Quantitative Methods and Analysis

Kevin's Guide to Political Science and Political Economy

Kevin Lingfeng Li

Table of contents

Preface

This book is part of a 5-part series: **Kevin's Guide to Political Science and Political Economy**:

1. **Quantitative Methods and Analysis** (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. **Game Theory and Formal 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. **Empirical Political Economy** 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 Quantitative Methods** 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. **Data Science and Political Analysis** 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 book on Quantitative Methods focuses on developing the core statistical toolkit of political science and social science research. We start with a brief review of probability and basic statistics. Then, we introduce and rigorously analyse the Linear Regression model, its form, assumptions, extensions, and model selection. Finally, we discuss how we can achieve causal inference through a variety of techniques, such as randomised experiments, selection on observables, instrumental variables, and quasi-experimental methods.

This book comes with a companion manual - Mathematical Methods for Political Science, which covers some calculus and linear algebra that may be helpful for understanding some of the concepts and math in this book.

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 Mathematical Methods 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.

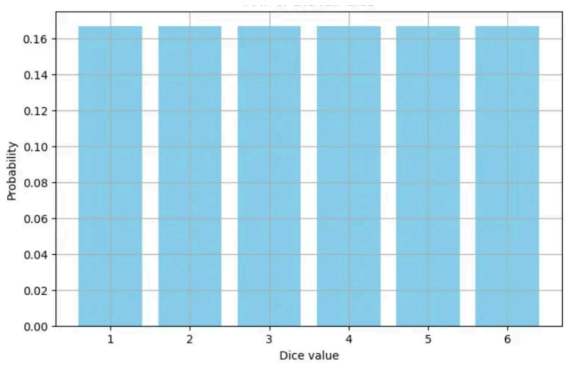
There are two types of random variables:

- **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 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:

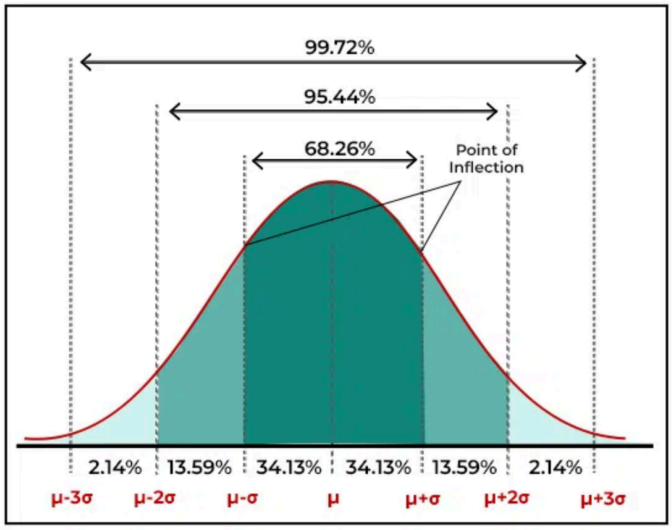
$$Var(X) = \sigma^2 = \frac{1}{n} \sum (X - \mu)^2 = E[(X - \mu)^2]$$

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.

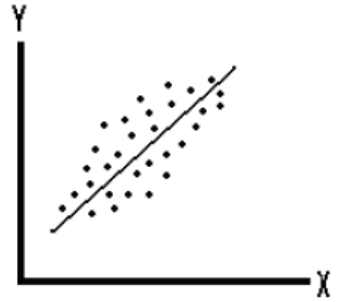
2.5 Covariance and Correlation

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:



Positive Correlation

Negative Correlation

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.

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

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?

- We can interpret the sign of β: a positive β is a positive relationship, a negative β is a negative relationship, and $\beta = 0$ is no relationship.
- We can also interpret the magnitude of β: as X increases by 1 unit, Y is expected to increase by β units.

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:

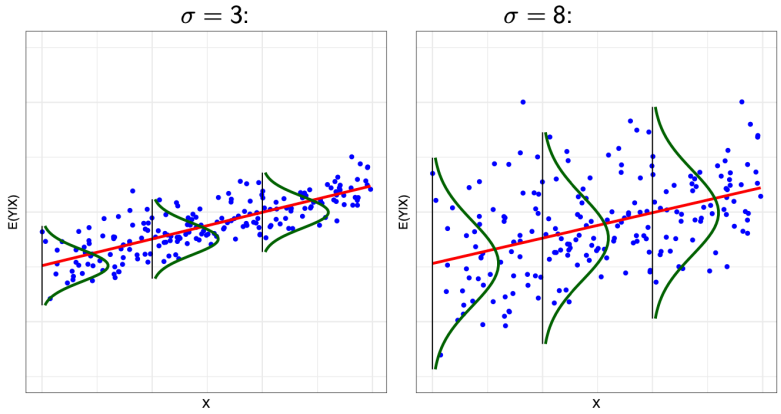
$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - k - 1}$$

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):

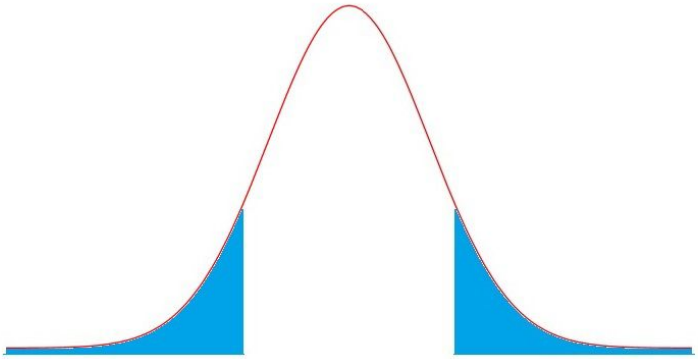
$$t = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$$

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 F-Tests of Nested Models

An F-test is a variation on the coefficient significance tests. The standard F-test is quite simple - it tests for the significance of a model across multiple coefficients:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_1 : \text{ at least one of } \beta_1, \beta_2, ..., \beta_k \neq 0$$

Thus, we assume that all explanatory variables in the model have no relationship to the outcome variable, unless we can reject the null hypothesis and show that at least one coefficient is statistically significant.

For a standard Linear Regression, if we have any significant individual coefficients, then the F-test will also become significant. So what is the point of F-tests? Well, for some models that we will see later, such as categorical explanatory variables, the coefficient's significance levels mean something else. Thus, to see if the model is statistically significant, we need to use the F-test.

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

$$M_0 : E[Y] = \alpha + \beta_1 X_1 + ... + \beta_g X_g$$

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

Importantly, note how the null hypothesis model is entirely contained within the alternate hypothesis model. This must be the case - all explanatory variables in model M_0 must also be in M_a, along with additional explanatory variables in M_a.

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

Let SSE_a and R_a^2 denote the sum of squared errors and R^2 values of model M_a, and SSE_0 and R_0^2 for model M_0. The total number of coefficients of M_a are k_a, and for M_0 is k_0. Mathematically, the F-test statistic is as follows:

$$F = \frac{(SSE_0 - SSE_a)/(k_a - k_0)}{SSE_a/[n - (k_a + 1)]}$$

$$F = \frac{\frac{R_a^2 = R_0^2}{[n-(k_0+1)]-[n-(k_a+1)]}}{(1 - R_a^2)/[n - (k_a + 1)]}$$

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

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

The p-values of the F-statistic show the following:

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

F-tests of nested models can help us determine if we should include certain extra explanatory variables. If the addition of the explanatory variables does not statistically significantly improve the performance of the model, there is little reason to include them, unless we have some other theoretical reason to include them.

4.6 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

Sometimes, a linear (straight-line) best-fit line is a poor description of a relationship. For example, the relationship between two variables could be curved, not straight.

We can model more flexible relationships that are not straight lines, by including a transformation of the variable X that we are interested in. The most common transformations for non-linearity are polynomial transformations, including quadratic and cubic transformations.

Quadratic Transformations

Quadratic transformations of X take the following form:

$$E[Y] = \alpha + \beta_1 X + \beta_2 X^2$$

If you recall from high-school algebra, an equation that takes the form of $y = ax^2 + bx + c$ creates a *parabola*. Indeed, this transformation fits a parabola as the best-fit line. However, there are a few things to consider:

- A true parabola has a domain of $(-\infty, \infty)$. However, our model often does not need to do this. The best-fit parabola is only used for the range of plausible X values, given the nature of our explanatory variable. For example, if X was age, a negative number would make no sense.
- Because the parabola's domain often exceeds our plausible range of X values, the vertex of the parabola (where it changes directions) may not be in our data.
- We always include lower degree terms in our model. For example, in this quadratic (power 2) model, we also include the X term without the square.

To fit a model like this, we simply do the same process of minimising the sum of squared errors.

When we run a quadratic model in R, we get two coefficients: β_1 is attached to the X term, while β_2 is attached to the X^2 term. How do we interpret these coefficients?

- β_1's value is no longer directly interpretable. This is because we cannot "hold all other coefficients constant", since β_2 also contains the same X variable. Thus, we cannot isolate the effect of X and β_1.
- β_2's value also cannot be directly interpreted. However, β_2 can tell us two things. First, if the coefficient of β_2 is statistically significant, we can conclude that there is a non-linear relationship between X and Y. Second, if β_2 is negative, the best-fit parabola will open downwards, and if β_2 is positive, the best-fit parabola will open upwards.

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

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

$$Y = \hat{\alpha} + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

$$\frac{dY}{dX} = 0 + \hat{\beta}_1 + 2\hat{\beta}_2 X$$

Then, set the derivative equal to 0.

$$0 = \hat{\beta}_1 + 2\hat{\beta}_2 X$$

$$-\hat{\beta}_1 = 2\hat{\beta}_2 X$$

$$X = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$$

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

General Polynomial Models

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

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

Each higher order coefficient, if statistically significant, indicates that the relationship between X and Y, is not of the previous highest power.

- For example, if the cubic term β_3 is statistically significant, we can reject a quadratic relationship between X and Y

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

5.2 Logarithmic Transformations

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

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

Thus, the logarithmic transformation takes the following form:

$$E[Y] = \alpha + \beta \log(X)$$

Interpretation of the β coefficient can be a little bit trickier for logarithmic transformations. We could interpret it in the same way we interpret linear regressions: given a one unit increase in the log of X, there is an expected β change in Y.

However, this issue is that this does not really say much - I mean, who knows what a *one unit increase in the log of* X even means?

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

$$\log(X) + A = \log(X) + \log(e^A)$$

$$\log(X) + \log(e^A) = \log(e^A \times X)$$

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

$$E[Y|X] = \alpha + \beta \log(X)$$

$$E[Y|e^A \times X] = \alpha + \beta \log(e^A \times X)$$

$$= \alpha + \beta[\log(X) + A] = \alpha + \beta A + \beta \log(X)$$

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

$$E[Y|e^A \times X] - E[Y|X] = [\alpha + \beta A + \beta \log(X)] - [\alpha + \beta \log(X)]$$

$$E[Y|e^A \times X] - E[Y|X] = \beta A$$

Thus, we can seen that when we multiply X by e^A, we get an expected βA change in Y.

We can make this interpretation more useful by purposely choosing some value A that makes e^A make more sense. For example, if $A = 0.095$, then $e^A = 1.1$. Why is that A value useful? Well, that means when we multiply $X \times e^A$, we are actually doing $X \times 1.1$, which if you remember your percentages, means a 10% increase in X. Thus, increasing X by 10% is associated with an expected change of 0.095β units of Y.

5.3 Binary Explanatory Variable

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

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

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

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

From the above, we can see the following interpretations of our coefficients:

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

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

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

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

Where $\bar{Y}_1$ is the sample mean of Y for observations in category $X = 1$, and $\bar{Y}_0$ is the sample mean of Y for observations in category $X = 0$.

5.4 Polytomous Explanatory Variable

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

How do we run a regression with polytomous explanatory variables? What happens is that we divide the variables into a set of dummy binary variables.

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

Thus, for the n number of categories in X, we would create $n-1$ dummy variables, and input it into a regression equation as follows:

$$E[Y] = \alpha + \beta_{x=0}X_{x=0} + \beta_{x=1}X_{x=1} + ... + \beta_{x=n-1}X_{x=n-1}$$

For example, take the following polytomous variable: *company*, which contains the categories *microsoft, google,* and *apple*.

- Let us create dummy variables for 2 of the 3 categories.
- *Google* will become the first dummy variable X_g. When $X_g = 1$, that observation is part of the *google* category. When $X_g = 0$, that observation is NOT a part of the *google* category.
- *Apple* will become the second dummy variable X_a. When $X_a = 1$, that observation is part of the *apple* category. When $X_a = 0$, that observation is NOT a part of the *apple* category.
- *Microsoft* will not get its own dummy variable. This is because when both *apple* and *microsoft* $X_g = X_a = 0$ that is referring to the *microsoft* category (since these are the only observations not a part of either previous category).

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

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

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

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

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

$$E[Y|X = \text{Microsoft}] = E[Y|X_g = 0, X_a = 0] = \alpha + \beta_g(0) + \beta_a(0) = \alpha$$

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

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

More generally, the β_j of category j's dummy variable, represents the expected difference in Y between category j and the reference/baseline category.

Notice how the coefficient p-values are a difference-of-means test between two categories, and not a statistical significance test of the entire categorical variable *company* that has 3 different categories. To test the statistical significance of the entire categorical variable, we use an F-test.

An F-test, simply, tests a model's explanatory power against the explanatory power of a model where all coefficients $\alpha, \beta_1, ..., \beta_n$ all are equal to 0. Thus, an F-test would allow us to test the significance of the effect of the variable *company* on Y, which our individual β coefficients do not tell us.

5.5 Interaction Effects

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

This is quite common in the real world. For example, imagine outcome variable Y to be the severity of a car crash. X_1 can represent the darkness of the road at the time of the car crash. X_2 can represent the slipperiness of the road at the time of the car crash. We could quite reasonably expect that as the road is more slippery, i.e. X_2 increases, the darkness of the road X_1's effect on the severity of a car crash Y might be stronger, since slippery further enhances the danger of dark roads.

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

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

$$E[Y] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

We can mathematically show that the effect of X_1 on Y is not constant - and varies due to the

value of X_2. We show this through finding the partial derivative of X_1 on Y, since the derivative is, by definition, the function of the rate of change between X_1 and Y.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

$$\frac{\partial \hat{Y}}{\partial X_1} = 0 + \hat{\beta}_1 + 0 + \hat{\beta}_3 X_2$$

$$\frac{\partial \hat{Y}}{\partial X_1} = \hat{\beta}_1 + \hat{\beta}_3 X_2$$

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

We can also find the effect of X_2 on Y:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

$$\frac{\partial \hat{Y}}{\partial X_2} = \hat{\beta}_2 + \hat{\beta}_3 X_1$$

With these equations, we can interpret the coefficients of our model:

- $\hat{\beta}_1$ is the relationship between X_1 and Y, given $X_2 = 0$.
- $\hat{\beta}_2$ is the relationship between X_2 and Y, given $X_1 = 0$.
- $\hat{\beta}_3$ represents two things. For every one unit increase of X_2, the magnitude of the relationship between X_1 and Y changes by $\hat{\beta}_3$. Similarly, for every one unit increase of X_1, the magnitude of the relationship between X_2 and Y changes by $\hat{\beta}_3$
- α is still the expected value of Y when all explanatory variables equal 0.

The coefficient β_3's significance level tells us if there is a statistically significant interaction. If β_3 is not statistically significant, we can often remove the interaction term. However, if β_3 is statistically significant, that means we have found two terms that interact.

Often times, our moderating effect X_2 is a binary variable (for example, developed/developing country, true/false, yes/no). In this scenario:

- $\hat{\beta}_1$ is the relationship between X_1 and Y when X_2 is in the category $X = 0$.
- $\hat{\beta}_1 + \beta_3$ is the relationship between X_1 and Y when X_2 is in the category $X = 1$.
- $\hat{\beta}_3$ is the difference in the magnitude of the relationship between X_1 and Y, between the categories $X = 1$ and $X = 0$.

You can get all of these interpretations above simply by plugging in $X_2 = 0$ and $X_2 = 1$ into the previous equations we have found.

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

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