Introductory Econometrics for Political Economy

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

Kevin Lingfeng Li

Table of contents

Preface **2**

I Probability and Statistical Theory **3**

1 Concepts in Probability **4**

2 Random Variables **8**

3 Basics of Inference and Correlation **12**

II Regression Anlaysis **17**

4 Linear Regression Model **18**

5 Extensions of the Linear Model **19**

6 Binomial Logistic Regression **20**

7 Multinomial and Ordinal Logistic Regression **21**

8 Regression for Counts **22**

III Causal Inference **23**

9 Causal Frameworks **24**

10 Random Experiements **25**

11 Selection on Observables **26**

12 Instrumental Variables **27**

13 Regression Discontinuity **28**

14 Differences-in-Differences **29**

15 Survey Experiments **30**

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

Econometrics depends on many aspects of probability. This chapter will go through what I consider the "essential" probability you must know to understand the later chapters. It is highly recommended to read more about probability theory, either through a math textbook, or through consulting the "Essential Mathematics for Political Economy" companion manual.

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.

- 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.

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

Chapter 2

Random Variables

Random Variables are the core of statistical techniques. This section builds on what we have learned in probability from above, and applies it to random variables and distributions.

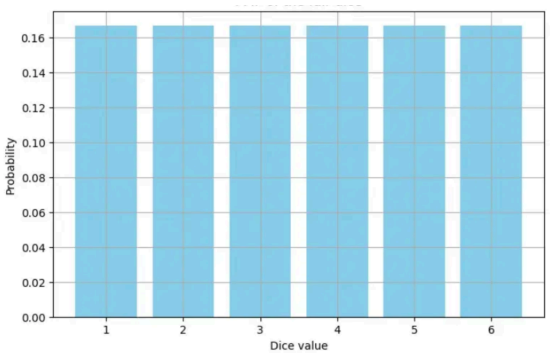
2.1 Distributions and Probability Density Functions

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.

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 $f(1), f(2)... = 1/6$.

2.2 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:

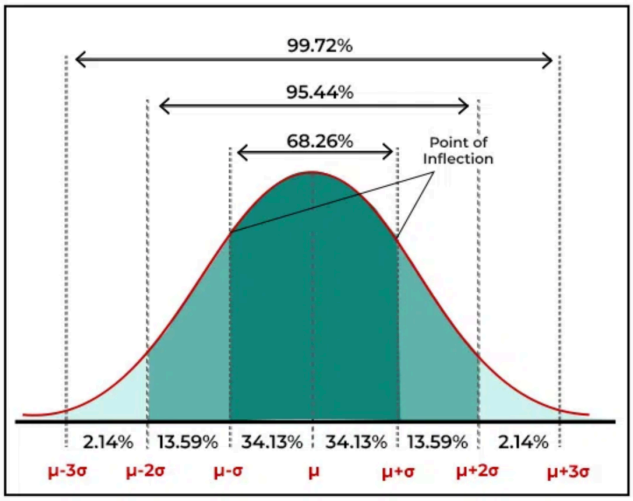
$$Var(X) = \sigma^2 = \frac{1}{n}\sum(X-\mu)^2 = E[(X-\mu)^2]$$

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

2.3 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

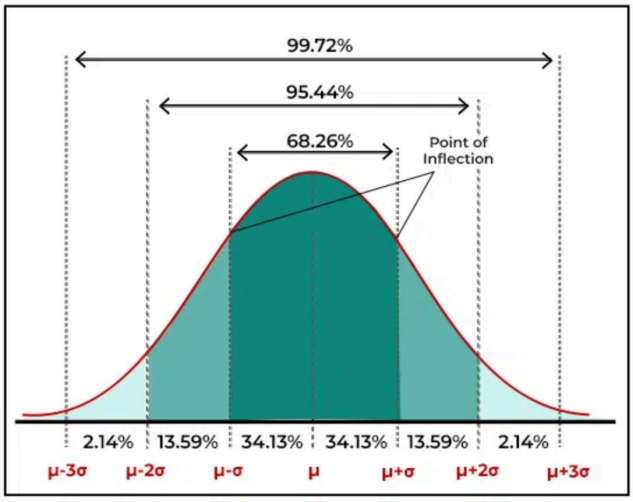
Remember how we explained that the quality of a sample depends not just on the sampling procedure, but also, luck? Well, Central Limit Theorem helps account for the luck aspect.

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

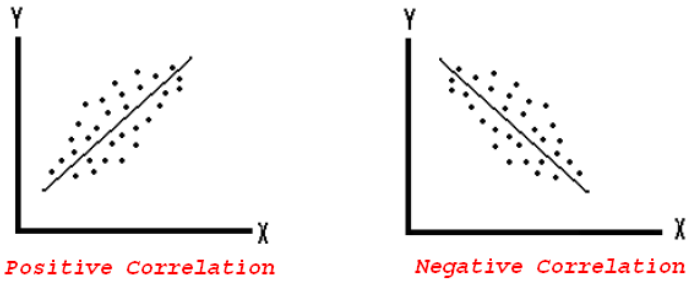
In political science, we are often interested in the relationship between two variables. For example, are oil producers more likely to be democratic? Are more educated voters more likely to turn out and vote?

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?

However, how do we draw a best fit line that fits the data? After all, if we can't figure out what is the best fit, we cannot get a β value to interpret.

- The solution is the Ordinary Least Squares estimator, a way to esimtate β, α ,and all other parameters of the linear line

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

Part II

Regression Anlaysis

Chapter 4

Linear Regression Model

Chapter 5

Extensions of the Linear Model

Chapter 6

Binomial Logistic Regression

Chapter 7

Multinomial and Ordinal Logistic Regression

Chapter 8

Regression for Counts

Part III

Causal Inference

Chapter 9

Causal Frameworks

Chapter 10

Random Experiements

Chapter 11

Selection on Observables

Chapter 12

Instrumental Variables

Chapter 13

Regression Discontinuity

Chapter 14

Differences-in-Differences

Chapter 15

Survey Experiments