional probability
- $Pr(B|A)$ is the posterior probability
- $Pr(A)$ is the prior probability
- $Pr(B)$ is the marginal probability

1.6 Law of Total Probability

The law of total probability helps calculate the probability of an event, by considering mutually exclusive events that completely partition the sample space. For example, if we know the probability of a male (or non-female) being a smoker, and a female being a smoker, we can calculate the probability of all humans for smoking. This is because male (or non-female) and female together contain the entire sample space of humans.

More mathematically, the relationship of the probability of A, given conditional probabilities $Pr(A|B_i)$ is as follows:

$$Pr(A) = \sum Pr(A|B_i) \times Pr(B_i)$$

For a more intuitive example, imagine you have 3 ways, and only 3 ways, you can get to work. That essentially means these 3 ways partition the sample space of getting to work.

Let us define $Pr(A)$ as the probability you arrive on time. $Pr(B_1), Pr(B_2), Pr(B_3)$ are the probabilities you take route 1, route 2, or route 3 to work. $Pr(A|B_1)$ is the probability that you are on time, given you take route 1. A similar thing applies for $Pr(A|B_2), Pr(A|B_3)$

Thus, the total probability of arriving on time $Pr(A)$, given we know the probability of arriving on time for route 1, route 2, and route 3, is:

$$Pr(A) = \sum Pr(A|B_i) \times Pr(B_i)$$

$$Pr(A) = Pr(A|B_1) \times P(B_1) + Pr(A|B_2) \times Pr(B_2) + Pr(A|B_3) \times Pr(B_3)$$

Chapter 2

Random Variables

2.1 Randomness

Random variables are variables that represent unobserved events that have some randomness - a set of potential outcomes, with each outcome having a probability of occurring.

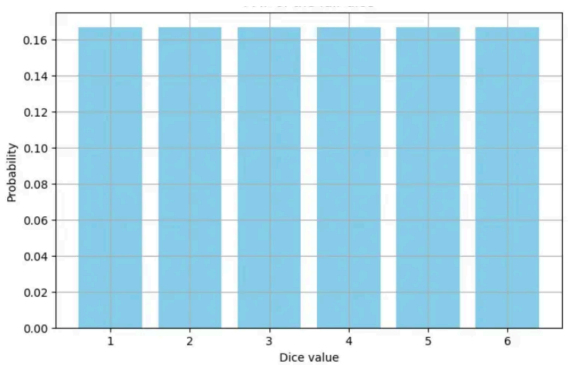
For example, if you flip a coin 10 times, and count the number of heads you get, you could get 5 heads, 6 heads, 4 heads, or any amount between 0 and 10. We are not sure what will happen - however, some outcomes are more likely than others, because of the probabilities associated with each outcome.

There are two types of random variables. **Discrete Random Variables** are random events which have a distinct number of outcomes. For example, rolling a dice has 6 outcomes. A **Continuos Random Variable** are random events which have an infinite amount of outcomes. For example, my drive to work tomorrow could take 5 minutes, 5.123 minutes, 5.234237847 minutes, and so on... there is no distinct outcomes since you can continuously subdivide the gaps between outcomes by adding more decimal points.

2.2 Distributions and Probability Density Functions

Random variables are often called distributions - because there are a distribution of outcomes, with associated probabilities for each outcome. We can actually graph this - put potential outcomes on the x axis, and the probability that each outcome occurs on the y axis.

For example, take this probability distribution of a die - there are 6 sides that you could land on, and each has an equal probability of occurring:



The **probability density function** $f(y)$ takes a potential outcome of an event as an input, and outputs the respective probability.

For example, the probability density function of a dice is $f(y) = 1/6$. This is because every outcome y has the same probability of occurring: $1/6$. So $f(1), f(2)... = 1/6$.

2.3 Expectation and Variance

Expectation and Variance are two ways we can summarise the distributions of random variables.

The **expectation**, often called the expected value or mean, is a measurement of the centre of a probability distribution. The expected value is statistically, the best guess of an outcome of a random variable, given no other information except its distribution. We notate expected value of a variable Y as either $E[Y]$, $\bar{Y}$, or μ.

A **discrete random variable** is one that has a countable number of distinct outcomes/categories, like the outcome of rolling a dice (6 distinct outcomes). The expected value for discrete variables is calculated by multiplying each outcome value by its associated probability, then doing that for all outcomes, and summing everything together. In other words, it is a weighted average of the outcomes, with the weights being the probability of each outcome.

$$E[Y] = y_1 \times f(y_1) + y_2 \times f(y_2)... = \sum [y_j \times f(y_j)]$$

For a **continuous random variable**, it is a little more complicated. This is because continuous variables have an infinite number of potential outcomes. For example, if you drive to school, your driving time could be 23 minutes, or 23.12 minutes, or 23.123324 minutes… basically, an infinite amount. As a result, we have to alter the expected value formula a little:

$$E[Y] = \int_{-\infty}^{\infty} y \times f(y) dy$$

Variance σ^2 is a measure of how spread out our distribution is. Variance basically measures how far values are, on average, from the mean of the variable. Mathematically:

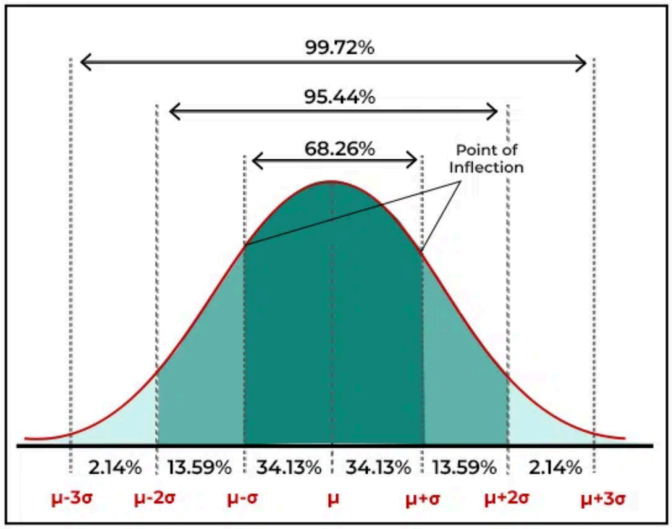
$$Var(X) = \sigma^2 = \frac{1}{n}\sum(X-\mu)^2 = E[(X-\mu)^2]$$

Where σ^2 is the variance, n is the number of observations, and μ is the mean of X

Standard Deviation σ is the square root of variance σ^2

2.4 Normal Distribution and T Distribution

A **normal distribution** is in the shape of a bell curve. The mean μ, mode, and median are all the same value at the centre, and the distribution is symmetrical on both sides. The figure below shows the typical shape of a normal distribution



All Normal Distributions, as shown in the image above, follow the 68-95-99.7 rule:

- Within one standard deviation σ of the mean μ, lies 68.26% of the total area under the curve
- Within 2 standard deviations 2σ of the mean μ, lies 95.44% of the total area under the curve
- In fact, any amount of standard deviations σ, including decimals, is related to a specific percent of total area under the curve, for all normal distributions.

This is important, because the area under the distribution curve is actually the probability. Thus, the normal distribution tells us there is a relationship between the standard deviation and the probability of an action occurring.

Any normal distribution can be described with 2 features: mean μ and variance σ^2 in the following form: $X \sim \mathcal{N}(\mu, \sigma^2)$. For example, $X \sim \mathcal{N}(30, 4)$ means a normal distribution with mean 30 and variance 4.

The T distribution is a distribution very similar to the shape and size of the normal distribution, however, generally has thicker tails and a lower peak. The key difference is that t-distributions are defined with only one parameter - degrees of freedom DF.

Chapter 3

Basics of Inference and Correlation

3.1 Samples and Population

In political science and the social sciences, we are often interested in studying large groups of people and entities. For example, we might be interested some feature regarding all people in a country, such as the average income, or average working hours, or average education level.

However, if we are dealing with large population sizes, it is often impossible to ask every single individual in the population. For example, if we wanted to study the average educational level of the UK, we would need to ask nearly 70 million people. This is completely impractical.

A **sample** is a subset of a population, which ideally, can tell us something about the population. If our sample can reflect the greater population, then we can use the sample in our study, instead of the large population.

Sampling is the process by which we select a sample from a larger population, and we want the sample to be representative of the population. The quality of a sample depends on two major factors:

1. The sampling procedure which we decide to implement
2. Luck

Let us first talk about sampling procedure. The gold standard of sampling procedure is a **random sample** - where individuals in the sample are selected at random from the population. This is because in a random sample, every possible individual has an equal chance of being selected, and thus, the resulting sample is likely to be reflective of the common traits of the population.

3.2 Central Limit Theorem

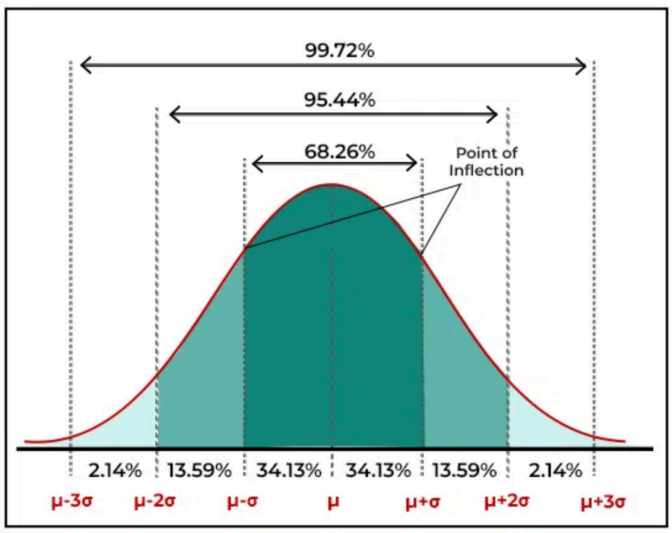
Remember how we explained that the quality of a sample depends not just on the sampling procedure, but also, luck? Well, Central Limit Theorem helps account for the luck aspect.

Before we introduce the Central Limit Theorem, we need to explain a **distribution of sample means**.

- Imagine that we take a random sample from a population. Then, we find the mean of the variable we are interested in the sample. That is a sample mean.
- Then, let us take another sample from the same population, and find the mean. This will be slightly different than the first sample, since we are randomly sampling. That is another sample mean. We keep taking samples from the same population, and getting more and more sample means.
- Now, let us plot all our sample means into a "histogram" or density plot. The x axis labels the possible sample means values, and the y axis is how frequently a specific sample mean occurs. We will get a distribution, just like a random variable distribution.
- That distribution is the d**istribution of sample means** - it basically measures the frequency of different sample means that we get, given we keep drawing samples from the same population and calculating their means.

The **Central Limit Theorem** states that the distribution of sample means of a variable, will be approximately normally distributed. This is regardless of the variable's population distribution shape.

From chapter 2, we know of the 68-95-99.7 rule of normal distributions, and how normal distributions have a systemic relationship between the standard deviation σ and the probabilities. The figure below reminds us of the properties of normal distributions:



Since the distribution of sample means tells us the probability of getting some sample mean, and Central Limit Theorem tells us that distribution is normally distributed, we can now tell how likely

a sample mean is to occur if a sample was drawn from the population.

- For example, if a certain sample mean is located 2 standard deviations above the mean, there is only a 2.14% chance that that sample mean would be that value or higher (see figure above)
- For any sample mean, we calculate how many standard deviations away from the mean of the distribution of sample means. Then, we can calculate how likely that sample mean is likely to occur.

This goes back to the "luck" aspect of sampling. What if we are unlucky in sampling, and end up randomly drawing all the tall people? All the smartest people? Well, we don't have to worry, since Central Limit Theorem tells us the likelihood of drawing a certain sample.

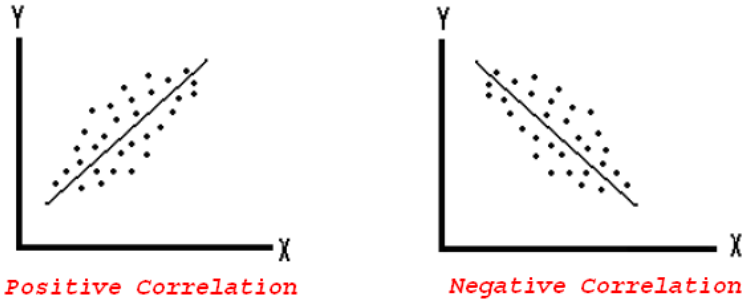
3.3 Covariance and Correlation

In political economy, we are often interested in the relationship between two variables. For example, are oil producers more likely to be democratic? Are more educated voters more likely to turn out and vote?

The relationship between two features, also called correlation, is the extent to which they tend to occur together.

- A positive correlation/relationship is when we are more likely to observe feature Y, if feature X is present
- A negative correlation/relationship is when we are less likely to observe feature Y, if feature X is present
- No correlation/relationship is when we see feature X, that does not tell us anything about the likelihood of observing Y

We can also visualise these graphically:



Covariance is a way to measure the relationship between two variables. Covariance is the extent that X and Y vary together. Mathematically:

$$Cov(X,Y) = \sigma_{XY} = \frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

Or more simply:

- In our data, we have many different pairs of data points (X_i, Y_i)
- X_i is some value of X, and $\bar{X}$ is the mean of X. Same goes for Y_i and $\bar{Y}$
- Thus, $X_i - \bar{X}$ is the distance between any point X_i and the mean $\bar{X}$. Same goes for $Y_i - \bar{Y}$
- n is the number of observations (data points) in our data

We can interpret the sign of the covariance: if it is positive, we have a positive relationship. if it is negative, we have a negative relationship. However, we cannot interpret the size of the covariance.

If we want to see the strength of a relationship/correlation, we have to find the **correlation coefficient**. We calculate this by taking the covariance, and dividing it by the product of the standard deviation of X and the standard deviation of Y. Mathematically:

$$Corr(X, Y) = r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

The correlation coefficient is always between -1 and 1.

- The direction is the same as the covariance - if the coefficient is positive, then we have a positive relationship, vice versa.
- If the correlation coefficient is closer to -1 or 1, it means a strong correlation. If the correlation is closer to 0, then it is a weak correlation

3.4 Best Linear Predictor

While the correlation coefficient tells us the strength of a correlation, it does not say anything about the magnitude of the relationship. For example, if X increases by one unit, how much does Y increase by? The correlation coefficient does not say.

Magnitude is quite an important concept. After all, even if two values are very highly correlated, if an increase of one unit in X only leads to a miniscule increase in Y, this relationship might not be very important for understanding the world.

A way to estimate the magnitude of the relationship between X and Y is the **best linear predictor.** The best linear predictor is a best fit line for the data, that takes the form of a linear equation: $Y = \alpha + \beta X$.

In this equation, the β term in the best fit line is the slope of the linear equation. Essentially, it tells us for every increase in one unit of X, how much do we expect Y to increase by?

- In a linear model, the X variable is considered the **explanatory** or **independent** variable, while the Y is the **response** or **dependent** variable.

The Best Linear Predictor is a form of Linear Regression, the primary topic we will cover in the next two chapters.

Part II

Linear Regression Anlaysis

Chapter 4

Linear Model, Estimation, and Interpretation

4.1 Specification of the Linear Model

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

- The **response variable** (dependent 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 $\overrightarrow{X}$

In formal statistical terminology, a regression model is the specification of the conditional distribution of Y, given $\overrightarrow{X}$. Essentially, it is stating that the distribution of possible Y outcomes depends on the value of $\overrightarrow{X}$.

- I say conditional **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 $X = 30$, and a different distribution with different mean for $X = 20, X = 25$, etc.

The linear regression model focuses on the **expected value** or mean of the conditional distribution of Y given $\overrightarrow{X}$.

Suppose we have 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:

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

- Where $E[Y_i|\overrightarrow{X_i}]$ is the expected value of the conditional distribution $Y_i|\overrightarrow{X_i}$
- The distribution of $Y_i|\overrightarrow{X_i}$ has a variance $Var(Y_i|\overrightarrow{X_i}) = \sigma^2$.
- 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 point. We will go into detail on this later.
- A key assumption (that we will discuss later) is that the error function overall is normally distributed: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- 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).

4.2 Estimation of Parameters

In our model, we have parameters $\alpha, \beta_1, ..., \beta_k$ and variance σ^2 that need to be estimated in order to create a best-fit line we can actually use. Our best estimates of these parameters will be denoted with a hat: $\hat{\alpha}, \hat{\beta}_1, ...$ 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 because of our prior assumption that the error term $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, which says the expected value of ϵ_i is $E[\epsilon_i] = 0$. So on average, error is 0, so our predictions do not include the error term.

However, how do we determine the estimates of our parameters $\alpha, \beta_1, ..., \beta_k$? **The Ordinary Least Squares Estimator (OLS).** OLS estimation attempts to **minimise the sum of squared errors** of our predicted line to our actual observed data. The estimation processes is as follows.

1. First, we propose some coefficient values to test. Let $\tilde{\beta}$ represent the vector of our proposed coefficient values $\tilde{\alpha}, \tilde{\beta}_1, ..., \tilde{\beta}_k$
2. Then, we use these proposed coefficients in our prediction line: $\hat{Y}_i(\tilde{\beta}) = \tilde{\alpha} + \tilde{\beta}_1 X_{1i}, ..., \tilde{\beta}_k X_{ki}$
3. Then, we calculate the residuals e_i of our predictions of Y using our proposed coefficients, compared to the actual values of Y_i : $e_i(\tilde{\beta}) = Y_i - \hat{Y}_i(\tilde{\beta})$
4. Then, we calculate the sum of squared errors (SSE) for all residuals: $SSE(\tilde{\beta}) = \sum(Y_i - \hat{Y}_i(\tilde{\beta}))^2$

The set of proposed $\tilde{\beta}$ coefficients that produces the lowest SSE is chosen as our estimated parameters.

Of course, testing every possible set of proposed coefficients $\tilde{\beta}$ is quite time-consuming. Luckily, mathematicians have derived a formula for the parameters that minimise the SSE of a simple linear regression:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

For more than one X variable, it is nearly impossible to hand calculate the estimated parameters. Luckily, the computer does this very quickly for us (we will show how to implement this in R and STATA later).

We can also derive the estimate of the **residual variance** σ^2 (variance of the error term) with this formula:

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

The standard deviation of the residuals is simply the square root of the variance.

4.3 Interpretations of Coefficients

Simple Linear Regression

Simple Linear Regression is a special case of the linear model, when there is only one explanatory variable X. How do we interpret parameters $\hat{\alpha}$ and $\hat{\beta}$ that we have calculated?

$\hat{\beta}$ is the slope of the the linear model. More formally, $\hat{\beta}$ is the expected change in Y, corresponding to a one-unit increase in X. Or in other words, as X increases by 1 unit, the expected value Y increases by $\hat{\beta}$ units.

- A positive $\hat{\beta}$ means a positive relationship, a negative $\hat{\beta}$ means a negative relationship, and $\hat{\beta} = 0$ means no relationship.

$\hat{\alpha}$ is the y-intercept of the linear model. More formally, $\hat{\alpha}$ **is the expected value of** $\hat{Y}$**, given** $X = 0$.

Multiple Linear Regression

Multiple linear regression is when there are multiple explanatory variables $X_1, X_2, ..., X_k$. How do we interpret these parameters $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ that we have calculated?

- Note, I will define $\hat{\beta}_j$ as any one of $\hat{\beta}_1, ..., \hat{\beta}_k$ (the interpretation is all the same). In the model, $\hat{\beta}_j$ will be multiplied to variable X_j.

Formally, any coefficient $\hat{\beta}_j$ is the expected change in Y, corresponding to a one unit increase in X_j, holding all other explanatory variables $X_1, ..., X_k$ that are not X_j constant.

- A positive $\hat{\beta}_j$ means a positive relationship between X_j and Y, a negative $\hat{\beta}_j$ means a negative relationship, and $\hat{\beta}J = 0$ means no relationship, given we hold all other explanatory variables $X_1, ..., X_k$ that are not X_j constant.
- Essentially, we do the same interpretation as the single linear regression, but adding the phrase "**holding all other explanatory variables constant**".

$\hat{\alpha}$ is the expected value of $\hat{Y}$, given all explanatory variables $X_1, ..., X_k$ equal 0.

You might wonder, why is $\hat{\beta}_j$ interpreted in the way it is: "holding other variables constant"? Well, we can actually mathematically prove this by taking the partial derivative in regard to X_j to see the rate of change of $\hat{Y}$ in regards to X_j:

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

$$\frac{\partial \hat{Y}_i}{\partial X_{ji}} = 0 + 0 + ... + \hat{\beta}_j + ... + 0$$

$$\frac{\partial \hat{Y}_i}{\partial X_{ji}} = \hat{\beta}_j$$

So we can see, the partial derivative of $\hat{Y}_i$ in regard to explanatory variable X_{ji}, holding all others constant, is $\hat{\beta}_j$, as we interpreted above.

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 democracy is measured on a 100 point scale, what does a 5 point change in democracy mean? Is it a big change, or a small change?

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

So, instead of the expected change of Y given one unit increase of X, we instead do the expected standard deviation change of Y, given a one standard deviation increase in X.

How do we calculate this? There is a formula! Y changes by $(SD_{X_j} \times \hat{\beta}_j)/SD_Y$, where SD represents standard deviation, and X_j is the variable whose coefficient we are interpreting.

Binary X Variable Interpretation

If our explanatory variable(s) are binary, (i.e. X_j only has two values, 0 and 1), then our interpretation differs slightly.

We essentially treat the explanatory variable as a variable of 2 different categories, $X_j = 0$, and $X_j = 1$

Now, $\hat{\alpha}$ is the expected value of Y for an observation in category $X = 0$.

- If you have more explanatory variables, remember to add - holding all other explanatory variables constant.

$\hat{\beta}_j$ is the expected difference in the value of Y, between the categories $X_j = 1$ and $X_j = 0$

- If you have more explanatory variables, remember to add - holding all other explanatory variables constant.

To find the expected value of Y for an observation in $X = 1$, you would need to do $\hat{\alpha} + \hat{\beta}_j$.

You might wonder, why are these interpretations the way they are? Well, let us take a simple linear regression that has a binary X variable. The fitted model takes the following form:

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

If we want the predicted Y for category $X = 0$, we simply set $X = 0$:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}(0) = \hat{\alpha}$$

Now, what would be the predicted Y for category $X = 1$? Let us simply set $X = 1$:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}(1) = \hat{\alpha} + \hat{\beta}$$

Then the predicted difference between the predicted value of Y for categories $X = 1$ and $X = 0$ would simply be:

$$(\hat{\alpha} + \hat{\beta}) - (\hat{\alpha}) = \hat{\beta}$$

4.4 Estimated Residual Standard Deviation

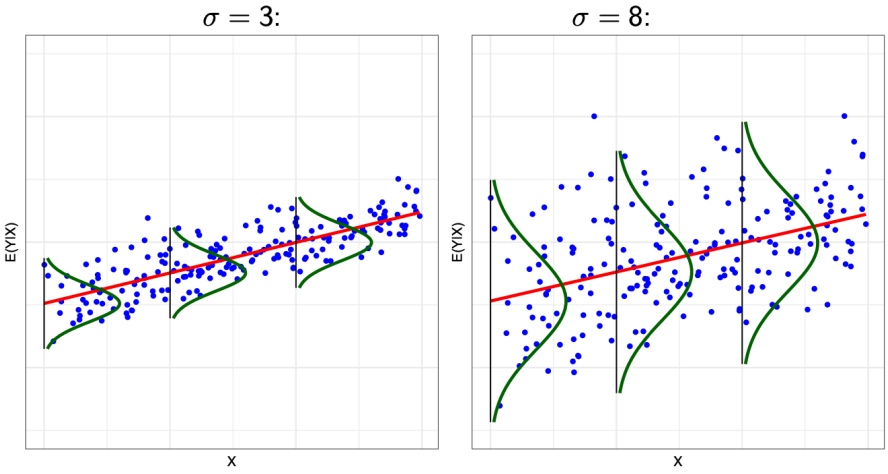
In section 4.2, we also discussed the estimate of the **residual variance** $\hat{\sigma}^2$ (variance of the error term) of the model. The residual standard deviation is just the square root of that.

But what is the residual variance and residual standard deviation. Recall the way we write our regression model:

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

The residual variance refers to the error term ϵ_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:



One important thing to note is that the residual standard deviation $\hat{\sigma}$ 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.

4.5 Model Statistics

Total Sum of Squares

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

$$TSS = \sum (Y_i - \bar{Y})^2$$

We can also rewrite the total sum of squares as the sum of two different sections:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$TSS = SSM + SSE$$

Where TSS is the total sum of squares, SSM is the model sum of squares, and SSE 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 R^2 is a measure of the percentage of variation in Y, that is explained by our model (with our chosen explanatory variables).

As we just explained before, SSM is the the amount of variation in Y that is explained by Y, and the TSS is the total amount of variation in Y. Thus, naturally, 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 (as is sometimes the case in many social sciences), it is just one metric with its benefits and drawbacks.

Interestingly, R^2 for a simple linear regression is also equal to the square of the correlation coefficient r.

4.6 Linear Regression in R

As for all of the examples in this book, we will use the "tidyverse" package. Let us also load our dataset that we will be using.

```
# if you haven't installed tidyverse, do: install.packages('tidyverse')
library(tidyverse)
```

Chapter 5

Hypothesis Testing

5.1 Confidence Intervals

5.2 Parameter Hypothesis Testing

5.3 F-Tests

5.4 Hypothesis Testing in R

Chapter 6

Interactions and Transformations

6.1 Binary Interaction Effect

6.2 Continuous Interaction Effect

6.3 Interactions in R

6.4 Logarithmic Transformations

6.5 Polynomial Transformations

6.6 Transformations in R

Chapter 7

Panel and Clustered Data

Chapter 8

Model Selection for Inference

Part III

Further Regression Models

Chapter 9

Binomial Logistic Regression

Chapter 10

Multinomial and Ordinal Logistic Regression

Chapter 11

Regression for Counts

Part IV

Causal Inference for Experiments

Chapter 12

Causal Frameworks

Chapter 13

Random Experiements

Part V

Causal Inference for Observational Studies

Chapter 14

Selection on Observables

Chapter 15

Instrumental Variables

Chapter 16

Regression Discontinuity

Chapter 17

Differences-in-Differences

Chapter 18

Survey Experiments