Political Econometrics

Using Statistics for Political Science and Political Economy

Kevin Lingfeng Li

Table of contents

Preface

This book is part of a 5-part series on **Political Science and Political Economy**:

1. **Political Econometrics: Using Statstics for Political Science and Political Economy** (This book) introduces many of the core statistical concepts and methods used in Political Science and Social Science research. This book mainly covers Linear Regression models and techniques of Causal Inference.
2. **Political Game Theory: Introduction to Formal Mathematical Models** introduces key concepts in the field of Game Theory, which has become a very popular tool to model and study political situations. This book starts off with Static Games, before moving to Dynamic Games and Bayesian Games.
3. **Political Economy: Using Economic Models to Study Politics** applies the econometric and game theory techniques in the previous two books to the study of current issues in Political Science and Political Economy.
4. **Advanced Statistical Methods: Further Regression, Time Series, and Latent Variables** provides higher-level statistical techniques for Political Science and Political Economy, building on the first book. This book mainly covers logit and count regression models, time series models, and latent variable models.
5. **Social Data Science: Using Machine Learning for Political Science and Political Economy** introduces modern statistical learning techniques that have improved prediction accuracy. The book starts off with prediction and classification methods, then covers model validation, before ending on text mining and quantitative text analysis.

This series is meant to be a relatively approachable introduction to the field. However, as Political Science and Political Economy depends on many economic tools, an rough understanding of Algebra, Single Variable Calculus, and simple Linear Algebra is recommended. 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 - 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.

This book will use the R language for some examples. This is not a coding course, so I will not introduce the basics of R.

Part I

Basics of Probability and Statistics

Chapter 1

Basic Probability

See Mathematics for Political Economy for more detailed explanations.

1.1 Sets and Set Operators

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 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.

1.2 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.
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

1.3 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:

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

1.4 Bayes' Theorem

Bayes' Theorem states the following relationship is true:

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

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

Basic Statistics

See Essential Mathematics for Political Economy for more detailed explanations.

2.1 Random Variables

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.

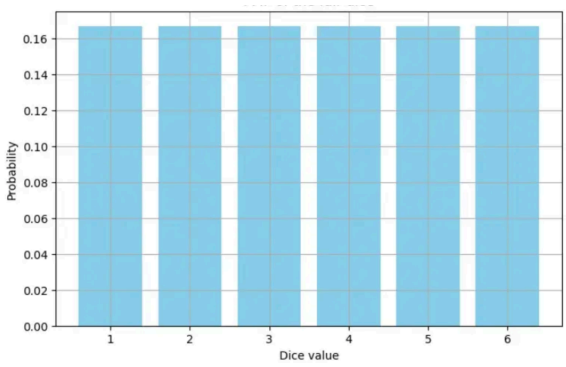
There are two types of random variables:

- **Discrete Random Variables** are random events which have a distinct number of outcomes. For example, rolling a dice has 6 outcomes.
- **Continuos Random Variables** 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 mass/density function** $f(y)$ takes a potential outcome of an event as an input, and outputs the respective probability.

For example, the probability mass/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.

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:

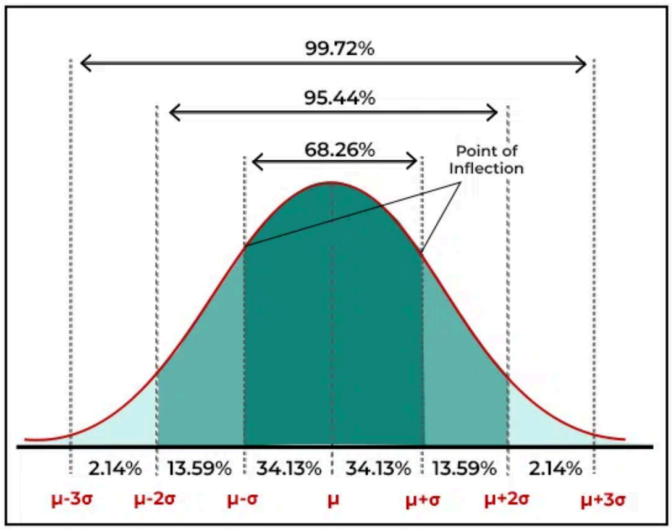
$$Var(X) = \sigma^2 = \frac{1}{n} \sum (X - \mu)^2 = E[(X - \mu)^2]$$

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 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.

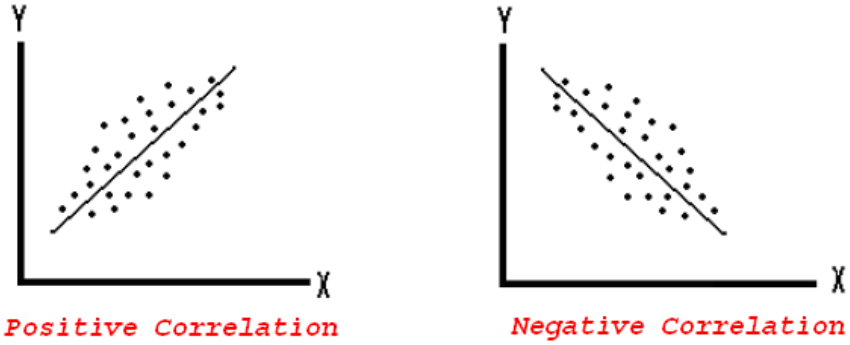
The T distribution is a distribution very similar to the shape and size of the normal distribution, however, generally has thicker tails and a lower peak. The key difference is that t-distributions are defined with only one parameter - degrees of freedom DF.

2.5 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 numerical value of the covariance.

To do that, 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

2.6 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

Multiple Regression Anlaysis

Chapter 3

Linear Model, Estimation, and Interpretation

3.1 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. 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 $\overrightarrow{X}$

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 **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 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).

3.2 Estimation of Parameters

In our model, we have parameters $\alpha, \beta_1, ..., \beta_k$ that need to be estimated 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 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. 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. 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. 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. 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 estimates. Testing every possible set of proposed coefficients $\tilde{\beta}$ is quite time-consuming. 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 the R programming language later).

3.3 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. $\hat{\beta}$ is the expected change in Y, given a one-unit increase in X.

- A positive $\hat{\beta}$ means a positive relationship, a negative $\hat{\beta}$ means a negative relationship, and $\hat{\beta} = 0$ means no relationship.
- This is only for continuous X explanatory variables. See Chapter 5 for categorical/binary explanatory variables.

$\hat{\alpha}$ is the y-intercept of the linear model. $\hat{\alpha}$ 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.

- 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.

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 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 $Y = 1$.

$\hat{\beta}_j$ is the expected change in the probability of getting category $Y = 1$, given a one-unit increase in X_j. We could multiply this by 100 to get the expected percentage point change.

$\hat{\alpha}$ is the expected probability of getting category $Y = 1$, when $X_j = 0$.

An important note is that, in theory, for a linear regression model, Y should be continuous, not binary. In theory, we should be using logistic regression instead of linear regression (which is covered in the Advanced Statistical Methods book).

- However, in practice, researchers often do use linear regression for binary Y, simply because linear regression is easier to interpret, and easier to incorporate into more causal inference methods.

3.4 Model Summary Statistics

Estimated Residual Standard Deviation

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

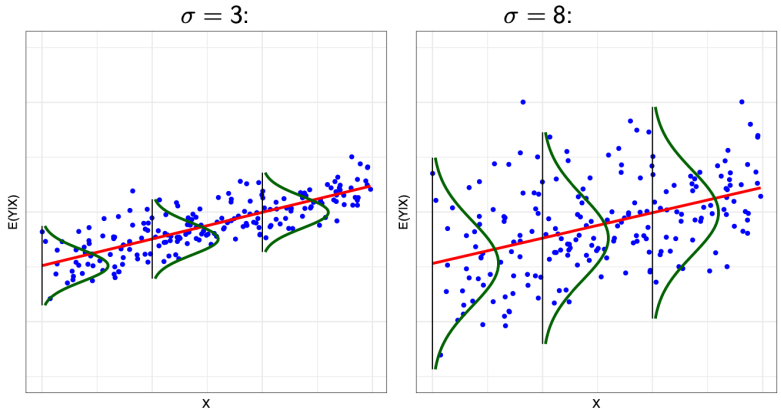
$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - k - 1}$$

But what is the residual variance? Recall the way we write 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

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

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, it is just one metric with its benefits and drawbacks.

3.5 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)
democracy_data <- read_csv("data/democracy.csv")
```

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)
```

For example, let us run a multiple linear regression with two explanatory variables:

```
model2 <- lm(polity_2 ~ GDP_Per_Cap_Haber_Men_2 + Total_Oil_Income_PC,
    data = democracy_data)
summary(model2)
```

```
Call:
lm(formula = polity_2 ~ GDP_Per_Cap_Haber_Men_2 + Total_Oil_Income_PC,
    data = democracy_data)

Residuals:
    Min      1Q  Median      3Q     Max
-51.604  -5.790  -0.162   6.246  40.458

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.775e+00  8.554e-02  -20.75   <2e-16 ***
GDP_Per_Cap_Haber_Men_2  4.764e-04  1.084e-05   43.93   <2e-16 ***
Total_Oil_Income_PC     -1.103e-03  2.850e-05  -38.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.639 on 10267 degrees of freedom
  (6936 observations deleted due to missingness)
Multiple R-squared:  0.1729,    Adjusted R-squared:  0.1727
F-statistic:  1073 on 2 and 10267 DF,  p-value: < 2.2e-16
```

We can see the output. In the coefficients table:

- The *Intercept - Estimate* is $\hat{\alpha}$
- The *GDP_Per_Cap_Haber_Men_2 - Estimate* is the coefficient $\hat{\beta}_1$.
- The *Total_Oil_Income_PC - Estimate* is the coefficient $\hat{\beta}_2$.
- For interpretation, see the preceding section on interpretation.

Underneath, the *Residual Standard Error* is the Residual Standard Deviation $\hat{\sigma}^2$, and the *Multiple R-Squared* is the R^2 value.

Chapter 4

Confidence Intervals and Hypothesis Testing

4.1 Samples and Population

In the social sciences, we are often interested in studying large groups of people and entities. However, it is often impossible to ask every single individual in the population. For example, if we wanted to study the effect of educational level on voter turnout in the UK, we would need to ask nearly 70 million people.

A **sample** is a subset of a population, which ideally, can tell us something about the population. For example, we could use a sample to estimate the relationship between two features X_j and Y with a coefficient β_j. If our sample can reflect the population, then we can use the sample to learn about the true population β_j

The quality of a sample depends on two major factors: the sampling procedure, and luck.

- Sampling procedure can create **bias**: systematic reasons why our samples are not representative of the population.
- Luck in sampling is called **noise** or **variance**: essentially, even if we have a perfect sampling procedure, every sample will differ slightly just due to luck.

The gold standard of sampling procedure is a **random sample** - where individuals in the sample are selected at random from the population. 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 population.

4.2 Sampling Distributions and Standard Error

A sampling distribution is a hypothetical construct that is useful to understanding how many of our statistical techniques work. A sampling distribution is as follows.

- Imagine that we take a sample from a population. Then, we find the $\widehat{\beta}_j$ coefficient between some X_j and Y. That is a **sample estimate**.

- Then, let us take another sample from the same population, 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, and getting more and more sample estimates.
- Now, let us plot all our sample estimates (different $\hat{\beta}_j$ values) into a "histogram" or density plot. The x axis labels the possible β_j 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 **sampling distribution**

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 $se(\hat{\beta}_j)$. Our software will calculate an estimate of this value, which will be notated as $\hat{se}(\hat{\beta}_j)$

- A smaller standard error generally means a more accurate estimate, and is usually due to a larger sample size n or lower residual standard deviation $\hat{\sigma}^2$

4.3 Confidence Intervals

Our OLS estimation (from chapter 3) has produced a $\hat{\beta}_j$ based on the data in one sample. However, remember, the sample is just a fraction of the population - and the true $\hat{\beta}_j$ is unlikely to be exactly the one in our sample, due to sampling variation.

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

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

- $\hat{se}(\hat{\beta}_j)$ is the standard error of our estimate of how precisely we have estimated the true value of β_j, introduced in the previous section.
- The 1.96 can slightly deviate depending on the sample size and number of variables (do not worry, the computer will solve this for us).
- Why 1.96? It is because in a normal distribution, 95% of the data is contained within 1.96 standard deviations (see section 2.4), and Central Limit Theorem states that sampling distributions are normally distributed.

What does the confidence interval represent? Essentially, it means 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 β_j.

Every value in a given confidence interval is a plausible value of the true β_j in the population.

The most important thing is if 0 is included within the confidence interval. $\beta_j = 0$ means that there is no relationship between X_j and Y. If our confidence interval contains 0, that means we cannot be confident that there is no relationship between X_j and Y.

4.4 Hypothesis Testing of Parameters

In academia, we are conservative - this means 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 relationship between variables X_j and Y* (since it is very rare we are studying something that has already been proven). No relationship means that $\beta_j = 0$

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 relationship between variables X_j and Y*, or $\beta_j \neq 0$

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 (the new theory we proposed). Why 95% confidence? It is just tradition - there have been several studies showing there is nothing special about this value.

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

How do we actually test these hypotheses? First, we have to calculate a t-test statistic. The formula for such a statistic is the parameter divided by its standard error (calculated by the computer):

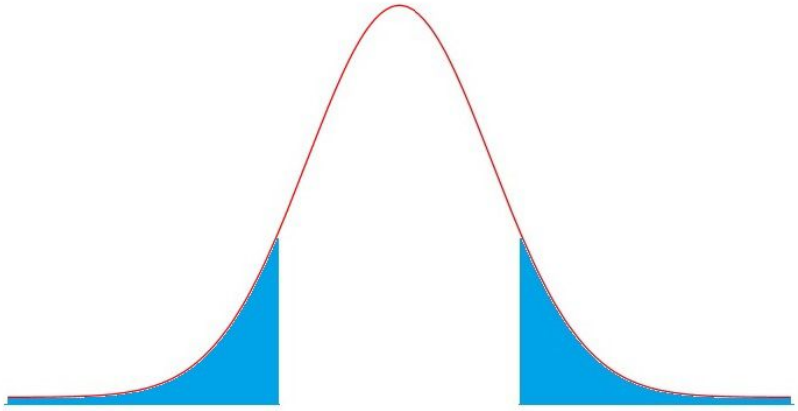
$$t = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$$

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

Then, we have to consult a t-distribution (see section 2.4). T-distributions only has one parameter - degrees of freedom, which is calculated by the 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:



The area highlighted is our 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 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.

4.5 Hypothesis Testing in R

For hypothesis testing, we just run the regression like we would normally (see section 4.5). For example, let us run a model prediction *polity_2* with 2 explanatory variables

```
model <- lm(polity_2 ~ GDP_Per_Cap_Haber_Men_2 + Total_Oil_Income_PC,
    data = democracy_data)
summary(model)
```

```
Call:
lm(formula = polity_2 ~ GDP_Per_Cap_Haber_Men_2 + Total_Oil_Income_PC,
    data = democracy_data)

Residuals:
    Min      1Q  Median      3Q     Max
-51.604  -5.790  -0.162   6.246  40.458

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.775e+00  8.554e-02  -20.75   <2e-16 ***
GDP_Per_Cap_Haber_Men_2   4.764e-04  1.084e-05   43.93   <2e-16 ***
Total_Oil_Income_PC      -1.103e-03  2.850e-05  -38.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.639 on 10267 degrees of freedom
  (6936 observations deleted due to missingness)
Multiple R-squared:  0.1729,    Adjusted R-squared:  0.1727
F-statistic:  1073 on 2 and 10267 DF,  p-value: < 2.2e-16
```

Now, let us look at the coefficients table:

- *GPD_Per_Cap_Haber_Men_2* is our first explanatory variable, *Total_Oil_Income_PC* is our second explanatory variable
- Under the *Std. Error* column is the respective standard errors for both β coefficients. This is used to calculate the t-statistic
- Under the *t value* column is the t-test statistics for both β coefficients. - Under the *Pr(>|t|)* column is the p-values for each test statistic.
- The stars after the p-values are the significance level. One star * means $p < 0.05$, two stars ** means $p < 0.01$, and three stars *** means $p < 0.001$
- For more details of interpretation, see the previous sections in this chapter.

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

```
confint(model)
```

```
                               2.5 %        97.5 %
(Intercept)             -1.9422959405 -1.6069289543
GDP_Per_Cap_Haber_Men_2  0.0004551738  0.0004976894
Total_Oil_Income_PC     -0.0011586240 -0.0010468834
```

As we can see, this command gives us 95% confidence intervals (both lower and upper bounds), for all parameters in our model.

Chapter 5

Explanatory Variable Analysis

5.1 Polynomial Transformations

5.2 Logarithmic Transformations

5.3 Binary Explanatory Variable

5.4 Polytomous Explanatory Variable

5.5 Interaction Effects

5.6 Explanatory Variables in R

Chapter 6

Panel and Clustered Data

6.1 Hierarchical Data

6.2 Fixed Effects

6.3 Further Approaches

Chapter 7

Model Selection for Inference

Part III

Causal Inference for Political Economy

Chapter 8

Causal Frameworks

Chapter 9

Random Experiements

Chapter 10

Selection on Observables

Chapter 11

Instrumental Variables Estimator

Chapter 12

Regression Discontinuity

Chapter 13

Differences-in-Differences

Chapter 14

Survey Experiments