Introduction to Econometrics

Michael Bar

November 26, 2018

San Francisco State University, department of economics.
## Contents

1 Preliminaries  
   1.1 Probability Spaces ........................................... 1  
   1.2 Random Variables ........................................... 3  
   1.3 Random Vectors ........................................... 7  
      1.3.1 Marginal pdf ........................................... 7  
      1.3.2 Conditional pdf ....................................... 8  
      1.3.3 Independence of random variables ...................... 11  
   1.4 Moments .................................................. 12  
      1.4.1 Expected value ......................................... 12  
      1.4.2 Variance and standard deviation ...................... 15  
      1.4.3 Covariance and correlation ........................... 19  
   1.5 Normal Distribution and Its Relatives .................... 25  
      1.5.1 Normal distribution .................................. 25  
      1.5.2 Chi-square distribution ............................... 30  
      1.5.3 t-distribution ......................................... 31  
      1.5.4 F-distribution ......................................... 33  
   1.6 Sampling and Estimators ................................... 33  
      1.6.1 Unbiased estimators ................................... 35  
      1.6.2 Efficient estimators ................................... 37  
      1.6.3 Conflicts between bias and low variance (MSE criterion)  ...................... 38  
      1.6.4 Consistent estimators .................................. 40  
      1.6.5 Distribution of estimators ............................ 42  

2 Simple Regression Analysis  
   2.1 Deriving the OLS Estimators ................................ 46  
   2.2 Graphical Illustration of OLS ............................... 49  
   2.3 Interpretation of Estimated Regression .................... 49  
   2.4 Results About the Estimated Model ......................... 50  
   2.5 Goodness of Fit: $R^2$ ...................................... 51  
   2.6 Impact of Changes in Units ................................. 52  
      2.6.1 Multiplying the data by constant ...................... 52  
      2.6.2 Adding a constant to the data ......................... 53  
      2.6.3 Multiplying by constant and adding a constant to the data ...................... 53  
   2.7 Restricted Models ......................................... 54  
      2.7.1 Model with no intercept, $\beta_1 = 0$ .................. 54
CONTENTS

2.7.2 Model with no slope, $\beta_2 = 0$ ................................................. 55

3 Properties of Regression Coefficients and Hypothesis Testing ................................................. 57
  3.1 Assumptions ................................................................................. 57
  3.2 Mean and Variance of OLS Estimators ........................................... 58
    3.2.1 Rewriting The OLS Estimators .................................................. 59
    3.2.2 Unbiasedness of OLS estimators ............................................... 60
    3.2.3 Efficiency of OLS estimators .................................................. 61
    3.2.4 Consistency of OLS estimators .................................................. 67
  3.3 General Idea of Statistical Hypothesis Tests ...................................... 67
  3.4 Hypothesis Testing of Regression Coefficients ..................................... 72
    3.4.1 Formulating hypotheses about the regression coefficients ............... 72
    3.4.2 Deriving the t-test for regression coefficients .............................. 73
    3.4.3 Performing the t-test ............................................................... 76

4 Multiple Regression Analysis ................................................................. 79
  4.1 OLS Estimators ............................................................................. 79
    4.1.1 Deriving the OLS estimators ..................................................... 79
    4.1.2 Properties of OLS estimators ................................................... 81
    4.1.3 Interpretation of regression estimates ........................................... 81
  4.2 Hypothesis Testing ....................................................................... 83
    4.2.1 t-tests ................................................................. 83
    4.2.2 p-value ................................................................. 83
    4.2.3 Confidence intervals ............................................................ 86
    4.2.4 Goodness of fit: $R^2$ and F-tests ............................................. 87
  4.3 Multicollinearity ............................................................................ 89

5 Transformations of Variables and Nonlinearities .................................................. 95
  5.1 Non-linearity in Variables ............................................................... 95
  5.2 Nonlinearity in Variables and Parameters ........................................... 96
  5.3 Models with Logarithmic Terms ....................................................... 97
    5.3.1 Log-log model ................................................................. 97
    5.3.2 Log-lin model ................................................................. 98
    5.3.3 Lin-log model ................................................................. 99
    5.3.4 Lin-lin model ................................................................. 99
    5.3.5 Summary of models with log forms ......................................... 100
  5.4 Estimating Elasticity in General Models ........................................... 100
  5.5 Quadratic Form ........................................................................... 101

6 Qualitative Regressors and Dummy Variables ................................................. 103
  6.1 Using Dummies to Represent Qualitative Variables .............................. 103
  6.2 Dummy Variables in Regression Analysis ........................................... 104
    6.2.1 One dummy variable ......................................................... 105
    6.2.2 Several dummy variables ..................................................... 106
    6.2.3 Interaction between two dummy variables .................................. 107
7 Specification of Regression Models

6.2.4 Interaction between a dummy and a quantitative variable . . . . . . . 108
6.2.5 Dummy variables in log-lin model . . . . . . . . . . . . . . . . . . . . 109

8 Heteroscedasticity

7.1 Omitted Variable Bias . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
7.1.1 Two variable model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
7.1.2 General k-variable model . . . . . . . . . . . . . . . . . . . . . . . . . . 113
7.2 Inclusion of Irrelevant Variable . . . . . . . . . . . . . . . . . . . . . . . . . 114
7.3 Proxy Variables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115
7.4 Steps of Applied Research . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116

8.1 Illustrating Heteroscedasticity . . . . . . . . . . . . . . . . . . . . . . . . . 117
8.2 Consequences of Heteroscedasticity . . . . . . . . . . . . . . . . . . . . . . 118
8.3 Detection of Heteroscedasticity . . . . . . . . . . . . . . . . . . . . . . . . . 118
8.3.1 Eyeballing (plotting the residuals) . . . . . . . . . . . . . . . . . . . . . 118
8.3.2 The Goldfeld-Quand test . . . . . . . . . . . . . . . . . . . . . . . . . . 121
8.3.3 The White test . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122
8.4 Remedies for Heteroscedasticity . . . . . . . . . . . . . . . . . . . . . . . . . 122
8.4.1 Heteroscedasticity-robust standard errors . . . . . . . . . . . . . . . . 122
8.4.2 Weighted least squares (WLS) . . . . . . . . . . . . . . . . . . . . . . . 123

9 Discrete Choice Models

9.1 The Probit . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 127
9.2 Logit Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 128
9.3 Estimation of the Probit and Logit Models . . . . . . . . . . . . . . . . . . 129

10 Time Series

10.1 Regression with Time Series Data . . . . . . . . . . . . . . . . . . . . . . . 133
10.1.1 Autocorrelation (serial correlation) . . . . . . . . . . . . . . . . . . . . 134
10.1.2 Trends . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136
10.2 Panel Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 140
Chapter 1

Preliminaries

In this chapter we review some basic concepts in probability and statistics. This chapter is not a substitute for an introductory course in statistics, such as ECON 311. My goal here is to review some key concepts for those who are already familiar with them, but might need a speedy refresher.

1.1 Probability Spaces

The most fundamental concept in probability theory is a random experiment.

Definition 1 A random experiment is an experiment whose outcome cannot be predicted with certainty, before the experiment is run.

For example, a coin flip, tossing a die, getting married, getting a college degree, investing in financial asset, are all random experiments. Although we cannot predict with certainty the exact outcome of a coin flip, but we know that it can be either "heads" or "tails". Similarly, in a toss of a die we know that the outcome is one of the following: \{1, 2, 3, 4, 5, 6\}.

Definition 2 A sample space, \(\Omega\), is the set of all possible (and distinct) outcomes of a random experiment. Specific outcomes are denoted by \(\omega \in \Omega\).

Thus, the sample space for the random experiment of a coin flip is \(\Omega = \{\text{"heads"}, \text{"tails"}\}\), and the sample space of the random experiment of a die toss is \(\Omega = \{1, 2, 3, 4, 5, 6\}\). It is less clear what is the sample space for the random experiments of getting married or getting a college degree. In the first, one can look at the number of kids as an outcome, or the duration the marriage will last. In the second, one can look at the types of jobs the person gets, his wage, etc.

As another example, what is the sample space of flipping two coins? Let heads and tails be denoted by H and T. Then the sample space is \(\Omega = \{HH, HT, TH, TT\}\), that is, the sample space consists of 4 outcomes.

Exercise 1 Write the sample space of flipping 3 coins.

Exercise 2 Write the sample space of tossing two dice.
Sometimes we don’t care about all the particular outcomes of an experiment, but only about groups of outcomes. For example, consider the random experiment of taking a course. The set of all possible numerical grades is \( \Omega = \{0, 1, \ldots, 100\} \). However, if a letter grade of "A" is assigned to all grades of 92 and above, then we might be particularly interested in the subset of \( \Omega \) which gives an "A": \( \{92, 93, \ldots, 100\} \).

**Definition 3** A random event is a subset of the sample space, denoted \( A \subseteq \Omega \).

In the previous example of a grade in a course, we can define an event \( A = \{92, 93, \ldots, 100\} \). If one of the outcomes in \( A \) occurs, we say that event \( A \) occurred. Notice that \( A \subseteq \Omega \), which reads "\( A \) is a subset of \( \Omega \)". As another example, suppose that a student fails a course if his grade is below 60. We can define an event \( F = \{0, 1, \ldots, 59\} \). Again, if one of the outcomes in \( F \) occurs, we say that event \( F \) occurred.

**Exercise 3** Let the random experiment be flipping two coins, and \( E \) be the event in which the two coins have identical outcomes. List the outcomes in \( E \).

**Exercise 4** Let the random experiment be a toss of a die, and \( O \) be the event in which the outcome is odd. List the outcomes in \( O \).

Finally, we would like to calculate the probability of an event. Intuitively, probability of an event is the "chance" that an event will happen. The precise definition of probability is beyond the scope of this course.

**Definition 4** The probability of a random event \( A \), denoted \( P(A) \), is the relative frequency of occurrence of outcomes in that event, when repeating the experiment many times.

For example, in a toss of a balanced coin, if we repeat the experiment many times, we expect to get "heads" half of the time, and "tails" half of the time. We then say that "the probability of heads is 1/2 and the probability of tails is also 1/2". Similarly, in tossing a balanced die, we expect each of the outcomes to occur with probability 1/6. Let the event \( E \) denote an even outcome in a toss of a die. Thus, \( E = \{2, 4, 6\} \). What is the probability of \( E \)? Intuitively, the probability of this event is the sum of the probabilities of each outcome, i.e. \( P(E) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \). In general, the probability of event \( E \) is calculated by adding up (integrating) the probabilities of the distinct outcomes in \( E \).

It should be obvious that \( P(\Omega) = 1 \) from the definition of the sample space. Since the sample space contains all possible outcomes of a random experiment, the probability that some outcome will occur is 1. It should also be obvious that probability of any event must be between 0 and 1, and your instructors will be very alarmed if you calculate negative probabilities or probabilities that are greater than 1.

In the examples of a coin flip and a toss of a die, each distinct outcome occurs with equal probability. This need not always be the case. Consider the experiment of picking a person at random and recording his/her education level (highest degree earned). The sample space can be for example \( \Omega = \{"less-than-high school", "high school", "BA", "MA", "Ph.D."\} \). If the probabilities of BA, MA and Ph.D. are: \( P(\text{BA}) = 0.25 \), \( P(\text{MA}) = 0.1 \), \( P(\text{Ph.D.}) = 0.02 \), then the probability of the event "education is at least BA" is:

\[
P(\text{Educ} \geq \text{BA}) = P(\text{BA}) + P(\text{MA}) + P(\text{Ph.D.}) = 0.25 + 0.1 + 0.02 = 0.37
\]
1.2 Random Variables

Notice that some random experiments have numerical outcomes (e.g. as a toss of a die) while others have outcomes that we described verbally (e.g. "heads" and "tails in a flip of a coin, or education level of a person, "high school", "college degree",...). Calculating probabilities and performing statistical analysis becomes much simpler if we could describe all outcomes in terms of numbers.

**Definition 5** A random variable is a function which assigns a real number to each outcome in the sample space. Formally, \( X : \Omega \to \mathbb{R} \) is the notation of a function (named \( X \)), which maps the sample space \( \Omega \) into the real numbers \( \mathbb{R} \). The particular number assigned to a given outcome is denoted by \( x \), and is called a realization of random variable \( X \).

From the above definition, note that if \( X \) is a random variable and \( g(\cdot) \) is some real valued function, then \( g(X) \) is another random variable. It is conventional to use capital letters to denote the random variable’s name, and small case letters to denote particular realizations of the random variable. For example, we had seen that the sample space of flipping two coins is \( \{HH, HT, TH, TT\} \). One can define a random variable \( X \), which counts the number of times that "heads" is observed. The possible values of \( X \) (called the support of \( X \)) are \( x \in \{0, 1, 2\} \). We can then calculate the probabilities \( P(X = 0) \), \( P(X = 1) \) and \( P(X = 2) \), in other words we can find the distribution of the random variable \( X \).

**Example 1** Calculate the probability distribution of the random variable \( X \), which counts the number of "heads" in a two coin flip. That is, find \( P(X = 0) \), \( P(X = 1) \) and \( P(X = 2) \).

The probability of zero heads is the probability that both coins are "tails", which is \( P(X = 0) = P(TT) = P(coin1 = T) \cdot P(coin2 = T) = 0.5 \cdot 0.5 = 0.25 \). The reason for multiplication will become clear later, after we define independence. In general, \( P(coin1 = T \text{ and coin2} = T) = P(coin1 = T) \cdot P(coin2 = T | coin1 = T) \), where the last term is the probability of coin2 = T, given that coin1 = T. Since the coins are independent, i.e. the outcome of coin2 does not depend on outcome of coin1, we have \( P(coin2 = T | coin1 = T) = P(coin2 = T) \). Similarly, \( P(X = 2) = P(HH) = P(coin1 = H) \cdot P(coin2 = H) = 0.5 \cdot 0.5 = 0.25 \). Finally, \( P(X = 1) = P(HT) + P(TH) = 0.25 + 0.25 = 0.5 \). In summary, the distribution of \( X \) is presented in the next table:

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = x) )</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>

As another example, consider the random experiment to be a course with 20 students, with the relevant outcomes being success or failure. Suppose that I want to calculate the probability that all students pass the course, or that 1 student fails and 19 pass, or that 2 students fail and 18 pass, etc. I can define a random variable \( X \) to be the number of students who fail the course. Thus, the possible values of \( X \) are \( \{0, 1, 2, ..., 20\} \). If I know from previous experience that with probability \( p \) a student fails my course, we can calculate the distribution of \( X \). Roughly speaking, we can calculate \( P(X = 0) \), \( P(X = 1), ..., P(X = 20) \).
Definition 6 A discrete random variable is one that has a finite or countable number of possible values. This means that the possible values can be enumerated with integers 1, 2, ...

Definition 7 A continuous random variable is one that has a continuum of possible values. This means that all possible values are real numbers in some intervals.

The distribution of a random variable is usually described by its probability density function (pdf). The probability density function for a discrete random variable assigns probabilities to each value of the random variable:

\[ f(x) = P(X = x) \]

Here \( X \) is the name of the random variable, and \( x \) is one possible realization. Another popular notation for discrete pdf is

\[ p_i = P(X = x_i) \]

where the outcomes of \( X \) are indexed by \( i \) (i.e., the list of outcomes is \( x_1, x_2, ..., x_n \), and generic outcome is \( x_i \), \( i = 1, 2, ..., n \)). The probability density function of a continuous random variable can be used to calculate the probability that the random variable gets values in a given interval \([a, b]\), by integrating the pdf over that interval:

\[ \int_a^b f(x) \, dx = P(a \leq X \leq b) \]

We can now formally define the support of a random variable.

Definition 8 The support of a random variable is the set of all the values for which the pdf is positive. Written mathematically,

\[ \text{support}(X) = \{x \mid f(x) > 0\} \]

which reads "the support of \( X \) is the set of values \( x \) such that the pdf is positive". This is the same as saying that the support of \( X \) is the set of all possible values of \( X \), because values for which the pdf is zero are not possible.

Example 2 The random variable that counts the number of failing students, among 20 who take the course, is an example of a discrete random variable. It has 21 possible values (the support is \( x \in \{0, 1, 2, ..., 20\} \)) and a binomial distribution. We write \( X \sim \text{Bin}(n, p) \) to indicate that there are \( n \) students, each fails with probability \( p \). If there are 20 students and \( p = 5\% \), we write \( X \sim \text{Bin}(20, 0.05) \). In case you are curious, the pdf of \( X \) is

\[ f(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \]

where

\[ \binom{n}{x} = \frac{n!}{x! (n-x)!} \]

\(^1\)The pdf of a discrete random variable is often called the probability mass function (pmf).
The next figure illustrates the pdf of $\text{Bin}(20, 0.05)$:

![Distribution of # of failed students in a class of 20](image)

The height of the bars indicates the probabilities of each number of failed students. Notice that a high number of failures is very unlikely, which is a good news.

**Exercise 5** Suppose that 20 students enroll in a course, and their instructor knows from experience that each of them fails with probability $0.05$. Calculate the probability that exactly 7 students fail the course.

An example of a continuous random variable is the **continuous uniform distribution**. If $X \sim U[a, b]$, we say that $X$ has a uniform distribution on the interval $[a, b]$. The probability density function of $X$ is $f(x) = \frac{1}{b-a} \forall x \in [a, b]$, and $f(x) = 0$ otherwise. Thus, the support of this random variable is $x \in [a, b]$. The next figure plots the pdf of a continuous uniform distribution.
Suppose that we have two numbers, $g, h$, such that $a \leq g \leq h \leq b$. The probability that $X \in [g, h]$ is

$$P(g \leq X \leq h) = \int_g^h \frac{1}{b-a} \, dx = \frac{h-g}{b-a}$$

In other words, the probability that a continuous uniform random variable falls within some interval $[g, h]$ is equal to the relative length of that interval to the support.

It should be obvious that sum (in discrete case) or integral (in continuous case) of all the values of a probability density function is 1. This has to be true, because a random variable assigns a number to any outcome in the sample space, and since $P(\Omega) = 1$, the probability of the random variable getting some number is also 1. Formally,

- $[X$ is discrete$]$ : $\sum_x f(x) = 1$, or $\sum_i p_i = 1$
- $[X$ is continuous$]$ : $\int_{-\infty}^{\infty} f(x) \, dx = 1$

In fact, any function which integrates (or sums) to 1 and is nonnegative (never attains negative values) is a probability density function of some random variable.

Definition 9 The cumulative distribution function (cdf) of a random variable $X$ gives the probability that the value is less than or equal to some $x$:

$$F(x) = P(X \leq x)$$

Calculation of the cdf follows from the definitions of the pdf:

- $[X$ is discrete$]$ : $F(x) = \sum_{t \leq x} f(t)$
- $[X$ is continuous$]$ : $F(x) = \int_{-\infty}^{x} f(t) \, dt$

The complementary probability is given by the survival function $S(x) = 1 - F(x)$. This is the probability that a random variable $X$ attains a value greater than $x$.

Exercise 6 Suppose that $X$ is a continuous uniform random variable, with support $[0, 1]$. Find the probability $F(0.12) = P(X \leq 0.12)$.

Exercise 7 Suppose that $X$ is a continuous uniform random variable, with support $[0, 1]$. Find the probability $S(0.97) = P(X > 0.97)$.

Exercise 8 Suppose that $X$ is a continuous uniform random variable, with support $[a, b]$. Find the critical value $h$ such that $P(X \geq h) = 0.05$.

---

\[ ^2 \] Mathematicians often don't even bother using the word sum, because integral is also a sum. The integral symbol, $\int$, comes from an elongated letter $s$, standing for summa (Latin for "sum" or "total").
Exercise 9 Suppose that $X$ is a continuous uniform random variable, with support $[a, b]$. Find the critical value $g$ such that $P(X \leq g) = 0.05$.

Exercise 10 Suppose that $X$ has exponential distribution, denoted $X \sim EXP(\beta), \beta > 0$, i.e. the pdf is:

$$f(x) = \begin{cases} \beta e^{-\beta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

(a) Prove that the above function is indeed a probability density function.
(b) Calculate the critical value $x^*$ such that $P(X \geq x^*) = 0.05$.

1.3 Random Vectors

Often in economics, we want to explore several random variables at the same time, i.e. jointly. For example, suppose that we choose people at random and look at how their education and wages are distributed together. Letting $X$ denote the years of education and $Y$ denote hourly wage, we can describe the joint distribution of $X$ and $Y$ with a joint probability density function (joint pdf). If $X$ and $Y$ are discrete, their joint pdf is

$$f(x, y) = P(X = x, Y = y)$$

If $X$ and $Y$ are continuous, the probability that $X \in [a, b]$ and $Y \in [c, d]$ is

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) \, dxdy$$

Obviously this discussion can be generalized to any number of random variables, say $X_1, \ldots, X_n$, with joint pdf $f(x_1, \ldots, x_n)$. For example, when the unit of observation is a randomly chosen household, we might be interested in the joint distribution of number of people in the household, their genders, their incomes, their education levels, their work experience, their age, their race, health, etc.

The joint pdfs must also integrate (or sum) to 1, just as in the single random variable, because we know with certainty that $X$ and $Y$ will attain some values. Formally,

[X and Y are discrete] : $\sum_y \sum_x f(x, y) = 1$

[X and Y are continuous] : $\int_{-\infty}^\infty \int_{-\infty}^\infty f(x, y) \, dxdy = 1$

Again, the above can be generalized to joint pdf of any number of random variables.

1.3.1 Marginal pdf

In relation to the joint pdf $f(x, y)$, the functions $f(x)$ and $f(y)$ are called marginal or (individual) probability density functions. The marginal pdfs are obtained by

[X and Y are discrete] : $f(x) = \sum_y f(x, y)$

f(y) = \sum_x f(x, y)

[X and Y are continuous] : $f(x) = \int_{-\infty}^\infty f(x, y) \, dy$

$f(y) = \int_{-\infty}^\infty f(x, y) \, dx$
Intuitively, $X$ can attain a given value of $x$ when $Y$ attains the value $y_1$ or $y_2$ or any other possible value. Thus, to get $P(X = x)$ we need to sum (integrate) over all these cases of $Y$ getting any of its possible values.

A point of caution is needed here. The marginal densities of $X$ and $Y$ do not need to be the same functions. Some texts are using the notations $f_X(x)$ and $f_Y(y)$ to make it clear in equations (1.1) and (1.2) that the marginal pdfs are in general not the same function; $X$ has pdf $f_X$ and $Y$ has pdf $f_Y$. These notes follow other texts, which give up some mathematical precision in favor of ease of notation.

For example, suppose that I am looking at the joint distribution of gender, $X$, attaining the value of 1 for female and 0 for male, and GPA, $Y$, attaining values $y_1, \ldots, y_n$, of students in this course. Suppose I want to obtain the marginal distribution of the gender $X$. Then, using equation (1.1), we have the probability of female and male:

\[
\begin{align*}
  f(1) &= P(X = 1) = f(1, y_1) + \cdots + f(1, y_n) = \sum_y f(1, y) \\
  f(0) &= P(X = 0) = f(0, y_1) + \cdots + f(0, y_n) = \sum_y f(0, y)
\end{align*}
\]

Written in another way, a female or a male student can have any of the GPA values $y_1, \ldots, y_n$, and therefore

\[
\begin{align*}
  P(\text{female}) &= P(\text{female and GPA} = y_1) + \cdots + P(\text{female and GPA} = y_n) \\
  P(\text{male}) &= P(\text{male and GPA} = y_1) + \cdots + P(\text{male and GPA} = y_n)
\end{align*}
\]

Exercise 11 Consider the function:

\[
f(x, y) = \begin{cases} 
  2 - x - y & 0 \leq x \leq 1; \ 0 \leq y \leq 1 \\
  0 & \text{otherwise}
\end{cases}
\]

(a) Prove that

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1
\]

In other words, prove that $f(x, y)$ is a probability density function.

(b) Find the marginal pdfs $f(x)$ and $f(y)$.

1.3.2 Conditional pdf

In econometrics, we are often interested in the behavior of one variable, conditional on given values of another variable. For example, I might want to study how well female students are doing in statistics courses, compared with male students. This is the study of grades pdf, given that the student is female, and grades pdf, given that the student is male. As another example, one might want to study the distribution of wages, given that the level of education is "college" v.s. the distribution of wages given that the education level is "MBA". Again, this is the study of the pdf of wages, conditional on education level.

\[\text{Exercise 11 Consider the function:}\]

\[
f(x, y) = \begin{cases} 
  2 - x - y & 0 \leq x \leq 1; \ 0 \leq y \leq 1 \\
  0 & \text{otherwise}
\end{cases}
\]

(a) Prove that

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1
\]

In other words, prove that $f(x, y)$ is a probability density function.

(b) Find the marginal pdfs $f(x)$ and $f(y)$.
1.3. RANDOM VECTORS

**Definition 10** Conditional probability density function of $Y$ given $X = x$ is given by

$$f (y|x) = \frac{f (x,y)}{f (x)} \quad (1.3)$$

In cases when $f (x) = 0$, we define $f (y|x) = 0$.

The left hand side of equation (1.3) is sometimes written as $f (y|X = x)$, to emphasize that this is the conditional density of $Y$ given that $X$ is fixed at a given value of $x$. The vertical bar with $x$ following it, $|x$, means that we are conditioning on $x$ (or it is given that $X = x$). For example, the conditional pdf of wages ($Y$), given that education is at level 3, ($X = 3$), is written

$$f (y|X = 3) = \frac{f (x,y)}{P(X = 3)}$$

or in short

$$f (y|3) = \frac{f (x,y)}{f_X (3)}$$

One can think of whatever is written after the bar as information. If two variables, $X$ and $Y$ are related, then the information on one of them should restrict the possible values of other. For example, consider the random experiment of tossing a die, and let $Y$ be the result of a toss. Thus, the possible values of $Y$ (the support of $Y$) are $\{1, 2, 3, 4, 5, 6\}$. The pdf of $Y$ is $f (y) = 1/6$ for $y = 1, ..., 6$. Now suppose that we provide an information that the result of a toss is an even number. We can provide this information by defining a random variable $X = 1$ if $y$ is even and $X = 0$ if $y$ is odd. What is the conditional pdf of $Y$ given that $X = 1$? We know that conditional on $X = 1$, $Y$ can only attain values $\{2, 4, 6\}$. Intuitively, these values can be attained with equal probabilities, so $P(Y = 2|X = 1) = P(Y = 4|X = 1) = P(Y = 6|X = 1) = 1/3$. This pdf can be obtained using the definition of conditional density in equation (1.3), and the fact that even numbers in a toss of a die occur with probability 1/2:

$$f (2|X = 1) = \frac{f (1,2)}{P(X = 1)} = \frac{\left(\frac{1}{6}\right)}{\left(\frac{1}{2}\right)} = \frac{1}{3}$$

$$f (4|X = 1) = \frac{f (1,4)}{P(X = 1)} = \frac{\left(\frac{1}{6}\right)}{\left(\frac{1}{2}\right)} = \frac{1}{3}$$

$$f (6|X = 1) = \frac{f (1,6)}{P(X = 1)} = \frac{\left(\frac{1}{6}\right)}{\left(\frac{1}{2}\right)} = \frac{1}{3}$$

**Exercise 12** Let $Y$ be the result of a die toss, and $X$ an indicator of even toss: $X = 1$ if $y$ is even and $X = 0$ if $y$ is odd. Calculate the conditional probabilities $f (1|X = 1)$, $f (3|X = 1)$, $f (5|X = 1)$.

**Exercise 13** Consider the following joint pdf of Gender ($X = 1$ if female, 0 if male) and GPA ($Y$, with possible values 1,2,3,4).

<table>
<thead>
<tr>
<th>Gender</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.05</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>M</td>
<td>0.1</td>
<td>0.15</td>
<td>0.125</td>
<td>0.075</td>
</tr>
</tbody>
</table>
(a) Calculate the marginal pdfs of $X$ and $Y$.
(b) Calculate the conditional pdfs of GPA ($Y$), given the Gender ($X$): $f(y|X = 1)$ and $f(y|X = 0)$.

Solution 1 (a) The marginal pdfs are obtained by

$$f(x) = P(X = x) = \sum_y f(x, y), \quad f(y) = P(Y = y) = \sum_x f(x, y)$$

Thus, the pdf of $X$ is:

$$P(X = 0) = 0.1 + 0.15 + 0.125 + 0.075 = 0.45$$
$$P(X = 1) = 0.05 + 0.1 + 0.3 + 0.1 = 0.55$$

The pdf of $Y$ is:

$$P(Y = 1) = 0.05 + 0.1 = 0.15$$
$$P(Y = 2) = 0.1 + 0.15 = 0.25$$
$$P(Y = 3) = 0.3 + 0.125 = 0.425$$
$$P(Y = 4) = 0.1 + 0.075 = 0.175$$

(b) The conditional pdfs are obtained by

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Thus, the distribution of GPA, given that the student is male, is:

$$P(Y = 1|X = 0) = \frac{0.1}{0.45}$$
$$P(Y = 2|X = 0) = \frac{0.15}{0.45}$$
$$P(Y = 3|X = 0) = \frac{0.125}{0.45}$$
$$P(Y = 4|X = 0) = \frac{0.075}{0.45}$$

The distribution of GPA, given that the student is female, is:

$$P(Y = 1|X = 1) = \frac{0.05}{0.55}$$
$$P(Y = 2|X = 1) = \frac{0.1}{0.55}$$
$$P(Y = 3|X = 1) = \frac{0.3}{0.55}$$
$$P(Y = 4|X = 1) = \frac{0.1}{0.55}$$

Thus, the information about gender reveals something about the students’ GPA. For example, without information on student’s gender, the probability of GPA = 3 is $P(Y = 3) = 0.425$. However, if we know that the student is male, we guess that the probability of him having GAP = 3 is $P(Y = 3|X = 0) = 0.125 + \frac{0.075}{0.45} = 0.27778$, while for a female student we predict that her probability of GPA = 3 is $P(Y = 3|X = 1) = \frac{0.3}{0.55} = 0.54545$. 
1.3. RANDOM VECTORS

1.3.3 Independence of random variables

What if we are given an information which is not useful? Intuitively, the distribution of a random variable with no information at all should be the same as the conditional distribution given some useless information. For example, suppose that I toss a die, and let $Y$ be the random variable which is equal to the result. Suppose that at the same time, my friend flips a coin, and let $X = 1$ if heads and $X = 0$ if tails. What is the conditional pdf of $Y$ given that $X = 1$? The information about the result of the coin flip does not reveal anything about the likely outcomes of the die toss. In this case we say that $Y$ and $X$ are independent. Intuitively, when information is useless, we expect that $f(y|x) = f(y)$. Thus, from the definition (1.3) it follows that when $X$ and $Y$ are independent, we have:

$$f(y|x) = \frac{f(x,y)}{f(x)} = f(y)$$

$$\Rightarrow f(x,y) = f(x)f(y)$$

**Definition 11** Two random variables are **statistically independent** if and only if

$$f(x,y) = f(x)f(y)$$

that is, the joint pdf can be written as the product of marginal pdfs.

In general though, if $X$ and $Y$ are dependent, the definition in (1.3) says that the joint pdf can be factored out in two ways:

$$f(y|x) = \frac{f(x,y)}{f(x)} \Rightarrow f(x,y) = f(x)f(y|x)$$

$$f(x|y) = \frac{f(x,y)}{f(y)} \Rightarrow f(x,y) = f(y)f(x|y)$$

In other words, the joint pdf is the product of a conditional and marginal pdfs. Comparing the two gives the relationship between the two conditional densities, known as the **Bayes rule** for pdfs:

$$f(y|x) = \frac{f(y)f(x|y)}{f(x)}$$

The concept of independence is crucial in econometrics. For starters, a random sample must be such that the units of observation are independent of each other.

**Exercise 14** Consider the pdf

$$f(x,y) = \begin{cases} 2 - x - y & 0 \leq x \leq 1; \ 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Check whether $X$ and $Y$ are statistically independent.
1.4 Moments

Occasionally the distribution of a random variable can be summarized in terms of a few of its characteristics, known as the moments of the distribution. The most widely used moments are mean (or expected value or expectation) and variance. While the distribution of all your grades on your transcript give a complete picture of your academic progress, the GPA is a single number, which is easy to process and which reveals important (although incomplete) information. In characterizing a random vector, in addition to mean and variance, we also look at covariance and correlation.

1.4.1 Expected value

Intuitively, expected value or the mean of a random variable, is a weighted average of all the values that the random variable can attain. In particular, the values of the random variable should be weighted by their probabilities.

Definition 12 The expected value (or mean) of a random variable \( X \) is

\[
[\text{If } X \text{ is discrete}] : \quad E(X) = \mu_X = \sum_x x f(x)
\]

\[
[\text{If } X \text{ is continuous}] : \quad E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) \, dx
\]

In the discrete case, the sum is over all the possible values of \( X \), and each \( x \) is weighted by the discrete pdf. No differently, in the continuous case, the integral is over all the possible values of \( X \) weighted by the continuous pdf.

For example of discrete case, consider \( X \) to be the result of a die toss, with pdf \( f(x) = 1/6 \) for all \( x \in \{1, 2, 3, 4, 5, 6\} \). This is called the discrete uniform distribution. The expected value of \( X \) is:

\[
E(X) = \sum_x x f(x) = \frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) = 3.5
\]

For example of a continuous case, consider \( X \sim U[a, b] \) with pdf \( f(x) = \frac{1}{b-a} (f(x) = 0 \text{ outside of the interval } [a, b]) \). This is the continuous uniform distribution. The expected value of \( X \) is:

\[
E(X) = \int_{-\infty}^{\infty} x f(x) \, dx = \int_a^b x \frac{1}{b-a} \, dx = \frac{1}{b-a} \left( \frac{x^2}{2} \right)_a^b
\]

\[
= \frac{1}{b-a} \left( \frac{b^2 - a^2}{2} \right) = \frac{1}{b-a} \left( \frac{(b-a)(b+a)}{2} \right)
\]

\[
= \frac{b+a}{2}
\]

This result is not surprising, when you look at the graph of the continuous uniform density and notice that the mean is simply the middle of the support \([a, b]\).

Recall that any function of a random variable, is also a random variable, and we often need to calculate the expected value of some function \( g(\cdot) \) of \( X \).
Definition 13 The **expected value of a function** $g(X)$ of a random variable $X$ is

If $X$ is discrete: $E[g(X)] = \sum_x g(x) f(x)$

If $X$ is continuous: $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) \, dx$

For example of discrete case, consider $X$ to be the result of a die toss, with pdf $f(x) = \frac{1}{6}$ for all $x \in \{1, 2, 3, 4, 5, 6\}$. The expected value of $g(X) = X^2$ is:

$$E[g(X)] = \sum_x g(x) f(x) = \frac{1}{6} (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = 15 \frac{1}{6}$$

Similarly, we can define the expected value of a function of two or more random variables.

Definition 14 Let $X$ and $Y$ be random variables, and let $g(X, Y)$ be some function of these random variables. The **expected value of a function** $g(X, Y)$ is

If $X$ and $Y$ are discrete: $E[g(X, Y)] = \sum_y \sum_x g(x, y) f(x, y)$

If $X$ and $Y$ are continuous: $E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, dxdy$

Any function of 3 random variables will involve joint 3-variable pdfs, and triple integrals (in continuous case) or triple sums (in discrete case).

It is important to realize that expectation of a random variable is always a number. While the toss of a die is a random variable which attains 6 different values, its expectation (mean) is a single number $3.5$. Next, we list some general properties of expected values, i.e. properties that apply to discrete and continuous random variables.

**Rules of Expected Values**

1. Expected value of a constant is the constant itself. Thus, if $b$ is a constant, then

   $$E(b) = b$$

2. Constants factor out. If $a$ is a constant number, then

   $$E(aX) = aE(X)$$

3. Expected value of a sum is the sum of expected values. That is, for any random variable $X$ and $Y$

   $$E(X + Y) = E(X) + E(Y)$$

   Together with rule 2, this generalizes to any linear combination of random variables. Let $X_1, ..., X_n$ be random variables, and let $a_1, ..., a_n$ be numbers. Then,

   $$E(a_1X_1 + ... + a_nX_n) = a_1E(X_1) + ... + a_nE(X_n)$$
4. If $X$ and $Y$ are independent, then

$$E(XY) = E(X)E(Y)$$

**Proof.** (Rule 1). The first rule does not really need a proof. But nevertheless, $E(b) = b \cdot f(b) = b \cdot 1 = b$. For example, the number 5 cannot be any number other than 5, so it is 5 with probability 1.

(Rule 2). We prove the second rule for continuous random variable $X$.

$$E(aX) = \int_{-\infty}^{\infty} axf(x) \, dx = a \int_{-\infty}^{\infty} xf(x) \, dx = aE(X)$$

The proof for discrete random variable is the same, with summation instead of integration.

(Rule 3). We prove the third rule for continuous random variables $X$ and $Y$.

$$E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y) \, dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [xf(x, y) + yf(x, y)] \, dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y) \, dxdy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y) \, dxdy$$

$$= \int_{-\infty}^{\infty} x \left[ \int_{-\infty}^{\infty} f(x, y) \, dy \right] \, dx + \int_{-\infty}^{\infty} y \left[ \int_{-\infty}^{\infty} f(x, y) \, dx \right] \, dy$$

$$= \int_{-\infty}^{\infty} xf(x) \, dx + \int_{-\infty}^{\infty} yf(y) \, dy \quad \text{(from definition of marginal pdf)}$$

$$= E(X) + E(Y)$$

(Rule 4). The product $XY$ is another random variable, since any function of random variables is another random variable. Thus, if $X$ and $Y$ are continuous, the expected value of the product is

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) \, dxdy$$

By definition of independence of random variables,

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x) f(y) \, dxdy$$

$$= \int_{-\infty}^{\infty} yf(y) \left[ \int_{-\infty}^{\infty} xf(x) \, dx \right] \, dy$$

$$= E(X) \int_{-\infty}^{\infty} yf(y) \, dy$$

$$= E(X) E(Y)$$

The proof is exactly the same for discrete $X$ and $Y$ with summation instead of integration.
Exercise 15 (Proof of rules 2 and 3 of expected values). Let $X$ and $Y$ be a random variables and $a, b$ are numbers.

(a) Prove that $E(aX) = aE(X)$ when $X$ is discrete.

(b) Prove that $E(X + Y) = E(X) + E(Y)$ when $X$ and $Y$ are discrete.

(c) Prove that $E(XY) = E(X)E(Y)$ when $X$ and $Y$ are independent discrete random variables.

1.4.2 Variance and standard deviation

While the expected value tells us about some average of a distribution, the variance and standard deviation measure the dispersion of the values around the mean. Imagine one uniform random variable with values $\{9, 10, 11\}$ and another with $\{0, 10, 20\}$. They both have the same mean of 10, but in the first the values are concentrated closer to the mean (smaller variance).

**Definition 15** Let $X$ be a random variable with mean $E(X) = \mu$. The variance of $X$ is

$$\text{var}(X) = \sigma_X^2 = E[(X - \mu)^2]$$

**Definition 16** The positive square root of variance, $s.d.(X) = \sigma_X = \sqrt{\text{var}(X)}$, is the standard deviation of $X$.

Thus, the variance is the expected value of the squared deviation of the random variable from its mean. We already know the definitions of expected value of discrete and continuous random variables, and therefore we can write the definition of variance as follows:

[If $X$ is discrete] : $\text{var}(X) = \sigma_X^2 = \sum_x (x - \mu)^2 f(x)$

[If $X$ is continuous] : $\text{var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx$

As example of a discrete case, lets compute the variance of a toss of a die, $X$, with pdf $f(x) = \frac{1}{6}$ for all $x \in \{1, 2, 3, 4, 5, 6\}$. We already computed the mean $\mu = 3.5$. Thus, the variance is:

$$\text{var}(X) = \frac{1}{6} \left[ (1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2 \right] = 2.9167$$

For example of a continuous case, consider $X \sim U[a, b]$ with pdf $f(x) = \frac{1}{b-a} \quad (f(x) = 0 \text{ outside of the interval } [a, b])$. This is the continuous uniform distribution. We already found
that the mean is $\mu = \frac{a+b}{2}$, and therefore the variance of $X$ is:

$$
\text{var} (X) = \int_a^b \left( x - \frac{a+b}{2} \right)^2 \frac{1}{b-a} \, dx \\
= \frac{1}{b-a} \int_a^b \left( x^2 - (a+b) x + \left( \frac{a+b}{2} \right)^2 \right) \, dx \\
= \frac{1}{b-a} \left. \left( \frac{x^3}{3} - (a+b) \frac{x^2}{2} + \left( \frac{a+b}{2} \right)^2 \right) \right|_a^b \\
= \frac{1}{b-a} \left[ \frac{b^3-a^3}{3} - (a+b) \frac{b^2-a^2}{2} + \left( \frac{a+b}{2} \right)^2 (b-a) \right]
$$

Using the rule, $b^3 - a^3 = (b-a) (b^2 + ab + a^2)$, we get

$$
\text{var} (X) = \left[ \frac{b^2 + ab + a^2}{3} - \frac{(a+b)^2}{2} + \left( \frac{a+b}{2} \right)^2 \right] \\
= \left[ \frac{b^2 + ab + a^2}{3} - \frac{(a+b)^2}{4} \right] \\
= \frac{4 (b^2 + ab + a^2) - 3 (a^2 + 2ab + b^2)}{12} \\
= \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}
$$

The derivation was pretty messy, but the result makes intuitive sense. Notice that the variance of $X \sim U [a, b]$ depends positively on the length of the interval $b-a$.

**Exercise 16** Consider the continuous uniform random variable $X \sim U [a, b]$.

(a) If $a = 0$ and $b = 1$, find the pdf of $X$, its expected value and variance.

(b) If $a = 0$ and $b = 2$, find the pdf of $X$, its expected value and variance.

For computational convenience, the definition of variance can be manipulated to create an equivalent, but easier to use formula.

$$
\text{var} (X) = E \left[ (X - \mu)^2 \right] = E \left[ X^2 - 2\mu X + \mu^2 \right] = E \left( X^2 \right) - 2\mu E (X) + \mu^2 = E \left( X^2 \right) - 2\mu^2 + \mu^2 = E \left( X^2 \right) - \mu^2
$$

Thus, you will often use the formula:

$$
\text{var} (X) = E \left( X^2 \right) - E (X)^2
$$
Having found the expected value, the last term is just its square. So we only need to compute the mean of \( X^2 \). For example of discrete case, consider again a toss of a die, \( X \), with pdf \( f(x) = 1/6 \) for all \( x \in \{1, 2, 3, 4, 5, 6\} \). We found that \( E(X) = 3.5 \), so \( E(X)^2 = 3.5^2 = 12.25 \). The first term is

\[
E(X^2) = \sum_x x^2 f(x) = \frac{1}{6} \left( 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 \right) = 15.1667
\]

Thus,

\[
var(X) = E(X^2) - E(X)^2 = 15.167 - 12.25 = 2.9167
\]

For example of a continuous case, consider \( X \sim U[a,b] \) with pdf \( f(x) = \frac{1}{b-a} \) (\( f(x) = 0 \) outside of the interval \([a,b]\)). We already found that \( E(X) = \frac{b+a}{2} \), so \( E(X)^2 = \left( \frac{b+a}{2} \right)^2 = \frac{(b+a)^2}{4} \). The first term is:

\[
E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \left( \frac{x^3}{3} \right)_a^b = \frac{1}{b-a} \left( \frac{b^3 - a^3}{3} \right) = \frac{b^2 + ab + a^2}{3}
\]

Thus, the variance is

\[
var(X) = E(X^2) - E(X)^2 = \frac{b^2 + ab + a^2}{3} - \frac{(b + a)^2}{4} = \frac{(b - a)^2}{12}
\]

Notice that expressions become less messy with the definition of variance in (1.4).

**Rules of Variance**

1. Variance of constant is zero. Thus, if \( b \) is a constant, then
   \[
   \text{var}(b) = 0
   \]

2. Constants factor out squared. If \( a \) is a constant number, then
   \[
   \text{var}(aX) = a^2 \text{var}(X)
   \]
   Thus, for example, if you multiply a random variable by 2, its variance increases by a factor of 4.
3. Adding a constant to a random variable, does not change its variance. Thus, if \( b \) is a constant number, then
\[
\text{var}(X + b) = \text{var}(X)
\]
Adding a constant to a random variable, just shifts all the values and the mean by that number, but does not change their dispersion around the mean.

4. Variance of the sum (difference) is NOT the sum (difference) of the variances. For random variables \( X \) and \( Y \), we have
\[
\begin{align*}
\text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) \\
\text{var}(X - Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y)
\end{align*}
\]
where \( \text{cov}(X,Y) \) is the covariance between \( X \) and \( Y \), to be defined in the next section.

The first 3 rules can be easily proved directly from the definition of variance \( \text{var}(X) = E[(X - \mu)^2] \). We’ll get back to the 4th rule in the next section, after we discuss covariance.

**Proof.** (Rules 1,2,3 of variance).

1. If \( X = b \), then from the rules of expected values \( E(b) = b \). Thus
\[
\text{var}(b) = E[(b - b)^2] = 0
\]

2. From the rules of expected values, \( E(aX) = aE(X) \). Thus,
\[
\begin{align*}
\text{var}(aX) &= E[(aX - aE(X))^2] \\
&= E[(aX - a\mu)^2] \\
&= a^2(E[(X - \mu)^2]) \\
&= a^2 \text{var}(X)
\end{align*}
\]

3. From rules of expected values, \( E(X + b) = E(X) + b = \mu + b \). Thus,
\[
\begin{align*}
\text{var}(X + b) &= E[(X + b - (\mu + b))^2] \\
&= (X - \mu)^2 \\
&= \text{var}(X)
\end{align*}
\]

**Rules of Standard Deviation**

Usually, textbooks do not even bother presenting the rules of standard deviations, because they follow directly from rules of variances. For example, taking square roots of the first 3 rules of variances, gives the following rules of standard deviations:

1. \( \sqrt{\text{var}(b)} = \sqrt{0} \Rightarrow \text{s.d.}(b) = 0 \)

2. \( \sqrt{\text{var}(aX)} = \sqrt{a^2 \text{var}(X)} \Rightarrow \text{s.d.}(aX) = |a| \cdot \text{s.d.}(X) \)
3. $\sqrt{\text{var} (X + b)} = \sqrt{\text{var} (X)} \Rightarrow \text{s.d.} (X + b) = \text{s.d.} (X)$

Notice that rule 2 implies that if we double a random variable, we also double its standard deviation.

Besides the importance of variance and standard deviation is statistics (we will be using them everywhere in this course), they are extremely important in finance. Any financial asset (or a portfolio of assets) is a random variable, which has expected return (mean return) and risk (standard deviation of returns). A great deal of modern portfolio theory deals with finding portfolios which maximize expected return for any given level of risk, or alternatively, minimizing the level of risk for any given expected return. These are called efficient portfolios. If you take Finance 350 course (which is highly recommended), these notes will be useful as well.

Exercise 17 Let $X$ be random variable with mean $\mu_X$ and variance $\sigma_X^2$. Let $Z$ be the following transformation of $X$:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

Thus $Z$ is obtained by subtracting the mean and dividing by standard deviation (this operation called standardization). Show that $Z$ has mean zero and standard deviation 1 (i.e. $\mu_Z = 0$, and $\sigma_Z = 1$). Such random variable is called standard.

Exercise 18 Can the variance of any random variable be negative? Prove it.

Exercise 19 Can the standard deviation of any random variable be negative? Prove it.

Exercise 20 Let $X$ be random variable with mean $\mu_X$ and variance $\sigma_X^2$. What is the standard deviation of $Y = 2X$?

Exercise 21 Let $X$ be random variable with mean $\mu_X$ and variance $\sigma_X^2$. What is the standard deviation of $Y = -2X$?

1.4.3 Covariance and correlation

When analyzing the joint behavior of two random variables, we often start by asking if they are correlated. For example, we would like to know if education level and wages are positively correlated (otherwise, what are we doing here?). As another example, in macroeconomics we often want to know if some variable is correlated with real GDP or not (say unemployment). What this means in plain language, is that we want to investigate whether two variables are moving around their respective means in the same direction, opposite direction or there is no connection between their movements.

Definition 17 Let $X$ and $Y$ be two random variables with means $\mu_X$ and $\mu_Y$ respectively. The covariance between $X$ and $Y$ is

$$\text{cov} (X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$
Notice that if $X$ and $Y$ are mostly moving together around their mean, then when $X$ is above $\mu_X$, $Y$ also tends to be above $\mu_Y$, and when $X$ is below $\mu_X$, $Y$ also tends to be below $\mu_Y$. In this case the terms $(X - \mu_X)$ and $(Y - \mu_Y)$ are either both positive or both negative, and $\text{cov}(X,Y) > 0$. We then say that $X$ and $Y$ are positively correlated. On the other hand, suppose that $X$ and $Y$ are for the most part moving in opposite directions. Then when $X$ is above $\mu_X$, $Y$ tends to be below $\mu_Y$, and when $X$ is below $\mu_X$, $Y$ tends to be above $\mu_Y$. In such case the terms $(X - \mu_X)$ and $(Y - \mu_Y)$ will have opposite signs and $\text{cov}(X,Y) < 0$. We then say that $X$ and $Y$ are negatively correlated. We will see later, that when $X$ and $Y$ are independent, then $\text{cov}(X,Y) = 0$, and we say that $X$ and $Y$ are uncorrelated.

Notice that just as the definition of variance, the definition of covariance involves computing expected value. In order to compute these expected values, we once again have two formulas for discrete and for continuous random variables.

\[
\begin{align*}
\text{If } X \text{ and } Y \text{ are discrete} & : \quad \text{cov}(X,Y) = \sum_x (x - \mu_X)(y - \mu_Y) f(x,y) \\
\text{If } X \text{ and } Y \text{ are continuous} & : \quad \text{cov}(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x,y) \, dx \, dy
\end{align*}
\]

Also as we did with variance, we can manipulate the definition of covariance to derive a more convenient formula for computation.

\[
\begin{align*}
\text{cov}(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\
&= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X \mu_Y \\
&= E(XY) - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y
\end{align*}
\]

Thus, you will often use the formula

\[
\text{cov}(X,Y) = E(XY) - E(X)E(Y) \quad (1.5)
\]

In words, the covariance between $X$ and $Y$ is equal to the expectation of a product, minus product of expectations.

**Rules of Covariance**

1. Covariance of a random variable $X$ with itself is the variance of $X$:

\[
\text{cov}(X,X) = \text{var}(X)
\]

Notice that we could have discussed covariance before variance, and then define variance as the covariance of a variable with itself. In fact, this way we can prove the rules of variances in a more elegant and easy way. These notes follow the order of coverage in most textbooks, i.e. mean $\rightarrow$ variance $\rightarrow$ covariance.

2. Covariance between a random variable and a constant is zero. Thus, if $X$ is a random variable, and $b$ is a constant, then

\[
\text{cov}(X,b) = 0
\]
3. Constants factor out as product. Let $X$ and $Y$ be random variables and $a$ and $b$ be constant numbers. Then,

$$cov(aX, bY) = ab \cdot cov(X, Y)$$

4. Distributive property. Let $X$, $Y$ and $Z$ be random variables. Then,

$$cov(X, Y + Z) = cov(X, Y) + cov(X, Z)$$

5. Adding constants to random variables, does not change the covariance between them. Let $X$ and $Y$ be random variables and $c$ and $d$ be constant numbers. Then,

$$cov(X + c, Y + d) = cov(X, Y)$$

6. Covariance between two independent variables is 0. That is, suppose that random variables $X$ and $Y$ are independent. Then,

$$cov(X, Y) = 0$$

or

$$E(XY) = E(X)E(Y)$$

Recall that variables with zero covariance are called **uncorrelated**. This rule says that independence implies lack of correlation.

**Proof.** For all the proofs of the covariance rules it is easier to use the definition in equation (1.5), that is, $cov(X, Y) = E(XY) - E(X)E(Y)$.

(Rule 1). This just follows from comparing the definition of covariance (1.5), when $X = Y$, with the definition of variance (1.4).

(Rule 2). It should be intuitive, that covariance of any random variable with a constant is zero, because covariance measures the comovement of the two around their means, and a constant simply does not move. Formally,

$$cov(X, b) = E(Xb) - E(X)E(b)$$

$$= bE(X) - bE(X)$$

$$= 0$$

(Rule 3). Constants factor out because they do so in expected values. Formally,

$$cov(aX, bY) = E(aX \cdot bY) - E(aX)E(bY)$$

$$= abE(XY) - abE(X)E(Y)$$

$$= ab[E(XY) - E(X)E(Y)]$$

$$= ab \cdot cov(X, Y)$$
(Rule 4). Distributive property.

\[
\text{cov}(X, Y + Z) = E(X \cdot (Y + Z)) - E(X) E(Y + Z)
= E(XY + XZ) - E(X) [E(Y) + E(Z)]
= E(XY) + E(XZ) - E(X) E(Y) - E(X) E(Z)
= \left[ E(XY) - E(X) E(Y) \right] + \left[ E(XZ) - E(X) E(Z) \right]
\]

(Rule 5). Adding constants does not affect the covariance.

\[
\text{cov}(X + c, Y + d) = E[(X + c) \cdot (Y + d)] - E(X + c) E(Y + d)
= E[XY + Xd + cY + cd] - [E(X) + c][E(Y) + d]
= [E(XY) + dE(X) + cE(Y) + cd] - [E(X) E(Y) + dE(X) + cE(Y) + cd]
= E(XY) - E(X) E(Y)
= \text{cov}(X, Y)
\]

(Rule 6). Covariance between two independent variables is 0. This is exactly rule 4 of expected values, which we already proved and will not repeat here. 

Now we are in a position to prove rule 4 of variance, which is a formula of variance of sums:

\[
\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)
\]

**Proof.** (Rule 4 of variance). Recall that variance is covariance between the variable and itself. Therefore,

\[
\text{var}(X + Y) = \text{cov}(X + Y, X + Y)
\]

By distributive property of covariance (rule 4), we have

\[
\text{cov}(X + Y, X + Y) = \text{cov}(X, X + Y) + \text{cov}(Y, X + Y)
= \left\{ \text{var}(X) + \text{cov}(X, Y) \right\} + \left\{ \text{var}(X) + \text{cov}(Y, Y) \right\}
= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)
\]

**Exercise 22** Prove that for any two random variables \( X \) and \( Y \), we have

\[
\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)
\]

**Exercise 23** Prove that for two **independent** random variables \( X \) and \( Y \), the variance of a sum or difference is the sum of the variances:

\[
\begin{align*}
\text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) \\
\text{var}(X - Y) &= \text{var}(X) + \text{var}(Y)
\end{align*}
\]
Recall that covariance measures the degree of comovement of two random variables around their means. The main disadvantage of this measure is that it is NOT unit-free. Suppose researcher A studies the comovement of years of schooling $X$ and annual wages measured in thousands of dollars, $Y$. He finds a positive covariance:

$$\text{cov}(X, Y) = 7$$

Researcher $B$ has the same goal and the same data on wages and schooling, but he chooses to measure wages in dollars (instead of thousands of dollars). Thus, he finds

$$\text{cov}(X, 1000 \cdot Y) = 1000 \cdot \text{cov}(X, Y) = 7000$$

Should researcher $B$ conclude that wages and years of schooling have stronger comovement than what was previously found by researcher $A$? NO, if researcher $A$ and $B$ use the same data, we want them to reach the same conclusion about the degree of comovement between wages and years of schooling. In order to make the results comparable, researchers report correlation instead of covariance.

**Definition 18** The correlation between random variables $X$ and $Y$ is

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

We define $\rho_{XY} = 0$ whenever $\sigma_X = 0$, or $\sigma_Y = 0$ or both. In other words, whenever the covariance between two random variables is zero, we define the correlation to be also zero, and say that the two variables are uncorrelated.

Thus, the correlation is the covariance divided by the product of standard deviations. How does this resolve the problem of covariance depending on units of the random variables? Let $X$ and $Y$ be random variables, and let $a$ and $b$ be positive numbers. We know that $\text{cov}(aX, bY) = ab \cdot \text{cov}(X, Y)$, so if $a$ and $b$ represent different units used by different researchers, we will have different results. Correlation however is unit free, and does not change when we scale the random variables by positive numbers.

**Properties of Correlation**

1. For any two random variables $X$ and $Y$, the sign of the correlation between them is the same as the sign of the covariance between them.

$$\text{sign}(\rho_{XY}) = \text{sign}(\text{cov}(X, Y))$$

In other words, correlation gives us the same information as covariance about the qualitative nature of the comovement between $X$ and $Y$ around their mean. If $\text{cov}(X, Y) > 0$, meaning that $X$ and $Y$ tend to be above and below their respective means at the same time, then we will also have $\rho_{XY} > 0$.

2. Correlation is unit free. Let $X$ and $Y$ be random variables, and let $a$ and $b$ be numbers, both of the same sign. Then,

$$\rho_{aX, bY} = \rho_{XY}$$
3. Correlation between any two random variables can be only between \(-1\) and \(1\). Let \(X\) and \(Y\) be random variables. Then,
\[
-1 \leq \text{corr}(X,Y) \leq 1
\]

**Proof.** (Property 1). Since standard deviation of any random variable is always positive, dividing by them does not change the sign.

(Property 2). Let \(X\) and \(Y\) be random variables, and let \(a\) and \(b\) be numbers with the same sign.

\[
corr(aX; bY) = \frac{\text{cov}(aX, bY)}{\sqrt{\text{var}(aX)} \cdot \sqrt{\text{var}(bY)}} = \frac{ab \cdot \text{cov}(X,Y)}{\sqrt{a^2 \text{var}(X)} \cdot \sqrt{b^2 \text{var}(Y)}} = \frac{ab \cdot \text{cov}(X,Y)}{|a| \cdot |b| \cdot \text{sd}(X) \cdot \text{sd}(Y)} = \text{corr}(X,Y)
\]

Notice that the last step follows because \(a\) and \(b\) have the same sign (so \(ab > 0\) and \(ab = |a| \cdot |b|\)).

We do not provide the proof of property 3, since it involves knowledge of inner products and Cauchy–Schwarz inequality. Instead, we discuss this property at intuitive level. The highest possible correlation that can be achieved between \(X\) and \(Y\), is when \(Y = X\). Then the two random variables \(X\) and \(Y\) are always guaranteed to be above or below their mean at the same time. The lowest possible correlation is achieved when \(Y = -X\). In this case, the two variables always deviate from their means in opposite directions. Thus, we will prove that \(\text{corr}(X, X) = 1\) and \(\text{corr}(X, -X) = -1\), and these will be the maximal and minimal correlations, so that in general, for any \(X\) and \(Y\) we have \(-1 \leq \text{corr}(X, Y) \leq 1\).

\[
\text{corr}(X, X) = \frac{\text{cov}(X, X)}{\sigma_X \sigma_X} = \frac{\text{var}(X)}{\text{var}(X)} = 1
\]
\[
\text{corr}(X, -X) = \frac{\text{cov}(X, -X)}{\sigma_X \sigma_X} = \frac{-\text{cov}(X, X)}{\sigma_X \sigma_X} = -1
\]

In the last equation we used rule 3 of covariance, that of constants factoring out, with the constant in this case being \(-1\).

**Exercise 24** Prove that if two random variables, \(X\) and \(Y\), are independent, then they are **uncorrelated**, i.e. \(\text{corr}(X, Y) = 0\).

The opposite statement is not true. Two random variables can be uncorrelated, but not independent. The next exercise gives such an example.

**Exercise 25** Let \(X\) be a random variable with pdf \(f(x) = \frac{1}{4}\) for \(x \in \{1, 2, -1, -2\}\). Let \(Y = X^2\).

(a) Show that \(X\) and \(Y\) are not correlated, i.e. show that \(\text{corr}(X,Y) = 0\).

(b) Show that \(X\) and \(Y\) are not independent. (Hint: compare the conditional pdf \(f(y|x)\) with \(f(y)\). If \(X\) and \(Y\) are independent, then we must have \(f(y|x) = f(y)\)).
Exercise 26 Let $X$ be a random variable, and $a, b$ be some numbers. Let $Y = aX + b$.

(a) Prove that if $a > 0$, then $\text{corr} (X, Y) = 1$.

(b) Prove that if $a < 0$, then $\text{corr} (X, Y) = -1$.

(c) Prove that if $a = 0$, then $\text{corr} (X, Y) = 0$.

Exercise 27 Prove that the correlation between two random variables $X$ and $Y$, is equal to the covariance between the standardized $X$ and $Y$:

$$\text{cov} \left( \frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y} \right) = \text{corr} (X, Y)$$

Solution 2

$$\text{cov} \left( \frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y} \right) = \frac{1}{\sigma_X \cdot \sigma_Y} \text{cov} (X - \mu_X, Y - \mu_Y) \quad \text{(constants factor out)}$$

$$= \frac{1}{\sigma_X \cdot \sigma_Y} \text{cov} (X, Y) \quad \text{(adding constants does not affect covariance)}$$

$$= \text{corr} (X, Y)$$

This is another way to see that correlation is unit free. When each variable is standardized, we effectively remove the units of the variables, and express both $X$ and $Y$ in terms of standard deviations from their respective means.

1.5 Normal Distribution and Its Relatives

In this section we briefly discuss the normal distribution, and its cousins, the Chi-square distribution, the Student’s t-distribution, and the F-distribution. These 4 distributions are the most important ones for this course.

1.5.1 Normal distribution

Definition 19 A random variable $X$ has normal distribution with mean $\mu$ and standard deviation $\sigma$, if its probability density function is

$$f (x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right], \quad -\infty \leq x \leq \infty$$ (1.6)

We then write $X \sim N (\mu, \sigma^2)$ as a short way of saying that $X$ has normal distribution with mean $\mu$ and standard deviation $\sigma$ (variance of $\sigma^2$).

The pdf of the normal distribution has a bell shape, which is symmetric around the mean. The next figure shows how the particular pdf changes when we change the parameters $\mu$ and
CHAPTER 1. PRELIMINARIES

The solid line is the pdf of the standard normal distribution, with \( \mu = 0 \) and \( \sigma = 1 \). The dotted curve plots the pdf of \( N(2, 1) \). Notice that the shape of this pdf is the same, but it is shifted by 2. The dashed curve is the pdf of \( N(2, 4) \), i.e. \( \mu = 2 \) and \( \sigma = 2 \). Notice that it is centered around the mean of 2, but because of the larger variance it is more flat than \( N(2, 1) \).

From the above graph, we learn that there are many normal distributions, characterized by \( \mu \) and \( \sigma \). Does this mean that to calculate probabilities that a normal random variable is in some interval \([a, b]\), we need to compute the integrals of different pdfs? Fortunately, once you see one normal distribution, you had seen all of them. It is therefore a common practice to use just one normal distribution - the standard normal.

**Definition 20** Let \( X \sim N(\mu, \sigma^2) \). The random variable

\[
Z = \frac{X - \mu}{\sigma}
\]

has **standard normal distribution**, with mean zero and standard deviation 1. We write \( Z \sim N(0, 1) \). The pdf of the standard normal is

\[
\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty \leq z \leq \infty
\]

Suppose we want to calculate the probability that \( X \sim N(\mu, \sigma^2) \) is between \( a \) and \( b \). We can always express this probability in terms of the standard normal distribution:

\[
P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right)
\]

\[
= P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)
\]

Therefore, for all practical purposes, we only need the standard normal distribution. The next figure, shows the pdf of \( Z \sim N(0, 1) \).
Properties of Normal Distribution

1. All normal distributions contain about 68% of the probability concentrated within one standard deviation from the mean, about 95% of the probability within 2σ of the mean, and about 99.7% of the probability within 3σ from the mean. Formally,

\[
P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68
\]
\[
P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95
\]
\[
P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997
\]

This property is illustrated in the next graph.

\[\text{In other words, it is highly unlikely to obtain random draws from the normal distribution, which are far from the mean.}\]

2. Linear combination of normal random variables is also normal. For example, suppose that \(X\) and \(Y\) are normal random variables. Then their sum is also normal. More general, let \(X_1, ..., X_n\) be normal random variables with mean \(\mu_1, ..., \mu_n\) and standard deviations \(\sigma_1, ..., \sigma_n\), and let \(a_1, ..., a_n\) be constant numbers. Then,

\[a_1X_1 + ... + a_nX_n \sim N(\mu, \sigma^2)\]

where \(\mu = a_1\mu_1 + ... + a_n\mu_n\), and the variance is something complicated to find in general.

There is no proof for the first property in the usual sense; one just needs to calculate the following integrals of the normal pdf in equation (1.6):

\[
\int_{\mu-\sigma}^{\mu+\sigma} f(x) \, dx, \quad \int_{\mu-2\sigma}^{\mu+2\sigma} f(x) \, dx, \quad \int_{\mu-3\sigma}^{\mu+3\sigma} f(x) \, dx
\]
These integrals cannot be calculated analytically and must be approximated numerically (using some computer algorithms).

The proof of the second property relies on familiarity with the concept of characteristic functions, and therefore I skip the proof of the second property. However, you already have the tools to show that the mean of the linear combination is \( \mu = a_1 \mu_1 + ... + a_n \mu_n \), and if \( X_1, ..., X_n \) are independent, you can also find the variance.

**Exercise 28** Let \( X_1, ..., X_n \) be normal random variables with mean \( \mu_1, ..., \mu_n \) and standard deviations \( \sigma_1, ..., \sigma_n \), and let \( a_1, ..., a_n \) be constant numbers.

(a) Show that the expected value of \( a_1 X_1 + ... + a_n X_n \) is \( a_1 \mu_1 + ... + a_n \mu_n \).

(b) Suppose that in addition \( X_1, ..., X_n \) are independent. Show that the variance of \( a_1 X_1 + ... + a_n X_n \) is \( a_1^2 \sigma_1^2 + ... + a_n^2 \sigma_n^2 \). Hint: use the property that for uncorrelated random variables, the variance of the sum is the sum of the variances.

**Central Limit Theorem**

The reason why the normal distribution is by far the most popular in statistics, is the remarkable result - the **Central Limit Theorem**.

**Theorem 1** (CLT). Let \( X_1, ..., X_n \) be \( n \) independent random variables, all having some distribution with mean \( \mu \) and variance \( \sigma^2 \). Let the simple average of those variables be

\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

Then as \( n \to \infty \), we have

\[
\bar{X}_n \xrightarrow{D} N \left( \mu, \frac{\sigma^2}{n} \right)
\]

The notation \( \xrightarrow{D} \) means "converges in distribution". Please notice the amazing generality of the theorem by paying attention to the words "some distribution". The CLT is one of the most important results in probability theory, and by far the most important result in statistics. The proof of such general result is much too advanced for this course (again requires knowledge of characteristic functions). But we can visualize the result using simulation.

We had seen the continuous uniform distribution in the section about random variables. Let \( X \sim U [a, b] \), with pdf \( f(x) = \frac{1}{b-a} \) if \( x \in [a, b] \). The pdf of \( X \) is a horizontal line and does not look anything like the normal distribution. However, the CLT states that if we draw samples and compute their averages, then the distribution of these averages will start looking more and more like the pdf of a normal random variable, as we increase the size of these samples.

The next figure shows the histograms of simulated sample means based on random draws from \( U [0, 5] \). The first panel shows the sample means, when the sample size is \( n = 1 \). This does not look anything like the normal pdf. The second panel shows the histogram of sample means, based on samples of size \( n = 2 \) from uniform distribution. We are beginning to see something that looks like a bell shape. The last panel shows the histogram of sample means,
when the sample size is \( n = 100 \). This histogram looks like it could have been generated by random draws from normal distribution.

Remember that the Central Limit Theorem is a limiting result that requires \( n \rightarrow \infty \). The approximation though can be very good even for \( n \) not too big. In other words, even when the sample size is \( n = 30 \), the distribution of the sample mean might be very close to normal.

**Exercise 29** Let \( X_1, \ldots, X_n \) be \( n \) independent random variables, all having some distribution with mean \( \mu \) and variance \( \sigma^2 \). Let the simple average of those variables be

\[
X_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

(a) Show that the mean of \( X_n \) is \( \mu \), i.e. \( E(X_n) = \mu \).

(b) Show that the variance of \( X_n \) is \( \frac{\sigma^2}{n} \), i.e. \( \text{var}(X_n) = \frac{\sigma^2}{n} \).
1.5.2 Chi-square distribution

**Definition 21** Let $Z_1, Z_2, ..., Z_n$ be independent standard normal random variables, i.e. $Z_i \sim N(0, 1)$ for $i = 1, 2, ..., n$. Then, the sum of squares of these random variables is called the **chi-square** distribution, with $n$ degrees of freedom. This is denoted

\[ X = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n \]

The only parameter of the chi-square distribution is the number of squared standard normal random variables that were added to generate it. Recall that any function of a random variable is another random variable. The chi-square is one of many random variables that we could have created from the normal distribution. The question you should be asking now is, what is the purpose of creating this particular distribution? The usefulness of chi-square is in the following result.

**Theorem 2** Let $X_1, ..., X_n$ be independent random variables, all with the same normal distribution: $X_i \sim N(\mu, \sigma^2) \forall i$. We define the sample variance as follows:

\[
[sample\ variance] : s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

Then,

\[
\frac{s_n^2 (n - 1)}{\sigma^2} \sim \chi^2_{n-1}
\]

This result allows us to test hypotheses about the population variance.

**Properties of chi-square Distribution**

1. Let $X \sim \chi^2_n$. Then $E(X) = n$.
2. Let $X \sim \chi^2_n$. Then $\text{var}(X) = 2n$.
3. Let $X \sim \chi^2_n$ and $Y \sim \chi^2_k$, then $X + Y \sim \chi^2_{n+k}$.

**Proof.** (Property 1).

\[
E(X) = E\left(\sum_{i=1}^{n} Z_i^2\right) = \sum_{i=1}^{n} E(Z_i) = n
\]

Recall that $Z_i \sim N(0, 1)$, so $E(Z_i) = 0$, and $\text{var}(Z_i) = E(Z_i^2) - E(Z_i)^2 = E(Z_i^2) = 1$.

(Property 2). Omitted. The proof requires knowledge of **moment generating functions**.
1.5. NORMAL DISTRIBUTION AND ITS RELATIVES

(Property 3). The chi-square is the sum of squared standard normal random variables. If $X$ is constructed by adding $n$ squared standard normal, and $Y$ is constructed by adding $k$, then when we add them all up, we get $n + k$.

$$X + Y = \sum_{i=1}^{n} Z_i^2 + \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{n+k} Z_i^2 \sim \chi^2_{n+k}$$

We do not present here the mathematical expression of the pdf of the chi-square, because it is too messy, but it exists and you can look it up if you want to. The next figure plots the pdfs of the chi-square with different number of degrees of freedom.

![Degrees of freedom](image)

1.5.3 t-distribution

**Definition 22** Let $Z$ be a standard normal random variable, $Z \sim N(0,1)$, and let $Y$ be a chi-square with $n$ degrees of freedom, $Y \sim \chi^2_n$, and $Z$ and $Y$ are statistically independent. Then the random variable $\frac{Z}{\sqrt{Y/n}}$ has a t distribution with $n$ degrees of freedom. This is denoted

$$T = \frac{Z}{\sqrt{Y/n}} \sim t_n$$

The pdf of the $t$ distribution looks very similar to the pdf of $Z \sim N(0,1)$, and in fact when $n$ becomes large, $T$ approaches $Z$. To prove this claim, one needs to show that the denominator in the above definition approaches 1 as $n \to \infty$.

$$E\left(\frac{1}{n}Y\right) = \frac{1}{n}E(Y) = \frac{1}{n} = 1$$

$$\text{var}\left(\frac{1}{n}Y\right) = \frac{1}{n^2}\text{var}(Y) = \frac{1}{n^2}2n = \frac{2}{n}$$

Notice that $\text{var}(Y/n) \to 0$ as $n \to \infty$. Thus, $Y/n \to 1$ as $n \to \infty$. 
CHAPTER 1. PRELIMINARIES

So why do we need a random variable which is so similar to something we already know, the standard normal? Indeed, for large \( n \) is doesn’t matter which distribution you use, the \( N(0, 1) \) or \( t_n \), but for \( n < 30 \) the two distributions are different. You probably encountered the \( t \) distribution in statistical tests about the mean of a normal population. We will state two theorems, which will become useful later, when we formally introduce the concept of random samples.

**Theorem 3** Let \( X_1, \ldots, X_n \) be independent random variables, all with the same normal distribution: \( X_i \sim N(\mu, \sigma^2) \) \( \forall i \). In other words, we have a random sample of size \( n \) from a normal population. We define the sample mean and the sample variance as follows:

\[
\text{[sample mean]} : \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

\[
\text{[sample variance]} : \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

Then,

1. \( Z = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1) \)
2. \( Y = \frac{s_n^2(n-1)}{\sigma^2} \sim \chi^2_{n-1} \)
3. \( T = \frac{Z}{\sqrt{Y / (n-1)}} = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2/n}} \sim t_{n-1} \)
4. \( Z \) and \( Y \) are independent

Note the similarity of \( T \) and \( Z \). When performing a test about a mean of a normal population, with known variance, we use the quantity in (1), which is called the \textbf{z-statistic}. However, when we don’t know the population variance \( \sigma^2 \), we replace it with its estimator \( s_n^2 \), and the resulting quantity in (3) is called the \textbf{t-statistic}.

We will only prove part (3) here, assuming that we already proved (2) and (4).

**Proof.**

\[
T = \frac{Z}{\sqrt{Y / (n-1)}} = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2(\sigma^2/n)}} = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2\sigma^2(n-1)}} \cdot \frac{1}{\sqrt{\sigma^2 / n}} = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2 / n}}
\]

Since \( Y \sim \chi^2_{n-1} \) and since \( Z \) and \( Y \) are independent, we have by definition of \( t \) distribution

\[
\frac{\bar{X}_n - \mu}{\sqrt{s_n^2 / n}} \sim t_{n-1}
\]

This means, that when we replace the population variance \( \sigma^2 \) is the z-statistic, with its estimator \( s_n^2 \), the resulting quantity has \( t \) distribution with \( n - 1 \) degrees of freedom.
1.5.4 F-distribution

**Definition 23** Let \( Y_1 \sim \chi^2_{n_1} \) and \( Y_2 \sim \chi^2_{n_2} \) be two independent chi-square random variables. Then the ratio \( \frac{Y_1/n_1}{Y_2/n_2} \) has \( F \) distribution with \( n_1 \) numerator degrees of freedom and \( n_2 \) denominator degrees of freedom. This is denoted

\[
F = \frac{Y_1/n_1}{Y_2/n_2} \sim F(n_1, n_2)
\]

The usefulness of the \( F \) distribution is that it allows us to compare the variability of two samples around their respective means, when the population is normal. Suppose we have two random samples from the same normal population: \( X_1, \ldots, X_{n_1} \) and \( Z_1, \ldots, Z_{n_2} \). The sample variances of these samples are

\[
s_{n_1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2
\]
\[
s_{n_2}^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Z_i - \bar{Z}_{n_2})^2
\]

Then, letting

\[
Y_1 = \frac{s_{n_1}^2}{\sigma^2} (n_1 - 1) \sim \chi^2_{n_1 - 1}
\]
\[
Y_2 = \frac{s_{n_2}^2}{\sigma^2} (n_2 - 1) \sim \chi^2_{n_2 - 1}
\]

If \( Y_1 \) and \( Y_2 \) are independent (which they are), we have:

\[
F = \frac{Y_1/(n_1 - 1)}{Y_2/(n_2 - 1)} = \frac{s_{n_1}^2}{s_{n_2}^2} \sim F(n_1 - 1, n_2 - 1)
\]

So the ratio of two sample variances, based on a normal population, follows \( F \) distribution. This allows us to test hypotheses related to **goodness of fit**, which will be discussed later.

1.6 Sampling and Estimators

In the section on random variables, in each example the distribution of a random variable, or at least its mean and variance were given. In reality though, nothing is given, and we don’t have for example the joint pdf of wages, schooling, experience, etc. The best we can hope for is collecting a **random sample** and learn from it about the **population**.

**Definition 24** A **random sample** of size \( n \) on a random variable \( X \) is a set of independent random variables \( X_1, \ldots, X_n \), each with the distribution of \( X \).

The random variable, from which the sample is collected, is sometimes called **population**. It is important to understand that **before** collecting the sample, each \( X_i \) is a random variable,
but after the sample is collected, we observe a particular realization $x_i$ of observation $i$. Also note that each $X_i$ has the same distribution as $X$, so if $X$ has mean $\mu$ and variance $\sigma^2$, so does each $X_i$. The process of learning from the sample about the population is the purpose of the field of statistics. In particular, a sample is used in order to estimate some parameters of the population, for example the mean or variance.

**Definition 25** An estimator is a function of a random sample $X_1, \ldots, X_n$.

For example, a sample average is the most popular estimator of the population mean:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The notation $\bar{X}_n$ indicates that the average was calculated over a sample of size $n$. Popular estimators of population variance $\sigma^2$ are

$$v_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

It is important to understand that an estimator is a random variable. Since every observation in the sample, $X_i$ is a random variable, and any function of random variables is also a random variable, then any estimator is a random variable which has a mean, variance and pdf. Each time we calculate an estimator from a given sample, we obtain an estimate, i.e. a particular realization of the estimator.

The next table shows 5 samples of size 10, obtained from a toss of a die $X$. We know that $E(X) = 3.5$, which is the population mean.

<table>
<thead>
<tr>
<th>Sample1</th>
<th>Sample2</th>
<th>Sample3</th>
<th>Sample4</th>
<th>Sample5</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>X2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>X3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>X4</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>X5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>X6</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>X7</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>X8</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>X9</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>X10</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>$\bar{X}$</td>
<td>4.4</td>
<td>3.9</td>
<td>4.4</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Notice however that none of the sample averages is equal to 3.5. Does this mean that something is wrong with $\bar{X}_n$ as an estimator of the population mean? In general, we would like to have some criteria by which to judge how "good" is a particular estimator. We will introduce 3 such criteria: (i) unbiasedness, (ii) efficiency, and (iii) consistency.
1.6.1 Unbiased estimators

The most important property of all estimators, is that they are all random variables. Thus, each estimator has a mean and a variance. Intuitively, we would like our estimators to be "correct" on average, which means that we would like the mean of an estimator to be equal to the population parameter that we estimate.

**Definition 26** Let \( \hat{\theta} \) be an estimator of some parameter \( \theta \) of population. We say that \( \hat{\theta} \) is **unbiased** estimator of \( \theta \) if

\[
E(\hat{\theta}) = \theta
\]

In words, an estimator of some population parameter is unbiased if its expected value is equal to the true population parameter. If this is not true, the estimator is said to be **biased**, and the bias is

\[
\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta
\]

Note that unbiasedness of estimator does not mean that each time it is used, the estimate equals the true population parameter. Instead, the unbiasedness property means that on average the estimator equals the true population parameter. Put differently, if we collect many samples \( s = 1, \ldots, N \), and for each sample we obtain an estimate \( \hat{\theta}_s \), then we will have \( \frac{1}{N} \sum_{s=1}^{N} \hat{\theta}_s \to \theta \) as \( N \to \infty \). It turns out that the sample average is an unbiased estimator of the population mean.

The next figure illustrates the distribution of two estimators of the parameter \( \theta \), when the true value is \( \theta = 2 \).

The dotted pdf is centered around the true parameter value, and therefore unbiased. Notice however, that even when we use the biased estimator (solid line), it is possible to get estimates close to the true value of \( \theta = 2 \). On the other hand, even if we use the unbiased estimator (dotted line), it is possible to obtain estimates that are far from the true value of \( \theta = 2 \). So the fact that we use an unbiased estimator, does not guarantee that the estimate will be close to the true value of the population parameter.
Chapter 1. Preliminaries

Theorem 4 Let \(X_1, \ldots, X_n\) be a random sample from some population which has mean \(\mu\), and let the sample average be \(\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i\). Then, \(\bar{X}_n\) is an unbiased estimator of \(\mu\), i.e.
\[
E(\bar{X}_n) = \mu
\]

Proof. The key element in this proof, is the fact that \(E(X_i) = \mu \forall i\). This is the property of a random sample, that the mean of each observation is equal to the population mean.

\[
E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right)
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mu
\]
\[
= \frac{1}{n} n\mu = \mu
\]

This means that our method of estimating the population mean is on average correct, but only by coincidence will we ever obtain an estimate which is exactly equal to the population parameter.

Exercise 30 Let \(X_1, X_2, X_3\) be a random sample. Consider the following estimators of the population mean \(\mu\):

(a) \(\hat{\theta}_1 = \frac{1}{3} X_1 + \frac{1}{3} X_2 + \frac{1}{3} X_3\)
(b) \(\hat{\theta}_2 = \frac{1}{4} X_1 + \frac{1}{2} X_2 + \frac{1}{4} X_3\)
(c) \(\hat{\theta}_3 = \frac{1}{2} X_1 + \frac{1}{2} X_2 + \frac{1}{2} X_3\)

Check whether these estimators are biased or not. If an estimator is biased, find its bias.

Theorem 5 Let \(X_1, \ldots, X_n\) be a random sample from some population which has variance \(\sigma^2\), and \(v_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2\). Then, \(v_n\) is a biased estimator of \(\sigma^2\).

Proof. By definition of \(\sigma^2\), we have for all \(i\):
\[
\sigma^2 = E[(X_i - \mu)^2] = E[(X_i - \bar{X}_n + \bar{X}_n - \mu)^2]
\]
\[
\sigma^2 = E[(X_i - \bar{X}_n)^2 + (\bar{X}_n - \mu)^2 + 2(X_i - \bar{X}_n)(\bar{X}_n - \mu)]
\]

Sum over all \(i\) and divide both sides by \(n\):
\[
\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + \frac{1}{n} \sum_{i=1}^{n} (\bar{X}_n - \mu)^2 + 2\frac{1}{n} (\bar{X}_n - \mu) \sum_{i=1}^{n} (X_i - \bar{X}_n)
\]

Notice that \(\sum_{i=1}^{n} (X_i - \bar{X}_n) = 0\), because
\[
\sum_{i=1}^{n} (X_i - \bar{X}_n) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X}_n = \sum_{i=1}^{n} X_i - n\bar{X}_n
\]
and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Thus, we have

$$\sigma^2 = E \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + \frac{1}{n} \sum_{i=1}^{n} (\bar{X}_n - \mu)^2 \right]$$

$$= E (v_n) + \frac{1}{n} \sum_{i=1}^{n} E \left[ \frac{(\bar{X}_n - \mu)^2}{\text{var}(\bar{X}_n)} \right]$$

$$= E (v_n) + \text{var} (\bar{X}_n)$$

$$= E (v_n) + \frac{\sigma^2}{n}$$

and

$$E (v_n) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \left( \frac{n-1}{n} \right)$$

Thus, the bias of $v_n$ is $-\frac{\sigma^2}{n}$, which means that $v_n$ underestimates the population variance.

**Theorem 6** Let $X_1, ..., X_n$ be a random sample from some population which has variance $\sigma^2$, and let $s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$. Then, $s_n^2$ is an unbiased estimator of $\sigma^2$.

**Proof.** Notice that

$$s_n^2 = \frac{n}{n-1} v_n$$

$$E (s_n^2) = \frac{n}{n-1} E (v_n) = \frac{n}{n-1} \sigma^2 \left( \frac{n-1}{n} \right) = \sigma^2$$

As a result of the last theorem, it is common to refer to $s_n^2$ as the sample variance or as the unbiased estimator of population variance.

### 1.6.2 Efficient estimators

The most important property of all estimators, is that they are all random variables. Thus, each estimator has a mean and a variance. In the last section we stated that it would be nice if an estimator is "correct" on average, i.e. unbiased. Suppose we have several estimators, all of which are unbiased, how do we choose between them? Intuitively, if two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased, and $\text{var} (\hat{\theta}_2) < \text{var} (\hat{\theta}_1)$, then we would prefer estimator $\hat{\theta}_2$ over $\hat{\theta}_1$. Smaller variance of an estimator means that the estimator is more accurate or more precise.
Definition 27 Suppose $\hat{\theta}_1, ..., \hat{\theta}_N$ are all unbiased estimators of the population parameter $\theta$, and suppose that $\hat{\theta}_i$ has the smallest variance among all of them. Then we say that $\hat{\theta}_i$ is efficient (or best unbiased or minimum-variance) estimator among the $N$ unbiased estimators which we compare.

The next figure illustrates two estimators of the population parameter, whose true value is $\theta = 2$. Both of these estimators are centered around the true population parameter $\theta = 2$, i.e. both are unbiased.

![PDF of estimators, true value $\theta = 2$](image)

Notice however, that the dashed line (labeled inefficient) has greater variance than that of the efficient estimator. Whether we use the efficient or inefficient estimator, there is a chance of getting an estimate far from the true value $\theta = 2$. However, when we use the efficient estimator the chances that we will obtain an estimate way off the true value are smaller than with inefficient estimator.

Exercise 31 Consider the following estimators of the population mean $\mu$:

(a) $\hat{\theta}_1 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$
(b) $\hat{\theta}_2 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3$
(c) $\hat{\theta}_3 = \frac{1}{8}X_1 + \frac{3}{4}X_2 + \frac{1}{8}X_3$

Verify whether these estimators are biased or not, and find the efficient estimator (among the unbiased).

1.6.3 Conflicts between bias and low variance (MSE criterion)

What if we have two estimators, $\hat{\theta}_1$ unbiased, and $\hat{\theta}_2$ which is biased, and $\text{var}(\hat{\theta}_2) < \text{var}(\hat{\theta}_1)$? Thus, the second estimator is biased, but it is more precise. How do we choose
between $\hat{\theta}_1$ and $\hat{\theta}_2$? The next figure illustrates such a dilemma.

The dashed line is the pdf of an unbiased estimator, while the solid line is the pdf of a biased estimator. However, the biased estimator has lower variance, which is a nice property.

One way to choose between estimators that might be biased, is to apply the **Mean Squared Error** criterion. Suppose that $\hat{\theta}$ is an estimator of some unknown parameter $\theta$. The error of an estimator, for a given sample, is $\hat{\theta} - \theta$, and the mean (expected value) of squared error is defined as:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

The reason for squaring is to equally treat the positive and negative errors of same size. The above can be written as

$$MSE(\hat{\theta}) = E(\hat{\theta}^2 + \theta^2 - 2\theta \hat{\theta}) = E(\hat{\theta}^2) + \theta^2 - 2\theta E(\hat{\theta})$$

Notice that $\theta$ is some constant, while $\hat{\theta}$ is a random variable. Recall that $var(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$. Substituting in the last equation,

$$MSE(\hat{\theta}) = var(\hat{\theta}) + E(\hat{\theta})^2 + \theta^2 - 2\theta E(\hat{\theta})$$

Recall the quadratic formula $(a - b)^2 = a^2 + b^2 - 2ab$, and notice that the last three terms are $E(\hat{\theta} - \theta)^2$, and the term in the brackets is the definition of the bias of $\hat{\theta}$. Thus,

$$MSE(\hat{\theta}) = var(\hat{\theta}) + bias(\hat{\theta})^2 \quad (1.7)$$

Equation (1.7) gives us a criterion for choosing among estimators that are not necessarily unbiased. According to the MSE criterion, we will choose an estimator which give the smallest $MSE$. We can think of $MSE(\hat{\theta})$ as a sum of two bad things: variance and bias (squared), and we are looking for the best compromise of the two.
Example 3 Consider two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$:

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\theta}_1$</th>
<th>$\hat{\theta}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>variance</td>
<td>1.5</td>
<td>3</td>
</tr>
<tr>
<td>bias</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>

Which estimator would we choose based on the MSE criterion?

\[
MSE(\hat{\theta}_1) = 1.5 + (-1)^2 = 2.5
\]
\[
MSE(\hat{\theta}_2) = 3 + 0^2 = 3
\]

The estimator $\hat{\theta}_1$ has lower MSE, and therefore we will prefer it to $\hat{\theta}_2$, despite the fact that $\hat{\theta}_1$ is biased. The smaller variance of $\hat{\theta}_1$ turns out to more than compensate for the bias.

Remark. In this course you will not be required to design estimators. Designing estimators that overcome various statistical problems is a very complicated task, and several Nobel prizes in economics were awarded for development of new estimators. What you will do in this course is analyze whether a given estimator is biased or not, finding an efficient estimator among a group of unbiased estimators, and applying the minimum $MSE$ criterion in cases of tradeoff between bias and variance.

1.6.4 Consistent estimators

In our previous discussion, we compared different estimators according to their bias and variance, but in all comparisons the sample size was fixed. In reality though it is sometimes possible (although costly) to increase the size of a sample. Intuitively, we realize that larger sample size is "better" in some sense, but what exactly happens to an estimator when the sample size goes up? For most estimators that we consider, the variance of an estimator decreases as the sample size increases. For example, consider the sample average be $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ as an estimator of the population mean $\mu$. The variance of this estimator vanishes, as the samples size goes to infinity:

\[
\text{var} \ (\bar{X}_n) = \text{var} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var} (X_i) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}
\]

\[
\lim_{n \to \infty} \text{var} \ (\bar{X}_n) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0
\]

This means that as the sample size becomes very large, the variance of this estimator converges to zero, which means that the estimator becomes a constant number. The next figure
1.6. **SAMPLING AND ESTIMATORS**

illustrates what happens to the distribution of $X_n$ as the sample size $n$ increases.

![PDF of the sample mean for various sample sizes](image)

All the curves in the above figure illustrate the pdf of the same estimator $X_n$, but for different sample sizes $n$.

This example is not unique, and in fact many estimators converge to some constant number, as the sample size becomes infinitely large. However, we would also like the estimator to converge to the true population parameter which we are trying to estimate. Such estimator is called **consistent**.

**Definition 28** An estimator $\hat{\theta}_n$, based on the sample $X_1, \ldots, X_n$, of a true population parameter $\theta$, is **consistent** if for all $\varepsilon > 0$

$$\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0$$

The definition requires that the probability that an estimator is more than $\varepsilon$ away from the true population parameter should become smaller when we increase the sample size. The limit in the definition is called probability limit, or $\text{plim}$ in short, so we can write instead of $\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0$,

$$\text{plim} \left(\hat{\theta}_n\right) = \theta$$

In practice, we don’t use the definition to check whether an estimator is consistent or not. Instead we have the following test.

**Theorem 7** An estimator $\hat{\theta}_n$ is consistent if (i) $\text{var} \left(\hat{\theta}_n\right) \to 0$ as $n \to \infty$, and (ii) $\text{bias} \left(\hat{\theta}_n\right) \to 0$ as $n \to \infty$.

In other words, both the bias and the variance of an estimator should vanish as the sample size becomes infinitely large. From the above theorem, we learn that if an estimator is unbiased, for it to be consistent we only need to check that its variance goes to zero when the sample size goes to infinity.
For example, we had seen that the sample mean, \( \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \), is an unbiased estimator of the population mean. We also showed that \( \lim_{n \to \infty} \text{var} (\bar{X}_n) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0 \). This means that \( \bar{X}_n \) is a consistent estimator of the sample mean. We write:

\[
\text{plim} (\bar{X}_n) = \mu
\]

This is a very general result, called the law of large numbers. In words it means that the sample mean, regardless of the underlying population, converges (in probability) to the population mean.

**Exercise 32** Let \( \tilde{X}_n = \frac{1}{n+1} \sum_{i=1}^{n} X_i \) be an estimator of the population mean based on the sample \( X_1, \ldots, X_n \).

(a) Show that \( \tilde{X}_n \) is biased estimator of the population mean, and derive the bias.
(b) Show that \( \tilde{X}_n \) is consistent estimator of the population mean.

Why would anybody care about consistency of an estimator? After all, in reality there are no cases where a researcher can use an infinite sample size, so all samples in practice are finite. The first reason is that in econometrics, we are often forced to use a biased estimator. However, when the sample size is big enough, the bias becomes negligible. For example, see the last exercise. The second reason is that sometimes it is hard or impossible to check whether an estimator is biased or not in finite samples. The math is too complicated sometimes. On the other hand, it is much easier to check whether a given estimator is consistent. It is better to be able to say that the bias is becoming small when the sample becomes big, than not being able to say anything about the bias.

### 1.6.5 Distribution of estimators

So far we focused attention on the mean and variance of estimators. Can we find out the entire distribution of an estimator (i.e. its pdf)? In this section we plot the pdfs of estimators in order to illustrate concepts such as bias, efficiency, consistency. Indeed, sometimes we can figure out the entire distribution of an estimator, if we are willing to make assumptions about the distribution of population from which the sample is collected. At other times, we are able to find the asymptotic distribution of an estimator, i.e. its limiting distribution as the sample size becomes large. The advantage of knowing the distribution of estimators, is that we can conduct statistical tests. The next two results give the distribution of estimators based on samples from normal population.

**Theorem 8** Suppose that \( X_1, \ldots, X_n \) is a random sample from normal distribution, i.e. \( X_i \)s are independent and each \( X_i \sim N (\mu, \sigma^2) \). Then

\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N \left( \mu, \frac{\sigma^2}{n} \right)
\]

This means that the sample mean, based on a sample from normal population, is also normally distributed. Notice that the variance of \( \bar{X}_n \) declines to zero as the sample size gets large.
Theorem 9 Suppose that \( X_1, ..., X_n \) is a random sample from normal distribution, i.e. \( X_i \)s are independent and each \( X_i \sim N(\mu, \sigma^2) \).

\[
\frac{s_n^2 (n-1)}{\sigma^2} = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}_n}{\sigma} \right)^2 \sim \chi^2_{n-1}
\]

This means that the sample variance, based on a sample from normal population, when scaled by \((n - 1) / \sigma^2\), has a chi-square distribution, with \( n - 1 \) degrees of freedom.

Exercise 33 Suppose that \( X_1, ..., X_n \) is a random sample from normal distribution, and let

\[ Y = \frac{s_n^2 (n-1)}{\sigma^2} \]

What is the mean and variance of \( Y \)?

Obviously in reality we never have samples from exactly normal populations. In fact, most populations that you can imagine are not normal. In such cases, we can still say something about the **limiting distribution** of some estimators. This is accomplished with the help of the **Central Limit Theorem**, presented above. The CLT says that the sample mean based on **any population** converges to a normal distribution when the sample size gets large:

\[ \bar{X}_n \xrightarrow{D} N \left( \mu, \frac{\sigma^2}{n} \right) \]

Using extensions of the above CLT, statisticians often can derive an **asymptotic** distribution of their estimators, i.e. an approximate distribution, with the approximation improving as the sample size gets large.
Chapter 2

Simple Regression Analysis

In section 1.4.3 we introduced covariance and correlation, which measure the degree of co-
movement between a pair of random variables. Sometimes economic theory suggests that
one variable, usually known as the dependent variable, is determined by other variables,
usually known as explanatory variables, independent variables, or regressors. For
example, economic theory suggests that the quantity demanded of some good $X$ depends on
its price, income of buyers, prices of other related goods:

$$Q^D_X = f(P_X, I, P_{other})$$

The purpose of regression analysis is to quantify this relationship. The correlation analy-
sis might tell us that the price of a good $P_X$ and the quantity demanded $Q^D_X$ are negatively
correlated, but regression analysis attempts to estimate the function $f$ itself. Consequently,
regression analysis allows researchers to predict (i) the percentage change in quantity de-
manded when the price goes up by 1%, (ii) the percentage change in quantity demanded
when the buyers’ income goes up by 1%, (iii) the percentage change in quantity demanded
when a price of related good goes up by 1%.

The importance of answering such questions is immense. For example, our faculty Sudip
Chattopadhyay, studies the effectiveness of Tobacco Control Programs and the impact of
cigarette taxes on the demand for cigarettes in "State-Level Tobacco Control Programs
and the Demand for Cigarettes: A Retrospective Analysis Using Panel Data" (With David
Pieper). Another area where regression analysis is widely used is in estimation of wage equa-
tions. This field attempts to quantify the effect of schooling, experience, gender, race, and
other demographics, on the wages earned by individuals. For example, our faculty Don Mar,
studies whether there is discrimination against Asian American women in "Asian American
Women Professionals: Does Discrimination Exist?".

In general, regression analysis attempts to estimate the relationship between a dependent
variable $Y$ and some regressors (explanatory variables) $X_1, ..., X_k$, i.e. $Y = f(X_1, ..., X_k)$. We start with simple regression analysis, where we assume only one explanatory variable
$X$. We also assume that $f (\cdot)$ is a linear function, and all other factors that affect $Y$, whether
they are observed or unobserved, will be summarized by a random variable $u$, known as the
disturbance term or error term. Thus, the simple regression model is:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$  \hspace{1cm} (2.1)
where \( Y_i \) is the value of the dependent variable in observation \( i \), \( X_i \) is the value of the independent variable in observation \( i \) (also called explanatory variable or regressor) \( \beta_1 \) and \( \beta_2 \) are two unknown population parameters (which we will estimate) and \( u_i \) is the disturbance term of observation \( i \). We can view the right hand side of (2.1) as the sum of (i) \( \beta_1 + \beta_2 X_i \), i.e. the nonrandom component (given the value of \( X_i \)), and (ii) \( u_i \), the random component. The slope parameter \( \beta_2 \) represents the effect of a 1 unit increase in \( X_i \) on \( Y_i \). The intercept parameter represent the expected value of \( Y_i \) when \( X_i = 0 \). The next section shows how we can obtain estimates of \( \beta_1 \) and \( \beta_2 \), given a sample of \( Y_1, ..., Y_n; X_1, ..., X_n \), using the method of Ordinary Least Squares (OLS).

### 2.1 Deriving the OLS Estimators

This section derives the Ordinary Least Squares estimators for the unknown parameters \( \beta_1 \) and \( \beta_2 \). At this point we assume that we don’t know anything about the distribution of the disturbance term \( u_i \), and we definitely don’t know the population parameters \( \beta_1 \) and \( \beta_2 \). All we have is a sample of values for dependent variable \( Y \) and regressor \( X \), i.e. we have \( 2n \) numbers: \( Y_1, ..., Y_n; X_1, ..., X_n \) (data). For any estimates \( b_1 \) and \( b_2 \) we define the fitted equation (or estimated equation):

\[
\hat{Y}_i = b_1 + b_2 X_i
\]

where \( \hat{Y}_i \) is the fitted value (or predicted value) of \( Y_i \).

The question is, how do we obtain "good" estimates \( b_1 \) and \( b_2 \)? Intuitively, we want the fitted values \( \hat{Y}_i \) to fit "well" the actual values \( Y_i \). We define \( e_i \) to be the residual associated with each observation:

\[
e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i
\]

These \( e_i \)s can be thought of as prediction errors. According to the OLS method, the estimates \( b_1^{OLS} \), \( b_2^{OLS} \) should be chosen so that we minimize the Residual Sum of Squares, RSS:

\[
RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2
\]

**Definition 29** The OLS estimators \( b_1^{OLS}, b_2^{OLS} \) of the true parameters \( \beta_1, \beta_2 \) are the values of \( b_1 \) and \( b_2 \) which minimize the RSS. Formally,

\[
b_1^{OLS}, b_2^{OLS} = \arg\min_{b_1, b_2} \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2
\]

The OLS estimators can be derived using standard calculus techniques.

\[
[b_1] : \quad \frac{\partial RSS}{\partial b_1} = -\sum_{i=1}^{n} 2 (Y_i - b_1 - b_2 X_i) = 0
\]

\[
[b_2] : \quad \frac{\partial RSS}{\partial b_2} = -\sum_{i=1}^{n} 2 (Y_i - b_1 - b_2 X_i) X_i = 0
\]
Rearranging the above equations, after canceling $-2$, gives the so-called normal equations:

\begin{align*}
\sum_{i=1}^{n} Y_i - nb_1 - b_2 \sum_{i=1}^{n} X_i &= 0 
\tag{2.2} \\
\sum_{i=1}^{n} Y_i X_i - b_1 \sum_{i=1}^{n} X_i - b_2 \sum_{i=1}^{n} X_i^2 &= 0 
\tag{2.3} 
\end{align*}

From (2.2) we have

\begin{align*}
\frac{1}{n} \sum_{i=1}^{n} Y_i - b_2 \frac{1}{n} \sum_{i=1}^{n} X_i &= b_1 \\
&= \bar{Y} - b_2 \bar{X}
\end{align*}

Substituting this into (2.3), gives:

\begin{align*}
\sum_{i=1}^{n} Y_i X_i - (\bar{Y} - b_2 \bar{X}) n \bar{X} - b_2 \sum_{i=1}^{n} X_i^2 &= 0 \\
\sum_{i=1}^{n} Y_i X_i - n \bar{X} \bar{Y} + b_2 n \bar{X}^2 - b_2 \sum_{i=1}^{n} X_i^2 &= 0
\end{align*}

Thus,

\[ b_2^{OLS} = \frac{\sum_{i=1}^{n} Y_i X_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2} \]

This formula can be rewritten in a way which is more intuitive and more useful. We need two results:

\begin{align*}
[\text{Result 1}] & : \sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X} \bar{Y}) \\
&= \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i - \bar{X} \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \bar{X} \bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y} - n \bar{X} \bar{Y} + n \bar{X} \bar{Y} \\
&= \sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}
\end{align*}
CHAPTER 2. SIMPLE REGRESSION ANALYSIS

[Result 2] : \[ \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \]
\[ = \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}^2 \]
\[ = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \]
\[ = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \]

Thus, using result 1 for the numerator of \( b_2^{OLS} \) and result 2 for the denominator, we have the OLS estimators:

\[ b_2^{OLS} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad (2.4) \]
\[ b_1^{OLS} = \bar{Y} - b_2^{OLS} \bar{X} \quad (2.5) \]

Thus, the slope estimator \( b_2^{OLS} \) can be viewed as the ratio of the sample covariance between \( X \) and \( Y \) to the variance of \( X \)\(^1\)

\[ b_2^{OLS} = \frac{s_{XY}}{s_X^2} = \frac{r_{XY} \cdot s_X \cdot s_Y}{s_X^2} = r_{XY} \frac{s_Y}{s_X} \]

Thus, if \( X \) and \( Y \) are positively correlated in the sample, the slope estimate \( b_2^{OLS} \) will be positive. The magnitude of the slope depends on the variation of \( Y \) around its mean relative to the variation of \( X \) around its mean. From equation (2.5) we see that the estimated regression line passes through the sample averages point \( (\bar{X}, \bar{Y}) \). So after obtaining an estimate for the slope \( b_2 \), we use \( \bar{X}, \bar{Y} \) and \( b_2^{OLS} \) to find \( b_1^{OLS} \). For example, if \( \bar{X} = 12 \), \( \bar{Y} = 50 \) and \( b_2^{OLS} = 3 \), then

\[ b_1^{OLS} = \bar{Y} - b_2^{OLS} \bar{X} = 50 - 3 \cdot 12 = 14 \]

Of course, in practice, we do not use equations (2.4) and (2.5) to calculate the estimates by hand. Instead, computer software does that for us. To obtain OLS estimates in R we use the command `model <- lm(Y ~ X, data = data_name)`, and to display the results use `summary(model1)`. In Stata the command `regress Y X` estimates the model in (2.1) and displays the results.

\(^1\)The variance of the sample \( \{X_i\}_{i=1}^{n} \) is \( s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \) and the covariance between \( \{X_i\}_{i=1}^{n} \) and \( \{Y_i\}_{i=1}^{n} \) is \( s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \). With standard deviations \( s_X = \sqrt{s_X^2} \) and \( s_Y = \sqrt{s_Y^2} \), the correlation between the two samples is defined as \( r_{XY} = \frac{s_{XY}}{s_X s_Y} \).
2.2 Graphical Illustration of OLS

The next figure illustrates the OLS estimation from the previous section. The diamond dots represent the true data points \((Y_i, X_i)\), and the fitted values are given by \(\hat{Y}_i = b_1 + b_2 X_i\).

\[RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2\]

The estimators \(b_1\) and \(b_2\) chosen in such a way are called the OLS estimators.

2.3 Interpretation of Estimated Regression

After estimating the regression model, you get the fitted equation (also called predicted equation, or estimated equation):

\[\hat{Y}_i = b_1 + b_2 X_i\]

Always remember that the chances are that the estimates \(b_1\) and \(b_2\) are not going to be exactly the same as the true population parameters \(\beta_1\) and \(\beta_2\), and if by any chance they
are exactly equal, you will never find that out. To interpret the slope coefficient, we say that based on our estimates, we predict that an increase in $X$ by one unit will change $Y$ by $b_2$ units. When interpreting the slope coefficient it is important to know exactly what are the units of $X$ and $Y$. Suppose that $Y$ is hourly wage in dollars, and $X$ is schooling in years, and suppose that $b_2 = 2.5$. Then we say that an increase in schooling by 1 year is predicted to increase hourly wage by $2.5$ per hour.

The estimate of the constant, $b_1$, usually does not have a sensible interpretation. Mathematically, $b_1$ is the predicted value of the dependent variable when $X_i = 0$. Often, this mathematical interpretation does not make sense. For example, suppose that we regress persons’ weight on their height (i.e. $Y_i$ is person’s $i$ weight and $X_i$ person’s $i$ height). In this case, the estimate $b_1$ is the predicted weight of a person with zero height. Obviously, in reality it is not possible to have zero height of a person. In addition, the value of $b_1$ can be negative, which means that the predicted weight of a person is negative - which is impossibility. Therefore, when regression results are interpreted, we will focus primarily on the slope coefficients.

### 2.4 Results About the Estimated Model

**Theorem 10** Suppose that the true regression model is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and we fit an equation $\hat{Y}_i = b_1 + b_2 X_i$ using OLS based on the sample $Y_1, \ldots, Y_n; X_1, \ldots, X_n$. Then:

(a) $\bar{e} = 0$,

(b) $\bar{Y} = \bar{\hat{Y}}$,

(c) $\sum_{i=1}^{n} X_i e_i = 0$,

(d) $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$.

**Proof.** (a) The average (and the sum) of residuals in the sample is zero: $\bar{e} = 0$.

\[
\bar{e} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i) \\
= \frac{1}{n} (n\bar{Y} - nb_1 - b_2 n\bar{X}) \\
= \bar{Y} - b_1 - b_2 \bar{X} = 0
\]

The last equality follows from the first order condition for optimal $b_1$, equation (2.2).

(b) The average of the predicted values for the dependent variable in the sample is equal to the average of the true values of the dependent variable in the sample: $\bar{\hat{Y}} = \bar{Y}$.

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n} \sum_{i=1}^{n} (Y_i - e_i) \\
= \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{n} \sum_{i=1}^{n} e_i = \bar{Y}
\]
(c) $\sum_{i=1}^{n} X_i e_i = 0$. This says that the sample covariance between the explanatory variable and the residual is zero.

$$
\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i (Y_i - b_1 - b_2 X_i)
= \sum_{i=1}^{n} (X_i Y_i - b_1 X_i - b_2 X_i^2)
= \sum_{i=1}^{n} X_i Y_i - b_1 \sum_{i=1}^{n} X_i - b_2 \sum_{i=1}^{n} X_i^2 = 0
$$

The last equality follows from equation (2.3).

(d) $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$. This says that the sample covariance between the predicted values of dependent variable and the residual is zero.

$$
\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} (b_1 + b_2 X_i) e_i
= b_1 \sum_{i=1}^{n} e_i + b_2 \sum_{i=1}^{n} X_i e_i = 0
$$

We know that for each observation $i$ we have

$$
Y_i = \hat{Y}_i + e_i
$$

intuitively, "good" fit of the model means that the predicted (or fitted) values $\hat{Y}_i$ are close to the true values $Y_i$. Thus,

$$
Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + e_i
$$

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

$$
\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} + e_i\right)^2
$$

$$
\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left[\left(\hat{Y}_i - \bar{Y}\right)^2 + 2 \left(\hat{Y}_i - \bar{Y}\right) e_i + e_i^2\right]
$$

$$
\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2 \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) e_i + \sum_{i=1}^{n} e_i^2
$$

Notice that from theorem 10 we have
\[ 2 \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) e_i = 2 \left( \sum_{i=1}^{n} \hat{Y}_i e_i - \bar{Y} \sum_{i=1}^{n} e_i \right) = 0 \]

Thus, we obtain the following.
\[ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2 \] (2.6)

The term labeled \( TSS \) represents the **Total Sum of Squares**, i.e. the variation of the dependent variable around its sample mean. We showed that this variation is a sum of two parts: (i) \( ESS \), i.e. **Explained Sum of Squares** and (ii) \( RSS \), i.e. **Residual Sum of Squares** (or unexplained). Basically, equation (2.6) tells us that the total variation of the dependent variable from its sample mean, is in part explained by the fitted model (\( ESS \)), and in part unexplained (\( RSS \)).

\[ TSS = ESS + RSS \]

Dividing both sides by \( TSS \), gives
\[ 1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \]

This gives rise to the measure of **goodness of fit**:
\[ R^2 = \frac{ESS}{TSS} \]

The \( R^2 \) therefore gives the fraction of the total variation of the dependent variable from its sample mean, which is explained by the fitted model. Notice that \( 0 \leq R^2 \leq 1 \), and the maximal \( R^2 = 1 \) is obtained when \( RSS = 0 \) (when there is no unexplained part).

### 2.6 Impact of Changes in Units

Suppose that two researchers are studying the relationship between the same variables, and even use the same sample, but the units are different. For example, one researcher uses inches to measure length, and another uses centimeters, or one researcher uses pounds to measure weight, while another uses kilograms. The question we study in this section is, what happens to the OLS estimates when we switch to different units.

#### 2.6.1 Multiplying the data by constant

Suppose the sample of one researcher is \( X = \{X_1, ..., X_n\} \) and \( Y = \{Y_1, ..., Y_n\} \), and another researcher uses \( X^* = aX \), and \( Y^* = cY \), where \( a \) and \( c \) are some constants. This situation
arises when the first researcher uses inches for height and pounds for weight, and the second uses centimeters and kilograms. Let the estimates of the first researcher be \( b_1, b_2 \) and of the second be \( b_1^*, b_2^* \). Thus,

\[
\begin{align*}
\hat{b}_2^* &= \frac{\sum_{i=1}^{n} (aX_i - a\hat{X}) (cY_i - c\hat{Y})}{\sum_{i=1}^{n} (aX_i - a\hat{X})^2} = \frac{c}{a} \frac{\sum_{i=1}^{n} (X_i - \hat{X}) (Y_i - \hat{Y})}{\sum_{i=1}^{n} (X_i - \hat{X})^2} = \frac{c}{a} \hat{b}_2 \\
\hat{b}_1^* &= \hat{Y}^* - \hat{b}_2^* \hat{X}^* = c\hat{Y} - \frac{c}{a} \hat{b}_2 \cdot a\hat{X} = c (\hat{Y} - b_2 \hat{X}) = cb_1
\end{align*}
\]

To summarize, the OLS estimators based on the transformed data set are related to those based on the original data set as follows:

\[
\begin{align*}
\hat{b}_2^* &= \frac{c}{a} \hat{b}_2 \\
\hat{b}_1^* &= cb_1
\end{align*}
\]

### 2.6.2 Adding a constant to the data

Suppose the sample of one researcher is \( X = \{X_1, \ldots, X_n\} \) and \( Y = \{Y_1, \ldots, Y_n\} \), and another researcher uses \( X^* = X + h \), and \( Y^* = Y + k \), where \( h \) and \( k \) are some constants. Let the estimates of the first researcher be \( b_1, b_2 \) and of the second be \( b_1^*, b_2^* \). Thus,

\[
\begin{align*}
\hat{b}_2^* &= \frac{\sum_{i=1}^{n} (X_i + h - (\hat{X} + h)) (Y_i + k - (\hat{Y} + k))}{\sum_{i=1}^{n} (X_i + h - (\hat{X} + h))^2} = b_2 \\
\hat{b}_1^* &= \hat{Y}^* - \hat{b}_2^* \hat{X}^* = Y + k - b_2 (\hat{X} + h) = \hat{Y} - b_2 \hat{X} + k - b_2 h = b_1 + k - b_2 h
\end{align*}
\]

To summarize, the OLS estimators based on the transformed data set are related to those based on the original data set as follows:

\[
\begin{align*}
\hat{b}_2^* &= b_2 \\
\hat{b}_1^* &= b_1 + k - b_2 h
\end{align*}
\]

### 2.6.3 Multiplying by constant and adding a constant to the data

Suppose the sample of one researcher is \( X = \{X_1, \ldots, X_n\} \) and \( Y = \{Y_1, \ldots, Y_n\} \), and another researcher uses \( X^* = aX + h \), and \( Y^* = cY + k \), where \( h \) and \( k \) are some constants. An example of such case is temperature measured in Fahrenheit v.s. Celsius: \( F = \frac{9}{5}C + 32 \). Let the estimates of the first researcher be \( b_1, b_2 \) and of the second be \( b_1^*, b_2^* \). Thus,

\[
\begin{align*}
\hat{b}_2^* &= \frac{\sum_{i=1}^{n} (aX_i + h - (a\hat{X} + h)) (cY_i + k - (c\hat{Y} + k))}{\sum_{i=1}^{n} (aX_i + h - (a\hat{X} + h))^2} = \frac{c}{a} \hat{b}_2 \\
\hat{b}_1^* &= c\hat{Y} + k - \frac{c}{a} \hat{b}_2 (a\hat{X} + h) = c (\hat{Y} - b_2 \hat{X}) + k - \frac{c}{a} b_2 h = cb_1 + k - \frac{c}{a} b_2 h
\end{align*}
\]

To summarize, the OLS estimators based on the transformed data set are related to those based on the original data set as follows:

\[
\begin{align*}
\hat{b}_2^* &= \frac{c}{a} \hat{b}_2 \\
\hat{b}_1^* &= cb_1 + k - \frac{c}{a} b_2 h
\end{align*}
\]
CHAPTER 2. SIMPLE REGRESSION ANALYSIS

The purpose of these exercises with transformed data is to increase awareness to the fact that changes in units do affect the estimated coefficients and their interpretation.

2.7 Restricted Models

Recall that the original model that we estimate in this chapter has both intercept and slope:

\[ Y_i = \beta_1 + \beta_2 X_i + u_i \]

In this section we consider special cases of this model, (i) \( \beta_1 = 0 \), and (ii) \( \beta_2 = 0 \).

(i) : \( Y_i = \beta_2 X_i + u_i \)

(ii) : \( Y_i = \beta_1 + u_i \)

2.7.1 Model with no intercept, \( \beta_1 = 0 \)

The model we wish to estimate is

\[ Y_i = \beta_2 X_i + u_i \]

and the fitted equation is

\[ \hat{Y}_i = b_2 X_i \]

Therefore, the OLS estimator of \( \beta_2 \) minimizes the Residual Sum of Squares:

\[ \min_{b_2} \sum_{i=1}^{n} (Y_i - b_2 X_i)^2 \]

The first order condition:

\[ -2 \sum_{i=1}^{n} (Y_i - b_2 X_i) X_i = 0 \]

\[ \sum_{i=1}^{n} Y_i X_i - b_2 \sum_{i=1}^{n} X_i^2 = 0 \]

\[ b_2 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \]

There isn’t much we can say about this model. For example, the fitted equation does not necessarily pass through the sample means point \((\bar{X}, \bar{Y})\). In order to prove that a claim is false, we only need one counter example. So here is a counter example Suppose that the data consists of two observations on \( X \) and \( Y \):

\[
\begin{array}{cc}
X & Y \\
0 & 1 \\
1 & 2 \\
\end{array}
\]
Notice that $\bar{X} = 0.5$, and $\bar{Y} = 1.5$. The slope estimate is

$$b_2 = \frac{0 \cdot 1 + 1 \cdot 2}{0^2 + 1^2} = 2$$

The fitted equation is

$$\hat{Y}_i = 2 \cdot X_i$$

When plugging $X_i = \bar{X}$, we get

$$\hat{Y}_i = 2 \cdot \bar{X} = 2 \cdot 0.5 = 1 \neq \bar{Y} = 1.5$$

The general practice in applied research is to never set the intercept to zero. Therefore, the model $Y_i = \beta_2 X_i + u_i$ has no practical interest for us, and only serves as an exercise.

### 2.7.2 Model with no slope, $\beta_2 = 0$

The model we wish to estimate is

$$Y_i = \beta_1 + u_i$$

and the fitted equation is

$$\hat{Y}_i = b_1$$

Therefore, the OLS estimator of $\beta_1$ minimizes the Residual Sum of Squares:

$$\min_{b_1} \sum_{i=1}^{n} (Y_i - b_1)^2$$

The first order condition:

$$-2 \sum_{i=1}^{n} (Y_i - b_1) = 0$$

$$\sum_{i=1}^{n} Y_i - n b_1 = 0$$

$$n \bar{Y} - nb_1 = 0$$

$$b_1 = \bar{Y}$$

Intuitively, we are trying to predict (or fit) the dependent variable with just one number. According to OLS, this number should be the sample average of $Y$. This model is obviously useless, because it defies the purpose of regression analysis - estimating a relationship between a dependent variable $Y$ and some other explanatory variables. The model $Y_i = \beta_1 + u_i$ does not utilize any dependence of $Y$ on any other variable, and is presented here only as an exercise.
Chapter 3

Properties of Regression Coefficients and Hypothesis Testing

In the previous chapter we studied how to estimate the simple regression model:

\[ Y_i = \beta_1 + \beta_2 X_i + u_i \]

The OLS (Ordinary Least Squares) estimators of \( \beta_1 \) and \( \beta_2 \) are:

\[
\begin{align*}
    b_2 & = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
    b_1 & = \bar{Y} - b_2 \bar{X}
\end{align*}
\]

This chapter discusses the properties of these estimators. In order to study the properties of these estimators, we need to make several assumptions about the simple regression model being estimated. Then we will show that with these assumptions, the OLS estimators \( b_1 \) and \( b_2 \) are unbiased, efficient, and consistent estimators of \( \beta_1 \) and \( \beta_2 \).

3.1 Assumptions

A.0 The regressor \( X \) is non-random. In other words, the values of \( X_i \) for each observation \( i \) were hand-picked. For example, if \( X \) is schooling in years, with values 0, 1, 2, ..., \( S_{\text{max}} \), then we include in our sample a certain number of people from each education level. Relaxing this assumption will not change the results, but the math is a bit simpler with A.0 assumption.

A.1 The model is linear in parameters and has the following form:

\[ Y_i = \beta_1 + \beta_2 X_i + u_i \]

An example of a model which is nonlinear in parameters, is \( Y_i = \beta_1 X_i^{\beta_2} + u_i \), and such model sometimes cannot be estimated by OLS (Ordinary Least Squares). Nonlinear models can be estimated using some other technique, for example NLS (Nonlinear Least Squares).
A.2 **There is some variation in the regressor in the sample.** For example, if $X$ is schooling we must make sure that our sample consists of at least two different values for $X$. Imagine if we studied the relationship between schooling and wages, and everybody in our sample had 16 years of schooling. Intuitively, such sample would not allow us to estimate the relationship between schooling and wages. Indeed, when you look at the formula for the slope estimator, you see that when $X_i = X \forall i$, we cannot even calculate a value of $b_2$ because the formula gives us $\frac{0}{0}$. If there is no variation in a regressor in your sample, Stata and R (or any statistical package) will drop that regressor from your estimated equation.

A.3 **The disturbance (error) term has zero expectation.**

$$E(u_i) = 0 \text{ for all } i$$

This assumption must hold from the definition of $\beta_1$ (the intercept), which is the conditional mean of the dependent variable given that the regressors are zero.

$$\beta_1 = E(Y_i|X_i = 0) = E(\beta_1 + u_i) = \beta_1 + E(u_i)$$

This implies that $E(u_i) = 0$ for all $i$.

A.4 **The disturbance (error) term is homoscedastic.**

$$var(u_i) = \sigma_u^2 \text{ for all } i$$

Chapter 8 in these notes discusses what happens when this assumption is violated, and we call this problem **heteroscedasticity**.

A.5 **The disturbance (error) terms are independent.** This means $u_i$ and $u_j$ are statistically independent random variables for all $i \neq j$. Remember that the error term $u_i$ combines all the factors besides $X_i$ that affect the dependent variable $Y_i$. For example, if $Y$ is wage and $X$ is schooling, then the error term reflects all factors (observed and unobserved) that affect wages, besides schooling. For example, $u_i$ might reflect luck, motivation, ambition, work experience, IQ, gender, race, etc. of individual $i$. The current assumption says that these factors for individual $i$ are independent of the same factors possessed by individual $j$. If this assumption is violated, the resulting problem is called **autocorrelation**.

### 3.2 Mean and Variance of OLS Estimators

Recall that our goal is to show that, under some assumptions (A.0 - A.5), OLS estimators $b_1$ and $b_2$ are unbiased, efficient, and consistent estimators of $\beta_1$ and $\beta_2$. We therefore need to find the mean and variance of OLS estimators. We start with deriving a very useful result about any sample.
3.2. MEAN AND VARIANCE OF OLS ESTIMATORS

**Theorem 11** Let $X_1, ..., X_n$ be a sample of numbers, and let the sample average be $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

In words, this theorem says that the sum of deviations of all the observations from the sample average is always zero (in any sample).

**Proof.**

$$\sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = n\bar{X} - n\bar{X} = 0$$

This result should not be confused with the sum of squared deviations of observations from the sample mean. Intuitively, the mean of a sample is such a number that is located in the "middle" of the sample. So some observations are below the mean and some are above it. When you add up all the deviations from the mean, the negative deviations and positive ones cancel each other. This is not the case with sum of squared deviations from the mean, because all the squared deviations are positive and they don't cancel out with each other. The only time we have $\sum_{i=1}^{n} (X_i - \bar{X})^2 = 0$ is when all the $X_i$s are the same, so $X_i = \bar{X} \forall i$.

**Corollary 1** For any fixed number $a$, we have $\sum_{i=1}^{n} (X_i - \bar{X}) a = 0$.

### 3.2.1 Rewriting The OLS Estimators

We now rewrite the formulas of the OLS estimators, in such a way that would allow us to calculate their mean and variance, starting with $b_2$.

$$b_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X}) [(\beta_1 + \beta_2 X_i + u_i) - (\beta_1 + \beta_2 \bar{X} + \bar{u})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X}) [\beta_2 X_i + u_i - \beta_2 \bar{X} - \bar{u}]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X}) [\beta_2 (X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \beta_2 \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X}) (u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \beta_2 + \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right) u_i$$

$$= \beta_2 + \sum_{i=1}^{n} a_i u_i$$
Thus, we expressed the slope estimator $b_2$ as the sum of the true slope $\beta_2$ and a weighted average of the error terms, with weights

$$a_i = \frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$$

Since the observations on the regressor, $X$, are non-random (assumption A.0), it follows that the $a_i$-s are constant numbers.

Similarly, we now rewrite the intercept estimator $b_1$.

$$b_1 = \bar{Y} - b_2 \bar{X}$$

$$= \beta_1 + \beta_2 \bar{X} + \bar{u} - \left( \beta_2 + \sum_{i=1}^{n} a_i u_i \right) \bar{X}$$

$$= \beta_1 + \frac{1}{n} \sum_{i=1}^{n} u_i - \bar{X} \sum_{i=1}^{n} a_i u_i$$

$$= \beta_1 + \sum_{i=1}^{n} \left( \frac{1}{n} u_i - \bar{X} a_i u_i \right)$$

$$= \beta_1 + \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{X} a_i \right) u_i$$

$$= \beta_1 + \sum_{i=1}^{n} c_i u_i$$

Thus, we expressed the intercept estimator $b_1$ as the sum of the true intercept $\beta_1$ and a weighted average of the error terms, with weights

$$c_i = \frac{1}{n} - \bar{X} a_i$$

### 3.2.2 Unbiasedness of OLS estimators

In the last section we rewrote the OLS estimators as a sum of the true parameter and a weighted average of the error terms.

$$b_2 = \beta_2 + \sum_{i=1}^{n} a_i u_i \quad (3.1)$$

$$b_1 = \beta_1 + \sum_{i=1}^{n} c_i u_i \quad (3.2)$$
3.2. MEAN AND VARIANCE OF OLS ESTIMATORS

Remember that since the observations on \( X \) are non-random, the weights \( a_i \) and \( c_i \) are also non-random. Taking expectation of \( b_2 \):

\[
E (b_2) = \beta_2 + E \left( \sum_{i=1}^{n} a_i u_i \right)
\]
\[
= \beta_2 + \sum_{i=1}^{n} a_i E (u_i) \quad \text{(by assumption A.0, } X_i \text{s are fixed)}
\]
\[
= \beta_2 \quad \text{(by assumption A.3, } E (u_i) = 0 \ \forall i )
\]

This result proves that \( b_2 \) is an unbiased estimator of \( \beta_2 \).

The proof that \( b_1 \) is an unbiased estimator of \( \beta_1 \) is similar:

\[
E (b_1) = \beta_1 + E \left( \sum_{i=1}^{n} c_i u_i \right)
\]
\[
= \beta_1 + \sum_{i=1}^{n} c_i E (u_i)
\]
\[
= \beta_1 \quad \text{(because of assumption A.3, } E (u_i) = 0 \ \forall i )
\]

This proves that due to assumption A.3 (\( E (u_i) = 0 \ \forall i \)), both \( b_1 \) and \( b_2 \) are unbiased estimators of \( \beta_1 \) and \( \beta_2 \).

3.2.3 Efficiency of OLS estimators

Unbiasedness is one good property of estimators, but even when you use unbiased estimators, you can obtain estimates that are far away from the true values of the population parameters. We therefore would like to know how **precise or accurate** the OLS estimators are. In other words, we would like the OLS estimators to have the smallest possible variance. We start with the variance of \( b_2 \), using formula (3.1):

\[
\text{var} (b_2) = \text{var} \left( \beta_2 + \sum_{i=1}^{n} a_i u_i \right) = \sum_{i=1}^{n} a_i^2 \text{var} (u_i)
\]

The last step uses assumptions A.5 (\( u_i \)-s are independent), and therefore \( u_i \)-s are uncorrelated, and the variance of the sum is the sum of the variances. Next, by assumption A.4 (\( \text{var} (u_i) = \sigma_u^2 \) for all \( i \)), we have

\[
\text{var} (b_2) = \sigma_u^2 \sum_{i=1}^{n} a_i^2
\]

Substituting the expression of \( a_i \), gives:

\[
\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\left( \sum_{j=1}^{n} (X_j - \bar{X})^2 \right)^2} = \frac{1}{\sum_{j=1}^{n} (X_j - \bar{X})^2}
\]
Finally, the variance of $b_2$ is:

$$\text{var} \ (b_2) = \frac{\sigma_u^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$$

Similarly, we derive the variance of $b_1$, using formula (3.2).

$$\text{var} \ (b_1) = \text{var} \left( \beta_1 + \sum_{i=1}^{n} c_i u_i \right) = \sigma_u^2 \sum_{i=1}^{n} c_i^2$$

Again, the last step used assumption A.4 and A.5. Now we use the expression of $c_i$:

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{X} a_i \right)^2$$

$$= \sum_{i=1}^{n} \left( \frac{1}{n^2} - \frac{2\bar{X} a_i}{n} + \bar{X}^2 a_i^2 \right)$$

$$= \sum_{i=1}^{n} \frac{1}{n^2} - \frac{2\bar{X}}{n} \sum_{i=1}^{n} a_i + \bar{X}^2 \sum_{i=1}^{n} a_i^2$$

$$= \frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$$

Thus,

$$\text{var} \ (b_1) = \sigma_u^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right)$$

To summarize, we derived the variance of OLS estimators:

$$\text{var} \ (b_2) = \frac{\sigma_u^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \quad (3.3)$$

$$\text{var} \ (b_1) = \sigma_u^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right) \quad (3.4)$$

First notice that the smaller is the variance of the error term, $\sigma_u$, the more precise are the OLS estimators $b_1$ and $b_2$. Also notice that that $b_1$ and $b_2$ are more precise (have smaller variance) when the variation of the regressor in the sample is greater. Recall that assumption A.2 requires that there would be some variation of regressors in the sample; otherwise we won’t even be able to obtain the OLS estimates. Now we had shown that more variation of the regressors in the sample is better. To gain intuition for this result, consider the regression of earnings (dependent variable) on schooling (regressor). If the sample has only one level of schooling, say 12, we will not be able to estimate the relationship between earnings and schooling. In other words, the sample does not contain enough information on schooling, to be able to determine how earnings vary with schooling, since schooling does not vary in
3.2. MEAN AND VARIANCE OF OLS ESTIMATORS

this sample. Next consider a sample $S_1$ with two values of schooling, say 12 years and 16 years and sample $S_2$ with schooling of 7, 10, 12, 14, 16, 20 years. Intuitively, sample $S_2$ contains more information on schooling, which enables us to more accurately estimate the relationship between earnings and schooling.

Finally notice that $\text{var}(b_1)$ gets higher when the average values of the regressor are farther from zero, i.e. $\bar{X}^2$. This means that it is harder to obtain a precise value of the intercept when the data is far away from zero. For example, recall that in the earnings regression $E\text{ARNINGS}_i = b_1 + b_2 S_i$, we obtained a negative value of $b_1$, which is supposed to be the estimated wage of workers with zero schooling. The reason it happened is that in the sample we had individuals with at least 7 years of schooling. Now, after seeing equation (3.4), this makes sense. If our sample of workers included enough individuals with close to zero years of schooling, then we would have estimated more realistic $b_1 > 0$.

The Gauss-Markov Theorem

We have shown that, under assumptions 3.1, OLS estimators are unbiased, and we derived their variance (equations (3.3) and (3.4)). Next we ask whether the OLS estimators are efficient, i.e. do they have the smallest possible variance among unbiased estimators? The Gauss-Markov Theorem states that the OLS estimators are BLUE, i.e. Best among all Linear Unbiased Estimators. So we don’t compare the OLS estimators to all the possible estimators in the world, but only to other linear estimators. We call an estimator linear if it can be expressed as a linear function of the $Y_i$s (the observations on the dependent variable). Indeed, the OLS estimators are linear functions of $Y_i$s:

$$b_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \sum_{i=1}^{n} a_i Y_i$$

$$b_1 = \bar{Y} - b_2 \bar{X} = \sum_{i=1}^{n} \frac{1}{n} Y_i - \sum_{i=1}^{n} \bar{X} a_i Y_i = \sum_{i=1}^{n} c_i Y_i$$

Therefore, the OLS estimators are a linear combinations of the $Y_i$s.

Furthermore, we proved that OLS estimators are unbiased. Thus, the Gauss-Markov theorem states that in the class of linear and unbiased estimators, the OLS estimators have the smallest variance, so they are efficient. This is the meaning of the word Best, i.e. smallest variance.

Before we prove the Gauss-Markov theorem we prove two properties of $a_i$-s, which will be used in the proof.

**Lemma 12** The OLS weights

$$a_i = \frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$$
satisfy the following properties:

\[ [1] \quad : \quad \sum_{i=1}^{n} a_i = 0 \]
\[ [2] \quad : \quad \sum_{i=1}^{n} a_i X_i = 1 \]

**Proof.** Property 1.

\[
\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} = 0
\]

Property 2.

\[
\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n} \left( \frac{(X_i - \bar{X}) X_i}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right) = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) (X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}
\]

The second step uses the fact that \( \sum_{i=1}^{n} (X_i - \bar{X}) X = 0 \).

\[
\sum_{i=1}^{n} a_i X_i = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} = 1
\]

Now we are ready to prove the Gauss-Markov theorem.

**Theorem 13 Gauss-Markov.** Consider a class of estimators for \( \beta_1 \) and \( \beta_2 \), which are both (i) linear, and (ii) unbiased. Under assumptions A.0 - A.5 \[3\]. OLS estimators are the best in that class, i.e. \( b_1 \) and \( b_2 \) have the smallest variance among linear and unbiased estimators.

**Proof.** We will prove that the variance of \( b_2 \) is the smallest among all other linear and unbiased estimators of \( \beta_2 \). Let \( \tilde{b}_2 \) be another linear and unbiased estimator of \( \beta_2 \):

\[
\tilde{b}_2 = \sum_{i=1}^{n} (a_i + h_i) Y_i
\]

where \( a_i \) is the weight on \( Y_i \) in OLS:

\[
a_i = \frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}
\]

We need to prove that \( \text{var} \left( \tilde{b}_2 \right) \geq \text{var} \left( b_2 \right) \). First, we rewrite \( \tilde{b}_2 \) as follows:

\[
\tilde{b}_2 = \sum_{i=1}^{n} (a_i + h_i) Y_i = \sum_{i=1}^{n} (a_i + h_i) (\beta_1 + \beta_2 X_i + u_i)
\]

\[
= \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i + \sum_{i=1}^{n} (a_i + h_i) u_i
\]
The mean of $\tilde{b}$ is

$$E\left( \tilde{b}_2 \right) = E \left( \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i + \sum_{i=1}^{n} (a_i + h_i) u_i \right)$$

$$= \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i + E \left( \sum_{i=1}^{n} (a_i + h_i) u_i \right)$$

$$= \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i + \sum_{i=1}^{n} (a_i + h_i) E u_i$$

$$= \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i$$

In the last step we used assumption A.3 that $E(u_i) = 0$ for all $i$. The variance of $\tilde{b}_2$ is

$$\text{var}\left( \tilde{b}_2 \right) = \text{var} \left( \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i + \sum_{i=1}^{n} (a_i + h_i) u_i \right)$$

$$= \text{var} \left( \sum_{i=1}^{n} (a_i + h_i) u_i \right) \quad \text{(added constants don’t affect the variance)}$$

$$= \sum_{i=1}^{n} \text{var} ((a_i + h_i) u_i) \quad \text{($u_i$-s are independent, A.5)}$$

$$= \sum_{i=1}^{n} (a_i + h_i)^2 \text{var} (u_i)$$

$$= \sigma_u^2 \sum_{i=1}^{n} (a_i + h_i)^2 \quad \text{(var} \,(u_i) = \sigma_u^2 \forall i, \, A.4)$$

$$= \sigma_u^2 \left( \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2 + 2 \sum_{i=1}^{n} a_i h_i \right)$$

To summarize, the mean and variance of $\tilde{b}_2$ are:

$$E\left( \tilde{b}_2 \right) = \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i \quad (3.5)$$

$$\text{var} \left( \tilde{b}_2 \right) = \sigma_u^2 \left( \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2 + 2 \sum_{i=1}^{n} a_i h_i \right) \quad (3.6)$$

The unbiasedness of $\tilde{b}_2$ implies certain restrictions on the $h_i$-s:

$$E\left( \tilde{b}_2 \right) = \beta_1 \sum_{i=1}^{n} (a_i + h_i) + \beta_2 \sum_{i=1}^{n} (a_i + h_i) X_i = \beta_2$$
For this equation to hold for all possible values of the parameters $\beta_1$ and $\beta_2$, we must have

$$\sum_{i=1}^{n} (a_i + h_i) = 0, \sum_{i=1}^{n} (a_i + h_i) X_i = 1 \quad (3.7)$$

Together with lemma 12

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

the restrictions in (3.7) imply

$$\sum_{i=1}^{n} h_i = 0, \sum_{i=1}^{n} h_i X_i = 0 \quad (3.8)$$

Using the restriction on the $h_i$-s, we can prove that the last term in the brackets of (3.6) is zero:

$$\sum_{i=1}^{n} a_i h_i = \frac{\sum_{i=1}^{n} \left( X_i - \bar{X} \right)}{\sum_{j=1}^{n} \left( X_j - \bar{X} \right)^2} \sum_{j=1}^{n} h_i X_j - \bar{X} \sum_{i=1}^{n} h_i = 0$$

The last step is a direct result of the restrictions on $h_i$-s in (3.8), implied by the unbiasedness of $\tilde{b}_2$. Thus, the variance of $\tilde{b}_2$ in (3.6), becomes:

$$\text{var} \left( \tilde{b}_2 \right) = \sigma_u^2 \left( \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2 \right)$$

$$= \sigma_u^2 \sum_{i=1}^{n} a_i^2 + \sigma_u^2 \sum_{i=1}^{n} h_i^2$$

$$= \text{var} \left( b_2 \right) + \sigma_u^2 \sum_{i=1}^{n} h_i^2$$

Notice that because the second term includes squared terms, there is no way for $\text{var} \left( \tilde{b}_2 \right)$ to be smaller than $\text{var} \left( b_2 \right)$. The only way the two can be equal, is when $h_i = 0$ for all $i$, and in such case $\tilde{b}_2 = b_2$. Thus, $\text{var} \left( \tilde{b}_2 \right) \geq \text{var} \left( b_2 \right)$. □

The above proof contains many details. Let's briefly summarize the logic of it. We wanted to prove that the OLS estimator $b_2$ has the smallest variance among all linear and unbiased estimators of $\beta_2$. We examined a generic estimator $\tilde{b}_2$, which is linear and also unbiased estimator of $\beta_2$. We then showed that the unbiasedness of this generic estimator $\tilde{b}_2$ imposed some restrictions on its formula, and these restrictions in turn implied that the variance of $\tilde{b}_2$ can only be at least as large as that of the OLS estimator $b_2$. Thus, we proved that the OLS estimator $b_2$ has the smallest variance among linear and unbiased estimators of $\beta_2$. 
3.2.4 Consistency of OLS estimators

We have proved that OLS coefficients are unbiased. Therefore, in order to prove consistency, we only need to show that their variance converges to zero as the sample size becomes very large. The variances of OLS estimators in (3.3) and (3.4) can be rewritten as follows:

\[
\text{var}(b_2) = \frac{\sigma_u^2/n}{\frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^2}
\]

\[
\text{var}(b_1) = \frac{\sigma_u^2}{n} \left( 1 + \frac{\bar{X}^2}{\frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^2} \right)
\]

The term \( v_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \) was introduced in section 1.6 as the variance of the sample. It turns out from the Laws of Large Numbers theorems, that \( v_n \to \text{var}(X) \) as \( n \to \infty \) and also, \( \bar{X}^2 \to \mu_X^2 \). In other words, the sample moments converge to the true population moments, as the sample size grows infinitely large. Thus,

\[
\lim_{n \to \infty} \text{var}(b_2) = \lim_{n \to \infty} \frac{\sigma_u^2/n}{\sigma_X^2} = 0
\]

\[
\lim_{n \to \infty} \text{var}(b_1) = \lim_{n \to \infty} \frac{\sigma_u^2}{n} \left( 1 + \frac{\mu_X^2}{\sigma_X^2} \right) = 0
\]

To summarize, with assumption A.0 - A.5, the OLS estimators \( b_1 \) and \( b_2 \) are:

1. Unbiased estimators of \( \beta_1 \) and \( \beta_2 \),

2. Efficient (have smallest variance) among unbiased and linear estimators of \( \beta_1 \) and \( \beta_2 \),

3. Consistent estimators of \( \beta_1 \) and \( \beta_2 \).

3.3 General Idea of Statistical Hypothesis Tests

In the next section we demonstrate how to perform statistical tests about the unknown true regression coefficients \( \beta_1 \) and \( \beta_2 \). But first, lets try to develop some understanding about statistical tests in general, not only tests related to regression analysis.

Suppose that you wish to learn about some unknown population parameter \( \theta \). You collect a random sample \( X_1, ..., X_n \), and use an estimator \( \hat{\theta} \), which then gives you some estimate \( \hat{\theta} (X_1, ..., X_n) \) - a particular number. It is almost impossible that your estimate is equal to the true population parameter \( \theta \), and even if it is equal, there is no way for you to know that because the true parameter is unknown. For example, if your estimate of the population mean is 5, this does not mean that the true population mean is also 5. Maybe the true population parameter is 4 or 6? As far as we know, an estimator is a random variable, and different samples will give us a different estimate. So what do you learn from the value of your estimate? Not much, really. However, your intuition tells you that if a sample gave you an estimate of 5, it is very unlikely that the true population parameter is 0, or 10, or any other value which is "very far" from 5. This intuition lead to the development of
statistical hypothesis tests - a procedure that uses a random sample to learn about the true population parameter.

A statistical hypothesis test about the unknown population parameter \( \theta \), consists of 5 steps:

1. Forming a null hypothesis, \( H_0 \), which assigns one particular value to the unknown population parameter. For example, the null hypothesis can be \( H_0 : \theta = \theta_0 \), or \( H_0 : \theta = 5 \) or \( H_0 : \beta_2 = 0 \). It will become clear (hopefully) later, why the null hypothesis in a statistical test must assign only one value to the unknown population parameter.

2. Forming an alternative hypothesis, \( H_1 \), which is the logical conclusion that we will accept if the null hypothesis is not true. For example, if the null hypothesis is \( H_0 : \theta = 5 \), then we can have 3 possible alternatives:
   
   i. \( H_1 : \theta \neq 5 \), if we reject \( H_0 \), we conclude that \( \theta \) can be any value other than 5.
   ii. \( H_1 : \theta > 5 \), if we reject \( H_0 \), we conclude that \( \theta \) can be any value greater than 5.
   iii. \( H_1 : \theta < 5 \), if we reject \( H_0 \), we conclude that \( \theta \) can be any value smaller than 5.

In general, if the null hypothesis is \( H_0 : \theta = \theta_0 \), then the only 3 alternative hypotheses (from which the researcher can choose one) are: (i) \( H_1 : \theta \neq \theta_0 \) or (ii) \( H_1 : \theta > \theta_0 \) or (iii) \( H_1 : \theta < \theta_0 \). It will become clear (hopefully) later, why the alternative hypothesis must be only one of the 3 types listed here, and why we cannot have for example \( H_0 : \theta = 5 \) and \( H_1 : \theta = 4 \).

3. Specifying the test statistic, which is a function of the random sample (just like an estimator), but in addition, (i) the distribution of \( T \) must depend on the hypothesized value in the null hypothesis, \( \theta_0 \), and (ii) the distribution of \( T \) should not depend on any other unknown parameters. For example, suppose that we test hypotheses about the mean \( \mu \) of some normally distributed population \( X \sim N(\mu, \sigma^2) \), and we do not know neither \( \mu \) nor \( \sigma^2 \). The random sample is \( X_1, ..., X_n \) and our estimator of \( \mu \) is the population average \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \). Suppose that the null hypothesis is \( H_0 : \mu = \mu_0 \). The following is NOT a test statistic:

\[
Z = \frac{\bar{X} - \mu_0}{\sigma^2/n}
\]

The above quantity does depend on the null value \( \mu_0 \), and if indeed \( H_0 \) is true (i.e. \( \mu = \mu_0 \)), then this test statistic would have standard normal distribution: \( Z \sim N(0, 1) \). However, \( Z \) also depends on the unknown parameter \( \sigma^2 \), and therefore we are unable to calculate the value of \( Z \) based on the sample and the hypothesized value \( \mu_0 \).\(^1\) However,

\(^1\)If somehow we had knowledge about the population variance \( \sigma^2 \), then \( Z \) would be a test statistic. However, in reality, we do not know any of the population parameters. Furthermore, it seems unreasonable not to know the population mean, but nevertheless know its variance.
if we replace the unknown parameter $\sigma^2$ with its estimator (sample variance) $s_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, then we get a valid test statistic:

$$T = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2/n}}$$

Notice that the value of $T$ depends on the null value $\mu_0$, and on the sample, but it does not depend on any other unknown population parameter (such as $\sigma^2$). Therefore, $T$ is a valid test statistic, and it has $t$-distribution with $n - 1$ degrees of freedom: $T \sim t_{n-1}$.\(^2\)

Typically, a test statistic is designed to measure the distance between the hypothesized value of the unknown parameter, and the evidence. In the above example, notice that $T$ measures the distance between the hypothesized value $\mu_0$ and the estimator of $\mu$ (i.e. $\bar{X}$). The distance is measured in units of standard errors of the estimator.

4. Determine the rejection region - the range of values of the test statistic, that lead to the rejection of the null hypothesis. These are supposed to be unlikely values (have low probability) of $T$, when the null hypothesis is true, so that if $T$ attains one of these values, we reject the null hypothesis. If the test statistic does not attain a value in the rejection region, then we do not reject the null hypothesis. There is no scientific way to determine what "unlikely or "low probability" means. In practice, we choose this probability to be $\alpha = 5\%$ or $\alpha = 1\%$ and call it the significance level of the test. The rejection region also depends on which of the 3 types of the alternative hypothesis is chosen. If the alternative is $H_1 : \theta \neq \theta_0$, then the rejection region will contain "unusually high" and "unusually low" values of the test statistic. If the alternative is $H_1 : \theta > \theta_0$, then the rejection region will contain only "unusually high" values of the test statistic. Finally, if the alternative is $H_1 : \theta < \theta_0$, then the rejection region will contain only "unusually low" values of the test statistic.

5. Write the conclusion of the hypothesis test and put it in the context of your economic (or other) theory, which motivated the hypothesis test. For example, the conclusion is usually either "we reject the null hypothesis at significance level of $\alpha$" or "we fail to reject the null hypothesis at significance level of $\alpha$". Putting the conclusion in the context of our theory means that we add a statement such as "we conclude that schooling level has a positive effect on earnings of individuals" if $\beta_2$ is the regression coefficient on schooling and we rejected $H_0 : \beta_2 = 0$, against the alternative $H_1 : \beta_2 > 0$.

In the next example, we illustrate the 5 steps of hypothesis testing. Suppose that you want to conduct tests about some unknown population parameter $\theta$. Let the test statistic be $T(\theta_0)$, which depends on the sample and the hypothesized value of $\theta$ under the null hypothesis. It might help if for the sake of this example, we pretend that the test statistic is an unbiased estimator $\hat{\theta}$ of the true population parameter $\theta$, i.e. $T(\theta_0) = \hat{\theta}$.\(^3\) If the null hypothesis is $H_0 : \theta = 0$ is true, then the pdf of the test statistic will be centered around 0,

\(^2\)For partial proof of this result, see section 1.5 and specifically theorem 3.

\(^3\)In practice, most test statistics are indeed constructed based on unbiased or consistent estimators.
CHAPTER 3. PROPERTIES OF REGRESSION COEFFICIENTS AND HYPOTHESIS TESTING

as shown in the next figure by the solid curve.

![PDF of T(θ)](image)

However, if \( H_0 : \theta = 2 \) is true, then the pdf of the test statistic will be centered around 2, as shown in the above figure by the dashed curve.

Suppose that we want to test the null hypothesis \( H_0 : \theta = 0 \) against one of the three possible alternatives. If the realized value of the test statistic \( T(\theta_0 = 0) \) is "close" to 0, then this means that the sample supports the null hypothesis. But what if the realized test statistic, under the null hypothesis \( H_0 : \theta = 0 \), is \( T(\theta_0 = 0) = 3 \)? Is this value "unlikely" under the null hypothesis, in which case we should reject it? How do we measure what is "unlikely"? This depends on what do we want to "accept" in case that we reject the null hypothesis, or what is the alternative hypothesis, denoted by \( H_1 \). If we reject \( H_0 : \theta = 0 \), the alternatives could be:

1. \( H_1 : \theta \neq 0 \)
2. \( H_1 : \theta > 0 \)
3. \( H_1 : \theta < 0 \)

Usually, the theory should determine the alternative hypothesis. The first one is called a two-sided alternative, and the other two are one-sided alternatives. Thus, if the alternative is two-sided, \( H_1 : \theta \neq 0 \), we reject the null hypothesis when the test statistic is either far enough to the right, or far enough to the left of its pdf under the null hypothesis. If the alternative hypothesis is \( H_1 : \theta > 0 \), then we reject the null when the realized test statistic is far enough to the right of its pdf under the null hypothesis. Finally, if the alternative is \( H_1 : \theta < 0 \), we reject the null when the test statistic is far enough to the left of its pdf under the null hypothesis.

Suppose that under the null hypothesis \( H_0 : \theta = 0 \), the test statistic has standard normal
3.3. GENERAL IDEA OF STATISTICAL HYPOTHESIS TESTS

distribution, \( T (\theta_0) \sim N(0,1) \), as depicted in the following figure.

![Normal Distribution](image)

The shaded areas indicate rejection regions, and they illustrate what is considered "far enough" in practice. The common practice is for rejection areas to be 0.05 (or 5%).

In the case of a two-sided alternative (left panel) if the sample results in the value of the test statistic which falls into either of the two shaded areas, the null hypothesis is rejected. In this case, when the test statistic has standard normal distribution, the rejection region consists of all values greater than 1.96 or smaller than \(-1.96\). Under the null hypothesis, the probability of obtaining a test statistic is either greater than 1.96 or smaller than \(-1.96\), is 0.05. The numbers 1.96 and 1.96 are called critical values for the test, which separate the acceptance and rejection regions.

In the case of a one-sided positive alternative \( H_1 : \theta > 0 \), the null hypothesis is rejected only if the realized value of the test statistic is far enough to the right (middle panel). If the test statistic has standard normal distribution, this means rejecting the null hypothesis when the realized value of the test statistic is greater than \(1.65\). Again, if the null hypothesis was true, the chances of getting an estimate greater than \(1.65\) is 0.05, and in practice we say that this chance is too small and therefore we reject the null hypothesis.

Finally, the right panel shows the rejection area for the case of a one-sided negative alternative hypothesis \( H_1 : \theta < 0 \). In this case we reject the null hypothesis when the realized value of the test statistic is "far enough" to the left (or small enough). If the estimator has a standard normal distribution, this means that we reject the null hypothesis whenever the realized value of the test statistic is less than \(-1.65\). If the null hypothesis is true, the chances that a particular sample gives an estimate which is less than \(-1.65\) is 0.05, and in practice this is considered too small and therefore we reject the null hypothesis.

In general, the size of the rejection area is denoted by \( \alpha \), and is called the significance level of the statistical test. Thus, \( \alpha \) is the probability of rejecting the null hypothesis when it is true, a Type I error. In our example, the significance level was chosen to be \( \alpha = 0.05 \). This means that in a random sample, there is a 5% chance of rejecting the null hypothesis even when it is true. This chance is due to the fact that random samples are random, and each sample gives a different estimate. Thus, even if the true parameter value is \( \theta = 0 \), it is possible to have a random sample which will produce an estimate very far from the true value, say \( \hat{\theta} = 3 \). This kind of risk is unavoidable in statistics, but we want it to be as small as possible. Therefore, sometimes we choose the significance level \( \alpha = 0.01 \), which means that there is only 1% chance of falsely rejecting a true null hypothesis (or only one out 100 samples will lead to the false rejection of a true null hypothesis).

In terms of language, if we choose significance level \( \alpha \), and the estimate falls into a rejection region (far enough), we say that "we reject the null hypothesis at 100 \cdot \alpha\% significance
level”. We do not say that the null hypothesis is false, because there is always a chance that we rejected a true null hypothesis. If the estimate does not fall into the shaded areas (rejection regions), then we say that "we cannot reject the null hypothesis at 100 · α% significance level”.

3.4 Hypothesis Testing of Regression Coefficients

So far we assumed that the true model (data generating process) is:

\[ Y_i = \beta_1 + \beta_2 X_i + u_i \]

Here \( Y_i \) is the dependent variable, for example earnings, and \( X_i \) is an independent variable, for example schooling. The parameters \( \beta_1 \) and \( \beta_2 \) are unknown, and we learned how to estimate them using OLS (ordinary least squares). Application of OLS procedure to a given sample provides us with the estimates \( b_1 \) and \( b_2 \), i.e. two numbers. In this section we add an important ingredient to statistical research - theory (economic or any other).

Our theory could for example suggest that earnings and schooling are related, or \( \beta_2 = 0 \). If our theory is wrong, then \( \beta_2 = 0 \). Since we don’t know \( \beta_1 \) and \( \beta_2 \), we can never find out with certainty whether our theory is correct or not. Instead, we can use the sample and estimation procedure to reach conclusion such as "our theory is very likely to be correct", or "there is not enough evidence to accept our theory as correct”.

3.4.1 Formulating hypotheses about the regression coefficients

Suppose that our theory suggests that earnings and schooling are related, i.e. \( \beta_2 \neq 0 \). Statistical testing procedure starts with the assumption that our theory is wrong, i.e. \( \beta_2 = 0 \), usually called the null hypothesis. A researcher who believes that the theory is correct puts the null hypothesis to a test with the hope of rejecting it. The null hypothesis is paired with an alternative hypothesis, which usually means that the theory is correct. The null and the alternative hypotheses are then written as follows:

\[ H_0 : \beta_2 = 0 \]
\[ H_1 : \beta_2 \neq 0 \]

The statistical test related to the above hypotheses is called the two-sided test. Sometimes the theory tells us not only that there should be a relationship between \( Y \) and \( X \), but also the sign of the relationship. For example, our theory might tell us that earnings \( (Y) \) should depend positively on schooling \( (X) \), i.e. \( \beta_2 > 0 \). Then the null and alternative hypotheses are:

\[ H_0 : \beta_2 = 0 \]
\[ H_1 : \beta_2 > 0 \]

and the statistical test related to these hypotheses is called one-sided upper-tail test. Similarly, If the theory claims that the effect of \( X \) on \( Y \) is negative, the corresponding null
and alternative hypotheses are:

\[ H_0 : \beta_2 = 0 \]
\[ H_1 : \beta_2 < 0 \]

and the statistical test related to these hypotheses is called **one-sided lower-tail** test.

Since the alternative hypothesis usually represents some theory, it is clear that \( H_1 \) is different depending on the theory being tested. It is less intuitive however, that the \( H_0 \) is the same in all the above tests. In particular, it is very tempting in the one-sided upper-tail test to write \( H_0 : \beta_2 \leq 0 \). As we will see below, the null hypothesis always equates the true parameter to **one particular value**. Then, if the null hypothesis is true, the researcher obtains a particular distribution of the estimator and is able to carry out the test. If the null hypothesis was stated as \( H_0 : \beta_2 = 0 \), then it consists of many hypotheses such as \( H_0 : \beta_2 = -3 \), \( H_0 : \beta_2 = -11.7 \), \( H_0 : \beta_2 = -100 \), etc. Obviously, it is not feasible to test them all, but fortunately for us, if we reject \( H_0 : \beta_2 = 0 \) against \( H_1 : \beta_2 > 0 \), then we also would reject all the hypotheses of the form \( H_0 : \beta_2 = \text{negative number} \). Intuitively, if we conclude that \( \beta_2 > 0 \) (beyond reasonable doubt), then we reject \( H_0 : \beta_2 = 0 \) and \( H_0 : \beta_2 = -3 \), and \( H_0 : \beta_2 = -17 \) or any \( H_0 : \beta_2 = \beta_2^0 \leq 0 \).

**Exercise 34** Suppose that in a simple regression model, \( Y_i = \beta_1 + \beta_2 X_i + u_i \), the dependent variable \( Y \) is gas mileage of a car and the independent variable \( X \) is the cars weight.

(a) Suppose that theory (of physics) claims that the car weight and gas mileage are related. Write the null and alternative hypothesis for testing this theory.

(b) Suppose that theory claims that car weight has **negative** impact on gas mileage. Write the null and alternative hypothesis for testing this theory.

### 3.4.2 Deriving the t-test for regression coefficients

In order to perform tests about \( \beta_1, \beta_2 \), we need to know the distribution of their estimators \( b_1 \) and \( b_2 \). We therefore add another assumption about the regression model.

**A.6 The disturbance (error) terms are normally distributed.** Together with assumptions A.3 and A.4, this implies that for all \( i \),

\[ u_i \sim N(0, \sigma_u^2) \]

Assumption A.6 implies that the estimators \( b_1 \) and \( b_2 \) have **normal distribution**. This follows from the property that linear combination of normal random variables is also normal (see section 1.5 properties of the normal distribution). Recall that we were able to rewrite the OLS estimators as linear combinations of the error terms (equations (3.1)-(3.2)):

\[
\begin{align*}
    b_2 &= \beta_2 + \sum_{i=1}^{n} a_i u_i \\
    b_1 &= \beta_1 + \sum_{i=1}^{n} c_i u_i
\end{align*}
\]

\[
\begin{align*}
    b_2 &= \beta_2 + \sum_{i=1}^{n} a_i u_i \\
    b_1 &= \beta_1 + \sum_{i=1}^{n} c_i u_i
\end{align*}
\]
where \(a_i\)-s and \(c_i\)-s are some weights. Thus, assumption A.6 implies that:

\[
\begin{align*}
\beta_2 &\sim N(\beta_2, \sigma_{b_2}^2) \\
\beta_1 &\sim N(\beta_1, \sigma_{b_1}^2)
\end{align*}
\]

Recall that we proved in section 3.2.2 that OLS estimators are unbiased, so the means of \(\beta_1\) and \(\beta_2\) are \(\beta_1\) and \(\beta_2\) respectively. Also recall that in section 3.2.3 we derived the variance of \(\beta_1\) and \(\beta_2\):

\[
\begin{align*}
\text{var}(b_2) &= \frac{\sigma_u^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \\
\text{var}(b_1) &= \sigma_u^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right)
\end{align*}
\]

If we knew \(\sigma_u^2\), then we could have used the z-test to test hypotheses about \(\beta_1\) and \(\beta_2\). In that case, we would have

\[
\begin{align*}
\frac{b_2 - \beta_2}{\text{s.d.}(b_2)} &\sim N(0, 1) \\
\frac{b_1 - \beta_1}{\text{s.d.}(b_1)} &\sim N(0, 1)
\end{align*}
\]

where the standard deviations of \(b_2\) and \(b_1\) are:

\[
\begin{align*}
\text{s.d.}(b_2) &= \sqrt{\frac{\sigma_u^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}} \\
\text{s.d.}(b_1) &= \sqrt{\sigma_u^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right)}
\end{align*}
\]

The reality is however that we don’t know \(\sigma_u^2\), and must estimate it, just as we estimate the unknown intercept and slope coefficients \(\beta_1\) and \(\beta_2\).

**Theorem 14** Let

\[
s_u^2 = \frac{1}{n - 2} \sum_{i=1}^{n} e_i^2
\]

Then, \(s_u^2\) is an unbiased estimator of \(\sigma_u^2\).

In other words, in order to estimate the variance of the disturbance term, we use something similar to the sample variance of the residuals. Recall that \(\bar{e} = 0\) by theorem 10.

The variance of the residuals in the sample is

\[
\frac{1}{n} \sum (e_i - \bar{e})^2 = \frac{1}{n} \sum e_i^2
\]

but this turns out to be a biased estimator of \(\sigma_u^2\). However, division by \(n - 2\) instead of by \(n\) gives an unbiased
estimator of \( \sigma_u^2 \). We then replace the unknown \( \sigma_u^2 \) by its estimator \( s_u^2 \) in (3.9)-(3.10), to get the standard errors of the estimators

\[
s.e.(b_2) = \sqrt{\frac{s_u^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}} \tag{3.11}
\]

\[
s.e.(b_1) = \sqrt{s_u^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right)} \tag{3.12}
\]

It is important to compare (3.9)-(3.10) with (3.11)-(3.12) and notice the replacement of \( \sigma_u^2 \) by its estimator \( s_u^2 \).

**Theorem 15**

\[
\frac{b_2 - \beta_2}{s.e.(b_2)} \sim t_{n-2}
\]

\[
\frac{b_1 - \beta_1}{s.e.(b_1)} \sim t_{n-2}
\]

We will not present the details of the proof, but it is related to theorem 3. We have seen that

\[
Z = \frac{b_2 - \beta_2}{s.d.(b_2)} \sim N(0, 1)
\]

The standard deviation of \( b_2 \) needs to be replaced somehow by its standard error (estimator of standard deviation). First, one needs to prove two properties.

(1) \( Y = \frac{s.e.(b_2)^2 (n - 2)}{s.d.(b_2)^2} \sim \chi_{n-2}^2 \)

(2) \( Z \) and \( Y \) are independent

Suppose we proved these two properties. We know that \( t \)-distribution is constructed by dividing standard normal by a square root of independent chi-square, divided by its degrees of freedom, i.e.

\[
T = \frac{Z}{\sqrt{\frac{Y}{n-2}}} \sim t_{n-2}
\]

where \( n - 2 \) is the number of degrees of freedom of the chi-square. Thus, plugging \( Z \) and \( Y \) into \( T \) gives:

\[
T = \frac{Z}{\sqrt{\frac{Y}{(n-2)}}} = \frac{\frac{b_2 - \beta_2}{s.e.(b_2)}}{\sqrt{\frac{s.e.(b_2)^2 (n-2)}{s.d.(b_2)^2}} / (n-2)} = \frac{b_2 - \beta_2}{s.e.(b_2)} \sim t_{n-2}
\]
### 3.4.3 Performing the t-test

Having found the distribution of the estimators, $b_1$ and $b_2$, (or more precisely we found that $\frac{b_2-\beta_2}{\text{s.e.}(b_2)} \sim t_{n-2}$) we are able to identify the rejection regions using our "far enough" intuition. The rejection regions are determined by critical values of the appropriate $t$-distribution, and will depend on the significance level $\alpha$ and the number of degrees of freedom ($df$). For example, when we perform a two-sided test of the null hypothesis $H_0 : \beta_2 = 0$, with $\alpha = 0.05$ and $df = 300$, we know that the test statistic is:

$$\frac{b_2 - 0}{\text{s.e.}(b_2)} \sim t_{300}$$

The critical values of the test statistic are:

$$t_{crit} = t_{300,0.025} \approx 1.968$$

The area to the right of 1.968 and to the left of $-1.968$ is 5%, as shown in the next figure.

If in our sample, if we get $\frac{b_2 - 0}{\text{s.e.}(b_2)} > 1.968$ or $\frac{b_2 - 0}{\text{s.e.}(b_2)} < -1.968$, we reject the null hypothesis $H_0 : \beta_2 = 0$ at 5% significance level.

In general, the number of degrees of freedom for the $t$-test is equal to the number of observations, minus the number of estimated regression coefficients $k$. In the simple regression model we estimate two regression coefficients, $(\beta_1$ and $\beta_2$) so $df = n - 2$. In general, $df = n - k$. The next table summarizes the $t$-test for general multiple regression with sample size $n$, with $k$ regression coefficients (including the intercept), and significance level $\alpha$. 
### 3.4. HYPOTHESIS TESTING OF REGRESSION COEFFICIENTS

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Hypothesis</th>
<th>Reject ( H_0 ) at sig. level ( \alpha ) if</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-sided</td>
<td>( H_0 : \beta = \beta_0 ) ( H_1 : \beta \neq \beta_0 )</td>
<td>( \frac{b_0 - \beta_0}{s.e.(b)} &gt; t_{n-k,\frac{\alpha}{2}} ) or ( \frac{b_0 - \beta_0}{s.e.(b)} &lt; -t_{n-k,\frac{\alpha}{2}} )</td>
</tr>
<tr>
<td>one-sided, upper tail</td>
<td>( H_0 : \beta = \beta_0 ) ( H_1 : \beta &gt; \beta_0 )</td>
<td>( \frac{b_0 - \beta_0}{s.e.(b)} &gt; t_{n-k,\alpha} )</td>
</tr>
<tr>
<td>one-sided, lower tail</td>
<td>( H_0 : \beta = \beta_0 ) ( H_1 : \beta &lt; \beta_0 )</td>
<td>( \frac{b_0 - \beta_0}{s.e.(b)} &lt; -t_{n-k,\alpha} )</td>
</tr>
</tbody>
</table>

The object \( t_{n-k,\frac{\alpha}{2}} \) denotes the critical value of the \( t \)-distribution, with \( n - k \) degrees of freedom, to the right of which the tail probability is \( \alpha/2 \). Formally, if \( T_{n-k} \) is a random variable having \( t \)-distribution with \( n - k \) degrees of freedom, then \( P(T_{n-k} \geq t_{n-k,\frac{\alpha}{2}}) = \alpha/2 \). By symmetry of the \( t \)-distribution, \( P(T_{n-k} \leq -t_{n-k,\frac{\alpha}{2}}) = \alpha/2 \). Similarly, \( t_{n-k,\alpha} \) is the critical value of the \( t \)-distribution with \( n - k \) degrees of freedom, to the right of which the tail probability is \( \alpha \). These critical values can be found in a statistical table for the \( t \)-distribution, which appears in any statistics textbook. Simply look up the tail probability in the column of the table, and find the degrees of freedom in the rows of the table.

In order to calculate the critical values in Stata, we use the commands `display invttail(df,p)`, which gives the right-tail critical value of a \( t \)-distribution with \( df \) degrees of freedom, and the probability in the right tail is \( p \). The display function tells Stata to display the result in the command window. The function `invttail` (abbreviation of inverse \( t \)-tail) calculates the critical value of the \( t \)-distribution, to the right of which the tail probability is \( p \). For example, `display invttail(300,0.025)` returns 1.967903, which is the critical value of the \( t \)-distribution with 300 degrees of freedom, to the right of which the tail probability is 0.025.

In order to calculate the critical values in R, we use the command `qt(p,df)`, which gives the left-tail critical value of a \( t \)-distribution with \( df \) degrees of freedom, and the probability in the left tail is \( p \). For example, `qt(0.025,300)` returns -1.967903, which is the critical value of the \( t \)-distribution with 300 degrees of freedom, to the left of which the tail probability is 0.025. Obviously, the right-tail critical value is the same as the left one, but with a sign change. If we want to calculate the right-tail critical value in R, we use `qt(0.025,300, lower.tail=FALSE)`, and this command returns 1.967903.

To summarize, the procedure of \( t \)-test about individual regression coefficients is simple. We calculate the critical values, which determine the rejection regions. If the realized value of test statistic in the sample falls into a rejection region, then we reject the null hypothesis at significance level \( \alpha \). If the test statistic does not fall into a rejection region, then we do not reject the null hypothesis at significance level \( \alpha \).
Chapter 4

Multiple Regression Analysis

The simple regression analysis studies the relationship between the dependent variable (say hourly earnings) and one regressor (say years of schooling). In reality, any dependent variable we study, is usually affected by several independent variables. For example, earnings might depend on years of schooling, work experience, gender, etc. Multiple regression analysis studies the relationship between a dependent variable, and several regressors. We assume that the true model (data generating process) is:

\[ Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_k X_{k,i} + u_i \]

The dependent variable in observation \( i \) is \( Y_i \). The \( X_2, X_3, \ldots, X_k \) are regressors (independent variables), and \( \beta_1, \beta_2, \ldots, \beta_k \) are unknown coefficients, which we want to estimate. Finally, the disturbance term combines all the influences on \( Y \), observed or unobserved, other than the \( X \)s.

4.1 OLS Estimators

4.1.1 Deriving the OLS estimators

As in the simple regression model, we define the fitted equation:

\[ \hat{Y}_i = b_1 + b_2 X_{2,i} + b_3 X_{3,i} + \ldots + b_k X_{k,i} \]

and residual (or prediction error) for observation \( i \):

\[ e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2,i} - b_3 X_{3,i} - \ldots - b_k X_{k,i} \]

The OLS estimators are defined as the \( b_1, b_2, \ldots, b_k \) which minimize the residuals sum of squares:

\[ RSS = \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2,i} - b_3 X_{3,i} - \ldots - b_k X_{k,i})^2 \]

Deriving the OLS estimators of the multiple regression model, without the use of linear algebra, is almost impossible. It is possible, but very messy, if we only have 2 regressors
(k = 3), but for the general case we need to use linear algebra. I will only present an outline of how this is done, without much details. First, all the data on regressors can be written as an n by k matrix:

\[ X = \begin{bmatrix}
1 & X_{2,1} & X_{3,1} & \cdots & X_{k,1} \\
1 & X_{2,2} & X_{3,2} & \cdots & X_{k,2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{2,n} & X_{3,n} & \cdots & X_{k,n}
\end{bmatrix}_{n \times k} \]

There are n observations (rows), and k regressors (columns), where the first regressor \( X_1 \) is just a vector of 1s. The data on the dependent variable is just an n by 1 vector, as well as the disturbance term:

\[ Y = \begin{bmatrix} Y_1 \\
Y_2 \\
\vdots \\
Y_n \end{bmatrix}_{n \times 1}, \quad u = \begin{bmatrix} u_1 \\
u_2 \\
\vdots \\
u_n \end{bmatrix}_{n \times 1} \]

Finally, the \( \beta \)s, and their OLS estimators are k by 1 vectors:

\[ \beta = \begin{bmatrix} \beta_1 \\
\beta_2 \\
\vdots \\
\beta_k \end{bmatrix}_{k \times 1}, \quad b = \begin{bmatrix} b_1 \\
b_2 \\
\vdots \\
b_k \end{bmatrix}_{k \times 1} \]

This notation allows us to write the multiple regression model in (4.1) in a compact way:

\[ Y = X \beta + u \]

The fitted model is

\[ \hat{Y} = Xb \]

The next few steps are for those of you who took linear algebra. The residual sum of squares:

\[ RSS = (Y - Xb)'(Y - Xb) = Y'Y - Y'Xb - b'X'Y + b'X'Xb = Y'Y - 2Y'Xb + b'X'Xb \]

The first step uses the fact that transpose of a product is product of transposes in reverse order. The second step might not be so obvious. Note that \( Y'Xb \) is a scalar, and \( b'X'Y \) is the same scalar transposed, and transpose of a scalar is the scalar itself. The first order condition:

\[ \frac{\partial RSS}{\partial b} = -2X'Y + 2X'Xb = 0 \]

Solving this system (of k equations and k unknowns)

\[ -2X'Y + 2X'Xb = 0 \]

\[ X'Xb = X'Y \]
The vector of OLS estimators is:

\[ b = (X'X)^{-1} X'Y \]  

(4.2)

The above derivation is not important for this course. It is enough that you know that to obtain OLS coefficients in a multiple regression, a statistical package uses the formula \( (4.2) \).

### 4.1.2 Properties of OLS estimators

We retain the assumptions we made about the simple regression model A.0 - A.6, and we add one more assumption:

A.7 There does not exists an exact linear relationship among the regressors in the sample. For example, suppose that \( X_2 \) and \( X_3 \) are two regressors. Assumption A.7 does not allow for example, that \( X_3 = \lambda_1 + \lambda_2 X_2 \), where \( \lambda_1, \lambda_2 \) are some constants.

If there is a linear relationship between two or more regressors in a model, the problem is called **multicollinearity**, and will be discussed later in this chapter. In the case of exact linear relationship, the regression coefficients cannot be estimated at all.

Given that all the assumptions are satisfied, just as in the case of simple regression model, the OLS coefficients are unbiased, and have the minimum variance among all unbiased estimators that are linear functions of \( Y_i \)'s. In other words, the **Gauss-Markov Theorem** holds for the OLS estimators of the multiple regression model. Assumption A.6, that \( u_i \sim N(0, \sigma_u^2) \), implies that the OLS estimators will also have normal distribution. The bottom line is that the properties that we proved for the simple regression model, extend to the multiple regression model, but in order to prove these properties in general, one needs to use linear algebra.

### 4.1.3 Interpretation of regression estimates

Suppose that we study the factors which affect the gas mileage of a car. To get the data in Stata, type:

```
use http://www.stata-press.com/data/r12/auto.dta, replace
```

Next we generate a new variable, which is the car's weight measured in hundreds of pounds:

```
generate weight100 = weight/100
```

Then we regress the gas mileage on weight100 and repair record (higher record means less repairs, more reliable car). In other words, we estimate the model

\[ mpg = \beta_1 + \beta_2 \cdot weight100 + \beta_3 \cdot rep78 + u \]

In Stata, type:

```
regress mpg weight100 rep78
```

The estimation results are presented in the next table.
The fitted equation is

\[
\hat{mpg} = b_1 + b_2 \cdot weight100 + b_3 \cdot rep78
\]

The interpretation of these results is as follows. The value of \( b_2 = -0.568 \) means that an increase in car’s weight by 100 pounds, holding all else constant, reduces the gas mileage by 0.568 miles per gallon. The value of \( b_3 = 0.564 \) means that an improvement in repair record by 1 point, holding all else constant, increases the gas mileage by 0.564 miles per gallon. The constant \( b_1 = 36.588 \) means that a car with weight 0 and repair record of 0, is predicted to have gas mileage of 36.588 miles per gallon. As usual, the intercept rarely makes any real sense, because there are no cars with weight zero.

To perform the same tasks in R, we use the following commands:

```r
library(foreign) #Needed to read data not in R format
auto <- read.dta("http://www.stata-press.com/data/r12/auto.dta")
auto$weight100 <- auto$weight/100 #Weight in 100s pounds
model1 <- lm(mpg ~ weight100 + rep78, data = auto)
summary(model1) #Basic regression output table
```

The resulting table in R is

```
Residuals:
   Min      1Q  Median      3Q     Max
-7.7204 -1.8951 -0.4827  1.2053  13.1785

Coefficients: Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 36.58773    2.84055   12.880  < 2e-16 ***
weight100   -0.56784    0.05825  -9.748   2.08e-14 ***
rep78        0.56355    0.46657   1.208    0.231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.49 on 66 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared: 0.6565,  Adjusted R-squared: 0.646
F-statistic: 63.06 on 2 and 66 DF,  p-value: 4.869e-16
```
4.2 Hypothesis Testing

4.2.1 t-tests

Recall that in the simple regression analysis the test statistic had \( n - 2 \) degrees of freedom, because in order to estimate the unknown variance of the disturbance term, \( \sigma_u^2 \), we used the 2 estimated regression coefficients, \( b_1 \) and \( b_2 \). In the multiple regression model, an unbiased estimator of \( \sigma_u^2 \) is:

\[
\hat{\sigma}_u^2 = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2
\]

It therefore follows that the test statistic in a multiple regression model has \( t_{n-k} \) distribution, i.e. \( t \)-distribution with \( n - k \) degrees of freedom:

\[
T = \frac{b - \beta}{s.e.(b)} \sim t_{n-k}
\]

There are \( n - k \) degrees of freedom because in order to estimate the unknown variance of the disturbance term, \( \sigma_u^2 \), we used \( k \) estimated regression coefficients, \( b_1, b_2, \ldots, b_k \).

We repeat the table which summarizes the \( t \)-tests of an individual regression coefficient \( \beta \), whose estimate is \( b \):

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Hypothesis</th>
<th>Reject ( H_0 ) at sig. level ( \alpha ) if</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-sided</td>
<td>( H_0 : \beta = \beta_0 )</td>
<td>( \frac{b - \beta_0}{s.e.(b)} &gt; t_{n-k, \frac{\alpha}{2}} ) or ( \frac{b - \beta_0}{s.e.(b)} &lt; -t_{n-k, \frac{\alpha}{2}} )</td>
</tr>
<tr>
<td>one-sided, upper tail</td>
<td>( H_0 : \beta = \beta_0 )</td>
<td>( \frac{b - \beta_0}{s.e.(b)} &gt; t_{n-k, \alpha} )</td>
</tr>
<tr>
<td>one-sided, lower tail</td>
<td>( H_0 : \beta = \beta_0 )</td>
<td>( \frac{b - \beta_0}{s.e.(b)} &lt; -t_{n-k, \alpha} )</td>
</tr>
</tbody>
</table>

Recall that the object \( t_{n-k, \frac{\alpha}{2}} \) denotes the critical value of the \( t \)-distribution, with \( n - k \) degrees of freedom, to the right of which the tail probability is \( \alpha/2 \). Once again, the test statistic measures the distance between the estimated value \( b \) and the hypothesized value of \( \beta \) (i.e. \( \beta_0 \)) in units of standard errors. If this distance is "large enough", then we reject the null hypothesis. In other words, we reject the null hypothesis if the estimated value is "far enough" away from the hypothesized value in the null hypothesis. In each test, far enough means beyond some critical value of the \( t \)-distribution (denoted \( t_{n-k, \frac{\alpha}{2}} \) for 2-sided test, and \( t_{n-k, \alpha} \) for one-sided tests), which depends on degrees of freedom and the significance level \( \alpha \). In one-sided tests, not only distance from \( \beta_0 \) matters, but also the direction. In the upper tail test the estimate \( b \) must be "far enough" above the hypothesized value \( \beta_0 \), and in the lower tail test the estimate \( b \) must be "far enough" below the hypothesized value \( \beta_0 \).

4.2.2 p-value

The fifth column in the Stata (labeled \( P>|t| \)) and R (labeled \( \text{Pr}(>|t|) \)) regression output, gives the p-value for each coefficient. This is the probability of obtaining a t-statistic at
least as large in absolute value as what we obtained for the sample, when $H_0 : \beta = 0$ is true. Put another way, the reported p-values in any statistical package, by default usually give the smallest significance level at which we can reject $H_0 : \beta = 0$ against the two-sided alternative $H_1 : \beta \neq 0$. Thus, we can reject $H_0$ at any significance level $\alpha \geq$ p-value. Recall that under this null hypothesis, we have:

$$T = \frac{b - 0}{\text{s.e.}(b)} \sim t_{n-k}$$

That is, the t-statistic has t-distribution with $n - k$ degrees of freedom. The p-value is defined mathematically as follows:

$$\text{p-value} = 2 \cdot P(t_{n-k} > |t|)$$

where $t$ is the realized t-value of the t-statistic in the sample.

For example, consider the p-value of the coefficient on rep78, which equals 0.231. Suppose that the null hypothesis is $H_0 : \beta_3 = 0$. The realized test statistic for this sample is then

$$t = \frac{b_3 - 0}{\text{s.e.}(b_3)} = \frac{0.564 - 0}{0.466} = 1.21$$

This t-statistic is reported in column 4, labeled t and in column labeled t value in R. Regression outputs of Stata and R report the realized value of the test statistic for the null hypothesis $H_0 : \beta = 0$. In general, the test statistic is a random variable, and its value will be different in different samples. The p-value reported in Stata and R is the probability that in repeated samples we obtain $T > 1.21$ or $T < -1.21$, when $H_0$ is true. To calculate this probability, we need to find the degrees of freedom $df = n - k = 69 - 3 = 66$. Then,

$$\text{p-value} = P(t_{66} > 1.21 \text{ or } t_{66} < -1.21) = 2 \cdot P(t_{66} > 1.21)$$

This probability is illustrated in the next figure.
The shaded area is 0.2306, which is the p-value reported in the regression output of Stata or R.

The advantage of the p-value, is that it gives us the smallest significance level at which we can reject the null hypothesis $H_0 : \beta = 0$, with the two-sided alternative: $H_1 : \beta \neq 0$. In the case of $H_0 : \beta_3 = 0$, and $H_1 : \beta_3 \neq 0$, we can only reject the null hypothesis at significance level of $\alpha = 0.2306$ or higher. We cannot reject the null hypothesis at $\alpha = 0.05$ for example or at any significance level smaller than 0.2306. For the coefficient on weight, we got p-value = 0.000. This means that the p-value for $\beta_2$ is smaller than 0.001 or 0.1% (one tenth of a percent). Thus, in the standard two-sided test

$$
H_0 : \beta_2 = 0 \\
H_1 : \beta_2 \neq 0
$$

we would reject the null hypothesis at significance level of $\alpha = 5\%$, and at significance level of $\alpha = 1\%$, and at 0.1%. Since Stata rounds the p-value to 3 decimal places, we cannot see from the table what is the minimal significance level at which we reject the null hypothesis in this case. Thus, reporting p-values gives an immediate test of significance for the regression coefficients.

To summarize, for the standard two-sided test

$$
H_0 : \beta = 0 \\
H_1 : \beta \neq 0
$$

the next table summarizes the interpretation of the p-value reported in Stata.

<table>
<thead>
<tr>
<th>p-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p-value &lt; 0.001$</td>
<td>reject $H_0$ at 0.1% significance level</td>
</tr>
<tr>
<td>$0.001 &lt; p-value &lt; 0.01$</td>
<td>reject $H_0$ at 1% significance level (but not at 0.1%)</td>
</tr>
<tr>
<td>$0.01 &lt; p-value &lt; 0.05$</td>
<td>reject $H_0$ at 5% significance level (but not at 1%)</td>
</tr>
<tr>
<td>$0.05 &lt; p-value$</td>
<td>fail to reject $H_0$ at 5% significance level</td>
</tr>
</tbody>
</table>

R presents the p-value in scientific format, so we simply conclude that the reported p-value is the smallest significance level at which we can reject the null hypothesis in the standard two-sided test. Thus, we can reject $H_0$ at any significance level $\alpha \geq p$-value.

Once again it should be emphasized that the reported p-values in Stata or R are only for the two sided test with the null hypothesis being that the coefficient is zero, $H_0 : \beta = 0$, and the alternative hypothesis is that its not zero, $H_1 : \beta \neq 0$. If the test is one-sided, with $H_0 : \beta = 0$ and the alternative is either $H_1 : \beta > 0$ or $H_1 : \beta < 0$, the p-value is just 1/2 of the p-value reported in Stata or R for the two-sided test, since the t-distribution is symmetric, and both tails are the same. If you need the p-value for a different hypothesis, you need to calculate it yourself. The next table shows how to calculate p-values in Stata and R for general tests, where the t-statistic in all cases is

$$
T = \frac{b - \beta_0}{s.e. (b)}
$$

and the realized value of $T$ for a given sample is denoted by $t$. Recall that Stata, by default, calculates the upper tail probablity, while R, by default, calculates the lower tail probability.
Test | p-value defined | p-value in Stata | p-value in R
---|---|---|---
$H_0 : \beta = \beta_0$
$H_1 : \beta \neq \beta_0$ | $2 \cdot P(t_{n-k} > |t|)$ | $2 \cdot t_{n-k}\cdot \text{abs}(t))$ | $2 \cdot \text{pt}(-\text{abs}(t), n-k)$
$H_0 : \beta = \beta_0$
$H_1 : \beta > \beta_0$ | $P(t_{n-k} > t)$ | $\text{ttail}(n-k, t)$ | $\text{pt}(-t, n-k)$
$H_0 : \beta = \beta_0$
$H_1 : \beta < \beta_0$ | $P(t_{n-k} < t)$ | $\text{ttail}(n-k, -t)$ | $\text{pt}(t, n-k)$

The null hypothesis is rejected at significance level $\alpha$ if p-value $\leq \alpha$.

### 4.2.3 Confidence intervals

So far we started a statistical test by stating the null and alternative hypotheses. Now we ask, what is the set of all null hypotheses which are consistent with the data at hand, i.e. the hypotheses that will not be rejected. We do not reject the null hypothesis ($H_0 : \beta = \beta_0$) against a two-sided alternative ($H_1 : \beta \neq \beta_0$) when

$$-t_{\frac{a}{2}} \leq \frac{b - \beta_0}{\text{s.e.}(b)} \leq t_{\frac{a}{2}}$$

The critical values are determined by the significance level $\alpha$ and degrees of freedom $n - k$. We can solve the above inequality to locate all the hypothesized values of $\beta$ that will not be rejected based on our sample.

$$t_{\frac{a}{2}} \geq \frac{-b + \beta_0}{\text{s.e.}(b)} \geq -t_{\frac{a}{2}}$$

This solution is:

$$b - \text{s.e.}(b) \times t_{\frac{a}{2}} \leq \beta_0 \leq b + \text{s.e.}(b) \times t_{\frac{a}{2}}$$

Thus, all the hypothesized values of $\beta \in \left[b - \text{s.e.}(b) \times t_{\frac{a}{2}}, b + \text{s.e.}(b) \times t_{\frac{a}{2}}\right]$ will not be rejected at significance level $\alpha$. This interval is called the $100(1 - \alpha)\%$ confidence interval for $\beta$.

For example, construct a 95% confidence interval for $\beta_2$, the coefficient on wight100. The positive critical value is 1.9965644, based on 66 degrees of freedom and $\alpha = 0.05$. They are obtained in Stata by: `disp invttail(e(df_r),0.05/2)`. Thus the 95% confidence interval for $\beta_2$ is:

$$[-0.568 - 0.558 \cdot 1.996, - 0.568 + 0.558 \cdot 1.996]$$
$$[-0.684154, -0.4515354]$$

This means that any null hypothesis $H_0 : \beta_2 \in [-0.684154, -0.4515354]$, will not be rejected at $\alpha = 5\%$ confidence level, while any null hypothesis $H_0 : \beta_2 \notin [-0.684154, -0.4515354]$ will be rejected in favor of the two sided alternative.
4.2. HYPOTHESIS TESTING

The last two columns of the Stata regression output, labeled [95% Conf. Interval], report the 95% confidence intervals for the estimated coefficients. The default is 95% confidence intervals, and the user can change this default to any other percentage. For example, in the regression used in this section, we can write:

```
regress mpg weight100 rep78, level(99)
```

This results is the same regression output as above, but with 99% confidence intervals.

The regression output of R does not automatically report confidence intervals, and in order to obtain them, we need to use to command:

```
confint(model1,level=0.99)
```

Without the part level=0.99, the confidence level is assumed to be 95%.

4.2.4 Goodness of fit: $R^2$ and F-tests

As in the simple regression model, we can decompose the Total Sum of Squares, $TSS$, into two parts: (i) Explained Sum of Squares, $ESS$, and (ii) residual Sum of Squares, $RSS$. See equation (2.6) which is repeated here for convenience.

$$
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2
$$

The definition of $R^2$ is also unchanged from the simple regression analysis:

$$
R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS},
$$

measures the proportion of the variation in the dependent variable $Y$, "explained" by the fitted model. Intuitively, higher $R^2$ means that the model fits the data better. Experience tells us that in some situations a value of $R^2$ of 0.5 is considered "high" while in other situations an $R^2$ of 0.7 might be considered "low" and indicate "poor" fit. It would be nice if we could test somehow whether our model has a poor fit overall.

Such a test exists, and is called the F-test of overall fit of the model. The null and alternative hypotheses are:

$$
H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \\
H_1 : \text{at least one of the } \beta \text{s is not 0}
$$

In other words, the null hypothesis is that none of the regressors included in the model help predict the dependent variable, and the alternative is that at least one of them does. The F-test is based on the following theorem.

**Theorem 16** Suppose that the sample size is $n$ and the model has $k$ parameters ($\beta_1, \beta_2, \ldots, \beta_k$). If the null hypothesis is true ($\beta_2 = \beta_3 = \ldots = \beta_k = 0$), then:

$$
\frac{ESS}{(k-1)} / \frac{RSS}{(n-k)} \sim F(k-1, n-k)
$$

\[1\] Recall that $0 \leq R^2 \leq 1$. 

In other words, the ratio of \( ESS \) to \( RSS \), each divided by their respective degrees of freedom, has an \( F \)-distribution with \( k - 1 \) numerator degrees of freedom and \( n - k \) denominator degrees of freedom.

Recall from section 1.5 that the \( F \)-distribution is constructed as ratio of two independent chi-square random variables. Thus, to prove this theorem, one needs to show that

\[
\begin{align*}
(1) & : V_1 = \frac{ESS}{\sigma_u^2} \sim \chi^2_{k-1} \\
(2) & : V_2 = \frac{RSS}{\sigma_u^2} \sim \chi^2_{n-k} \\
(3) & : V_1 \text{ and } V_2 \text{ are independent}
\end{align*}
\]

We will not prove these results here and will only apply the theorem. The next figure plots a typical shape of pdf of \( F(k - 1, n - k) \).

![pdf of F(3, 536)](image)

Intuitively, the larger \( ESS \) relative to \( RSS \), the better is the fit of the model. Note that if the model cannot explain any of the variation of the dependent variable, i.e. \( ESS = 0 \), then the \( F \)-statistic will be zero, and we will not reject the null hypothesis. Therefore, we will reject the null hypothesis when the \( F \)-statistic, \( F(k - 1, n - k) \), is "large enough". Usually, large enough means that the probability of obtaining such large \( F \)-statistic is less than 0.05 or less than 0.01. Stata and R report the \( F \)-statistic, as well as the probability of obtaining such large \( F \)-statistic when the null hypothesis is true. This is the p-value for the \( F \)-test of overall model fit. In Stata, the computed \( F \)-statistic for the sample is reported as \( F(d1, d2) \), and the corresponding p-value is \( \text{Prob} > F \). In R regression output, the bottom line reports \( F \)-statistic and the associated p-value. If the reported p-value is smaller than 0.05, we reject the null hypothesis at significance level of 0.05. When Stata reports \( \text{Prob} > F = 0.0000 \), it means that we can reject the null hypothesis at significance level of 0.0001 since the actual p-value is smaller than 0.0001. The smallest p-value reported in R regression output is \( 2.2e-16 \).
4.3. MULTICOLLINEARITY

It can be shown that the $F$-test of overall fit of the model is in fact testing how large is the $R^2$. Recall that

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Then, dividing the numerator and denominator of the $F$-statistic by $TSS$ gives:

$$\left( \frac{ESS}{TSS} \right) / (k - 1) = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

The $F$-statistic is therefore increasing in $R^2$, as shown in the next graph:

![F(3,500) graph]

Therefore, when we are performing the $F$-test of overall fit, we are actually testing whether $R^2$ is equal to zero or not.

4.3 Multicollinearity

Recall that assumption A.7 about the multiple regression model, states that there does not exists an exact linear relationship among the regressors in the sample. We mentioned that if such linear relationship exists, then we have a "problem" called Multicollinearity. In this section we define the concept of multicollinearity, discuss under what conditions it is indeed a "problem", and suggest some remedial steps to combat it.

**Definition 30** Suppose that we are trying to estimate the general multivariate regression model:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

We say that this model has **perfect multicollinearity** if the following equation holds for some numbers $\lambda_1, \lambda_2, \ldots, \lambda_k$, not all of them zero:

$$\lambda_1 + \lambda_2 X_2 + \lambda_3 X_3 + \ldots + \lambda_k X_k = 0$$

We say that this model has **imperfect multicollinearity** if

$$\lambda_1 + \lambda_2 X_2 + \lambda_3 X_3 + \ldots + \lambda_k X_k + v = 0$$

where $v$ is some random variable with mean zero and variance $\sigma_v^2$. 
The following models are examples of perfect multicollinearity:

1. \[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u, \quad X_4 = X_2 + X_3 \]
2. \[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u, \quad X_4 = 2X_2 - 0.8X_3 + 7 \]
3. \[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u, \quad X_3 = 3X_2 - 5 \]

Assumption A.7. does not allow perfect multicollinearity, and the reason is that it is impossible to obtain OLS estimates in the presence of perfect multicollinearity. Consider for example a model, where the dependent variable \( Y \) is family’s consumption, \( X_2 \) is labor income, \( X_3 \) is non-labor income (dividends, interest) and \( X_4 = X_2 + X_3 \) is total income.

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u \]

In this model, \( \beta_2 \) represents the effect on consumption of a $1 increase in \( X_2 \), holding \( X_3 \) and \( X_4 \) fixed. But if we increase \( X_2 \) by $1, we will necessarily also increase \( X_4 \) by $1, and the measured effect will be \( \beta_2 + \beta_4 \). Similarly, suppose we increase \( X_3 \) by $1. Instead of the effect on \( Y \) being \( \beta_3 \), it is going to be \( \beta_3 + \beta_4 \). In other words, instead of estimating \( \beta_2, \beta_3 \) and \( \beta_4 \), it is only possible to estimate \( \beta_2 + \beta_4 \) and \( \beta_3 + \beta_4 \). Substituting the linear relationship \( X_4 = X_2 + X_3 \) into the original model, we get:

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 + X_3) + u \]

\[ Y = \beta_1 + (\beta_2 + \beta_4) X_2 + (\beta_3 + \beta_4) X_3 + u \]

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \]

It should be emphasized that non-linear relationships among regressors do not constitute perfect multicollinearity. For example, consider the model:

\[ Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 \ln X + u \]

All the regressors are functions of \( X \), but they are not linear functions of \( X \), and this model therefore does not violate assumption A.7. The above form is questionable in the sense that it is hard to interpret the estimated coefficients, but OLS estimates can be at least obtained.

The bottom line is that with perfect multicollinearity, OLS estimates cannot be obtained. Stata automatically drops one of a group of variables which have exact linear relationship. Mathematically, obtaining OLS estimates in the presence of perfect multicollinearity, is the same as trying to solve for \( k \) unknowns from a system \( k - 1 \) linear equations. Perfect multicollinearity always arises when the econometrician makes a mistake, and does not notice that he included a variable, which is a linear combination of some other variables. For example, the econometrician included labor and non-labor incomes as regressors, and in addition also included the total income. Perfect multicollinearity is readily noticed when you estimate the model, and Stata reports that it dropped one of the regressors. You conclude that the dropped regressor can be obtained as a linear combination of one or more of other regressors. The solution is to reexamine your model, and avoid perfect multicollinearity by excluding one of the collinear variables.

In practice, unless the econometrician makes a mistake, one rarely encounters perfect multicollinearity. Instead, it is common to have a high correlation between two variables,
4.3. MULTICOLLINEARITY

or between a variable and a combination of other variables. For example, we consider a hypothetical data on consumption $c$, income $i$, and wealth $w$. Economic theory suggests that both current income and wealth are important determinants of consumption. We estimate 3 models; in the first we regress consumption on income only, in the second we regress consumption on wealth only, and in the third we include both income and wealth. The R commands and output:

```r
multi <- read.csv("http://online.sfsu.edu/mbar/ECON312_files/multi.csv")
model1 <- lm(c ~ i, data = multi)
model2 <- lm(c ~ w, data = multi)
model3 <- lm(c ~ i + w, data = multi)
```

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable:</strong> Consumption</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>0.509***</td>
<td>0.942</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.823)</td>
<td></td>
</tr>
<tr>
<td>Wealth</td>
<td></td>
<td>0.050***</td>
<td>-0.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.004)</td>
<td>(0.081)</td>
</tr>
<tr>
<td>Constant</td>
<td>24.455***</td>
<td>24.411***</td>
<td>24.775***</td>
</tr>
<tr>
<td></td>
<td>(6.414)</td>
<td>(6.874)</td>
<td>(6.752)</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>R²</td>
<td>0.962</td>
<td>0.957</td>
<td>0.964</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.957</td>
<td>0.951</td>
<td>0.953</td>
</tr>
<tr>
<td>Residual Std. Error</td>
<td>6.493 (df = 8)</td>
<td>6.938 (df = 8)</td>
<td>6.808 (df = 7)</td>
</tr>
<tr>
<td>F Statistic</td>
<td>202.868*** (df = 1; 8)</td>
<td>176.668*** (df = 1; 8)</td>
<td>92.402*** (df = 2; 7)</td>
</tr>
</tbody>
</table>

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 4.1: Illustration of Strong Multicollinearity

In order to produce the above table in R, we need to first install the stargazer package, `install.packages("stargazer")`, (you need to install it only once). Then use use the commands:

```r
library(stargazer)
stargazer(model1, model2, model3, type="text",
          dep.var.labels="Consumption",
          covariate.labels=c("Income","Wealth"),
          out="models_multi.htm")
```

The `out="models_multi.htm"` saves the table in htm format, and it can be can be pasted into MS Word.
Notice that when we regressed consumption only on income (model1) or wealth (model2), these regressors were highly significant. Once we included both income and wealth however (model3), they both became insignificant, and the sign on Wealth changed from positive to negative. Nevertheless, model3 has the highest $R^2$ and the F-statistic is very high as well. This outcome is a typical result of strong, but not perfect, multicollinearity.

The problem is that in our hypothetical data set, income and wealth are very similar. Stata command `correlate` generates a table of pairwise correlations:

```
. correlate i w
(obs=10)

          | i       | w
----------|---------|---------
i       | 1.0000
w       | 0.9990  | 1.0000
```

Notice that the correlation between income and wealth is 0.9990, which is close to 1. In addition, one can regress wealth on income, and plot the fitted equation with the actual data points, we see that there is almost exact linear relationship between wealth and income. Stata commands and graph are as follows:

```
regress w i
predict w_hat

twoway (scatter w i) (line w_hat i, sort title(Wealth v.s. Income))
```

![Wealth v.s. Income](image)
This is the reason why when both income and wealth are included, OLS procedure (or any other estimation procedure for that matter) has a hard time measuring the separate effect of income on consumption from that of wealth on consumption.

The above steps can be easily performed with R. Suppose the data frame containing the variables c, i, w is multi. Then \( \text{cor(multi}\_i,\text{multi}\_w) \) calculates the correlation between wealth and income in the sample. We can regress wealth on income, and plot the fitted equation with the actual data points, to produce similar graph with R:

```r
model <- lm(w ~ i, data=multi) #Regressing wealth on income, checking R-squared
plot(w ~ i, data=multi, main="Wealth vs Income", col="blue")
abline(coef(model), col="red")
legend("topleft", legend = c("data","fitted equation"), col=c("blue","red"), pch=c("o","_"))
```

Correlation test detects multicollinearity between pairs of variables, i.e. whether one variable is close to linear function of another variable. In a multiple regression model \( Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u \), we can detect strong multicollinearity by estimating the auxiliary regression of say \( X_2 \) on the rest of the regressors: \( X_2 = \gamma_1 + \gamma_3 X_3 + \ldots + \gamma_k X_k + \varepsilon \). High \( R^2 \) in this regression will indicate presence of strong multicollinearity.

What can we do after we have detected strong multicollinearity? The easiest thing to do with this sample is drop either income or wealth from the model, and regress consumption on one remaining variable. But what if the entire purpose of this research is to measure the separate effect of income and wealth on consumption? The next best thing to do is to try and increase the sample size. Recall that OLS estimators are consistent, which means that their variance tends to zero as the sample size becomes infinitely large. Some authors make a point that a non-perfect multicollinearity problem is exactly identical to the problem of small data set (micronumerosity: micro - small, numerosity - number).
Chapter 5

Transformations of Variables and Nonlinearities

So far we learned how to estimate the linear multiple regression model:

\[ Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u \]

This model is linear in parameters \((\beta_1, \ldots, \beta_k)\) and in variables \((Y, X_2, \ldots, X_k)\). We can look at the right hand side of the regression model as a function of the \(\beta\)s, and notice that it is a weighted average of the \(\beta\)s, with weights being the \(X\)s, and therefore linear in \(\beta\)s. This means that the model is linear in parameters. On the other hand, we can look at the right hand side as a function of the \(X\)s and notice that it is a weighted sum of the \(X\)s, with weights being the \(\beta\)s, and therefore linear in \(X\)s. The left hand side contains a linear function of \(Y\), and so we conclude that the model is linear in variables \(Y\) and the \(X\)s. In practice however, we often will encounter nonlinear models. In this chapter we demonstrate that if the model is not linear in variables, it can be easily made linear by transforming the variables. On the other hand, if the model is not linear in variables and parameters, we can sometimes make the model linear, but in some cases this is impossible. If the model is not linear in parameters, and cannot be transformed into a linear model in parameters, we are forced to estimate it using methods other than OLS (for example maximum likelihood, or non-linear least squares).

5.1 Non-linearity in Variables

The following is an example of a model that is nonlinear in variables, but it is linear in parameters.

\[ \frac{100}{100 - Y} = \beta_1 + \beta_2 \sqrt{X_2} + \beta_3 X_3^2 + \beta_4 \ln (X_4) + u \]

Notice that this model is clearly not linear in variables \((Y, X_2, X_3, X_4)\) but it is linear in parameters because the right hand side is a weighted average of the \(\beta\)s. In order to estimate the unknown parameters with OLS, we need to transform the nonlinear variables as follows:

\[ Y_{new} = \frac{100}{100 - Y}, \quad Z_2 = \sqrt{X_2}, \quad Z_3 = X_3^2, \quad Z_4 = \ln (X_4) \]
and estimate the linear model:

\[ Y_{new} = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + u \]

In Stata, use the command `generate` to create the transformed variables. The conclusion is that nonlinearities in variables can be easily resolved by a straightforward transformation. In R, we can create new variable `X_new <- f(X)`, where `f` is some function. For example, `ln_X <- log(X)` creates a new variable, with values equal to natural logs of `X`. A nice thing about R is that it is possible to use functions of variables, without creating new variables. For example, `lm(log(Y) ~ log(X), data=mydata)` estimates the linear model with dependent variable \( \ln (Y) \) and regression \( \ln (X) \).

### 5.2 Nonlinearity in Variables and Parameters

The following model is nonlinear in both variables and parameters:

\[ Y = \beta_1 X_2^{\beta_2} X_3^{\beta_3} X_4^{\beta_4} e^u \]

This model is nonlinear in \( X \)'s and \( \beta \)'s, but in this case it is easy to transform the model into a linear one by taking logarithms:

\[ \ln (Y) = \ln \beta_1 + \beta_2 \ln (X_2) + \beta_3 \ln (X_3) + \beta_4 \ln (X_4) + u \]

This is a linear model in parameters. It is nonlinear in variables, but we have already seen in the last section that this nonlinearity is resolved by transforming the variables: \( Y_{new} = \ln (Y) \), \( Z_2 = \ln (X_2) \), \( Z_3 = \ln (X_3) \), \( Z_4 = \ln (X_4) \).

This was one example in which nonlinearity in parameters can be eliminated. In many cases, there is no way to eliminate the nonlinearities. For example, consider the model:

\[ Y = \beta_1 + \frac{\beta_2}{\beta_3 + X} + u \]

There is no mathematical transformation that can be applied to the right hand side of this model, that would create a linear function in \( \beta \)'s. In this case, we typically apply non linear least squares, which works exactly as OLS, but the formulas of estimators are not linear functions of the \( Y \)'s and cannot be obtained analytically. Just as in OLS, we define the fitted (predicted) model:

\[ \hat{Y}_i = b_1 + \frac{b_2}{b_3 + X_i} \]

---

1. R does not recognize the function \( \ln (\cdot) \), so in order to compute the natural logarithm (base \( e \)) one needs to use \( \log (\cdot) \).

2. Recall the rules of logarithms:

\[ \log (XY) = \log (X) + \log (Y) \]
\[ \log (X^a) = a \log (X) \]
and residuals (prediction errors) for observation $i$:

$$e_i = Y_i - b_1 - \frac{b_2}{b_3 + X_i}$$

The nonlinear least square estimators minimize the Residuals Sum of Squares:

$$RSS = \sum_{i=1}^{n} \left( Y_i - b_1 - \frac{b_2}{b_3 + X_i} \right)^2$$

In Stata you can find these estimates using the command `nl`, while R has the `nls` function, which works as follows: `nls(ydata ~ b1 + b2/(b3+X), data = mydata, start=list(b1 = 1, b2 = 1, b3 = 1))`.

### 5.3 Models with Logarithmic Terms

Transformations of original variables changes the way we interpret the estimated coefficients. In this section we discuss how the interpretation of the estimated coefficients changes when we apply logarithmic transformation to a variable.

#### 5.3.1 Log-log model

Suppose that both the dependent variable and the independent variables are logs of some original variables as follows:

$$\ln(Y) = \beta_1 + \beta_2 \ln(X) + u$$

We see that the slope coefficient

$$\beta_2 = \frac{\partial \ln(Y)}{\partial \ln(X)} = \frac{1}{X} \frac{\partial Y}{\partial X} = \frac{\partial Y}{Y} \frac{X}{\partial X}$$

This expression is the familiar **point elasticity** formula for the elasticity of $Y$ with respect to $X$.

Recall that the definition of elasticity is percentage change in $Y$ divided by the percentage change in $X$:

$$\eta_{Y,X} = \frac{\%\Delta Y}{\%\Delta X},$$

which is the percentage change in $Y$ resulting from a 1% increase in $X$. Recall from rules of derivatives that

$$\frac{\partial \ln(Y)}{\partial Y} = \frac{1}{Y}$$

which implies that

$$\partial \ln(Y) = \frac{\partial Y}{Y}$$
The term \(\frac{\partial Y}{Y}\) is the rate of change in \(Y\), expressed as a fraction. Thus, the term \(100 \cdot \frac{\partial Y}{Y}\) is roughly the percentage change in \(Y\). Similarly, the term \(100 \cdot \frac{\partial X}{X}\) is the percentage change in \(X\). For example, suppose that \(X\) changed from 50 to 54.

\[
\text{rate of change} \approx \frac{54 - 50}{50} = \frac{4}{50} = \frac{8}{100} = 0.08
\]

\[
\% \text{ change} = (100 \cdot 0.08)\% = 8\%
\]

You can also see now why the point elasticity formula is consistent with the definition of elasticity:

\[
\eta_{Y,X} = \frac{\partial Y}{\partial X} \frac{X}{Y} = \frac{(100 \cdot \frac{\partial Y}{Y})\%}{(100 \cdot \frac{\partial X}{X})\%} = \frac{\%\Delta Y}{\%\Delta X}
\]

Equipped with the above results and understanding the meaning of elasticity, we can interpret the slope coefficient \(b_2\), or more precisely, its estimated value. Estimating the log-log model gives the fitted equation:

\[
\ln(Y) = b_1 + b_2 \ln(X)
\]

Thus, \(b_2\) is the estimated elasticity of \(Y\) with respect to \(X\), or the estimated percentage change in \(Y\) resulting from a 1% increase in \(X\). The advantage of the log-log model is exactly in the interpretation of the slope coefficients, which is easy and does not require knowledge of the exact units in which \(X\) and \(Y\) are measured. For example, if \(X\) is height and \(Y\) is weight, whether they are measured in inches and pounds, or kilograms and meters (or centimeters), the estimated slope coefficient will always be the same, and will be interpreted in the same way.

**Example 4** Suppose that we study the relationship between car’s gas mileage and weight, and our fitted equation is:

\[
\ln(mpg_i) = b_1 + b_2 \ln(weight_i)
\]

Suppose that we estimated \(b_2 = -0.83\). Thus \(b_2 = -0.83\) is the elasticity of mpg with respect to car’s weight. That is, an increase in car’s weight by 1% is predicted to lower it’s gas mileage by 0.83%.

### 5.3.2 Log-lin model

Suppose that the dependent variable is log-transformed, but not the regressor:

\[
\ln(Y) = \beta_1 + \beta_2 X + u
\]

Here

\[
\beta_2 = \frac{\partial \ln Y}{\partial X} = \frac{\partial Y}{\partial X} \frac{1}{\partial X Y}
\]

Multiplying by 100 transforms the rate of change into units of percentages:

\[
(100\beta_2)\% = \frac{(100 \cdot \frac{\partial Y}{Y})\%}{\Delta X} \approx \%\Delta Y
\]
which is the percentage change in $Y$ as $X$ changes by one unit. Thus, to interpret $b_2$ we say that $100 \cdot b_2$ is the percentage change in $Y$ resulting from a one unit increase in $X$.

The coefficient $\beta_2$ is not the same as the elasticity of $Y$ with respect to $X$. The elasticity of $Y$ with respect to $X$ is:

$$\eta_{Y,X} = \frac{\partial Y X}{\partial X Y} = \beta_2 \cdot X$$

and its estimate is:

$$\hat{\eta}_{Y,X} = b_2 \cdot X$$

Thus, this is a model of a non-constant elasticity of $Y$ with respect to $X$, and the elasticity depends on the values of $X$.

### 5.3.3 Lin-log model

Suppose that the regressor is log-transformed, but not the dependent variable:

$$Y = \beta_1 + \beta_2 \ln (X) + u$$

Here

$$\beta_2 = \frac{\partial Y}{\partial \ln (X)} = \frac{\partial Y}{\partial X} \frac{X}{X} = \frac{\partial Y}{\partial X} X$$

Dividing by 100 transforms the rate of change into units percentages:

$$\frac{\beta_2}{100\%} = \frac{\Delta Y}{100 \frac{\partial Y}{\partial X} X \%} \approx \frac{\Delta Y}{\% \Delta X}$$

which is the change in $Y$ resulting from a 1% increase in $X$. Thus, to interpret the estimated coefficient $b_2$ we say that $b_2/100$ is the change in $Y$ resulting from a 1% increase in $X$.

The coefficient $\beta_2$ is not the elasticity of $Y$ with respect to $X$. The elasticity of $Y$ with respect to $X$ is:

$$\eta_{Y,X} = \frac{\partial Y X}{\partial X Y} = \frac{\beta_2}{Y}$$

and its estimate is:

$$\hat{\eta}_{Y,X} = \frac{b_2}{Y}$$

Thus, this is a model of a non-constant elasticity of $Y$ with respect to $X$, and the elasticity depends on the values of $Y$.

### 5.3.4 Lin-lin model

For completeness, consider the linear model in variables - our original model, which is called lin-lin (both the dependent variable and the regressor are not in logarithmic form).

$$Y = \beta_1 + \beta_2 X + u$$

Here

$$\beta_2 = \frac{\partial Y}{\partial X}$$
which is the change in $Y$ resulting from a one unit change in $X$. The elasticity of $Y$ with respect to $X$ is:

$$\eta_{Y,X} = \frac{\partial Y}{\partial X} \frac{X}{Y} = \beta_2 \frac{X}{Y}$$

and its estimate is:

$$\hat{\eta}_{Y,X} = b_2 \frac{X}{Y}$$

Thus, in the lin-lin model, the point elasticity of $Y$ with respect to $X$ depends on the values of $X$ and $Y$.

### 5.3.5 Summary of models with log forms

The next table summarizes the interpretation of the estimated slope coefficient in the models discussed in this section, as well as the estimated elasticity of $Y$ with respect to $X$ in these models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Interpretation of slope $b_2$</th>
<th>Estimating $\eta_{Y,X}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(Y) = \beta_1 + \beta_2 \ln(X) + u$</td>
<td>$b_2 = \frac{\Delta Y}{\Delta X} = \eta_{Y,X}$</td>
<td>$\hat{\eta}_{Y,X} = b_2$</td>
</tr>
<tr>
<td>$\ln(Y) = \beta_1 + \beta_2 X + u$</td>
<td>$100b_2 = \frac{\Delta Y}{\Delta X}$</td>
<td>$\hat{\eta}_{Y,X} = b_2 X$</td>
</tr>
<tr>
<td>$Y = \beta_1 + \beta_2 \ln(X) + u$</td>
<td>$b_2/100 = \frac{\Delta Y}{\Delta X}$</td>
<td>$\hat{\eta}_{Y,X} = b_2 Y$</td>
</tr>
<tr>
<td>$Y = \beta_1 + \beta_2 X + u$</td>
<td>$b_2 = \frac{\Delta Y}{\Delta X}$</td>
<td>$\hat{\eta}_{Y,X} = b_2 X$</td>
</tr>
</tbody>
</table>

### 5.4 Estimating Elasticity in General Models

In general suppose you have a model when both the dependent and an independent variables appear transformed with some functions:

$$f(Y) = \beta_1 + \beta_2 g(X) + u$$

We see that

$$\beta_2 = \frac{\partial f(Y)}{\partial g(X)} = \frac{f'(Y)}{g'(X)} \frac{\partial Y}{\partial X}$$

In order to calculate the elasticity of $Y$ with respect to $X$ you need:

$$\eta_{Y,X} = \beta_2 \frac{X g'(X)}{Y f'(Y)}$$

Thus, the estimated elasticity is

$$\hat{\eta}_{Y,X} = b_2 \frac{X g'(X)}{Y f'(Y)}$$
5.5 Quadratic Form

In the previous examples we considered models where the regressor has a monotone effect on the dependent variable. That is, we considered models where a regressor $X$ can have either positive or negative effect on $Y$. For example, we postulated that schooling should always have a positive effect on earnings, and car’s weight should always have a negative effect on the car’s fuel consumption. Indeed, our estimates were consistent with these presumptions. However, sometimes we expect a non-monotone effect of a regressor on the dependent variable.

For example, experience may have a positive effect on earnings in the beginning of a career, but after certain number of years the worker’s productivity can decline due to health. For example, the relationship between earnings and experience could be like in the next figure.

![Graph showing the relationship between earnings and experience](image)

In the above figure we see that experience has positive effect on earnings up to 25 years, and after that each additional year of experience actually lowers earnings. In order to estimate such relationship, we can use a model such as:

$$ Y = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 EXP^2 + u $$

Here $Y$ is hourly earnings in dollars, $S$ is years of schooling, and $EXP$ is years of experience. If our presumption is correct, and indeed the effect of experience on earnings is as depicted in the figure above, then we expect that $\beta_3 > 0$ and $\beta_4 < 0$. The effect of experience on earnings is:

$$ \frac{\partial Y}{\partial EXP} = \beta_3 + 2\beta_4 EXP $$

Notice that with $\beta_3 > 0$, when the worker has little to no experience, the effect of experience on earnings is positive. However, if $\beta_4 < 0$, as experiences reaches certain level, the effect of experience on earnings become negative.
We can calculate the threshold level of experience above which the effect of experience on earnings switches from positive to negative:

$$\frac{\partial Y}{\partial EXP} = \beta_3 + 2\beta_4 EXP^* = 0$$

$$EXP^* = -\frac{\beta_3}{2\beta_4}$$

With the assumptions $\beta_3 > 0$ and $\beta_4 < 0$, the threshold level of experience we find, $EXP^*$, must be positive. For example, suppose that we estimated $b_3 = 0.27$ and $b_4 = -0.005$. The threshold level of experience, beyond which experience has negative effect on earnings is:

$$\overline{EXP}^* = -\frac{b_3}{2b_4} = -\frac{0.27}{2 \cdot (-0.005)} = 27$$

This means, that experience has positive effect on earnings, up to 27 years of experience, but each additional year of experience after 27 years, is reduces the hourly earnings. We can conjecture that this happens because of age, fatigue and lack of modern skills.

Non-monotone effects of one variable on another is abundant in economics. In the Solow growth model the effect of saving rate on consumption per worker is non-monotone; positive up to certain level of saving rate (the golden rule saving rate), and negative above this level. In business, we realize that spending on advertising can boost profit, but too much spending on advertising must decrease profit (consider the extreme case where the firm spends all its revenue on advertising, which results in a negative profit). Finally, macroeconomists study whether more government is good for the economy or not. We may argue that certain level of government is essential, because public goods cannot be provided efficiently (or at all) by competitive markets, and some regulation is necessary. However, governments that are too big create inefficiencies due to corruption and overregulation. In this section we developed a tool that allows estimating the optimal size of the government, by including the quadratic term of government size in a regression with real GDP or its growth rate as dependent variable.
Chapter 6

Qualitative Regressors and Dummy Variables

In many practical applications, the data is presented in the form of qualitative variables (words), as opposed to quantitative variables (numbers). Variables such as weight, height, earnings, are all quantitative because they assume numerical values. Other important variables, such as sex = \{female, male\}, race = \{white, black, Asian, Other\}, occupation = \{services, manufacturing, education, farming,...\}, religion = \{Christianity, Islam, Buddhism,...\}, political regime = \{Democracy, Communism, Dictatorship,...\}. Each word in the curly brackets is called a category, as opposed to value in the case of a quantitative variable. As you can see, many important variables in the real world are qualitative and do not have natural numerical values.

It should be mentioned that statistical agencies assign arbitrary numbers to categories of qualitative variables, for convenience. These are called codes. For example, the category male can be coded as 1, and the category female by 2. These codes however, do not represent numerical values in the usual sense, such that higher number represents greater value. For example, an income of $50k is greater than income of $40k, but we cannot state that "female is greater than male" in the usual mathematical sense. We need to be careful, and realize that just because a qualitative variable is coded with some numbers by a statistical agency, this does not mean that the variable is numerical, and we must remember that these codes do not represent any meaningful values.

6.1 Using Dummies to Represent Qualitative Variables

Qualitative variables cannot be used directly in statistical analysis, because estimation requires numerical values. However, each category of a qualitative variable can be represented by a dummy variable (also known as indicator variable), which attains the value of "1" if the observation belongs to that category, and "0" if it doesn’t. For example, consider the qualitative variable sex = \{female, male\}, which has two categories. We can create two
dummy variables, one for each category:

\[
[1]: \quad \text{female} = \begin{cases} 
1 \quad & \text{if female} \\
0 \quad & \text{otherwise}
\end{cases}
\]

\[
[2]: \quad \text{male} = \begin{cases} 
1 \quad & \text{if male} \\
0 \quad & \text{otherwise}
\end{cases}
\]

These variables are referred to as "female dummy" and "male dummy". Notice that the sum female + male = 1 for every observation in the sample, indicating that a person can be either female or male.

As another example, consider the qualitative variable race = \{white, black, Asian, Hispanic, Other\}. In this example it has 5 categories of race, so we can create 5 dummy variables, one for each category:

\[
[1]: \quad \text{white} = \begin{cases} 
1 \quad & \text{if white} \\
0 \quad & \text{otherwise}
\end{cases}
\]

\[
[2]: \quad \text{black} = \begin{cases} 
1 \quad & \text{if black} \\
0 \quad & \text{otherwise}
\end{cases}
\]

\[
[3]: \quad \text{Asian} = \begin{cases} 
1 \quad & \text{if Asian} \\
0 \quad & \text{otherwise}
\end{cases}
\]

\[
[4]: \quad \text{Hispanic} = \begin{cases} 
1 \quad & \text{if Hispanic} \\
0 \quad & \text{otherwise}
\end{cases}
\]

\[
[5]: \quad \text{Other} = \begin{cases} 
1 \quad & \text{if not white, black, Asian or Hispanic} \\
0 \quad & \text{otherwise}
\end{cases}
\]

These 5 dummy variables are referred to as "white dummy", "black dummy", "Asian dummy", "Hispanic dummy" and "other dummy". Notice once again that the sum of all the dummies for each qualitative variable is always 1. This is because we created dummy variable for all categories, and the categories are mutually exclusive. Therefore, each observation unit (each person here) can only be either white or black or Asian or Hispanic or other.

**Exercise 35** Consider the variable \textit{STATE} = \{Alabama, Alaska, Arizona,..., Wisconsin, Wyoming\}, i.e. the list of all the states in the united states.
(a) \textit{Is STATE} a quantitative or a qualitative variable?
(b) \textit{Demonstrate how you can represent the state of California with a dummy variable.}

### 6.2 Dummy Variables in Regression Analysis

Dummy variables can be used in regression analysis in the same way as other quantitative variables, and we can perform statistical hypothesis testing about the unknown coefficients, just like with regular quantitative variables. One has to keep in mind two important properties about using dummy variables. First, since all the dummy variables for any qualitative variable add up to 1, you must exclude one variable from the list of regressors. In the case
of sex = \{\text{female, male}\}, we created two dummy variables, one for female and the other for male. You cannot include both of them in any regression, because female + male = 1, so there is an exact linear relationship between them and that creates perfect multicollinearity. If you include both dummies, any statistical package will drop one of them and issue a warning. Similarly, in the case of the qualitative variable race, we created 5 dummies for 5 race categories. Again, you cannot include all 5 and must exclude one dummy. The excluded category is called reference category, or omitted category, or excluded category.

The second issue with dummy variables, is that the interpretation of the coefficients is different from regular numerical variables. Suppose that \( S \) is years of schooling, which is a regular quantitative variable. The estimated coefficient on \( S \) shows the predicted impact of one year increase in the schooling level on the dependent variable (say earnings). If instead we look at the coefficient on female dummy, what is the meaning of "one unit increase in female"? Since the female dummy attains the value of "0" for males, then, roughly speaking, the change from 0 to 1 captures the difference between male and female. The remaining subsections of this chapter are dedicated to the interpretation of the estimated coefficients on dummy variables, and interactions between them and other (dummy or numerical) variables.

### 6.2.1 One dummy variable

Suppose that we would like to measure how earnings (in dollars per hour) depend on schooling, work experience, and also on sex = \{\text{female, male}\}. The sex (gender) variable is qualitative variable with two categories, so we can create two dummy variables:

\[
F = \begin{cases} 
1 & \text{if female} \\
0 & \text{otherwise} 
\end{cases}, \quad M = \begin{cases} 
1 & \text{if male} \\
0 & \text{otherwise} 
\end{cases}
\]

We choose the "male" to be the reference category, so we do not include the male dummy in the regression. Therefore, we only include the female dummy and the model is:

\[
EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 F + u
\]

Next, we estimate the coefficients \( \beta_1, \beta_2, \beta_3 \) as usual, and obtain the fitted equation:

\[
EARNINGS = b_1 + b_2 S + b_3 EXP + b_4 F \quad (6.1)
\]

We have see that \( b_2 \) is the estimated effect on earnings resulting from 1 unit increase in schooling, holding all other regressors the same. In other words, additional year of schooling increases earnings by \( b_2 \) dollars per hour. Similarly, additional year of experience increases earnings by \( b_3 \) dollars per hour. The interpretation of \( b_4 \) is similar. The dummy variable female increases by one unit as we switch from observation on male to an observation on female. Thus, \( b_4 \) is the additional earnings of female over male, holding other regressors same for both (schooling and experience). To see this more vividly, notice that equation (6.1) can be broken into two fitted earnings equations, one for females and the other is for males.

\[
\begin{align*}
EARNINGS_F &= b_1 + b_2 S + b_3 EXP + b_4 \\
EARNINGS_M &= b_1 + b_2 S + b_3 EXP
\end{align*}
\]
Subtracting the male earnings from female earnings, holding $S$ and $EXP$ the same for both, we get:

$$EARNINGS_F - EARNINGS_M = b_4$$

Notice that we subtracted the reference category’s earnings from female’s earnings, which means that women earn $b_4$ dollars per hour more than men. If the estimate $b_4$ is negative, we conclude that women earn less than men by $|b_4|$ dollars per hour, given same schooling and experience.

The reason why male was chosen to be the reference category, is that we usually suspect that women are discriminated against, and we want to measure women’s earnings relative to male. It would not make a difference if we had chosen female to be the reference category, and fitted:

$$EARNINGS = b_1 + b_2S + b_3EXP + \tilde{b}_4M$$

This time $\tilde{b}_4$ would give us the excess earnings of male over female:

$$EARNINGS_M - EARNINGS_F = \tilde{b}_4$$

The two estimates are related as follows:

$$\tilde{b}_4 = -b_4$$

### 6.2.2 Several dummy variables

Suppose that in addition to gender, we are interested in the impact of race on earnings. Suppose that race = \{white, black, other\}, so for simplicity we only have 3 race categories.

We create race dummies as follows:

$$W = \begin{cases} 1 & \text{if white} \\ 0 & \text{otherwise} \end{cases}, \quad B = \begin{cases} 1 & \text{if black} \\ 0 & \text{otherwise} \end{cases}, \quad O = \begin{cases} 1 & \text{if not white or black} \\ 0 & \text{otherwise} \end{cases}$$

We choose the reference category for race to be white, so the estimated model is:

$$EARNINGS = \beta_1 + \beta_2S + \beta_3EXP + \beta_4F + \beta_5B + \beta_6O + u$$

and the fitted equation is:

$$EARNINGS = b_1 + b_2S + b_3EXP + b_4F + b_5B + b_6O \quad (6.2)$$

The reference category is therefore white male.

Equation \(6.2\) allows us to compare earnings not only by gender, but also by race. It can be broken into six equations (earnings of white males, black male, other males, and the

\footnote{In Census, there are many more categories of race, including Asian, Native American, and mixes between them. Hispanic is a separate variable, which has many sub categories.}
6.2. DUMMY VARIABLES IN REGRESSION ANALYSIS

same for females):

\[
\begin{align*}
E\text{ARNINGS}_{M,W} & = b_1 + b_2 S + b_3 EXP \\
E\text{ARNINGS}_{M,B} & = b_1 + b_2 S + b_3 EXP + b_5 \\
E\text{ARNINGS}_{M,O} & = b_1 + b_2 S + b_3 EXP + b_6 \\
E\text{ARNINGS}_{F,W} & = b_1 + b_2 S + b_3 EXP + b_4 \\
E\text{ARNINGS}_{F,B} & = b_1 + b_2 S + b_3 EXP + b_4 + b_5 \\
E\text{ARNINGS}_{F,O} & = b_1 + b_2 S + b_3 EXP + b_4 + b_6
\end{align*}
\]

The estimates \( b_5 \) and \( b_6 \) capture the race effects on earnings. For example, the estimated difference between black males and white males earnings, holding everything else the same, is:

\[
E\text{ARNINGS}_{M,B} - E\text{ARNINGS}_{M,W} = b_5
\]

A negative \( b_5 \) is evidence of discrimination against black men relative to white men.

Similarly, we can compare black women with white women:

\[
E\text{ARNINGS}_{F,B} - E\text{ARNINGS}_{F,W} = b_5
\]

This is the same as the difference between black men and white men. Thus, this model assumes that there is no difference between race effect among male and female. In other words, \( b_5 \) is the difference between earnings of black and earnings of white, holding all other characteristics fixed (schooling, experience, gender).

The interpretation of \( b_6 \) is similar:

\[
E\text{ARNINGS}_O - E\text{ARNINGS}_W = b_6
\]

That is, \( b_6 \) is the difference between earnings of other race and earnings of white, holding all other characteristics fixed (schooling, experience, gender).

It is possible however, that black women experience different degree of discrimination from black men. In order to test this hypothesis, we need to introduce interaction terms between the sex dummy and gender dummy.

6.2.3 Interaction between two dummy variables

Suppose that we want to test whether race effect is different for the two genders. We estimate a model with interaction effects as follows:

\[
E\text{ARNINGS} = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 F + \beta_5 B + \beta_6 O + \beta_7 F \cdot B + \beta_8 F \cdot O + u
\]

Interaction of two regressors is a regressor which is equal to the product of the two regressors. The fitted equation is:

\[
\hat{E\text{ARNINGS}} = b_1 + b_2 S + b_3 EXP + b_4 F + b_5 B + b_6 O + b_7 F \cdot B + b_8 F \cdot O
\]
The black-white earnings gap for male is:

\[ \text{EARNINGS}_{M,B} - \text{EARNINGS}_{M,W} = b_5 \]

The black-white earnings gap for female is:

\[ \text{EARNINGS}_{F,B} - \text{EARNINGS}_{F,W} = b_5 + b_7 \]

Thus, \( b_7 \) represents the difference between black-white earnings gap for women and men, both relative to white (the reference race category). To test if the black-white earnings gap is different for men and women, we test:

\[
H_0 : \beta_7 = 0 \\
H_1 : \beta_7 \neq 0
\]

If we do not reject the null hypothesis, this would mean that we did not find enough evidence that black-white earnings gap for women is any different from the black-white earnings gap for men (both relative to their white counterparts).

The same coefficient, \( b_7 \), also represents the difference between gender earnings gap for black and for white. The female-male earnings gap for white is

\[ \text{EARNINGS}_{F,W} - \text{EARNINGS}_{M,W} = b_4 \]

The female-male earnings gap for black is

\[ \text{EARNINGS}_{F,B} - \text{EARNINGS}_{M,B} = b_4 + b_7 \]

Thus, \( b_7 \), is the difference between the female-male earnings gap for black and for white.

### 6.2.4 Interaction between a dummy and a quantitative variable

So far we assumed that the contribution of each year of schooling and experience is the same for both genders, and for all races. Suppose that we want to test whether each year of schooling increases the earnings of female by the same amount as for male. We introduce an interaction term between female dummy and schooling variable, which is continuous. The model we estimate is:

\[ \text{EARNINGS} = \beta_1 + \beta_2 S + \beta_3 \text{EXP} + \beta_4 F + \beta_5 S \cdot F + u \]

The fitted model is

\[ \text{EARNINGS} = b_1 + b_2 S + b_3 \text{EXP} + b_4 F + b_5 S \cdot F \]

Again, it is helpful to break this equation into earnings for male and female:

\[
\text{EARNINGS}_M = b_1 + b_2 S + b_3 \text{EXP} \\
\text{EARNINGS}_F = b_1 + (b_2 + b_5) S + b_3 \text{EXP} + b_4
\]
Notice that for male, each year of schooling increases earnings by $b_2$ dollars, and for female by $b_2 + b_5$ dollars. Thus, $b_5$ is the extra effect of additional year of schooling on females’ earnings, compared with males’ earnings. If the coefficient $\beta_5$ turns out to be insignificant, then we conclude that additional year of schooling has the same effect on female as on male. In similar fashion, one can interact female with experience to test whether each year of experience adds the same amount to female earnings as to male earnings. Also, one can interact race dummies with schooling and experience to test whether the gains from schooling and experience are the same for different races.

6.2.5 Dummy variables in log-lin model

Suppose that our model is

$$\ln (Y) = \beta_1 + \beta_2 S + \beta_3 F + u$$

where $Y$ is earnings, $S$ is schooling, and $F$ is a dummy variable for female. The fitted equation is

$$\ln (\hat{Y}) = b_1 + b_2 S + b_3 F$$

Thus, $b_3$ is the estimated difference between log of earnings of female and male, holding all other regressors the same:

$$\ln (\hat{Y}_F) = b_1 + b_2 S + b_3$$
$$\ln (\hat{Y}_M) = b_1 + b_2 S$$

$$\Rightarrow \ln (\hat{Y}_F) - \ln (\hat{Y}_M) = b_3$$

We can show that $100 \cdot b_3$ is approximately the % difference between female and male earnings. Notice that we can write:

$$\ln (\hat{Y}_F) - \ln (\hat{Y}_M) = \ln \left( \frac{\hat{Y}_F}{\hat{Y}_M} \right)$$
$$= \ln \left( \frac{\hat{Y}_M + \hat{Y}_F - \hat{Y}_M}{\hat{Y}_M} \right)$$
$$= \ln \left( 1 + \frac{\Delta}{\hat{Y}_M} \right)$$
$$= \ln (1 + g)$$

where $\Delta = \hat{Y}_F - \hat{Y}_M$ is the predicted gap between female and male earnings, and $g$ is the relative gap between female and male earnings, expressed as a fraction of males earnings. One can prove that if the relative gap $g$ is small, then $\ln (1 + g) \approx g$, and we have approximately $g = b_3$ or $100g = 100b_3$. If the relative gap $g$ is large, say more that 0.1, this approximation is not very good and we can calculate the exact % difference between female and male earnings as follows:

$$\ln (1 + g) = b_3$$
$$g = \exp (b_3) - 1$$
Thus, the exact % difference between female and male earnings is:

$$100g = 100 \left[ \exp (b_3) - 1 \right]$$
Chapter 7

Specification of Regression Models

In our previous homework assignments and class exercises, the regression model was always given and you were asked to estimate it, interpret the coefficients and perform statistical tests. In practice, econometricians (researchers) must decide for themselves what model they want to estimate. The decision about what should be the dependent variable, the regressors, and the functional form is called the model specification. In principle, economic theory (or other theory) and common sense should provide guidance for what regressors need to be included in the estimated model. For example, economic theory and common sense suggest that schooling and work experience are important determinants of earnings. More schooling might indicate higher ability of workers, and therefore employees are willing to pay more to workers with more advanced degrees. Work experience increases the specialized skills of the worker through learning-by-doing. When you decide what variables should be included in any empirical model, you must justify each variable by reference to economic theory, other theory or common sense.

It is very common that researchers make specification errors. For example, a researcher might not be familiar enough with economic theory, and failed to include a relevant regressor in the model. It could also be the case that an important variable is missing because the researcher was unable to find data on that variable. For example, in the earnings equation, we included years of schooling and experience, but common sense also suggests that "quality of schooling" should matter. It is not obvious how to obtain data on the variable that represents quality of schooling, and we will get back to this issue in the section on proxy variables. In any case, the first specification error that we will discuss is omitted variable bias. Similarly, a researcher might include a variable that is not supposed to be in the model. For example, the number of squirrels in your state is probably of no consequence for your income. We will see that the consequences of omitting a relevant variable are far more severe than of including an irrelevant variable. In fact, the biggest problem in applied econometrics, is by far the omitted variable bias.
7.1 Omitted Variable Bias

7.1.1 Two variable model

Suppose that the true model is:

\[ \text{true model} : Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \]

but the researcher thinks that the model is:

\[ \text{misspecified model} : Y = \beta_1 + \beta_2 X_2 + \bar{u} \]

In other words, the researcher omits a relevant variable \( X_3 \) and fits \( \hat{Y}_{miss} = \bar{b}_1 + \bar{b}_2 X_2 \). Notice that the error term in the misspecified model is \( \bar{u} = \beta_3 X_3 + u \). We will show that the OLS estimators \( \bar{b}_2 \) and \( \bar{b}_1 \) are biased and inconsistent. The estimated standard errors are biased as well, so we are unable to test any hypotheses with the regression results.

Recall that the OLS estimator of \( \beta_2 \) in the misspecified model is:

\[
\hat{b}_2 = \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2}
\]

Substituting the true \( Y_i \)s:

\[
\hat{b}_2 = \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) \left[ (\beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i) - (\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u}) \right]}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2}
\]

\[
= \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) \left[ \beta_2 (X_{2,i} - \bar{X}_2) + \beta_3 (X_{3,i} - \bar{X}_3) + (u_i - \bar{u}) \right]}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2}
\]

\[
= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) (X_{3,i} - \bar{X}_3)}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2} + \sum_{i=1}^{n} a_i u_i
\]

where

\[
a_i = \frac{(X_{2,i} - \bar{X}_2)}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2}
\]

The expected value of \( \hat{b}_2 \) is therefore:

\[
E(\hat{b}_2) = \beta_2 + \beta_3 E \left( \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) (X_{3,i} - \bar{X}_3)}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2} \right) = \beta_2 + \beta_3 \left( \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) (X_{3,i} - \bar{X}_3)}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2} \right)
\]

Bias

\[
E(\hat{b}_2) = \beta_2 + \beta_3 \left( \frac{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2) (X_{3,i} - \bar{X}_3)}{\sum_{i=1}^{n} (X_{2,i} - \bar{X}_2)^2} \right)
\]

Bias, if A.0 assumed
Recall that in section 3.2.2 we showed that \( E(\sum_{i=1}^{n} a_i u_i) = 0 \). Also recall that assumption A.0. was made, that the regressors are non-random, and therefore the last expectation is gone. If we do not make assumption A.0., then the bias includes the expectation term. Thus, the bias term is the product of \( \beta_3 \) (the true slope coefficient on the omitted variable \( X_3 \)) and \( d_2 \), the OLS estimator of \( \delta_2 \) in the regression

\[
X_3 = \delta_1 + \delta_2 X_2 + \nu
\]

Therefore, the estimator \( \hat{b}_2 \) is a biased estimator of \( \beta_2 \) whenever \( d_2 \neq 0 \), with bias

\[
\text{bias} \left( \hat{b}_2 \right) = \beta_3 d_2
\]

Recall that the OLS estimator \( d_2 \) has the same sign as the sample correlation between \( X_2 \) and \( X_3 \), which allows us to figure out the direction of bias. Observe that \( \text{bias} \left( \hat{b}_2 \right) \) is positive (i.e. \( \hat{b}_2 \) overestimates the true parameter \( \beta_2 \)) if \( \beta_3 \) and \( d_2 \) have the same sign. The \( \text{bias} \left( \hat{b}_2 \right) \) is negative when \( \beta_3 \) and \( d_2 \) have different signs. In general, in a multiple regression model, the direction of omitted variable bias depends on the correlations between the omitted variable and all the included variables.

Important! The bias \( \text{bias} \left( \hat{b}_2 \right) = \beta_3 d_2 \) will never vanish, even when sample size increases indefinitely (as \( n \to \infty \)). This is because \( \beta_3 \neq 0 \) and \( d_2 \to \delta_2 \neq 0 \) as \( n \to \infty \).

**Exercise 36** Consider two models of earnings, with regressors \( S \) - years of schooling, and \( EXP \) - experience.

[true model] : \( \text{EARNINGS}_i = \beta_1 + \beta_2 S_i + \beta_3 \text{EXP}_i + u_i \)

[misspecified model] : \( \text{EARNINGS}_i = \beta_1 + \beta_2 S_i + \bar{u}_i \)

Suppose that a researcher estimates the misspecified model, and obtains an estimate \( \hat{b}_2 \) of \( \beta_2 \).

(a) Is \( \hat{b}_2 \) likely to overestimate or underestimate the true parameter \( \beta_2 \)?

(b) Provide intuition for the result in part (a).

To summarize, omitted variable bias is by far the most severe problem in applied econometrics. In general, omitted variables lead to biased and inconsistent estimators of coefficients on the included variables. Recall that biased estimators are not necessarily bad, if they are consistent and we have a large sample, in which case the bias vanishes. But inconsistency means that no matter how big our sample is, we can never get rid of the bias. Moreover, omitted variable bias also results in biased standard errors of estimators, which makes all statistical tests invalid.

### 7.1.2 General k-variable model

Suppose that the true model is:

[true model] : \( Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{k-1} X_{k-1} + \beta_k X_k + u \)
but the researcher thinks that the model is:

\[ \text{misspecified model} : Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_{k-1} X_{k-1} + \tilde{u} \]

where \( \tilde{u} = u + \beta_k X_k \). It can be shown that the OLS fitted equation of the misspecified model is:

\[ \hat{Y}_{\text{miss}} = (b_1 + b_k d_1) + (b_2 + b_k d_2) X_2 + (b_3 + b_k d_3) X_3 + ... + (b_{k-1} + b_k d_{k-1}) X_{k-1} \]

Where \( b_1, b_2, ..., b_k \) are OLS estimates from the true model, and \( d_1, d_2, ..., d_{k-1} \) are OLS estimates of the coefficients \( \delta_1, \delta_2, ..., \delta_{k-1} \), from the regression of the omitted variable on all the included variables:

\[ X_k = \delta_1 + \delta_2 X_2 + ... + \delta_{k-1} X_{k-1} + v \]

Thus, the difference between estimated coefficients in the misspecified model and the true model are \( b_k d_1, b_k d_2, ..., b_k d_{k-1} \), and the biases are \( \beta_k d_1, \beta_k d_2, ..., \beta_k d_{k-1} \). For example, the bias of \( b_2 \) is \( \beta_k d_2 \), i.e. the unknown coefficient on the omitted variable in the true model, times the estimated coefficient on \( X_2 \) in the regression of the omitted variable on all the included variables.

### 7.2 Inclusion of Irrelevant Variable

Suppose that the true model is:

\[ \text{true model} : Y = \beta_1 + \beta_2 X_2 + u, \]

but the researcher thinks that the model is:

\[ \text{misspecified model} : Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \]

In other words, the true coefficient on \( X_3 \) is \( \beta_3 = 0 \), but the researcher, not realizing this, fits the equation:

\[ \hat{Y} = b_1 + b_2 X_2 + b_3 X_3 \]

Including irrelevant variable does not introduce a bias of regression coefficients. The reason is that the assumptions which guarantee that OLS estimators are unbiased, are not violated. In particular, assumption A.3 \( \mathbb{E}(u_i) = 0 \) for all \( i \) was used to prove that OLS coefficients are unbiased, still holds. In contrast, when we excluded a variable which was relevant (part of the true model), the error term in the misspecified model included the omitted variable, and as a result the mean of the error was not the same for all observations.

The only adverse consequence of including an irrelevant variable is that the OLS coefficients are no longer efficient, i.e. do not have the smallest possible variance. To be specific, the variance of \( b_1, b_2 \) when fitting \( \hat{Y} = b_1 + b_2 X_2 + b_3 X_3 \) is higher than the variance of the same estimators when fitting the correct equation \( \hat{Y} = b_1 + b_2 X_2 \). Intuitively, this happens because in the true model we have \( \beta_3 = 0 \), which is a piece of information that the researcher ignores when fitting \( \hat{Y} = b_1 + b_2 X_2 + b_3 X_3 \). As usual, more information (which is relevant of course) improves the precision of the estimators. For example, larger sample size
increases the precision of estimators (lowers their variance). Including irrelevant variable is equivalent to not using some relevant information (that the coefficient on that variable is zero), and therefore the result is higher variance of the estimators. A larger variance means a less precise estimator, and this translates into larger confidence intervals and less accurate hypothesis tests.

That being said, the consequences of including irrelevant variables are not as severe as those of omitting relevant variable. Nevertheless, it is not a good practice to include arbitrary variables in your regression model. To avoid inclusion of irrelevant variables, get the habit of justifying each and every variable in the regression model with economic or other theory. You must have a good reason behind inclusion of any variable in the model.

7.3 Proxy Variables

As we have seen in the last two sections, omitting a relevant variable has far more severe consequences than including an irrelevant variable into your regression model. What should we do if we realize that based on economic theory a given variable should be in the regression, but the data on such variable does not exist? Sometimes it is possible to find a proxy variable, which is highly correlated with the omitted variable and therefore can substitute for it.

For example, when you might realize that quality of schools in a district is an important determinant of housing prices, and ideally you would include schools quality in a housing price equation. The problem is that there is no data on "quality of schools". One solution is to include proxy variable(s) in your regression model, such as teacher/student ratio, average spending per student, etc. Similarly, a researcher realizes that besides schooling and experience in the wage regression, one should also include "ability" variable, which perhaps captures innate ability, motivation, wisdom of the worker. The problem again is that there is no data on "ability". Instead, researchers sometimes use IQ tests, or other similar tests which hopefully measure various aspects of "ability". As another example, consider "health" level of population, which might be relevant in various studies related to economic performance of a country. Again, there is no such variable as "health", so researchers use life expectancy, mortality rates and prevalence of diseases as proxies for "health".

Researchers show great ingenuity in collecting data on proxy variables. As our last example, imagine that you would like to include a variable "honesty" in a model of economic development, with the idea that societies with higher degree of honesty tend to prosper more. Obviously, there is no data anywhere on the variable "honesty". In order to proxy for honesty in different countries, researchers "lost" a bunch of wallets in different cities (New York, Beijing, London, etc.). Each wallet had equivalent amount of money, adjusted for the purchasing power in that location. In addition, each wallet contained an address, so that the finder of the wallet could return it to the owner. This way, the researchers collected data on the fraction of wallets returned to their owner in each location, and that was their proxy for "honesty".
7.4 Steps of Applied Research

In this section I provide very brief guidelines for conducting empirical research.

1. Research (Topic) Question. The starting point of applied research project is a research topic/question. I advise to phrase the title as a question, and be very specific. A bad research title is "Labor Markets", because it is too general. A good research title "Does Higher Minimum Wage Increase Unemployment of Unskilled Workers?".

2. Literature Review. The starting point can be a textbook, which cites seminal papers on a particular topic. Try to identify the famous researchers who published influential article on your topic, and look at their webpages for further references. Focus on the introduction and conclusions, if articles that are too technical.

3. Economic (or other) Theory. Find out the relevant economic theory for the topic or the question at hand. For example, if the topic is Minimum Wage, then it is part of the field of Labor Economics, and the relevant theories are the labor markets, efficiency wage theory, search model of labor markets, etc.

4. Data Collection and Inspection. The economic theory should dictate what kind of data is needed for the project. Before performing any type of statistical analysis, you must summarize and visualize the data (summary statistics, visualization). You must document the data sources, and the units of the data.

5. Econometric (Statistical) Model. The economic theory should dictate the econometric model. Each variable included in your model, must be justified on the basis of theory. For example, to justify the inclusion of schooling in the earnings model, one can refer to the human capital theory (whereby schooling helps accumulate skills, which enhance to earning ability), signalling theory (whereby schooling serves as a signal to employers about the ability and potential of the candidate), or social networking theory (whereby attending college helps meet business partners and make connections).

6. Statistical Analysis. Typically regression analysis, binary choice models, etc. Present the results in publication quality tables.

7. Conclusions. You must summarize and interpret the results of your statistical analysis. For example "we found strong evidence that higher minimum wage increases unemployment among low-skilled workers, especially teenagers".

Chapter 8

Heteroscedasticity

So far we discussed the following problems, encountered in practical research: (i) Multicollinearity (perfect, imperfect), (ii) Omitted variable bias, and (iii) Inclusion of irrelevant variable. In this chapter we add another practical problem - Heteroscedasticity (from Ancient Greek ἕτερος is “different” and σκέδασις is “dispersion”). Recall assumption A.4 about the regression model, in section 3.1, which states that the disturbance (error) term is homoscedastic.

\[ \text{var}(u_i) = \sigma^2_u \quad \text{for all } i \]

In other words, we assumed that the variance of the error term is the same for all observations. In this chapter we discuss what happens when this assumption is violated. The resulting problem is called heteroscedasticity. Mathematically, heteroscedasticity can be defined as \( \text{var}(u_i) \neq \text{var}(u_j) \) for some \( i \neq j \).

8.1 Illustrating Heteroscedasticity

The next figure illustrates a simple regression model with homoscedastic disturbance term: \( Y_i = \beta_1 + \beta_2 X + u_i \). The bell shaped curves are plots of the pdf of \( Y_i \), which have a mean of \( \beta_1 + \beta_2 X \). You can see from these graphs that all \( u_i \) have mean of zero and the same variance.

---

1Homo means same in Greek.
CHAPTER 8. HETEROSCEDASTICITY

The next figure illustrates a simple regression model with heteroscedastic disturbance term, such that \( \text{var}(u_i) \) is increasing with the value of the regressor \( X \). The pdfs of the \( u_i \)s are still centered around zero, but they get more dispersed (or less precise) as the values of \( X \) increase.

\[ Y = \beta_1 + \beta_2 X \]

8.2 Consequences of Heteroscedasticity

In the presence of heteroscedasticity, the OLS estimators are still unbiased and consistent. However, with heteroscedasticity we have two adverse consequences: (i) OLS estimators are inefficient, i.e. no longer have the lowest variance among all unbiased linear estimators, and (ii) the estimated standard errors of the regression coefficients, \( s.e.(b) \), are biased and as a result the \( t \)-tests and the \( F \)-test are invalid.

8.3 Detection of Heteroscedasticity

There are several formal tests that detect heteroscedasticity in a regression model. Some are based on particular assumptions about the form of heteroscedasticity, while others are more general and do not depend on the particular form of heteroscedasticity.

8.3.1 Eyeballing (plotting the residuals)

We present first an informal method of detecting heteroscedasticity, which is based on simply observing the residuals. The following example uses artificial data available at [http://online.sfsu.edu/mbar/ECON312_files/hetero.csv](http://online.sfsu.edu/mbar/ECON312_files/hetero.csv). The file contains data on one regressor \( X \), and two dependent variables, \( Y_{\text{homo}} \) and \( Y_{\text{hetero}} \). The first one, \( Y_{\text{homo}} \), was generated with errors drawn from the same distribution \( u_i \sim N(0, \sigma_u^2) \), while the second one, \( Y_{\text{hetero}} \), was generated using error terms with standard deviation that is increasing in the values of the regressor \( X \). Formally, the two models are as follows:

\[
Y_{\text{homo},i} = \beta_1 + \beta_2 X_i + u_i, \quad \text{with } s.d.(u_i) = \sigma_u \quad \forall i
\]
\[
Y_{\text{hetero},i} = \beta_1 + \beta_2 X_i + v_i, \quad \text{with } s.d.(v_i) = 0.05 X_i \sigma_u
\]
Estimation of the homoscedastic model results in the following fitted equation: $\hat{Y}_{homo} = b_1 + b_2 X$. Using the command `predict` in Stata or the commands `predict` or `fitted` in R, we generate the fitted values $\hat{Y}_{homo}$, and plot them together with the actual values $Y_{homo}$ on the same diagram:

Notice that the actual values are spread evenly around the fitted line, which is expected when the error terms are homoscedastic. The difference between the actual and fitted values are the residuals ($e_i = Y_i - \hat{Y}_i$), and is generated with the command `residuals` in R. The next figure plots the residuals against the regressor $X$, and we see that the residuals are evenly spread around the zero (their mean).
Next we estimate the heteroscedastic model, and plot the actual and fitted values in the next diagram.

Notice that as the values of the regressor $X$ increase, the actual values are more dispersed around the fitted equation. This is an informal graphical evidence of the presence of heteroscedasticity. The next figure plots the residuals from the fitted model, and again notice how the residuals appear more spread as the values of the regressor $X$ increase.

The message from this section is that one informal way to detect heteroscedasticity is to plot the scatter of the residuals against regressors, and see if the variability looks constant, or changing in a systematic way with one of the regressors.
8.3. DETECTION OF HETEROSCEDASTICITY

8.3.2 The Goldfeld-Quand test

The figures in the previous section motivated the Goldfeld-Quand test, which we will describe now. The test is designed to detect one particular type of heteroscedasticity, the kind in which the standard deviations of the error terms are proportional to the values of a regressor \( X \) (increasing or decreasing). In other words, this is not a general test for detecting any kind of heteroscedasticity. The test starts with sorting the data by values of the regressor \( X \) in ascending order. Next we run the same regression as the original model on the first \( m \) observations and compute the residuals sum of squares (\( RSS_1 = \sum_{i=1}^{m} e_i^2 \)). Recall that \( RSS \) measures the variability of residuals in the sample, in this case in the sub sample containing the smallest \( m \) values of the regressor \( X \). Next we run the same regression using the last \( m \) observations (i.e. using the largest values of the regressor \( X \)). The residual sum of squares in this regression are \( RSS_2 \). It turns out that under the assumption of homoscedasticity, both \( RSS_1 \) and \( RSS_2 \) have a chi-square distribution, with \( m - k \) degrees of freedom and they are independent. Then the ratios have an \( F \)-distribution:

\[
\frac{RSS_2}{RSS_1} \sim F(m - k, m - k) \\
\frac{RSS_1}{RSS_2} \sim F(m - k, m - k)
\]

Let’s assume, without loss of generality, that we are testing for heteroscedasticity where the variance of the error terms is increasing in the values of the regressor \( X \). Thus, the residuals sum of squares in the regression on the last \( m \) observations should be greater than that from the first \( m \) observations. We therefore would like to perform the test:

\[
H_0 : \text{homoscedasticity} \\
H_1 : \text{heteroscedasticity, } \text{var}(u_i) \text{ increasing in } X_i
\]

In this case we reject the null hypothesis at significance level \( \alpha \), if the probability of \( F(m - k, m - k) > \frac{RSS_2}{RSS_1} \) is smaller than 5%. In other words, we reject the null hypothesis of homoscedasticity if the F-statistic \( \frac{RSS_2}{RSS_1} \) is large enough. In Stata, the above probability is calculated as follows: \( \text{Ftail}(m-k,m-k,RSS2/RSS1) \). In R, the same probability is calculated with \( \text{pf}(RSS2/RSS1,m-k,m-k,\text{lower.tail}=\text{FALSE}) \).²

If instead we want to test for heteroscedasticity where the variance of the error terms is decreasing in the values of the regressor \( X \), then the test is

\[
H_0 : \text{homoscedasticity} \\
H_1 : \text{heteroscedasticity, } \text{var}(u_i) \text{ decreasing in } X_i
\]

We reject the null hypothesis at significance level \( \alpha \) if the probability of \( F(m - k, m - k) > \frac{RSS_1}{RSS_2} \) is smaller than \( \alpha \).

The number of observations used in the two regression of the Goldfeld-Quand test, is usually set to be \( m = \frac{3}{8}n \), i.e. 3/8 of the total number of observations. This means that 1/4 of the observations from the middle are dropped, to increase the contrast between the first and last \( m \) observations. This number is recommended by Goldfeld and Quand based on their simulations.

²R, by default, calculates the lower tail probability.
8.3.3 The White test

The Goldfeld-Quandt test is designed to test for a particular type of heteroscedasticity— the kind in which the standard deviation of the error terms is proportional to the values of one of the regressors. Now we present a test (variation of a test developed by White 1980), which is aimed at detecting any unknown type of heteroscedasticity. Suppose that you estimate the model:

\[ Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_k X_{k,i} + u_i \]

The fitted equation is

\[ \hat{Y}_i = b_1 + b_2 X_{2,i} + b_3 X_{3,i} + \ldots + b_k X_{k,i} \]

and the associated residuals are:

\[ e_i = Y_i - \hat{Y}_i \]

The test requires estimating the following equation:

\[ e_i^2 = \delta_1 + \delta_2 \hat{Y}_i + \delta_3 \hat{Y}_i^2 + v_i \quad (8.1) \]

The dependent variable in equation (8.1), \( e_i^2 \), is an estimate of \( \text{var}(u_i) \). The idea is that if this variance does not depend on the regressors, then we should have \( H_0: \delta_2 = \delta_3 = 0 \) (homoscedasticity). Thus, the \( F \)-test of overall fit of the model (8.1) allows us to test for Heteroscedasticity of unknown kind. In particular, we reject \( H_0 \) at significance level \( \alpha \) if \( \text{Prob} > F \) after estimating (8.1) is smaller than \( \alpha \). In this case, we conclude that the original model suffers from heteroscedasticity of unknown kind.

8.4 Remedies for Heteroscedasticity

Recall that there are two adverse consequences of heteroscedasticity: (i) OLS estimators are inefficient, i.e. no longer have the lowest variance among all unbiased linear estimators, and (ii) the estimated standard errors of the regression coefficients, \( s.e.(b) \), are biased and as a result the \( t \)-tests and the \( F \)-test are invalid.

8.4.1 Heteroscedasticity-robust standard errors

The first solution we propose, corrects the estimated standard errors, which enables us to perform statistical tests based on the estimated model. Regardless of what form of heteroscedasticity we have, it is possible to obtain estimates of standard errors of estimates which are consistent. We will not discuss the details of this procedure, since it is beyond the scope of this course. We simply point that if testing hypothesis is very important, one simply needs to specify the option \texttt{robust} in the \texttt{regress} command, and Stata will compute robust standard errors which are imuned to any form of heteroscedasiticity.

For example, in Stata, \texttt{regress wage edu, robust} regresses the dependent variable wage on the regressor education, and the option \texttt{robust} following the comma indicates that you wish to obtain standard error of estimators which are robust to any heteroscedasticity. In R one needs to install packages \texttt{lmtest} and \texttt{sandwich}, and then the command...
coeftest(model,vcovHC(model,type="HC1")) estimates heteroscedasticity corrected (robust) standard errors.

This solution is very simple, and it always fixes the estimated standard errors of the OLS estimators. However, the estimated coefficients themselves are still inefficient, although they are unbiased and consistent. Most applied researchers however are not concerned too much with the fact that OLS estimators in the presence of heteroscedasticity are not efficient, especially if they have a large sample. This is why most researchers today use the robust option and only correct the standard errors (but not the estimators) for heteroscedasticity.

8.4.2 Weighted least squares (WLS)

The robust option is by far the most popular remedy for heteroscedasticity today. However, if we are able to figure out the form of heteroscedasticity, one can correct not only the standard errors of the coefficients, but also obtain efficient estimators for the regression coefficients themselves. The idea is to weight the observations in inverse proportion to the variance of the disturbance term. This way, we assign more weight to observations with smaller variance of the disturbance term, which is why this estimation is called weighted least squares (as opposed to ordinary least squares, which treats all observations the same).

Suppose that you estimate the model:

\[ Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + v_i \]  

(8.2)

where \( var(v_i) = \sigma_u^2 Z_i^2 \). This is a special case of heteroscedasticity, where the variance of the error term is proportional to some variable \( Z_i^2 \) (equivalently, the standard deviation of the error term is proportional to \( Z_i \)). Now, transform the original variables and the error term by dividing them by \( Z_i \):

\[ \frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_{2,i}}{Z_i} + \ldots + \beta_k \frac{X_{k,i}}{Z_i} + \frac{v_i}{Z_i} \]  

(8.3)

Notice that the variance of the transformed error term is homoscedastic:

\[ var\left(\frac{v_i}{Z_i}\right) = \frac{1}{Z_i^2} var(v_i) = \sigma_u^2 \]

Thus, estimating the model in (8.3) using OLS, gives not only unbiased estimators, but also efficient. The estimated standard errors of the coefficients are also unbiased, and all the tests are valid. In other words, we fixed all the problems of heteroscedasticity, but remember that we were able to do so only if we knew the particular form of the heteroscedasticity.

In practice, one way to estimate the model in (8.3), is to manually generating the transformed variables:

\[ Y_{i}^* = \frac{Y_i}{Z_i}, \quad X_{1,i}^* = \frac{1}{Z_i}, \quad X_{2,i}^* = \frac{X_{2,i}}{Z_i}, \ldots, \quad X_{k,i}^* = \frac{X_{k,i}}{Z_i} \]

and fitting the model:

\[ \hat{Y}_{i}^* = b_1 X_{1,i}^* + b_2 X_{2,i}^* + \ldots + b_k X_{k,i}^* \]  

(8.4)

Notice that this equation does not have an intercept term, unless \( Z \) is one of the regressors in the original model. Also, you must interpret the estimated coefficients based on the
original model only (8.2) and not based on the transformed model (8.4). In other words, in order to interpret the estimated coefficients $b_1, b_2, ..., b_k$, you must use the fitted equation corresponding to the original model (8.2):

$$\hat{Y}_i = b_1 + b_2 X_{2,i} + ... + b_k X_{k,i}$$

A more convenient way of estimating the model in (8.3) in Stata, is using **analytic weights**, for example, suppose that the variance in the earnings regression model is proportional to education squared. Then, one can correct for heteroscedasticity in Stata by:

```
regress EARNINGS S EXP [aw=1/S^2]
```

Analytic weights, are weights that are inversely proportional to the variance of an observation. Stata automatically creates the transformed variable, by multiplying the original model by the root of the analytic weights. One has to be careful to avoid dividing by zeros, so the variable $Z_i$ must be strictly positive for all $i$.

Similarly, in R adding the option `weights=1/X^2` in the `lm` command, will estimate the transformed model, and present the coefficients in the original model. For example:

```
model = lm(EARNINGS ~S + EXP, data=wage, weights=1/S^2)
```

Weighted least Squares are important in practice, not only as a way to fix the heteroscedasticity problem. Many official data sources (e.g. [https://www.ipums.org/](https://www.ipums.org/)) present data in weighted form, i.e. each observation represents a number of actual observations in the real world. In order to obtain representative statistics, researchers must use the provided weights.
Chapter 9

Discrete Choice Models

Some of the most important choices we make in our lives are "either-or". For example, after graduating with a B.A. in economics, you decide whether to look for a job, or continue to a graduate or professional school. Other examples include getting married or not, buying a house or renting, voting for proposition #N or not. Such choices are modeled with a binary (dummy) dependent variable. More formally, in such models the observed dependent variable (outcome) that we want to analyze takes on two values, as in the next examples:

[Marriage choice] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ marries} \\ 0 & \text{individual } i \text{ stays single} \end{cases} \)

[Voting for a proposition] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ votes "YES"} \\ 0 & \text{individual } i \text{ votes "NO"} \end{cases} \)

[Home ownership] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ owns a house} \\ 0 & \text{individual } i \text{ rents} \end{cases} \)

[Transportation mode] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ drives to work/school} \\ 0 & \text{individual } i \text{ uses public transportation} \end{cases} \)

[Smoking] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ smokes} \\ 0 & \text{individual } i \text{ does not smoke} \end{cases} \)

[Labor force participation] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ is in the labor force} \\ 0 & \text{individual } i \text{ is not in the labor force} \end{cases} \)

[Union membership] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ is a union member} \\ 0 & \text{individual } i \text{ is not a union member} \end{cases} \)

[Migration] : \( Y_i = \begin{cases} 1 & \text{individual } i \text{ decides to migrate} \\ 0 & \text{individual } i \text{ does not migrate} \end{cases} \)

In some cases, the word "choice" is replaced by response or outcome when the result is not literally a cognitive choice. For example, a firm going out of business or staying in business

---

1 Discrete choice models in general include choices between several alternatives, e.g. high school diploma, college degree, advanced degree contains 3 educational choices. In these notes we restrict attention to binary choice models - containing two alternatives only.
is a binary response. Examples of binary responses:

[Loan application] : \( Y_i = \begin{cases} 1 & \text{loan for individual } i \text{ approved} \\ 0 & \text{loan for individual } i \text{ denied} \end{cases} \)

[Bankruptcy] : \( Y_i = \begin{cases} 1 & \text{firm } i \text{ goes bankrupt} \\ 0 & \text{firm } i \text{ stays in business} \end{cases} \)

[Recession] : \( Y_{it} = \begin{cases} 1 & \text{economy } i \text{ is in recession at time } t \\ 0 & \text{economy } i \text{ is not in recession at time } t \end{cases} \)

In analyzing binary outcome variables, our goal is usually to predict the probability of a certain outcome, such as probability of individual voting for a candidate, based on that individual’s characteristics (income, race, gender,...). Suppose that the theory tells us that the response \( Y_i \) depends on some characteristics \( X_2, ..., X_k \). Our first instinct is to use the linear probability model

\[
Y_i = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + u_i \tag{9.1}
\]

How would one interpret the coefficients in the linear probability model, in (9.1)? Remember that the data on the left hand side is a sequence of 0’s and 1’s, so the values of \( Y \) do not change gradually. The standard interpretation of \( \beta_k \) is the effect of a one unit increase in \( X_k \) on the probability that \( Y_i = 1 \). In other words, we pretend that the model predicts \( P(Y_i = 1) \).

One problem with this interpretation is that the right hand side cannot be restricted to the \([0, 1]\) interval, and the predictions of this model can give negative probabilities or probabilities of more than 1. Therefore, although the linear probability model is easy to estimate using OLS, the inconsistency of the left hand side and the right hand side are the reason why it is no longer used in modern research.

One way out of this problem is to construct a model that is explicitly designed to predict probabilities. Such models involve the so called latent variable (unobserved variable) \( Y_i^* \), which can attain all values in \([-\infty, \infty]\), and represents the tendency or propensity of a positive response of \( Y_i \). For example, \( Y_i \) could be the indicator of a bank failure (which is observed), and \( Y_i^* \) is the tendency of a bank to fail, which is unobserved and depends on all kinds of observed characteristics, such as delinquency rate, leverage (risk), and fraction of real estate loans. Thus, we can postulate that

\[
Y_i^* = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + u_i
\]

So the tendency of the bank to fail depends on some observed characteristics \( X \)-s, parameters \( \beta \)-s and some unobserved characteristics represented by the error term \( u_i \). We cannot estimate this model with \( Y_i^* \) as dependent variable, because \( Y_i^* \) is not observed, and we only observe the binary outcome \( Y_i \). The solution is to postulate that we observe \( Y_i = 1 \) if \( Y_i^* \geq 0 \) and \( Y_i = 0 \) if \( Y_i^* < 0 \). The threshold of 0 is not restrictive because the constant term will adjust to any threshold we would choose. Thus, the probability model we will try to estimate is:

\[
P(Y_i = 1) = P(Y_i^* \geq 0) = P(\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + u_i \geq 0)
\]
9.1. The Probit

In other words, we will try to predict the probability of \( Y_i = 1 \) (say bank failure), which is the probability that the tendency (of a bank to fail) \( Y_i^* \) is "high enough" or bigger than some threshold.

### 9.1 The Probit

A common assumption about the error term is that for all \( i \),

\[ u_i \sim N(0, 1) \]

With this assumption, we have:

\[
P(Y_i = 1) = P(Y_i^* \geq 0) = P(\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + u_i \geq 0) = P(u_i \geq - (\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i})) = P(u_i \leq \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i})
\]

The last step follows from the fact that the distribution of \( u_i \) is symmetric around zero.

Thus, we can write:

\[
P(Y_i = 1) = \Phi(\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i})
\]

where \( \Phi \) is the cumulative distribution function of the standard normal distribution, defined as:

\[
\Phi(x) \equiv P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z^2}{2} \right) dz
\]

The model in (9.2) is called the **Probit model**. The integrand is the standard normal p.d.f.:

\[
\phi(z) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z^2}{2} \right)
\]

We postpone the discussion about how to estimate the unknown coefficients \( \beta_1, \beta_2, \ldots, \beta_k \) in the probit model, and suppose that we have obtained the fitted equation:

\[
P(Y_i = 1) = \Phi(b_1 + b_2 X_{2,i} + \ldots + b_k X_{k,i})
\]

Notice that the right hand side is restricted to be in \([0, 1]\), since \( \Phi \) is cumulative distribution function which by definition returns values that are probabilities.

The purpose of the model is to estimate the effects of the regressors on the probability of \( Y_i = 1 \). Therefore, with regular (non-dummy) variables, we are interested in:

\[
\frac{\partial P(Y_i = 1)}{\partial X_{k,i}} = \phi(b_1 + b_2 X_{2,i} + \ldots + b_k X_{k,i}) \cdot b_k
\]

which is the effect of a one unit increase in \( X_{k,i} \) on the probability of \( Y_i = 1 \). This is called the **marginal effect** of regressor \( k \), and notice that it depends on the values of all the regressors. Thus, the marginal effect of say one year of schooling on the probability of home ownership,
will depend on the level of schooling, as well as on the levels of other regressors such as income, family size, etc. In Stata marginal effects are computed using `mfx compute`, after estimating the probit model. The marginal effects are, by default, evaluated at the means of all the regressors. Instead of the default, the marginal effects can be computed at any point using the `mfx at()` command. For regressor $X_k$ that is a dummy variable, the marginal effect is the change in predicted probability as $X_k$ changes from 0 to 1, i.e. the marginal effect is $P(Y_i = 1|X_k = 1) - P(Y_i = 1|X_k = 0)$. In R one needs to install the `mfx` package and use `probitmfx()` in the same way we use `lm()` for OLS: `probitmfx(Y~X_2+...+X_k, data=dataname)`.

### 9.2 Logit Model

An alternative to the probit model, assumes that the error terms have some other distribution - the so called extreme value distribution. This leads to a very similar model to the one in (9.2):

$$P(Y_i = 1) = \Lambda(\beta_1 + \beta_2 X_{2,i} + ... + \beta_k X_{k,i})$$

(9.3)

where $\Lambda$ is the cumulative distribution function of the logistic distribution, defined as:

$$\Lambda(x) = \frac{\exp(x)}{1 + \exp(x)}$$

with p.d.f.:

$$\lambda(x) = \frac{\exp(x)}{[1 + \exp(x)]^2}$$

One might wonder why do we need another model, which is very similar to probit? The main advantage of the logit is that the estimation is easier in terms of the computation required. Notice that $\Lambda(\beta_1 + \beta_2 X_{2,i} + ... + \beta_k X_{k,i})$ is given explicitly, while in order to compute $\Phi(\beta_1 + \beta_2 X_{2,i} + ... + \beta_k X_{k,i})$, one needs to solve an integral. Stata and R have very efficient algorithms, so this advantage of logit over probit will matter only if you use huge samples. For most practical purposes, you won’t care which model to use, and researchers often report both results just to show that they are very similar.

As with probit, after estimation we obtain the fitted equation:

$$P(Y_i = 1) = \Lambda(b_1 + b_2 X_{2,i} + ... + b_k X_{k,i})$$

The marginal effects are similar to the probit case:

$$\frac{\partial P(Y_i = 1)}{\partial X_{k,i}} = \lambda(b_1 + b_2 X_{2,i} + ... + b_k X_{k,i}) \cdot b_k$$

which is the effect of a one unit increase in $X_{k,i}$ on the probability of $Y_i = 1$. In Stata, marginal effects are computed using the command `mfx compute` after estimating the logit model. In R one needs to install the `mfx` package and use `logittmfx()` in the same way we use `lm()` for OLS: `logittmfx(Y~X_2+...+X_k, data=dataname)`.
9.3 Estimation of the Probit and Logit Models

The probit and logit models can be written as

\[ P(Y_i = 1) = F(\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i}) \]

where \( F = \Phi \) in the probit model and \( F = \Lambda \) in the logit model. The unknown parameters cannot be estimated using OLS, and usually estimated using the maximum likelihood approach. Remember that we do not have data on the probability that \( Y_i = 1 \), but instead the dependent variable is binary (a bunch of zeros and ones). The model however defines the probability that for given observation \( i \) we have \( Y_i = 1 \). The probability (or the likelihood) that our sample was generated from our binary choice model is therefore (remembering that observations are independent):

\[ L = \prod_{Y_i=1} F(b_1 + b_2 X_{2,i} + \ldots + b_k X_{k,i}) \times \prod_{Y_i=0} (1 - F(b_1 + b_2 X_{2,i} + \ldots + b_k X_{k,i})) \]

The maximum likelihood estimation technique tries to find parameters \( b_1, b_2, \ldots, b_k \) that maximize the likelihood of obtaining the observed sample on the dependent variable from our model. Maximizing the likelihood is mathematically the same as maximizing the log-likelihood, \( l = \ln L \), but the latter is computationally more stable. Thus, in practice, the function being maximized is:

\[ l = \sum_{Y_i=1} \ln F(b_1 + b_2 X_{2,i} + \ldots + b_k X_{k,i}) + \sum_{Y_i=0} \ln (1 - F(b_1 + b_2 X_{2,i} + \ldots + b_k X_{k,i})) \]

It is understood that the first summation is over the observations with \( Y_i = 1 \) and the second summation is over the observations with \( Y_i = 0 \).

In Stata, one uses the command `probit` or `logit` just the way you would use `regress` to estimate a regression model. After these commands, you need to specify the dependent variable, which must be binary (dummy), followed by a list of regressors. For example, suppose that \( Y \) is the dependent variable, and \( X_2, \ldots, X_k \) are regressors. The Stata commands for logit and probit estimates are:

- `probit Y X_2 X_k`
- `logit Y X_2 X_k`

In R, the same is achieved with:

- `model_Probit <- glm(Y~X_2+...+X_k, data=dataname, family=binomial(link="probit"))`
- `model_Logit <- glm(Y~X_2+...+X_k, data=dataname, family=binomial(link="logit"))`

As usual, to present the output summary in R, we need the `summary(model_Probit)` or `summary(model_Logit)` commands.

The marginal effects in Stata:

- `probit Y X_2 X_k, mfx at()`
- `logit Y X_2 X_k, mfx at()`

Marginal effects in R:

- `probitmfx(Y~X_2+...+X_k, data=dataname)`
- `logitmfx(Y~X_2+...+X_k, data=dataname)`
Chapter 10

Time Series

All the data that you will ever encounter in econometrics or statistics can be classified into 3 types: (i) **cross-sectional data** refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time, (ii) **time-series data** follows one subject’s changes over the course of time, and (iii) **panel data** combines both cross-section and time series and looks at multiple subjects and how they change over the course of time.

An example of cross-sectional data would be Census, which contains income, schooling, age, sex, and many other characteristics of many individuals in a given year (say 2000). The next sheet displays a part of a cross-sectional data set. Each row corresponds to a particular individual, indexed by id number (1st column).

<table>
<thead>
<tr>
<th>id</th>
<th>sex</th>
<th>race</th>
<th>age</th>
<th>wage</th>
<th>schooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>white</td>
<td>45</td>
<td>53.41</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>other</td>
<td>40</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>white</td>
<td>38</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>white</td>
<td>43</td>
<td>29.5</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>white</td>
<td>38</td>
<td>32.05</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>black</td>
<td>39</td>
<td>14.73</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>white</td>
<td>40</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>white</td>
<td>37</td>
<td>11.25</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>white</td>
<td>39</td>
<td>18.01</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>white</td>
<td>42</td>
<td>11.78</td>
<td>12</td>
</tr>
</tbody>
</table>

So far in this course we focused on cross-sectional data, since the econometric theory for its analysis is the most basic.

An example of a time series would be NIPA (National Income and Product Accounts) for the U.S. So the subject is just one country, the U.S., and the sample contains data on GDP, Consumption, Investment, Government Spending, etc., for different years. The next sheet displays a part of a time-series data on the U.S. Notice that each row is a year in this case, but it could be a quarter or a month.
Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. This chapter presents a brief discussion of specific econometric problems arising in time-series models.

An example of a panel data would be World Development Indicators (WDI) [http://databank.worldbank.org/data/home.aspx](http://databank.worldbank.org/data/home.aspx), which contains many variables on countries since 1960. Another example is Panel Study of Income Dynamics (PSID) [http://psidonline.isr.umich.edu](http://psidonline.isr.umich.edu), which has information on more than 70,000 individuals in the U.S. spanning as many as 4 decades of their lives. It is important to understand that combining several census data, say for 1990, 2000 and 2010, does not constitute a panel. The census does not keep track of the same individuals over time, and an individual who had id number 9723 in 1990, might have a totally different id number in 2000 census and yet another number in 2010 census. A panel must track the same individuals over several time periods. For example, Rita Morgan in 1990 panel had age = 15, income = 0, schooling = 9, and the same Rita in the 2000 panel had age = 25, income = $35,000, and schooling = 16. Since panel data analysis contains methods from cross-section and time series, and also methods specific to panel data analysis, this topic is not covered in a one-semester intro to econometrics course. The next table shows an example of panel data from WDI.

<table>
<thead>
<tr>
<th>Year</th>
<th>GDP</th>
<th>CONS</th>
<th>INV</th>
<th>GOV</th>
<th>EX</th>
<th>IM</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>10285</td>
<td>6792.4</td>
<td>2033.8</td>
<td>1834.4</td>
<td>1096.8</td>
<td>1472.6</td>
<td>...</td>
</tr>
<tr>
<td>2001</td>
<td>10622</td>
<td>7103.1</td>
<td>1928.6</td>
<td>1958.8</td>
<td>1026.7</td>
<td>1395.4</td>
<td>...</td>
</tr>
<tr>
<td>2002</td>
<td>10978</td>
<td>7384.1</td>
<td>1925</td>
<td>2094.9</td>
<td>1002.5</td>
<td>1429</td>
<td>...</td>
</tr>
<tr>
<td>2003</td>
<td>11511</td>
<td>7765.5</td>
<td>2027.9</td>
<td>2220.8</td>
<td>1040.3</td>
<td>1543.9</td>
<td>...</td>
</tr>
<tr>
<td>2004</td>
<td>12275</td>
<td>8260</td>
<td>2276.7</td>
<td>2357.4</td>
<td>1181.5</td>
<td>1800.7</td>
<td>...</td>
</tr>
<tr>
<td>2005</td>
<td>13094</td>
<td>8794.1</td>
<td>2527.1</td>
<td>2493.7</td>
<td>1308.9</td>
<td>2030.1</td>
<td>...</td>
</tr>
<tr>
<td>2006</td>
<td>13856</td>
<td>9304</td>
<td>2680.6</td>
<td>2642.2</td>
<td>1476.3</td>
<td>2247.3</td>
<td>...</td>
</tr>
<tr>
<td>2007</td>
<td>14478</td>
<td>9750.5</td>
<td>2643.7</td>
<td>2801.9</td>
<td>1664.6</td>
<td>2383.2</td>
<td>...</td>
</tr>
<tr>
<td>2008</td>
<td>14719</td>
<td>10014</td>
<td>2424.8</td>
<td>3003.2</td>
<td>1841.9</td>
<td>2565</td>
<td>...</td>
</tr>
<tr>
<td>2009</td>
<td>14419</td>
<td>9847</td>
<td>1878.1</td>
<td>3089.1</td>
<td>1587.7</td>
<td>1983.2</td>
<td>...</td>
</tr>
<tr>
<td>2010</td>
<td>14964</td>
<td>10202</td>
<td>2100.8</td>
<td>3174</td>
<td>1852.3</td>
<td>2365</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>15518</td>
<td>10689</td>
<td>2239.9</td>
<td>3168.7</td>
<td>2106.4</td>
<td>2686.4</td>
<td>...</td>
</tr>
<tr>
<td>2012</td>
<td>16155</td>
<td>11051</td>
<td>2511.7</td>
<td>3158.6</td>
<td>2198.2</td>
<td>2763.8</td>
<td>...</td>
</tr>
<tr>
<td>2013</td>
<td>16663</td>
<td>11392</td>
<td>2665</td>
<td>3114.2</td>
<td>2263.3</td>
<td>2771.7</td>
<td>...</td>
</tr>
<tr>
<td>2014</td>
<td>17348</td>
<td>11866</td>
<td>2860</td>
<td>3152.1</td>
<td>2341.9</td>
<td>2871.9</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 10.1 Regression with Time Series Data

The linear model we used so far in analysis of cross-sectional data is the multiple regression models:

\[ Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_k X_{k,i} + u_i \]

The index \(i\) refers to subject, say \(i = 25\) is a particular individual with income \(Y_{25}\), schooling level \(X_{2,25}\), experience \(X_{3,25}\), and disturbance term (capturing all other influences on income) \(u_{25}\). To distinguish the regression for cross-sectional data from regression using time series data, the index becomes \(t\) instead of \(i\):

\[ Y_t = \beta_1 + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \ldots + \beta_k X_{k,t} + u_t \]
For example, suppose that we have data on average consumption, income and wealth of individuals in Japan, for the years $t = 1960, \ldots, 2000$. We might want to estimate the model:

$$C_t = \beta_1 + \beta_2 I_t + \beta_3 W_t + u_t, \quad t = 1960, \ldots, 2000$$

If the only difference between time series and cross-section was the index $t$ instead of $i$, then we would not need to study time series analysis as a separate subject. There are however several important differences between estimating the above consumption function and estimating the cross-sectional relationship

$$C_i = \beta_1 + \beta_2 I_i + \beta_3 W_i + u_i$$

where $i = 1, \ldots, 40$ are all individuals from the same year (say 2000). We will briefly discuss two issues that arise primarily in time series analysis: (i) **autocorrelation**, and (ii) **nonstationarity**.

### 10.1.1 Autocorrelation (serial correlation)

Assumption A.5 in section 3.1 states that the disturbance (error) terms for different observations are independent of each other. This means $u_i$ and $u_j$ are statistically independent random variables for all $i \neq j$, and since independence implies no correlation, this assumption implied also that $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$. The analog of A.5 for time series would be that $u_t$ and $u_{t'}$ are statistically independent, and therefore $\text{cov}(u_t, u_{t'}) = 0$ for all $t \neq t'$. When assumption A.5 is violated, we say that the error terms are subject to **autocorrelation**, or **serial correlation**.

Remember that the error term captures all the effects on the dependent variable, other than the effects of the included regressors. In cross sectional data it is reasonable to assume that A.5 holds, since we sample individuals independently and unobserved factors affecting my earnings (motivation, ambition,...) should not be correlated with those of another individual who is also included in the sample. In time series data however, the error terms represent the unobserved (or excluded) effects on the same object over different time periods. For example, factors affecting inflation rate, unemployment rate, interest rate, GDP, etc. in 2010, are likely to still be present in 2011. Therefore, it is often assumed that **autocorrelation** (or **serial correlation**) is a potential problem only in time series data (not in cross-sectional data).

The consequences of serial correlation are the same as those of heteroscedasticity: (i) the OLS estimators are still unbiased and consistent, but they are inefficient (do not have the smallest possible variance in the class of unbiased linear estimators), and (ii) the estimated standard errors of the OLS estimators are biased, which makes statistical hypothesis tests invalid. As in the case of heteroscedasticity, if we know exactly the form of the autocorrelation, we can perform some transformation of the original model and eliminate the problem.

The detection of serial correlation is based on residuals, just as heteroscedasticity. For example in Stata, consider quarterly data on inflation and unemployment:

```
use http://online.sfsu.edu/mbar/ECON312/phillips.dta, clear 
regress Inf Unemp //Estimating linear Phillips curve
predict resid, residual //Generating the residuals
```
The above graph plots the estimated autocorrelation function based on the residuals, i.e. the estimated $\rho_k = \text{corr}(u_t, u_{t+k})$ for $k = 1, \ldots, 40$. Let $\hat{\rho}_k$ be an estimate of $\rho_k$, so the vertical bars are values of $\hat{\rho}_k$ for $k = 1, \ldots, 40$. The grey area shows a 95% confidence bounds constructed for the null hypothesis (with two-sided alternative):

$$H_0 : \rho_k = 0$$

$$H_1 : \rho_k \neq 0$$

If $\hat{\rho}_k$ is outside the 95% confidence bounds, then we reject the null hypothesis $\rho_k = 0$ for that $k$.

To create similar autocorrelation plot (correlogram) in R, we execute the following steps:

```r
model1 <- lm(Inflation ~ Unemp, data=phillips)
resid <- resid(model1)
acf(resid, ci=0.95, ci.type = "ma", lag.max = 40, lwd=2, main="Correlogram of resid")
```
Notice that unlike the Stata correlogram, which starts at \( k = 1 \), R starts with \( k = 0 \), and of course the first bar is \( \rho_0 = 1 \) is redundant.

Other tests for serial correlation, that can be performed in Stata:

- `estat bgodfrey, lags(6)` // Breusch-Godfrey (1988) test
- `estat durbinalt, lags(6)` // Durbin (1970)
- `wntestq resid` // white noise test

The null and alternative hypothesis in all the above tests is:

\[
H_0 : \text{no serial correlation} \\
H_1 : \text{there is serial correlation}
\]

The Breuch-Godfrey and Durbin tests allow specifying the maximum number of lags (\( k \)) for which to test for serial correlation. All the tests are based on chi-square statistic, and report the p-value of the test as: Prob > chi2. In all the cases we reject the null hypothesis (\( H_0 : \text{no serial correlation} \)) at significance level \( \alpha \), if Prob > chi2 is smaller than \( \alpha \).

As with heteroscedasticity, it is possible to eliminate the problem of serial correlation if we know its exact form. The biased standard errors can be corrected in Stata or R in the same way as with heteroscedasticity.

### 10.1.2 Trends

Many economic time series exhibit a trend, due to some unobserved factors or omitted factors. For example, consumption per capita, investment per capita, income per capita or taxes per capita in a country, could all grow because of some underlying technological progress which makes everybody richer on average. Theory of trends is an enormous field in engineering, called filtering or signal processing. In these notes we only consider linear trends. If a time series grows at constant rate, then it is easy to show that its natural
10.1. REGRESSION WITH TIME SERIES DATA

logarithm has a linear trend (is a linear function of time). For example, the Gross National Product in the U.S.A. is plotted in the next graph.

![Real GNP per capita (USA)](image)

The natural logarithm of the above series looks like fluctuations around a linear trend (linear function of time).

![ln(Real GNP per capita)](image)

When using regression analysis, it is very important to be aware that the dependent variable \(Y_t\) and the regressors, might exhibit trends. If unobserved factors (e.g. technological advances) are responsible for the trend in the dependent variable, then ignoring these factors can cause an omitted variable bias (see section 7.1 for review of omitted variable bias). The simplest solution is to include the variable "time" as one of the regressors.

The following example uses time series on investment in housing and housing prices. The idea is to estimate the demand for housing investment as a function of housing prices. The dependent variable \(\text{linvpc}\) is \(\ln\) of housing investment per capita, and the regressor \(\text{lprice}\) is the \(\ln\) of housing price index.

use [http://online.sfsu.edu/mbar/ECON312/HSEINV.dta](), clear
regress linvpc lprice //Ignoring time trend

The resulting Stata output is:

```
Source | SS       df       MS                  Number of obs =  42
-------------+----------------------------------------------------
Model | .254364572  1  .254364572             F( 1,   40) = 10.53
Residual | .966255373  40  .024156384            Prob > F =    0.0024
          +----------------------------------------------------
Total | 1.22061994   41  .029771218            R-squared =    0.2084
          +----------------------------------------------------

|        Coef.     Std. Err.      t    P>|t|     95% Conf. Interval
-------------+----------------------------------------------------------------
  lprice | 1.240944    .3824192     3.24  0.002     .460455    2.013642
_cons   |  -.5502345   .0430266    -12.79  0.000    -.6371945   -.4632745
```

Notice that the estimated price elasticity of demand for housing investment is positive, suggesting that a 1% increase in housing prices increases the demand for housing investment by 1.24%. This result does not make sense, as we expect to have a negative relationship between price and quantity demanded.

A closer look at both variables, suggests that they might contain a (linear ?) time trend:

![Graph showing time trends in linvpc and lprice](image)

We can see that the positive relationship, previously estimated, does not represent a causal link from prices to demand, but instead both prices and demand are driven by the growth in income. The previous regression, which ignores the trends, is often called *spurious regression* (meaning false or fake), because it does not measure the causal effect of prices on demand.

Now we include a linear time trend in the above regression:

```
regress linvpc lprice year //Including time trend
```

```
The resulting Stata output is:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>0.415945135</td>
<td>2</td>
<td>0.207872568</td>
<td>F(2,39) = 10.08</td>
</tr>
<tr>
<td>Residual</td>
<td>0.804674809</td>
<td>39</td>
<td>0.020632687</td>
<td>Prob &gt; F = 0.0003</td>
</tr>
<tr>
<td>Total</td>
<td>1.22061994</td>
<td>41</td>
<td>0.029771218</td>
<td>R-squared = 0.3408</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.3070</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 0.14364</td>
</tr>
</tbody>
</table>

Notice that after accounting for the linear trend, the regression results completely changed. The price elasticity of demand is now negative, as economic theory suggests.

**Technical note**

Consider the simple regression model, with both the dependent variable and regressor exhibiting constant growth rate, and affected by random error terms. In particular, \( Y_t \) is growing at constant rate of \( g_y \) and \( X_t \) is growing at another rate \( g_x \):

\[
Y_t = Y_0 (1 + g_y)^t e^{z_{1t}} \\
X_t = X_0 (1 + g_x)^t e^{z_{2t}}
\]

This implies that the natural logarithms of both variables will fluctuate around linear functions of time.

\[
\ln(Y_t) = \ln Y_0 + t \ln(1 + g_y) + z_{1t} \\
\ln(X_t) = \ln X_0 + t \ln(1 + g_x) + z_{2t}
\]

Suppose that the relationship between the two variables is such that the deviations from linear trend of \( \ln(Y_t) \) are affected by the deviations from linear trend of \( \ln(X_t) \). In other words, we assume that the trends of both time series are affected by some external (exogenous) force, and we want to study the impact of \( z_{2t} \) on \( z_{1t} \), of the form

\[
z_{1t} = \beta_1 + \beta_2 z_{2t} + u_t
\]

Substituting the detrended variables \( z_{2t} \) and \( z_{1t} \):

\[
\ln Y_t - \ln Y_0 - t \ln(1 + g_y) = \beta_1 + \beta_2 [\ln X_t - \ln X_0 - t \ln (1 + g_x)] + u_t
\]

Rearranging the above:

\[
\ln Y_t = \left(\frac{\beta_1 + \ln Y_0 - \beta_2 \ln X_0}{\beta_1}\right) + \beta_2 \ln X_t + \frac{[\ln (1 + g_y) - \beta_2 \ln (1 + g_x)]}{\beta_3} t + u_t
\]

\[
\ln Y_t = \tilde{\beta}_1 + \beta_2 \ln X_t + \beta_3 t + u_t
\]
First, the above shows that a model without $t$ as one of the regressors \( \ln Y_t = \tilde{\beta}_1 + \beta_2 \ln X_t + \tilde{u}_t \) will suffer from omitted variable bias. The estimate of $\beta_2$ will be biased upward if $X_t$ is positively correlated with $t$ (growing over time) and downward if $X_t$ is negatively correlated with $t$ (decreasing over time). However, notice that despite the fact that $Y_t$ and $X_t$ exhibit different growth rates along their trends, including time as regressor allows estimating, without bias, the true effect of $X_t$ on $Y_t$ - i.e. $\beta_2$. In conclusion, when you suspect that the dependent variable and/or some of the regressors contain a time trend, the easiest remedy is to include time as a regressor.

Other informal remedies include (i) detrending, and (ii) normalizing. Detrending means that you remove the time trend from the variables, and use detrended variables only. For example, suppose the original variable $x_t$ contains a time trend $x_t$. We create a new variable $x_t^* = x_t - \bar{x}_t$, which is called detrended $x_t$. All variables that contain trends are transformed in this way, and replaced by detrended variables in regression analysis. Another approach, similar to detrending is normalizing. Suppose a bunch of macroeconomic variables exhibit the same trend as GDP. Examples include consumption, investment, saving, government spending, tax revenues, deficit, national debt, etc. All such variables, can be divided by GDP, and therefore called normalized. For example, if $C_t$ is aggregate consumption, then $c_t = C_t / GDP_t$ is the ratio of consumption to GDP or normalized consumption. If your analysis includes macroeconomic variables such as the ones mentioned above, and which are likely to exhibit growth trend similar to that of GDP, normalizing these variables will eliminate the trend.

### 10.2 Panel Data Analysis

**Fixed effects model**

\[
Y_{it} = \beta_{1i} + \beta_2 X_{2, it} + \beta_3 X_{3, it} + \cdots + \beta_k X_{k, it} + u_{it}
\]

**Random effects model**

\[
Y_{it} = (\beta_1 + \varepsilon_i) + \beta_2 X_{2, it} + \beta_3 X_{3, it} + \cdots + \beta_k X_{k, it} + u_{it}
\]

\[
= \beta_1 + \beta_2 X_{2, it} + \beta_3 X_{3, it} + \cdots + \beta_k X_{k, it} + (u_{it} + \varepsilon_i)
\]