Econometrics Lecture Notes

Max Heinze

March 24, 2026

Contents

Disclaimer

This document is mostly a more elaborated version of my own notes that I use to teach. Since they are based on the lecture slides, and the lecture slides are in part based on various textbooks, this document will have similarities to these textbooks. This is explicitly not a textbook, and explanations will be a mix of my own explanations and ones taken from the respective textbooks, although no content is copied verbatim, or even directly rephrased. How closely this document follows other work will vary, and in principle, no part is intended to present as original work. The primary textbooks used are those by Wooldridge (2020) for Econometrics I, Cunningham (2021) for Econometrics II and Stock and Watson (2019) for Applied Econometrics.

This version of the document was compiled on March 24, 2026. The current version is available on https://maxheinze.eu/assets/econometrics_script.pdf. Note that not every version of this document contains material on all three courses. All versions are preliminary, and I ask you to please tell me about any errors you find by sending an email to mheinze@wu.ac.at.

Preliminaries

0.1 Math Basics

0.1.1 Sums

Summation means adding together a set of values, called the *summands*. The summands can be numbers, functions, vectors or matrices. Summation is very simple and intuitive, and while often feared, the summation operator Σ can be used to greatly simplify repetitive sequences. See for example:

$$1+2+3+4+...+100=\sum_{n=1}^{100} n \ (=5050) \tag{0.1}$$

What did we do here? We added every number from 1 to 100 together. This is an enormously long operation to write down, so we use the summation operator to simplify things. The letter n you see below the Σ is called the *index of summation*. From the *lower bound of summation*, in our example defined as $n = 1$, you can imagine the index counting up by one until it reaches the *upper bound of summation*, 100. If the term on the right contains the summation index, it is correspondingly altered for each step. All of the steps are then added together. In our example, it is therefore easy to see that as the term to the right of Σ becomes first 1, then 2, then 3, ..., then 100, the notation using the summation operator is equal to the left hand side of the equation.

Now knowing the basic concept, we can easily understand why the following rules apply. For every constant c,

$$\sum_{i=1}^{N} c = c \cdot N \tag{0.2}$$

For every constant c,

$$\sum_{i=1}^{N} cx_i = c \cdot \sum_{i=1}^{N} x_i \tag{0.3}$$

When $N = M$,

$$\sum_{i=1}^{N} x_i + \sum_{j=1}^{M} y_j = \sum_{i=1}^{N} (x_i + y_i) \tag{0.4}$$

We will encounter the summation operator quite often when we discuss econometric concepts. It is therefore important to know how some expressions that we know well look when written using this notation. For example, if we have a list of data like the following: $\{x_1, x_2, x_3, x_4\}$, we know that the mean will be:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4}{4} \tag{0.5}$$

We can easily write this term using the summation operator:

$$\frac{1}{4} \sum_{i=1}^{4} x_i \tag{0.6}$$

Granted, this does not help much in terms of clarity and intuition, but this is largely due to the number of elements in our list being 4. When we deal with hundreds or thousands or even an unspecified number of observations, as we frequently do in econometrics, this changes. We therefore say, in more general terms, that the mean equals

$$\frac{1}{N}\sum_{i=1}^{N} x_i \tag{0.7}$$

Let us finally see what we do here in terms of econometrics. Assume we want to talk about the *sum of squared residuals*, a *residual* being the difference between an observed value y and the corresponding predicted value $\hat{y}$. Let us assume that we have N individuals, and for each of them, we have a predicted and an actual value. So the sum of squared residuals is

$$\text{SSR} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{0.8}$$

0.1.2 Derivatives

The basic concept of what differential calculus is should be known to all of us. This section is therefore deliberately kept brief. To start, suppose that we have a function that looks like the one in Figure 0.1. We can imagine the *first derivative* as being the *slope* of the function graph at every point x. Of course, this is more intuition than definition, and does not work well for more than two (or three) dimensions, but our plan was to keep this section brief. So for now, we will continue by adding an orange curve, the first derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$, to the graph, shown in Figure 0.2.

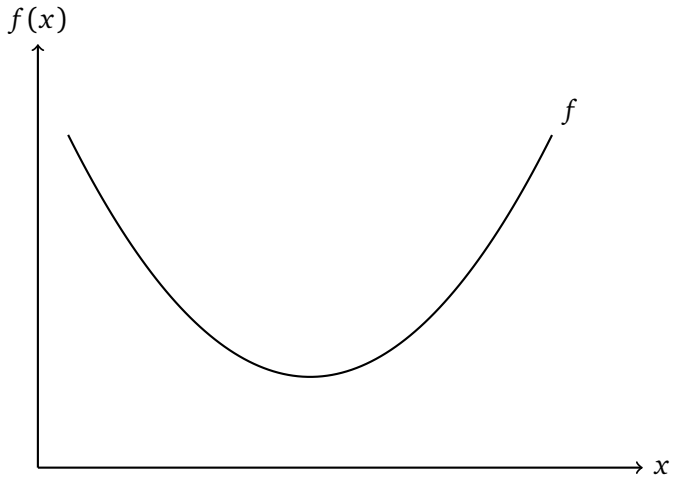


Figure 0.1: A function.

Take time to revisit the graph until you see that the value of the orange line always corresponds with the slope of the black line. First, the black line falls. This means that the orange line is below zero. Then, the black line hits a minimum, a point where a tangent would be horizontal, and at that point the orange line crosses zero. Later, the black line rises ever faster and the orange line grows away from zero, linearly. By the way: The first derivative of the first derivative is called the second derivative, and we can interpret it as the change of the slope. We can see that in our example, the second derivative must be constant and positive.
It makes sense that the derivative equals zero at most minima and maxima of the function (there are exceptions, such as minima and maxima that constitute endpoints of a function).

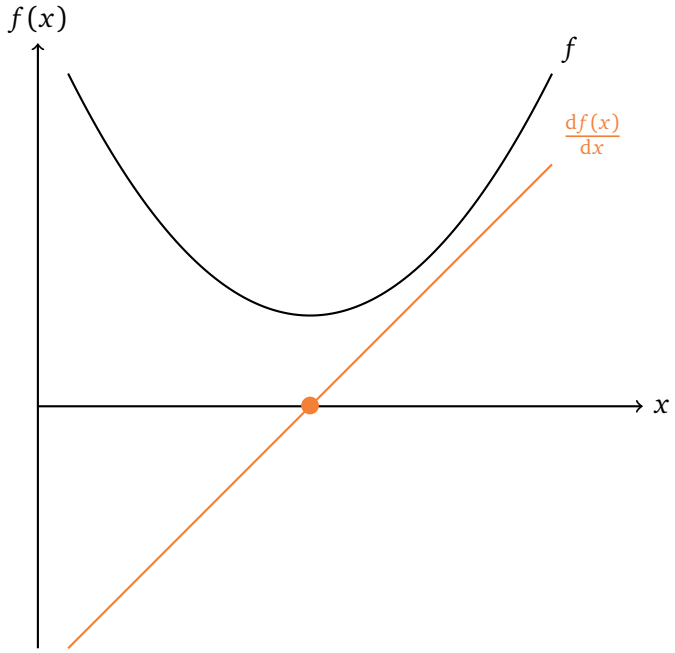


Figure 0.2: A function, and its first derivative.

We will use this fact later to *find* these minima and maxima.

There is a flurry of different notations for derivatives, among those the following:

$$f(x),\ f'(x),\ f''(x),\ \ldots \tag{0.9}$$

$$f(x),\ \frac{\mathrm{d}f(x)}{\mathrm{d}x},\ \frac{\mathrm{d}^2f(x)}{\mathrm{d}x^2},\ \ldots \tag{0.10}$$

Another notation you will encounter very often is this one:

$$f(x,y),\ \frac{\partial f(x,y)}{\partial x},\ \frac{\partial^2 f(x,y)}{\partial x^2},\ \ldots \tag{0.11}$$

Actually, this notation refers to a slightly different concept, that of *partial* differentiation. Imagine you have a function $f(x,y)$ that depends on two variables x and y, and you want to differentiate with respect to x only. Then you can indicate this using the ∂ sign: $\frac{\partial f(x,y)}{\partial x}$. Partial differentiation is intuitively very simple: You treat all other variables, here y, as constants, and only apply the known rules of differentiation to the variable in question, here x. When we work with functions that take in multiple variables, we need to broaden our understanding of what a first derivative is, as the term "slope" is now more difficult to understand intuitively. We can therefore say: The first derivative tells us how much the value of our function f changes when the value of the variable (be it x, y or whatever else) changes.

Be aware of the following basic rules for differentiation:

function $f(x)$	derivative $f'(x)$
c	0
x^a	$a \cdot x^{a-1}$
e^x	e^x
$\ln(x)$	$\frac{1}{x}$

When dealing with square roots, remember that $\sqrt{x} = x^{\frac{1}{2}}$. Also be aware that $\frac{1}{x} = x^{-1}$. If you are no longer familiar with basic differentiation rules, you should consult a dedicated math textbook and revisit them. But if you only need an overview of the most heavily used rules for differentiation, this paragraph is for you. The *summation rule* is:

$$\frac{d}{dx}(f(x) + g(x)) = \frac{df}{dx} + \frac{dg}{dx}. \tag{0.12}$$

The product rule is:

$$\frac{d}{dx}(f(x) \cdot g(x)) = \frac{df}{dx} \cdot g(x) + f(x) \cdot \frac{dg}{dx}. \tag{0.13}$$

Especially useful is the *chain rule*:

$$\frac{d}{dx}(f(g(x))) = \frac{df(g(x))}{dg(x)} \cdot \frac{dg}{dx}. \tag{0.14}$$

And finally, there is also a *quotient rule*:

$$\frac{d}{dx}(\frac{f(x)}{g(x)}) = \frac{g(x) \cdot \frac{df}{dx} - \frac{dg}{dx} \cdot f(x)}{(g(x))^2} \tag{0.15}$$

For econometric derivations, it is also particularly useful to know that the sum of derivatives equals the derivative of sums:

$$\sum \frac{\partial}{\partial x} = \frac{\partial \sum}{\partial x}. \tag{0.16}$$

0.1.3 Logarithms and Exponential Functions

Logarithms in general are simple to understand. When you have an exponentiation like

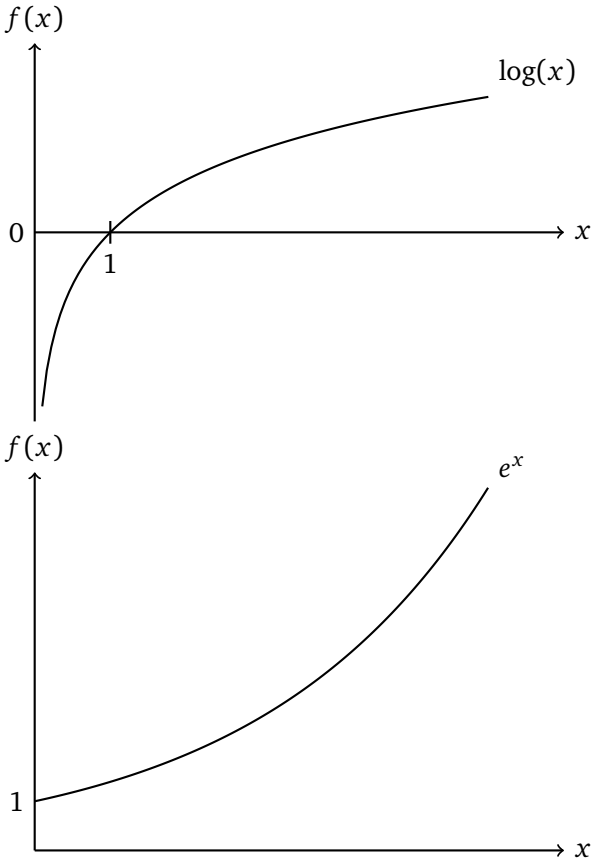
$$a^b = c, \tag{0.17}$$

then the logarithm is the inverse function:

$$\log_a(c) = b. \tag{0.18}$$

This operation is especially useful when you need to rearrange an expression and the variable you need to isolate is in the exponent. Another use case in Econometrics is that of *logarithmically transforming variables*. When you log-transform a variable, you usually take the logarithm with base e of every value (more on this and how to interpret this in the lecture). The logarithm with base e is called *natural logarithm*. Outside econometrics, it is often denoted as $\ln(x)$, but in econometrics, many people just use $\log(x)$. Another common logarithm is that with base 10.

This is how logarithms and exponential functions look like when graphed:



When working with logarithms, you should also be aware of the following rules.
Products:

$$\log_b(xy) = \log_b(x) + \log_b(y) \tag{0.19}$$

Quotients:

$$\log_b\left(\frac{x}{y}\right) = \log_b(x) - \log_b(y) \tag{0.20}$$

Exponentiation:

$$\log_b(x^p) = p \cdot \log_b(x) \tag{0.21}$$

Roots:

$$\log_b\left(\sqrt[p]{x}\right) = \frac{\log_b(x)}{p} \tag{0.22}$$

0.1.4 Matrices and Vectors

A *matrix* is a rectangular, two-dimensional array of values, arranged in rows and columns. These values can be numbers, functions, symbols or whatever else you desire. A $n \times m$ matrix has n rows and m columns. We denote matrices by bold uppercase letters ($\boldsymbol{X}$). Lowercase letters with subscripts denote individual elements of the matrix:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}. \tag{0.23}$$

Vectors are equally important in econometrics. You are probably more familiar with vectors than with matrices, but it does help to revisit them here. Think of a vector as something like a matrix with only one column. If it is arranged like this, we call it a *column vector* and denote it by a bold lowercase letter.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{0.24}$$

For both matrices and vectors, bolding them is not strictly necessary, but it helps to understand what we are actually talking about when we discuss an econometric equation. If we want to clarify in handwriting that something is a matrix or a vector, we can underline them instead of using bold text: $\underline{X}$ and $\underline{x}$. In most cases, this will not be necessary except if we deliberately want to point out the fact that something is not a scalar.

The matrix $\boldsymbol{X}$ above has dimensions $n \times k$, and the vector $\boldsymbol{x}$ has length n (or dimensions $n \times 1$). We can also write $\boldsymbol{x}$ as a *row vector*. To do this, we must *transpose* the vector. Simply put, rows become columns, and columns become rows. We denote the transposition with a small prime (like in $\boldsymbol{x}'$) or, alternatively, a small superscript capital T.

$$\boldsymbol{x}' = (x_1, x_2, \ldots, x_n). \tag{0.25}$$

We can also transpose matrices. The *transpose* of a $n \times k$ matrix $\boldsymbol{X}$, denoted as $\boldsymbol{X}'$, is the $k \times n$ matrix defined as

$$\boldsymbol{X}' \text{ where } (\boldsymbol{X}')_{ij} = x_{ji}. \tag{0.26}$$

If $\boldsymbol{X}$ is an $n \times k$ matrix and $\boldsymbol{Z}$ is an $k \times p$ matrix, then

1. $\boldsymbol{X}'' = \boldsymbol{X}$
2. $(\boldsymbol{XZ})' = \boldsymbol{Z}'\boldsymbol{X}'$.

An $n \times n$ matrix is called a *square matrix*. A square matrix $\boldsymbol{A}$ is called *symmetric* if $\boldsymbol{A}' = \boldsymbol{A}$. An $n \times n$ matrix with zeroes everywhere except in the main diagonal is called a *diagonal matrix* and can be denoted by $\text{diag}(x_1, x_2, \ldots, x_n)$. An $n \times n$ matrix with ones in the main diagonal and zeroes everywhere else is called *identity matrix* and is denoted by $\boldsymbol{I_n}$. An $m \times n$ matrix where every element is zero is called zero matrix and can be denoted by $\mathbf{0}_{mn}$.

The *rank* of a matrix $\boldsymbol{X}$, denoted by $\text{rank}(\boldsymbol{X})$, is defined as the dimension of the vector space spanned by the columns of a matrix. Simply put, it corresponds to the maximum number of *linearly independent* columns of $\boldsymbol{X}$. Two vectors (the columns are vectors) $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent when there are two scalars a and b, of which at least one is different from zero, that you could insert to make the following equation true:

$$a\boldsymbol{x} + b\boldsymbol{y} = 0. \tag{0.27}$$

Consider the following matrix:

$$\boldsymbol{A} = \begin{pmatrix} 12 & 2 & 10 \\ 3 & 1 & 2 \\ 7 & 4 & 3 \\ 8 & 6 & 2 \end{pmatrix} \tag{0.28}$$

This matrix has rank 2. There are three columns, but the third is just a linear combination of the first and second: $a_{i3} = a_{i1} - a_{i2}$. A matrix is said to have *full rank* if its rank equals the maximum possible for a matrix of the same dimensions (which is the lesser out of the number of rows and the number of columns), if not, the matrix is called *rank-deficient*.

Matrix addition is performed element-wise:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} + \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{pmatrix} = \begin{pmatrix} x_{11}+z_{11} & x_{12}+z_{12} & \cdots & x_{1k}+z_{1k} \\ x_{21}+z_{21} & x_{22}+z_{22} & \cdots & x_{2k}+z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}+z_{n1} & x_{n2}+z_{n2} & \cdots & x_{nk}+z_{nk} \end{pmatrix}. \tag{0.29}$$

Multiplication of a scalar and a matrix is also performed element-wise:

$$\alpha \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \alpha x_{11} & \alpha x_{12} & \cdots & \alpha x_{1k} \\ \alpha x_{21} & \alpha x_{22} & \cdots & \alpha x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha x_{n1} & \alpha x_{n2} & \cdots & \alpha x_{nk} \end{pmatrix}. \tag{0.30}$$

Multiplication of two matrices is a little more complicated. Let $\boldsymbol{X}$ be a $n \times k$ matrix and $\boldsymbol{Z}$ be a $k \times p$ matrix. Then, matrix multiplication can be performed by multiplying rows by columns. Look at the following example, where $\boldsymbol{X}$ is a 2×3 matrix and $\boldsymbol{Z}$ is a 3×2 matrix:

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} \cdot \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \end{pmatrix} = \begin{pmatrix} x_{11}z_{11} + x_{12}z_{21} & x_{11}z_{12} + x_{12}z_{22} & x_{11}z_{13} + x_{12}z_{23} \\ x_{21}z_{11} + x_{22}z_{21} & x_{21}z_{12} + x_{22}z_{22} & x_{21}z_{13} + x_{22}z_{23} \\ x_{31}z_{11} + x_{32}z_{21} & x_{31}z_{12} + x_{32}z_{22} & x_{31}z_{13} + x_{32}z_{23} \end{pmatrix} \tag{0.31}$$

The following table helps visualize the process:

	z_{11}, z_{21}	z_{12}, z_{22}	z_{13}, z_{23}
x_{11}, x_{12}	$x_{11}z_{11} + x_{12}z_{21}$	$x_{11}z_{12} + x_{12}z_{22}$	$x_{11}z_{13} + x_{12}z_{23}$
x_{21}, x_{22}	$x_{21}z_{11} + x_{22}z_{21}$	$x_{21}z_{12} + x_{22}z_{22}$	$x_{21}z_{13} + x_{22}z_{23}$
x_{31}, x_{32}	$x_{31}z_{11} + x_{32}z_{21}$	$x_{31}z_{12} + x_{32}z_{22}$	$x_{31}z_{13} + x_{32}z_{23}$

Written more compactly, the individual elements of the resulting matrix $\boldsymbol{S}$ are:

$$s_{ij} = \sum_{\ell=1}^{k} x_{i\ell} z_{\ell j} \tag{0.32}$$

We can see above that the number of columns of $\boldsymbol{X}$ and the number of rows of $\boldsymbol{Z}$ must be equal in order for us to be able to multiply. The resulting matrix has the same number of rows as $\boldsymbol{X}$ and the same number of columns as $\boldsymbol{Z}$. Importantly, matrix multiplication is not commutative: That means that $\boldsymbol{XZ}$ is not the same as $\boldsymbol{ZX}$. If $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{A}$ are all matrices of the appropriate size, then the following additional rules apply:

- $\boldsymbol{X}+\boldsymbol{Z}=\boldsymbol{Z}+\boldsymbol{X}$
- $(\boldsymbol{X}+\boldsymbol{Z})+\boldsymbol{A}=\boldsymbol{X}+(\boldsymbol{Z}+\boldsymbol{A})$
- $\boldsymbol{X}+\boldsymbol{0}=\boldsymbol{X}$
- $\boldsymbol{XZ}\neq\boldsymbol{ZX}$
- $(\boldsymbol{XZ})\boldsymbol{A}=\boldsymbol{X}(\boldsymbol{ZA})$
- $\boldsymbol{IX}=\boldsymbol{XI}=\boldsymbol{X}$
- $(\alpha\boldsymbol{X})\boldsymbol{Z}=\boldsymbol{X}(\alpha\boldsymbol{Z})=\alpha(\boldsymbol{XZ})$
- $\boldsymbol{A}(\boldsymbol{X}+\boldsymbol{Z})=\boldsymbol{AX}+\boldsymbol{AZ}$
- $(\boldsymbol{X}+\boldsymbol{Z})\boldsymbol{A}=\boldsymbol{XA}+\boldsymbol{ZA}$.

A square matrix $\boldsymbol{X}$ is called *invertible* if there exists a matrix $\boldsymbol{X}^{-1}$ such that

$$\boldsymbol{X}\boldsymbol{X}^{-1}=\boldsymbol{X}^{-1}\boldsymbol{X}=\boldsymbol{I}. \tag{0.33}$$

In this case, $\boldsymbol{X}^{-1}$ is the *inverse* matrix of $\boldsymbol{X}$. If there exists no such matrix, $\boldsymbol{X}$ is called singular. If $\boldsymbol{X}$ is invertible, then its inverse is uniquely defined. A linearly independent matrix is invertible. In addition:

1. $(\boldsymbol{X}^{-1})^{-1}=\boldsymbol{X}$
2. $(\boldsymbol{X}')^{-1}=(\boldsymbol{X}^{-1})'$.

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are two invertible matrices of the same size,

3. $(\boldsymbol{XZ})^{-1}=\boldsymbol{Z}^{-1}\boldsymbol{X}^{-1}$.

0.2 Probability Basics

0.2.1 Random Variables

Suppose we observe a random event, like a coin toss, throwing a die or playing a lottery and drawing a number. A *random variable* is a variable that takes on a numerical value that is determined by the event you observe. We (often) denote it with an uppercase letter:

$$X \tag{0.34}$$

We denote all possible outcomes with the corresponding lowercase letter:

$$x_i \tag{0.35}$$

A *discrete random variable* is a random variable that can only have a finite or countably infinite number of potential outcomes. If the variable is called X, we denote outcomes x_i and associated probabilities p_i. Note that the sum of all probabilities $\sum_i p_i$ must equal 1. An example for a discrete random variable would be throwing two dice. The potential outcomes are $\{2,3,4,5,6,7,8,9,10,11,12\}$, and the associated probabilities are

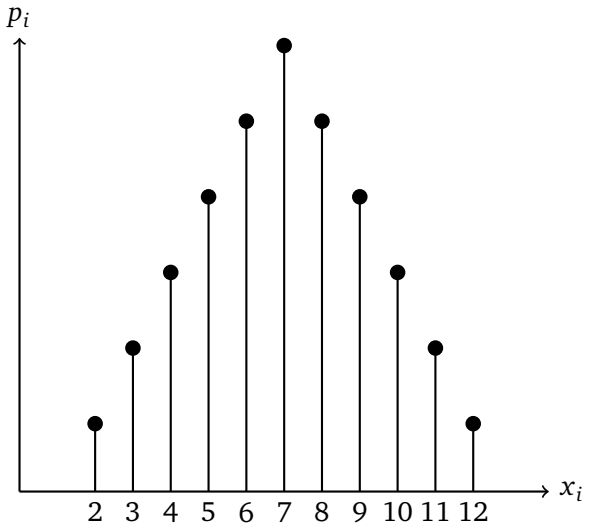


Figure 0.3: A representation of the probability mass function of the combined outcome when throwing two dice at once.

$\{\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}\}$. Figure 0.3 contains an illustration of the *probability mass function* (PMF), the function that associates outcomes with probabilities:
A *Bernoulli variable* is a discrete random variable that can only take two outcomes, like a coin toss.

A *continuous random variable* is a random variable that can take an uncountably infinite number of different outcomes. We know that there is an infinite number of outcomes and that the sum of all of those is 1. It follows that the probability of each exact outcome equals zero. Therefore, there is no probability *mass* function (PMF). What we can do, however, is plot a *probability density function* (PDF). It tells us the probability of the outcome falling within a certain interval. The total area below the entirety of the PDF equals 1. Figure 0.4 depicts a possible probability density function for body height, a nice example for a continuous random variable. It would make no sense to ask for the probability of a person being exactly 1.734681092536 meters tall. This probability is zero. But we can have a look at the PDF and determine how likely it is that the person's height is between 1.73 and 1.74 meters.

Whether an infinite set of numbers is *countably* or *uncountably* infinite can be intuitively answered. All *natural* numbers $\mathbb{N}$ are countably infinite. We can clearly lay out a path how to count them (start at 0, then 1, then 2, then 3, ...), we just don't know where and when the path *ends*. After all, it's still infinite. All *real* numbers $\mathbb{R}$, however, are uncountably infinite. We cannot even lay out the rules for a path to follow when we count. Say we start at 0, then 0.001 ... what about all numbers in between? And all numbers between those? There is no way to count them all.

In addition to the probability density function, we can plot the *cumulative distribution function* (CDF), an example of which is pictured in Figure 0.5. It tells us the probability that the outcome is equal to or smaller than a certain value. This function is strictly monotonically increasing. The dashed line in Figure 0.5 shows how to read the plot: the value of the density function at $X = 1.74$ represents the probability that a randomly selected person is shorter than or exactly 1.74 meters tall.

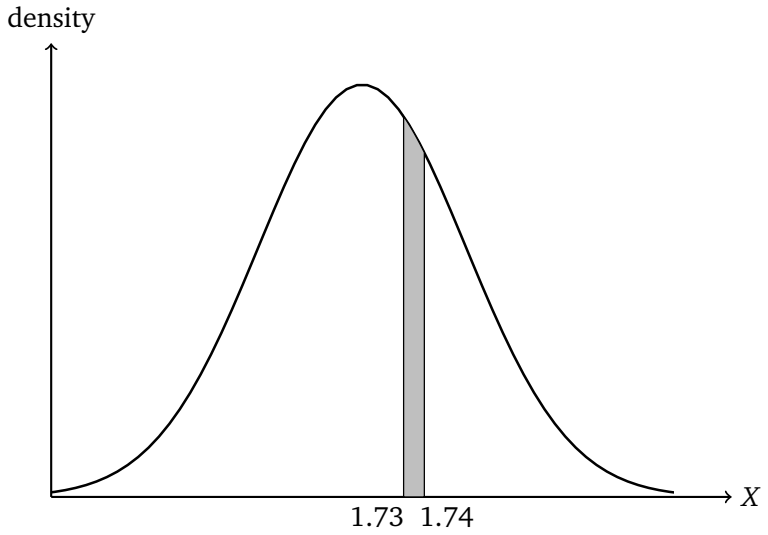


Figure 0.4: A representation of the probability density function of body height.

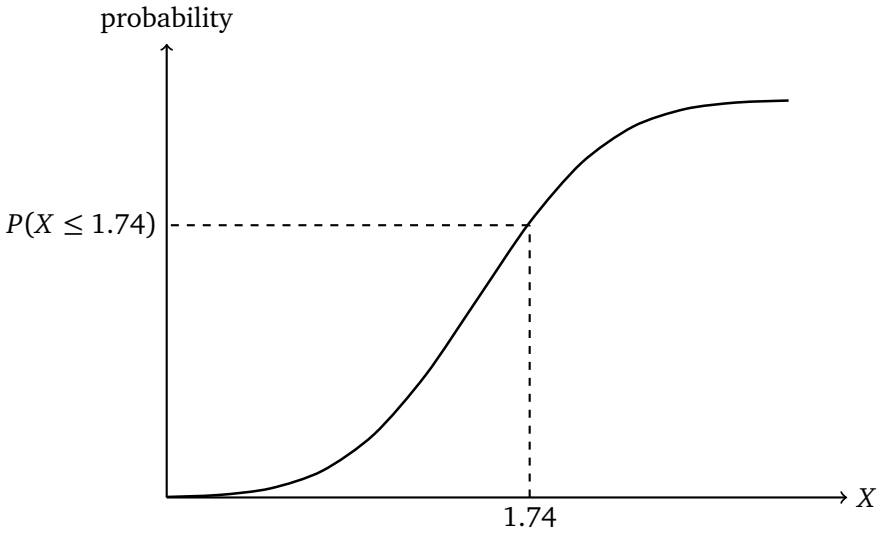


Figure 0.5: A representation of the cumulative density function of body height.

0.2.2 Analysis of One Random Variable

Let us go back to our example of throwing a die. The outcome of this fun but simple game is of course a discrete random variable with the following outcomes and associated probabilities:

Outcome	Probability
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Table 0.1: Outcomes and associated probabilities when throwing one die.

The *expected value* is a concept that allows us to analyze what value we can *expect* when throwing the die. We calculate it as the arithmetic mean of the outcomes, weighted by their respective probabilities. We denote the expected value, or *expectation*, with a capital E:

$$\mathrm{E}(X) := \sum_{i=1}^{N} x_i p_i \tag{0.36}$$

The expected value for a fair die is 3.5. As we draw more and more outcomes from this distribution, that is, we roll the die many, many times, the mean of all our rolls will move closer and closer to the expected value. All of this is quite easy and straightforward to interpret as long as we're dealing with discrete variables. It gets more difficult with continuous variables, but the general intuition of what an expected value is remains the same.

We deal with expectations a lot in Econometrics, so it is useful to know some rules for handling them. For any constant c,

$$\mathrm{E}(c) = c \tag{0.37}$$

For random variables X, Y and constants c, d:

$$\mathrm{E}(c \cdot X + d \cdot Y) = c \cdot \mathrm{E}(X) + d \cdot \mathrm{E}(Y) \tag{0.38}$$

For constants $c_1, \ldots, c_n$ and random variables $X_1, \ldots, X_n$:

$$\mathrm{E}\left(\sum_{i=1}^{N} c_i X_i\right) = \sum_{i=1}^{N} c_i \mathrm{E}(X_i) \tag{0.39}$$

For two *independent* random variables X, Y:

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) \tag{0.40}$$

Often times, the expected value is not enough to analyze a distribution. Imagine that you own a firm that produces screws. You have two machines that produce them. You advertise your screws as being 35 millimeters long each, but in reality, the length of the screws is randomly distributed: The expectation of the screw length is 35 mm for both machines. However, machine A usually produces screws that are really close to their desired length,

whereas machine B sometimes produces screws that are as short as 33 mm or as long as 37 mm. What is the difference between the two machines with identical expectations?

The answer is called *variance*. Simply put: The expectation tells us where the "center" of a distribution is. The variance, on the other hand, tells us how far the outcomes tend to deviate from that expectation. We denote it as $\text{Var}(X)$ and calculate it as follows:

$$\text{Var}(X) := \text{E}\left((X-\mu)^2\right), \tag{0.41}$$

where $\mu = \text{E}(X)$.

It makes a lot of sense that the variance of any constant is zero. Additionally, the following rule applies for any random variable X and constants a, b:

$$\text{Var}(aX+b) = a^2\text{Var}(X) + \text{Var}(b) = a^2\text{Var}(X) \tag{0.42}$$

The standard deviation, denoted as $\text{sd}(X)$, is simply the square root of the variance.

0.2.3 Analysis of Two Random Variables

Suppose that X and Y are two discrete random variables. In addition to their individual distributions, we can describe their *joint distribution*. For this, we use a *joint probability mass function*.

$$f_{X,Y}(x,y) = P(X=x, Y=y) \tag{0.43}$$

This function simply tells us what the probability for each combination of X and Y is. If X and Y are independent, then

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \tag{0.44}$$

where $f(x)$ and $f(y)$ are the probability mass functions for X and Y, respectively. Two random variables being *independent* means that knowing the outcome of X does not alter the probabilities associated with each possible outcome of Y. The concept exists similarly for continuous variables, called *joint probability density function*.

Another important concept is the *conditional distribution*. The *conditional probability density function* tells us how the outcome of X affects that of Y:

$$f_{Y|X}(y|x) = P(Y=y|X=x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \text{ for all } f_X(x) \text{ that are greater than 0.} \tag{0.45}$$

If X and Y are independent, the outcome of X does not affect Y and thus $f_{Y|X}(y|x) = f_Y(y)$.

The *covariance* is a concept that is similar to the variance, but is useful for jointly analyzing two distributions. It is defined as follows and denoted with $\text{Cov}(X,Y)$:

$$\text{Cov}(X,Y) := \text{E}\left((X-\mu_X)(Y-\mu_Y)\right), \tag{0.46}$$

where $\mu_X = \text{E}(X)$ and $\mu_Y = \text{E}(Y)$.

We can intuitively interpret the sign of the covariance. If the covariance is positive, we expect Y to be above its mean when X is too. If the covariance is negative, we expect Y to be below its mean when X is above its mean. In simple terms, a positive covariance tells us that two variables are positively associated with each other, and vice versa. A covariance

of 0 means that there is no relationship. If X and Y are independent, the covariance will therefore always be zero. Note that being *associated* in this sense does not necessarily allow for any conclusions on whether one of the variables causes, or does not cause, the other.

The following rules apply for the covariance:

$$\text{Cov}(X,Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) \tag{0.47}$$

For constants a, b, c, d:

$$\text{Cov}(aX + b, cY + d) = a \cdot c \cdot \text{Cov}(X,Y) \tag{0.48}$$

The covariance is useful for determining whether variables are associated. However, as it depends on how large the numbers of each of the variables are, we cannot use it to determine *how strongly* two variables are associated. For this, we can calculate the *correlation* between them, which we denote as $\text{Corr}(X,Y)$ and calculate as follows:

$$\text{Corr}(X,Y) := \frac{\text{Cov}(X,Y)}{\text{sd}(X) \cdot \text{sd}(Y)} \tag{0.49}$$

By dividing by the product of the standard deviations, we can get rid of units of measurement. The resulting number will always be between -1 and 1. The sign is to be interpreted the same way as that of the covariance, a value of 0 again means no association, and the larger the absolute value (i.e. the farther away from 0), the stronger the association between the variables.

Keep in mind the following rules. For constants a, b, c, d, where $ac > 0$:

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X,Y) \tag{0.50}$$

For constants a, b, c, d, where $ac < 0$:

$$\text{Corr}(aX + b, cY + d) = -\text{Corr}(X,Y) \tag{0.51}$$

Now suppose that we have two random variables X and Y that are in some way associated. We want to know what the expectation of Y is, *given* that X take a certain value. This is called a *conditional expectation* and is denoted $\text{E}(Y|X = x)$. The following rules apply for conditional expectations: If $f(x)$ is any function,

$$\text{E}(f(X)|X) = f(X) \tag{0.52}$$

For two functions $f(X)$ and $g(x)$:

$$\text{E}(f(X)Y + g(X) \mid X) = f(X)\text{E}(Y|X) + g(X) \tag{0.53}$$

If X and Y are independent:

$$\text{E}(Y|X) = \text{E}(Y) \tag{0.54}$$

One particularly important rule, which we will repeatedly need in our econometrics courses, is called the *law of iterated expectations*:

$$\text{E}(\text{E}(Y|X)) = \text{E}(Y) \tag{0.55}$$

Econometrics I

1.1 Introduction

1.1.1 What is Econometrics?

Let us start this econometrics course by defining what *econometrics* actually means, that is, which topics we will be dealing with. You may ask yourself, what is the difference between econometrics and applied statistics? Why do you have to take a course in both? You may have heard that we will be talking about *regression* a lot. You presumably have already done that in your statistics course, so why are you doing it again?

Actually, econometrics is not just plain applied statistics. One, it is a subfield of *economics*. This means that we deal with economic questions and are confined to the data environments these questions are usually answered in. Two, econometrics can be seen as a flavor of *applied statistics*. This means that we apply statistical methods to test our hypotheses. One key point in how econometrics differs from applied mathematical statistics is its focus on the peculiarities that arise from using *non-experimental data*. In other words, we cannot always conduct an experimental study, but have to rely on what we observe in the real world. This makes inferring conclusions from the data much more challenging.

We said that we use statistical methods to answer *economic questions*. What does that mean? You may be tempted to think of the word "economic" in a very narrow sense, but actually, we can answer a very broad set of questions using econometric methods. Let us consider one example to see what kind of question we need in order to perform an econometric analysis. Suppose, for instance that the government is interested in evaluating the effectiveness of government-funded educational leave. In order to make statements based on a quantitative analysis, we need to think of a *hypothesis*, and we need the *data* to test it.

For instance, we can formulate a *research question* like this:

If a worker takes advantage of educational leave*, does their* wage *increase over the course of their career?*

This question implies the hypothesis that taking advantage of educational leave impacts a person's later wage, possibly – and quite likely – among other things. We can now proceed by expressing this hypothesis as part of a mathematical model:

$$\text{wage} = f(\text{education}, \text{experience}, \text{talent}, \text{educational leave}, \dots) \tag{1.1}$$

This equation is to be read as: "the wage is a function of education, experience, talent, educational leave, and other factors," and nothing else. We deliberately do not impose any functional form at this stage, but use only a general function, $f(\cdot)$. Thereby, we leave open whether the relationship is positive or negative, and whether it is linear or not, among other things.

The variables inside the function in Equation 1.1 differ in a number of ways. One of those aspects which might not be immediately obvious concerns whether the variables are *observable* or not. Education, for instance, is a variable that we can easily *observe*. Which measurement we choose for it is by no means trivial, but there is no doubt about that we can observe at least some measure of education, such as the number of years a person went to school. This is more difficult for a variable like talent: Not only is it not immediately obvious which measurement one would choose to quantify a person's talent, it is also impossible to observe talent directly. This applies to a lot of latent personal characteristics, but also more broadly to variables that are inherently *unobservable*, or *unobserved* in at least a specific dataset.

Now that we know what econometrics is, we can discuss *why* we should care. What is the *purpose* of econometrics? Why should you learn all of this? What will it enable you to do in your later career? Actually, the answer to the latter question depends on what you choose to do. But we can at least try to answer what the purpose of econometrics is. Broadly, we can think of four use cases:

1. *Test and falsify economic theories*. The times when economic research consisted of somebody proposing a theoretical model and everyone just rolling with it have long been over. Today, we want to do our best to test models using the data we have. Do households actually save more when interest rates rise? Do countries eventually converge to a common equilibrium when they grow? There are compelling arguments for both, but ideally we want to test them empirically.

2. *Evaluate policy measures.* Good politicians propose and implement policies, and subsequently want to evaluate their effects. Or we could look at other places that have already implemented a policy and see how they fared. Actual politics, of course, is much more vibes-based, but our aim is different here. Does a minimum wage reduce or increase unemployment? Does a reduction in class size have different effects on male and female students? Of course, a "policy" can be anything and is not actually limited to actions taken by politicians. Think of a teacher changing their attendance policy, or a train company implementing new signage. Those are all policies, and they can be evaluated.

3. *Quantify relationships between economic variables*. Admittedly, there is a lot of overlap with the previous two use cases. But now, we explicitly talk about magnitudes. What is the causal effect of education on wages, and how large is it? How wide is the gender pay gap in a given industry? In all of these questions, the size of the effect actually matters.

4. *Predictions and forecasts*. The last three cases covered different flavors of causal questions. The second main reason of why we do econometrics is that we want to predict. That is, we want to know the outcome of something that either has not happened yet, or that we have not observed yet. How much will GDP grow next year? How volatile will stock markets be next week? How much does this customer want to pay for their flight? All of these questions concern predictions.

Going back to econometric questions, suppose we are tasked to answer the following:

Does the average *class size* in a district influence *test performance*? If so, by how much?

As before, we can assume that there are both *observed* and *unobserved* factors at play. We can think of the following observed factors: the average household income in the district, the rate of literacy, the share of students that do not speak the language of instruction at home, ...On the unobserved side, think of: average student motivation, average teacher motivation, ...We can easily think of a lot more variables.

This is a good time for our first code example. Consider the following:

```
# Load packages
library(AER) # Contains our dataset
library(dplyr) # Contains mutate()

# Load data
```

```
6  data("CASchools")
7
8  # Compute variables with mutate()
9  CASchools <- CASchools |>
10   mutate(student_teacher_ratio = students / teachers,
11          test_score = (read + math)/2)
```

`CASchools` is a dataset that contains data on test scores, class size, and twelve other variables for 420 school districts in California. We start by loading the `AER` package, which contains this dataset, and the `dplyr` package, which contains a bunch of functions that are useful when we need to transform data. Then, we load the dataset and create two variables using `mutate()` from the `dplyr` package: We divide the number of students by the number of teachers to compute the student-teacher ratio, and we compute the mean test score by averaging reading and math scores.

```
12 plot(CASchools$student_teacher_ratio, CASchools$test_score,
13      xlab="Class␣Size", ylab="Test␣Scores")
14 abline(lm(test_score~student_teacher_ratio, data = CASchools)
   , col="red")
```

Next, we plot the two variables we are interested in. First, we create a scatterplot with the student-teacher ratio on the x axis and the average test score on the y axis. Then, we fit a regression line through the dots. The output is shown in Figure 1.1: There are a lot of observations, and if we fit a line through them, it is negatively sloped. When we additionally compute means of two groups, those districts where the average class size is smaller than or equal to 22, and where it is larger than 22, we can see that the test score is higher in the group where class sizes are smaller.

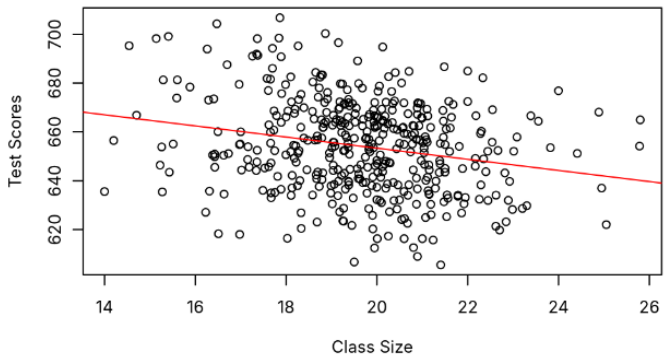


Figure 1.1: Output from Lines 12 to 14.

```
15 CASchools[CASchools$student_teacher_ratio > 22,]$test_score
   |>
16   mean()
17
18 CASchools[CASchools$student_teacher_ratio <= 22,]$test_score
   |>
19   mean()
```

```
[1] 647.4973
[1] 654.7999
```

What does all of this tell us? Actually, not much. The first reason for this is immediately obvious from the scatterplot: The data is very noisy, meaning that there is a lot of uncertainty in the relationship we estimate. Also, only 37 districts have more than 22 students per class, while 383 have less than that. We can be much more confident in the mean of the larger subsample than that of the smaller subsample. This is the first instance of a problem that we should keep in mind whenever we conduct econometric analyses: Every time we analyze samples, we deal with uncertainty.

Beyond uncertainty, there is a second caveat: We cannot say anything about causality. Even when we have accounted for all uncertainty in the sample, causal statements are non-trivial to make. We do not know why results are worse in high-ratio districts, and we do not know whether they would improve if we sent more teachers there. It may well be that students are different in those districts. If they are, they are also likely different in both observed and unobserved characteristics, which further complicates our analysis.

All of this tells us: It is not sufficient to simply analyze economic data with statistical tools. We also have to carefully think about *how* we analyze and interpret the data. That starts with data collection, involves deciding which methods we use and how we apply them, and includes the interpretation of our results, meaning that we should be explicit both about what we can say and what we cannot say. In Econometrics I, Econometrics II, and Applied Econometrics, we will learn step by step how to address all of these issues. By the end of these three courses, we are able to independently answer econometric research questions.

1.1.2 Causality

Think about what the following two statements have in common, and how they are different:

1. *One additional year of education leads to an average increase in wages of 20 percent.*
2. *People who have one more year of education earn on average 20 percent more.*

Quantitatively, the two statements capture the same relation, but they differ in their causal interpretation. The words “leads to” add a causal component to the first statement that fundamentally changes how we need to treat that hypothesis in our analysis.

As economists, we are often interested in *causal effects*, that is, situations where one variable affects another variable. Examples are:

- How does the price affect the demand for a product?
- How does a particular policy measure affect the rate of unemployment?
- How does the use of fertilizer affect agricultural yields?

Informally, we can say that we speak of a *causal effect* when the isolated change of one variable has a direct, measurable effect on another variable.

Let us consider the fertilizer example. Consider Figure 1.2: Imagine we have a square plot of land, which is divided into 100 subplots. We randomly choose 50 plots out of the 100 and apply fertilizer there, but not on the other plots. Then we wait for a bit, and finally record yields from every plot and compare whether they were higher where we applied fertilizer.

What we just described is an example for a *randomized controlled trial* (RCT). An RCT involves assigning an intervention to a randomly selected study group. A control group, which ideally is of comparable size, does not receive the intervention. This type of study

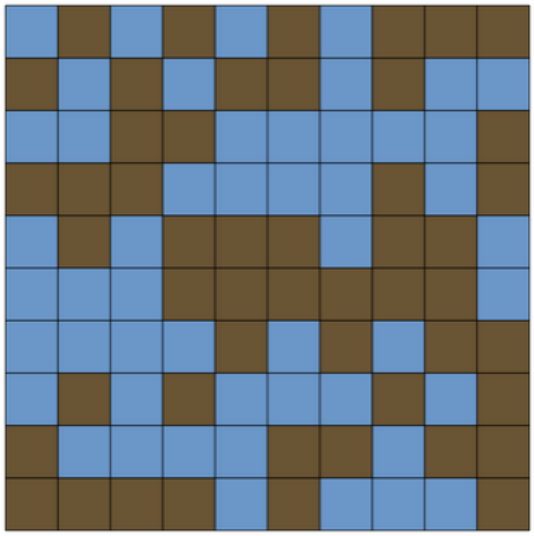


Figure 1.2: Stylized research design for a fertilizer experiment.

works well under a number of assumptions. The first is that because yields are also influenced by other variables, we assume that the expectation of those variables does not differ between treated and untreated fields. If our randomization was successful, this is a trivial assumption. If treatment assignment was not random, it may have been correlated with some of the variables that influence yields. In that case, we can no longer identify a causal effect. The second assumption is not guaranteed even under randomized treatment: Fertilizer should not have an effect on neighboring subplots. In this setting, the assumption is not particularly credible. Subplots are connected by groundwater flows, and the positive effects of fertilizer may spill over to adjacent plots. This means that we cannot cleanly identify an effect because the control plots may also be affected by the treatment.

But generally speaking, experiments (RCTs or lab experiments) offer the cleanest opportunities to identify causal effects. Why do we then even think about using observational data instead? The reason is that experiments often are infeasible for practical, financial, legal, or even ethical reasons. Consider the following examples:

- Coville et al. (2020) want to find out if people who have not paid their water bills pay faster when their water is shut off. For this, they randomly select from a sample of households with payment issues in Nairobi, Kenya, some whose access to water they shut off. Because all customers have initially signed a contract with the water company that includes cutting off water supply as a measure of last resort, they argue that they have obtained informed consent of the participants in the experiment.

- Cohen and Dupas (2008) examine whether co-payments for malaria nets reduce "wasteful" use of them. For this, they randomize the price at which malaria protection nets are distributed to pregnant women, from 0 to 40 Kenyan shillings. In the end, they find no evidence for that distributing them for free leads to wasteful use.

In these two instances, ethical issues with the experimental design are immediately apparent. But quite often, conducting an experiment may be infeasible for other, sometimes less apparent reasons.

One way or another, we are often confined to using *observational data*, even though economic experiments are becoming more common. Broadly speaking, observational data is everything that is not generated through a lab experiment or an RCT. We can obtain it by conducting a survey, from a survey that somebody else has conducted, from administrative data sources, from satellite measurements, as well as from many other sources. The main advantage is that this data is usually available on a much larger scale than experimental data, often covering entire countries. Also, observations reflect real behavior and are not prone to experiment-specific types of bias. The main disadvantage is that observational data was not generated with our particular research question in mind, which makes isolating the effect of interest much harder.

1.1.3 Structure of Econometric Data

Consider again our model for evaluating the wage effects of going on educational leave:

$$\text{wage} = f(\text{education}, \text{experience}, \text{talent}, \text{educational leave}, \dots) \tag{1.2}$$

We have repeatedly mentioned that we need appropriate *data* to evaluate this question, but we have not yet discussed what an appropriate dataset would look like. Suppose, for example, that we collect data on wage, education, experience, and educational leave. Then we end up with something like Table 1.1.

?	Wage	Education	Experience	Educational Leave
1	15	12	9	Yes
2	21	14	2	No
3	14	11	7	No
4	18	9	22	No
⋮	⋮	⋮	⋮	⋮

Table 1.1: A dataset.

In tables like these, we can refer to the columns as *variables* and to the rows as *observations*. Looking at Table 1.1, you will have noticed that the first column contains no actual data, but an incrementally increasing index of observations, and that the column is marked with a question mark. This is because how we index our observations, or rather, what the index refers to, determines what kind of data we have at hand and what we can do with it.

Individual	Wage	Education	Experience	Educational Leave
$i = 1$	15	12	9	Yes
$i = 2$	21	14	2	No
$i = 3$	14	11	7	No
$i = 4$	18	9	22	No
⋮	⋮	⋮	⋮	⋮

Table 1.2: A *cross-sectional* dataset.

Suppose, for instance, that each observation represents one *individual*. Then our table looks like Table 1.2. We call this a *cross-sectional* dataset, because observations represent a cross-section of the population, with all individuals surveyed (ideally) at the same time. We

denote individuals by the index i, and we call the number of observations, that is, the sample size, N. Individuals can be actual individuals, but also households, firms, cities, countries, ... As a rule, we will assume that the sample is randomly drawn from the population.

Time	Wage	Education	Experience	Educational Leave
$t = 2022$	0	8	0	No
$t = 2023$	0	9	0	No
$t = 2024$	12	10	1	No
$t = 2025$	14	10	2	Yes
⋮	⋮	⋮	⋮	⋮

Table 1.3: A dataset containing multiple *time series*.

But this is not the only possibility. Instead of different individuals, we can survey the *same individual at different points in time*. Then, we end up with something like Table 1.3. We call this a *time series*. It consists of a sequence of time points at which data is collected on the same unit. Here, we usually use the index t, and the number of observations is denoted as T. Since later observations mechanically depend on earlier ones, we cannot assume that a time series sample is random.

Individual	Time	Wage	Education	Experience	Educational Leave
$i = 1$	$t = 2024$	20	14	1	No
$i = 2$	$t = 2024$	12	10	1	No
$i = 1$	$t = 2025$	21	14	2	No
$i = 2$	$t = 2025$	14	10	2	No
⋮	⋮	⋮	⋮	⋮	⋮

Table 1.4: A *panel* dataset.

A third type of dataset combines the two dimensions. When we follow multiple units over time, we end up with something like Table 1.4, which we call a *panel dataset*. It includes both a cross-sectional and a time series component. Each observation is indexed by both i (for the individual) and t (for the point in time). We observe N units over T time periods, so the number of observations is NT. A major advantage of data that has both individual and time variation is that we can account for certain kinds of unobserved variation.

1.2 Simple Linear Regression

1.2.1 Conditional Expectations

Let us start thinking about regression by looking at the newspaper headlines in Figure 1.3 and asking ourselves: What do all of these headlines have in common?

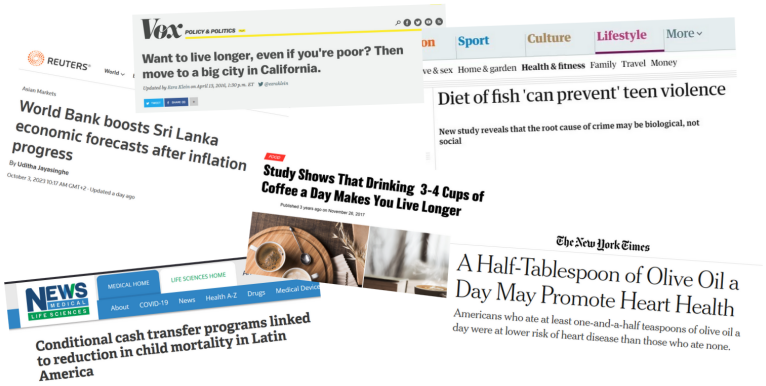


Figure 1.3: Some newspaper headlines

Actually, all of the headlines contain a statement about the *conditional expectation* of a variable, given another variable.

- Your life expectancy *given* low income and the place you live in.
- Expected teen violence *given* how much fish those teens have previously eaten.
- The predicted economic development of a country *given* its past inflation record.
- Again, your life expectancy, this time *given* the amount of coffee you consume.
- Expected child mortality *given* participation in a conditional cash transfer program.
- Expected heart health *given* the amount of olive oil you consume.

In somewhat more formal terms we can say: The statements concern the *conditional expectation* of a *dependent variable*, often denoted y, and an *explanatory variable*, often denoted x. More briefly:

$$\mathrm{E}(y \mid x). \tag{1.3}$$

We frequently encounter statements of this form, and it is important that we know how to interpret them. Some of the statements in Figure 1.3, for instance, are at the very least highly questionable. In this course, we will learn about how we can extract useful information from data we observe, how to correctly interpret what we learn, and importantly, how to avoid misleading interpretations. One implication of this is that we will be able to tell *why* some of the statements presented above are nonsense.

Back to conditional expectations for now. We can model a conditional expectation more explicitly, for example like this:

$$\mathrm{E}(y \mid x) = 0.4 + 0.5x. \tag{1.4}$$

We are relating the dependent variable y to the explanatory variable x by imposing a linear relationship between the expectation of y (which is different from just saying "y"!) and x. This way, we can divide all variation we observe in y into two parts:

1. Variation that stems from the explanatory variable x,
2. and other variation, which can either be random or explained by unobserved factors, that is, variables other than x.

To understand this better, you can think of a simple example. Imagine you collect data from a group of people, for example a class of school children, about their height and body weight. Say also you want to model the conditional expectation of body weight, given a person's height. To get a more intuitive understanding about what that means, let us start by thinking about the expectation of body weight, not conditional on anything:

$$\mathrm{E}(\text{weight}). \tag{1.5}$$

The measure in our sample that is analogous to the expected value is just the sample mean. In this first step, we are talking about a constant, just one single value. There is, of course, still variation in the variable y, and there is no variation in the expectation of y, since it is just one value. This means that using a plain expected value, we cannot explain any variation at all.

If we go one step further and consider a *conditional expectation*, we can explain some variation in the dependent variable. In our example, the explanatory variable is continuous, but if you have trouble imagining what all of this means, it may be easier to think of a discrete variable, say one that takes the values "short" and "tall." We now allow for two different expected values, *conditional* on being either short or tall. The variation between the short mean and the tall mean is now captured, and less unexplained variation remains.

If we now go back to a continuous explanatory variable, i.e. people's height as a number, we can write down an equation that looks somewhat like the following:

$$\mathrm{E}(\text{weight} \mid \text{height}) = \text{constant} + \text{slope} \cdot \text{height}. \tag{1.6}$$

It is important to remember that we do not have weight on the left hand side, but the *conditional expectation* of weight given height. The difference between the two is the unexplained variation, which is not captured by our simple linear model. This variation can be random, or it can be due to unobserved factors (in this case, you can think of age, gender, genetics, and so forth).

Actually, briefly dividing people into two height groups was not only useful as a thought experiment for understanding how conditional expectations work; it also serves as a nice introduction to one especially common use case for thinking in this framework. Often, we will be interested in *evaluating* a certain measure, for example a policy, a drug, or anything that lets us divide people into different groups. If we are evaluating the effect of a drug, say, in a randomized double-blind medical trial, we are interested in

$$\mathrm{E}(\text{Health} \mid \text{Drug} = 1) - \mathrm{E}(\text{Health} \mid \text{Drug} = 0), \tag{1.7}$$

where $\text{Drug} = 1$ means that somebody received a drug and $\text{Drug} = 0$ means they did not.

Another example is evaluating a conditional gender pay gap:

$$\mathrm{E}(\log(\text{wage}) \mid \text{Male} = 1, \dots) - \mathrm{E}(\log(\text{wage}) \mid \text{Male} = 0, \dots). \tag{1.8}$$

Here, the $\cdots$ indicate that we are also conditioning on other variables. If we are additionally conditioning on education, we are evaluating the gender pay gap for a given level of education. As before, Male $= 1$ means that somebody is male, and Male $= 0$ means that they are not.

The difference in both of these cases is what we will later call an *average treatment effect*, and the binary variable we condition on will be called a *treatment*. This is straightforward in the context of the drug trial, and less straightforward if the treatment is being male; but it is still the commonly used terminology.

1.2.2 The Bivariate Linear Model

We have seen that conditional expectations can be used to relate two variables. We now do this more explicitly: We are going to talk about how we can model the *conditional expectation function* of a given random variable, call it y, depending on another variable, here called x. There are many possible ways to do that, but we are going to start with the simplest one: a linear function. We can write:

$$\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i. \tag{1.9}$$

This is actually exactly what we already did, sneakily, in Equations 1.4 and 1.6. There are just a few new bits and pieces of notation that we need to clarify at this point. For one, both y and x are now indexed by i. i is an index for an *observation*. For example, say we are surveying a class of high school students, and there are 23 students in the class. Each of the students is then given an index $i = 1, \ldots, 23$, and we have 23 observations in our dataset. Using this index is a way of denoting that the equation is valid for each observation. You could, for example, replace i by 7, look up the values for y_7 and x_7 in the dataset you collected, and then plug them into the equation.

The remaining parts of the equation are not complicated, either. β_0 and β_1 are what we call *parameters*. You know how these work from learning about linear equations in high school: β_0 is the *intercept* or *constant parameter*, and β_1 is the *slope parameter*. So far, we have called y the *dependent variable*, but there is a myriad of other terms you will encounter. The most common are: *explained variable*, *outcome variable* or *regressand*. Likewise, x, which we called the *explanatory variable*, can also be called *independent variable* or *regressor*.

Back to what the conditional expectation function in Equation 1.9 actually does: It gives us information about the expected value of y_i for a given value of x_i. It is important that we read this statement precisely, because the conditional expectation function does only that: inform us about the expected value. We explicitly cannot infer the actual value of y_i (as opposed to what we expect) for a given x_i. We also gain no other information, such as the distribution of y_i and x_i. We only learn about the conditional expectation.

This is maybe best illustrated using an example. Say we want to learn about whether students in districts where school classes are smaller on average perform differently in an exam. We model the conditional expectation of test scores as follows:

$$\mathrm{E}(\text{Test Scores}_i \mid \text{Class Size}_i) = 720 - 0.6 \times \text{Class Size}_i. \tag{1.10}$$

What can we say about a new district (i.e., a district for which we have not recorded any test score data and know only its average class size) where classes have 20 students on average?

- The expected value for the test scores in that district is 708 points.

- The actual test scores can be higher or lower than that.
- This is because there is some *error*, an unobserved component, which the conditional expectation function has not accounted for.

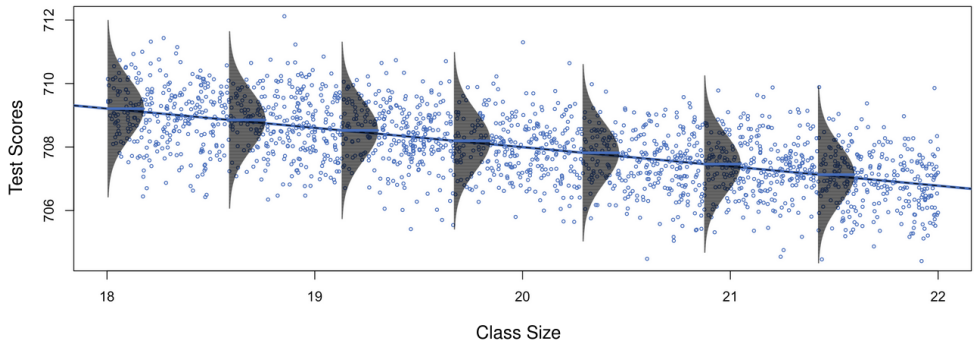


Figure 1.4: Class size, test scores and the conditional expectation function.

This is illustrated in Figure 1.4. The blue line is the conditional expectation function. At a class size of 20, the value of this function is 708. This means that the *expected value* of the test score is 708 conditional on a district having an average class size of 20. You can also see a large amount of small blue dots in the figure. These represent actual observations (in this case, the data is not real and those observations are simulated): Each dot is one combination of an average class size and a test score value. For example, you can see districts that have 20 students per class with a test score average as low as 706, and you can see others with equally large classes that score more than 710 points on average. This is reflected by the multiple small bell curves along the line: There is a good chance that some of the values we observe will be very close to the conditional expected value we modeled, but some of them will also be very far from that expectation.

We are now ready for a very important step: We can combine our thoughts about the *conditional expectation function* and the unobserved *error* that makes actual values stray a bit from our expectation. When we do that, we end up with a *linear regression model*:

$$y_i = \beta_0 + \beta_1 x_i + u_i. \tag{1.11}$$

This, again, looks very similar to what we had before. The only difference is that we added u_i, the unobserved component. We also call this unobserved component *error* or *error term* of the population. Sometimes, this error will be negative; sometimes, it will be positive – allowing for both values that exceed our conditional expectations or values that are lower than it. The introduction of this error term allows for the second important change in this equation compared to before: We now have y_i, the actual value, on the left-hand side, instead of its expectation. The rest of the equation is very similar to before: $\beta_0 + \beta_1 x_i$, which we call the *regression function* of the population, contains β_0, an intercept, and β_1, a slope parameter. The intercept represents the (theoretical) value we expect for $x_i = 0$, and the slope represents the expected change in predicted values for y_i if x_i changes by one unit.

We have used the term *population* a few times now. We will talk about what it means in more detail later on. But for now, you can think of it like this: Imagine a universe of data, for example, all high school students that exist anywhere. Then you choose some areas, and in those areas, you choose some students, for which you collect data. The students for

which you have data in your dataset will be your *sample*, and the entire universe of school students that exist anywhere will be called the *population*.

Let us look once more at the example from before to deepen our understanding of what the *slope* and *intercept* parameters do:

$$\text{Test Scores}_i = \beta_0 + \beta_1 \times \text{Class Size}_i + u_i, \tag{1.12}$$

or alternatively, at its conditional expectation function counterpart,

$$\text{E}(\text{Test Scores}_i \mid \text{Class Size}_i) = \beta_0 + \beta_1 \times \text{Class Size}_i. \tag{1.13}$$

In this case, the *slope* parameter is given by the following expression:

$$\beta_1 = \frac{\text{d E}(\text{Test Scores}_i \mid \text{Class Size}_i)}{d\ \text{Class Size}_i}. \tag{1.14}$$

In other words, β_1 is the expected difference in test scores when we change the average class size in a district by one unit.

The *intercept* parameter,

$$\beta_0 = \text{E}(\text{Test Scores}_i \mid \text{Class Size}_i = 0), \tag{1.15}$$

is the expected value for the test score in a district when the average amount of students per class there is zero. Of course, this makes no sense, there are no classes with zero people, especially not on average, and so on, and so forth. But you know what a linear function is, and so you also know that this is just a way by which we can shift the regression line up- or downwards.

Actually, while we are still at the beginning of our journey to understand linear regression, this is a good point to ask a first question about a simple mathematical property of those parameters: How do they change when we *scale* our variables? Consider for this the following regression:

$$\text{Test Scores}_i = \beta_0^\bullet - \beta_1^\bullet \times \frac{\text{Class Size}_i}{10} + u_i. \tag{1.16}$$

How do, all else kept constant, $\beta_0^\bullet$ and $\beta_1^\bullet$ change with respect to β_0 and β_1? A simple look at Equation 1.16 tells us: $\beta_0^\bullet$ is going to be the same as β_0 before, but $\beta_1^\bullet = 10 \times \beta_1$. If we change the scale of the independent variable, the associated slope parameter also needs to be scaled. If we instead scale the dependent variable like this:

$$\frac{\text{Test Scores}_i}{10} = \beta_0^\circ - \beta_1^\circ \times \text{Class Size}_i + u_i, \tag{1.17}$$

we can see that both $\beta_0^\circ = 10 \times \beta_0$ and $\beta_1^\circ = 10 \times \beta_1$, that is, all parameters are scaled correspondingly. If you wonder why we would want to scale variables, consider that sometimes we have lengths, distances, weights, temperatures, ... as variables, which are commonly measured in different units. If we deal with such types of variables, knowing about the effects of scaling one of them is very useful.

1.2.3 The OLS Estimator

Remember when we introduced the terms *sample* and *population*? This distinction becomes important now. All of what we discussed so far concerned relationships in the *population*. The regression model of the population describes a hypothetical relationship between different variables. To simplify things, we can think of the data being "generated" by the population regression function (PRF) and the error term. Together, we sometimes refer to this as a *data-generating process* (DGP). The core problem we are dealing with is that we do not know, and cannot observe, the parameters β_0 and β_1.

Therefore, we need to *estimate* those parameters. To do that, we also need data. The data we can use for estimating the parameters is what we call our *sample*. In the following, we will discuss how we can use our *sample* data to estimate the population parameters. This also means that we will encounter equations that look awfully similar to what we have seen before, but now they concern the sample, not the population. It is important to keep this in mind when discussing estimation of parameters. In short: There is a *population* and there are relationships between variables in that population which we are interested in. But we only have a *sample* of the data, and so we need to estimate these relationships based on the sample.

Every try at parameter estimation thus starts with an effort to collect a sample. Let us, for now, model data collection like this:

$$\left.\begin{matrix}\{y_1, x_1\}\\ \{y_2, x_2\}\\ \{y_3, x_3\}\\ \vdots\\ \{y_N, x_N\}\end{matrix}\right\} \{y_i, x_i\}_{i=1}^{N} \qquad \text{randomly drawn from a population } F_{y,x}(\cdot,\cdot). \tag{1.18}$$

Maybe this looks intimidating. But it actually is very simple if we read it from left to right. We have multiple pairs of data points, for example $\{y_1, x_1\}$, that all consist of one realization of the outcome variable, in this case y_i, and one realization of the explanatory variable, in this case x_1. Together, those two represent one *observation*. We can also see that there are multiple observations, from $1, 2, 3, \ldots$ all the way to N. N is the number of observations of our sample, also called the *sample size*. Because this is cumbersome to write, and takes up a lot of space, we can summarize this in one line using the observation index i: $\{y_i, x_i\}_{i=1}^{N}$. Here, $\{\}_{i=1}^{N}$ simply means, the index is i, it starts at 1, and we count in steps of 1 up to the number N. The remaining bit of Equation 1.18, $F_{y,x}(\cdot,\cdot)$, simply represents some joint distribution of the two random variables y and x.

We now want to approximate $\mathrm{E}(y \mid x)$, the conditional expectation of y given x in the population, using a linear conditional expectation function – and all of this only using the sample we have. We will do this by *fitting* a *regression line* with intercept $\tilde{\beta}_0$ and slope $\tilde{\beta}_1$:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \hat{u}_i \tag{1.19}$$

that minimizes the following prediction error:

$$\hat{u}_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i. \tag{1.20}$$

Very importantly, this prediction error $\hat{u}_i$, often called the *residual*, is not the same as the population error term. The *residual* is the difference between an actual observed value y_i and the predicted value $\hat{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$. The *error term* is the random or unobserved

component from the data-generating process of the population. Similarly, the estimated *regression coefficients* for intercept and slope, $\tilde{\beta}_0$ and $\tilde{\beta}_1$, are not the same as the *population parameters* β_0 and β_1.

There are, of course, many different ways we could construct an estimator. An *estimator*, strictly speaking, is any function of the data we have that yields an estimate for a population parameter. To decide on one of them, we need a metric of how "good" they are, or in other words, how well the regression line implied by their estimates fits the data. Since the residuals $\hat{u}_i$ are a measure of the difference between actual observed y_i and their prediction, they are a good starting point.

You can think of the situation like this: There are many candidate estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$, among which we want to find the best in terms of minimizing prediction errors. The only thing we need now is a metric we can minimize. Taking the sum of all residuals would make no sense, since positive and negative residuals would cancel each other out. Absolute values are also inconvenient, since we are going to take derivatives later on. Thus, *squaring* residuals and then *adding* them together is our best bet. In other words, we will find our estimator by minimizing the *sum of squared residuals*:

$$S\left(\tilde{\beta}_0, \tilde{\beta}_1\right) = \sum_{i=1}^{N} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2. \tag{1.21}$$

In a straightforward fashion, we will call the resulting estimator the *least squares estimator*, or more elaborately the *ordinary least squares* (OLS) estimator.

To find the OLS estimator for β_0 and β_1, we begin by taking the derivative of the sum of squared residuals with regard to $\tilde{\beta}_0$ and setting it equal to zero:

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_0} = -2 \sum_{i=1}^{N} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) \stackrel{!}{=} 0. \tag{1.22}$$

By rearranging and pulling $\tilde{\beta}_0$ out of the summation operator, we get

$$\sum_{i=1}^{N} y_i = N\tilde{\beta}_0 + \tilde{\beta}_1 \sum_{i=1}^{N} x_i. \tag{1.23}$$

In the next step, we differentiate with respect to $\tilde{\beta}_1$ and set equal to zero:

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^{N} x_i \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) \stackrel{!}{=} 0, \tag{1.24}$$

and rearrange in a similar way:

$$\sum_{i=1}^{N} x_i y_i = \tilde{\beta}_0 \sum_{i=1}^{N} x_i + \tilde{\beta}_1 \sum_{i=1}^{N} x_i^2. \tag{1.25}$$

Going forward, we will use the shorthand notations $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ for the sample means of x and y. Then, Equation 1.23 becomes (note that we can just multiply the sums with $\frac{N}{N}$):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{1.26}$$

This is already the OLS estimator for β_0. To also get the OLS estimator for β_1, we can now plug it into Equation 1.25. From this, we then get

$$\sum_{i=1}^{N} x_i (y_i - \bar{y}) = \tilde{\beta}_1 \sum_{i=1}^{N} x_i (x_i - \bar{x}). \tag{1.27}$$

By rearranging a little, we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i (y_i - \bar{y})}{\sum_{i=1}^{N} x_i (x_i - \bar{x})}. \tag{1.28}$$

This is already a valid expression for the OLS estimator for β_1. However, we can make use of the fact that $\sum_{i=1}^{N} x_i (x_i - \bar{x}) = \sum_{i=1}^{N} (x_i - \bar{x})^2$ and $\sum_{i=1}^{N} x_i (y_i - \bar{y}) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$[1] and rewrite it in a way that immediately allows for an intuitive interpretation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{\widehat{\mathrm{Cov}}(x, y)}{\widehat{\mathrm{Var}}(x)} \tag{1.29}$$

The OLS slope estimator equals the sample covariance between the explanatory and the dependent variable, divided by the sample variance of the explanatory variable. As a side note, you can also see that this requires that $\sum_{i=1}^{N} (x_i - \bar{x})^2 > 0$, or in other words, that the x_i values are not all the same. This is one of multiple assumptions that we will state more explicitly later on.

The same OLS estimator can also be derived using an alternative method called the *method of moments*. To discuss it, we first need to know what a *moment* is. Simply put, you can think of a moment as a quantity that summarizes the distribution of a random variable. The first moment describes its location, the second describes how much variation there is, the third describes whether it leans left or right, and the fourth describes the thickness of the distribution's tails. More formally, the k-th moment of a random variable is defined as

$$m_k(X) := \mathrm{E}\left(X^k\right), \tag{1.30}$$

and the k-th *central moment* of a random variable is defined as

$$\mu_k(X) := \mathrm{E}\left((X - \mu)^k\right), \qquad \text{where } \mu = \mathrm{E}(X). \tag{1.31}$$

The sample analogues, which you can think of as being "sample versions" of the population quantities, are

$$\hat{m}_k = N^{-1} \sum_{i=1}^{N} X_i^k \tag{1.32}$$

and

$$\hat{\mu}_k = N^{-1} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^k. \tag{1.33}$$

Sample analogues of raw moments are unbiased and consistent estimators of the population moments, but sample analogues of central moments are only consistent, and not generally unbiased.

[1] To see that $\sum_{i=1}^{N} x_i(x_i - \bar{x}) = \sum_{i=1}^{N} (x_i - \bar{x})^2$, substitute $x_i = (x_i - \bar{x}) + \bar{x}$ to get $\sum_{i=1}^{N} ((x_i - \bar{x}) + \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{N} (x_i - \bar{x})^2 + \bar{x} \sum_{i=1}^{N} (x_i - \bar{x})$. The second term is zero since $\sum_{i=1}^{N} (x_i - \bar{x}) = 0$. Similarly, $\sum_{i=1}^{N} x_i(y_i - \bar{y}) = \sum_{i=1}^{N} ((x_i - \bar{x}) + \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) + \bar{x} \sum_{i=1}^{N} (y_i - \bar{y}) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$.

We usually refer to some moments by their more common names. The first moment is also called the mean:

$$\mathrm{E}(X). \tag{1.34}$$

The second central moment is better known as the variance of a random variable:

$$\mathrm{E}\left((X-\mu)^2\right). \tag{1.35}$$

Similarly, the third and fourth central moments are also known as skewness and kurtosis.

After having discussed what moments actually are, we can go back to discussing the method of moments. The basic idea of this way to derive an estimator is that we choose parameters in a way that our model implies moments that equal the moments of the data we observe. We do this by following three simple steps:

1. The model implies certain moment conditions. A *moment condition* is a restriction stating that, if the model is correct, the expectation of a function of the data and the parameters equals zero.
2. Replace population moments in the moment conditions by sample moments.
3. Solve for the parameter.

The simple linear regression model implies two moment conditions. The first comes from the fact that the mean of the errors is zero,

$$\mathrm{E}(u_i) = 0 \tag{1.36}$$

This is very clearly a moment condition, since it is about the expected value of a function of the data and the parameter (since $u_i = y_i - \beta_0 - \beta_1 x_i$), and the condition implies that this should be zero if the model is correct. The second moment condition comes from the exogeneity assumption:

$$\mathrm{Cov}(x_i, u_i) = \mathrm{E}(x_i u_i) = 0. \tag{1.37}$$

This was the first step. The second step is to replace the population moments in those two moment conditions by the sample moments. For the first moment condition, we receive

$$N^{-1} \sum_{i=1}^{N} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0, \tag{1.38}$$

and for the second, we get

$$N^{-1} \sum_{i=1}^{N} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0. \tag{1.39}$$

Maybe you have noticed that these are equivalent to Equation 1.23 and Equation 1.25. This means that we can follow the same procedure as above for the third step, solving for the parameter, and that we will receive the same OLS estimators as when using the derivative method. In short, we have obtained the same estimator using two different methods.

1.2.4 Properties of the OLS Estimator

Having derived the OLS estimator, we are now ready to discuss some of its properties. For this, let us look again at one representation we had for the estimator:

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(x, y)}{\widehat{\text{Var}}(x)}. \tag{1.40}$$

For the first property of this estimator, we are going to take a closer look at the denominator. In order for us to be able to compute the estimator, $\widehat{\text{Var}}(x)$, the sample variance of the x_i values, is not allowed to be zero. This makes a lot of intuitive sense, too. To see why, consider Figure 1.5. In this setting, all x values are exactly the same, meaning that x has zero variance. There is no way for us to find a single "correct" value for the slope estimator. Each line we can fit through this data will make equally little sense.

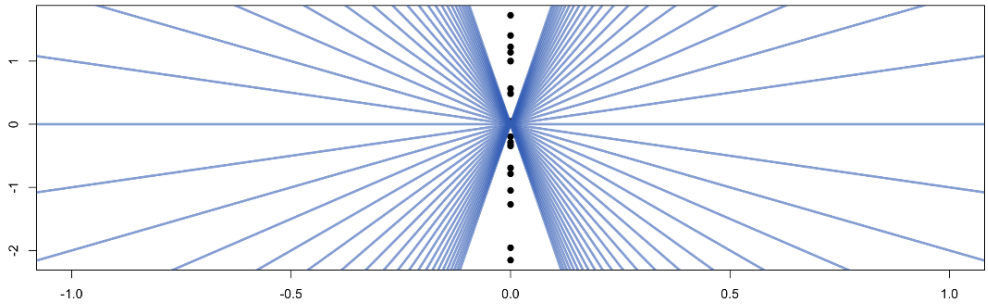


Figure 1.5: If all observations have the same x value, which line are we going to fit?

The second property we are going to discuss concerns the *residuals*. Again, it is important that we remember what the residuals are (the difference between the observed and the predicted value of the outcome) and what they are not (the errors). If you are unsure, go back to Equation 1.20 and re-read the explanation below. When we previously took the derivative of the loss function with respect to $\tilde{\beta}_0$, we got the following:

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_0} = -2 \sum_{i=1}^{N} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right) = 0. \tag{1.41}$$

Taking a closer look at this, the part within the parentheses is actually just the residual, implying that

$$\sum_{i=1}^{N} \hat{u}_i = 0. \tag{1.42}$$

In other words, this implies that the sum, and thus the *mean of the residuals is zero*. Intuitively, we can think of this property as follows: Imagine the mean of the residuals were not zero, but a small positive number. Then, observed values would on average be slightly higher than the fitted values. We could easily solve this by shifting our regression line upward. This would reduce the sum of squared residuals until the point where no systematic difference remains (the residual mean is zero). Note that the sum of *squared residuals* is not zero, as individual residuals are also not equal to zero, but that it is minimal at this point.

Which property can we get out of the other first order condition? Let us look at it one more time:

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^{N} x_i \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right) = 0. \tag{1.43}$$

This implies that $\sum_{i=1}^{N} x_i \hat{u}_i = 0$, which in turn implies[2] that

$$\sum_{i=1}^{N} (x_i - \bar{x})\, \hat{u}_i = 0. \tag{1.44}$$

This means that the correlation between the x_i and the residuals is zero. Again, an intuitive explanation in a two-dimensional scatterplot: Imagine the correlation between x and the residuals were positive. This would mean that higher x values would systematically have positive residuals, while lower x values would have negative residuals. We could easily fix this situation by just tilting the regression line a bit, thereby achieving a better fit.

Next, let us take a look at what we can say about the *variation in y values* we have:

$$\underbrace{\sum_{i=1}^{N} (y_i - \bar{y})^2}_{\text{variation in observed } y} = \underbrace{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}_{\text{variation in fitted } y} + \sum_{i=1}^{N} \hat{u}_i^2. \tag{1.45}$$

We can decompose[3] the variation in y values we observe into a part we can explain, which just means that this part of the variation also shows up in our fitted values, and another part we cannot explain, the sum of squared residuals. Often, this decomposition is written as follows:

$$\begin{aligned} &\text{Total sum of Squares (SST)} = \\ &\qquad \text{Explained Sum of Squares (SSE)} + \text{Residual Sum of Squares (SSR)} \end{aligned} \tag{1.46}$$

This gives us a chance to construct our first measure of *goodness of fit*, namely the *coefficient of determination*, often called by its shorter name R^2. It is defined as follows:

$$R^2 = \frac{\text{SSE}}{\text{SST}} \qquad = 1 - \frac{\text{SSR}}{\text{SST}}, \tag{1.47}$$

that is, R^2 indicates what proportion of the variation is explained by our model. If none of the observed variation is explained by our model, R^2 will be zero. If all of the variation is explained by the model, it will be one. In that latter case, all observations will lie exactly on the regression line.

Sometimes, R^2 is used to compare models. More often than not, this is a bad idea. Not only is there no "threshold" for what a "good" R^2 should be, having a high R^2 is also neither a necessary nor a sufficient condition for having a good model: There are bad models that fit a particular dataset well, but do not yield any useful conclusions; and there are models with a low R^2 that reveal important relationships. Anscombe's Quartet, depicted in Figure 1.6, illustrates this nicely: All four of the plotted models have an R^2 of 0.67, yet some are evidently better explanations of the data than others.

[2] Consider from the first FOC $\sum_{i=1}^{N} \hat{u}_i = 0$. Multiply with the constant $\bar{x}$ to get $\sum_{i=1}^{N} \bar{x}\hat{u}_i = 0$. Since $\sum_{i=1}^{N} x_i \hat{u}_i = 0$ (from the FOC w.r.t. $\tilde{\beta}_1$) and $\sum_{i=1}^{N} \bar{x}\hat{u}_i = 0$ (from the FOC w.r.t. $\tilde{\beta}_0$), $\sum_{i=1}^{N} (x_i - \bar{x})\, \hat{u}_i = 0$.

[3] Add and subtract $\hat{y}_i$ inside the square: $\sum (y_i - \bar{y})^2 = \sum ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 = \sum (\hat{u}_i + (\hat{y}_i - \bar{y}))^2 = \sum \hat{u}_i^2 + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 = \text{SSR} + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \text{SSE}$. The cross-term vanishes: $\sum \hat{u}_i (\hat{y}_i - \bar{y}) = \sum \hat{u}_i \hat{y}_i - \bar{y} \sum \hat{u}_i = \sum \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y} \sum \hat{u}_i = \hat{\beta}_0 \sum \hat{u}_i + \hat{\beta}_1 \sum \hat{u}_i x_i - \bar{y} \sum \hat{u}_i = 0$ by the FOCs. Thus SST = SSR+SSE.

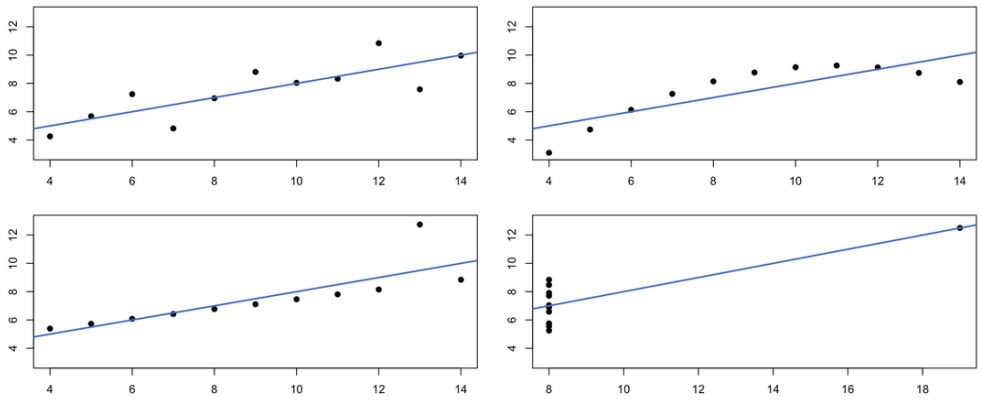


Figure 1.6: Anscombe's Quartet: In all of these, $R^2 = 0.67$.

1.2.5 Logarithmic Transformations

This section is motivated by a question you may have tacitly asked yourself before: Is it really sensible to assume that all relationships are linear? It is of course not, and this is why we will start discussing one of the ways to incorporate non-linearities relatively early on. It makes sense to use an example for this, so consider the following:

$$\text{wage}_i = f\left(\text{education}_i\right). \tag{1.48}$$

We are assuming that the wage a person earns depends on their education. Now consider the following: Which is more plausible? That an additional year of education always increases the wage by the same *amount*, or that it always increases the wage by the same *factor*? The latter is probably closer to the truth: Adding one year of education when you already have a lot, and thus earn a high wage, is probably going to lead to a comparatively larger absolute increase in your wage.

Fortunately, we can model this kind of relationship very easily. Consider the following model:

$$\text{wage}_i = e^{\beta_0} \cdot e^{\beta_1 \text{education}_i} \cdot e^{u_i}, \tag{1.49}$$

which can be written as

$$\text{wage}_i = \exp\left(\beta_0 + \beta_1 \text{education}_i + u_i\right). \tag{1.50}$$

This may look far from a linear relation that we can estimate using OLS, but it actually is not. To see this, we can logarithmize both sides of the equation:

$$\log\left(\text{wage}_i\right) = \beta_0 + \beta_1 \text{education}_i + u_i. \tag{1.51}$$

By defining $y_i^* = \log\left(y_i\right)$ and estimating

$$y_i^* = \beta_0 + \beta_1 x_i + u_i, \tag{1.52}$$

we are able to model a relationship that is non-linear in y (the wage) and x (the level of education), but is linear in $\log(y)$ and x. This relationship can then be interpreted as "an increase of x by some *amount* is associated to an increase of y by some *factor*."

We can apply the same transformation to the independent variable, x. We define $x_i^* = \log(x_i)$ and estimate

$$y_i = \beta_0 + \beta_1 x_i^* + u_i. \tag{1.53}$$

Analogously to before, this can be interpreted as "an increase of x by some *factor* is associated to an increase of y by some *amount*." We can also apply the logarithmic transformation to both the independent and the dependent variable. The resulting relationship can then be interpreted as "an increase of x by some *factor* is associated to an increase of y by some *factor*."

Most often, we use the *natural logarithm* for this because it has an interesting property that eases interpretation. To see this, consider the following:

$$\begin{aligned} \log(1.01x) &= \log(x) + \log(1.01) \\ &= \log(x) + 0.00995 \\ &\approx \log(x) + 0.01 \end{aligned} \tag{1.54}$$

An increase of x by 1 percent roughly corresponds to an increase in $\log(x)$ by 0.01. This means that we can interpret small changes in the logarithm as percentage changes of the untransformed variable. Figure 1.7 shows that this approximation is accurate for small percentage changes, but becomes less accurate for larger changes. For example, an increase in the logarithm by 0.5 corresponds to an increase in the untransformed variable of about 65 percent.

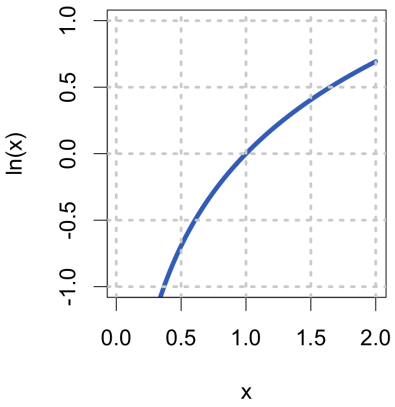


Figure 1.7: Plot of the relationship between the natural logarithm of x, and x.

Table 1.5 is useful for remembering how to interpret models where different variables were log-transformed:
Untransformed models allow us to make statements about the relationship between absolute changes in two variables. Models where one variable is transformed, but the other is not, allow us to make statements about *semi-elasticities*. The semi-elasticity we get out of the Log-Level model is referred to as the "semi-elasticity of y with respect to x," but the opposite term is not as frequently used for the Level-Log model, since it would be awkward to talk

Model	**Dep. Variable**	**Indep. Variable**	**Interpretation**
Level-Level	y	x	$+1$ in $x \Leftrightarrow +\beta_1$ in y
Level-Log	y	$\log(x)$	$+1\%$ in $x \Leftrightarrow +\beta_1/100$ in y
Log-Level	$\log(y)$	x	$+1$ in $x \Leftrightarrow +\beta_1 \times 100\%$ in y
Log-Log	$\log(y)$	$\log(x)$	$+1\%$ in $x \Leftrightarrow +\beta_1\%$ in y

Table 1.5: Interpretations of Level-Level, Level-Log, Log-Level, and Log-Log models.

about (semi-) elasticities of an independent variable. The fully logarithmized model allows us to make statements about *elasticities*.

1.2.6 The Gauss-Markov Theorem

In this section, we will talk about – and prove – one of the most basic econometric theorems, the *Gauss-Markov Theorem*. It is named after Gauss (1823), who published the first proof of the theorem, and Markov (1900), who published a slightly different version of the proof eighty years later. The Gauss-Markov Theorem states that the OLS estimator is the *best linear unbiased estimator*, often shortened as *BLUE*.

Let us break down that statement. First, we already know that the OLS estimator is a *linear estimator*. This just means that the OLS estimator is a linear function of the data. Now, we additionally need to prove that the estimator is *unbiased*. This means that the expected value of the estimator equals the true value of the parameter, in this case β. Finally, we then need to prove that among all linear unbiased estimators, the OLS estimator is the *best*. This is a way to express that it should have the lowest variance: Unbiasedness means that we are going to hit the target on average, and if we want to be best, we should ideally also stray as little as possible.

We are going to start by making explicit all assumptions that are needed for the proof to work. Then, in the next two sections, we are going to prove it step by step – first proving unbiasedness, and then showing that the OLS estimator has the lowest possible variance. To prove that the OLS estimator is BLUE, we need four assumptions:

1. Linearity in Parameters.
2. Random Sampling.
3. Variation in x.
4. Exogenous errors.

Assumption SLR.1 (Linearity in Parameters). The population regression function (PRF) must be linear in its parameters, that is, it should look like this:

$$y_i = \beta_0 + \beta_1 x_i + u_i. \tag{1.55}$$

What if we log-transform a variable? This is actually not a problem, as the only thing we require is linearity *in parameters*. If we have $\log(x_i)$ instead of x_i, y_i is not a linear function of x_i anymore, but it still is a linear combination of the parameters, β_0 and β_1. By saying "linear in parameters" instead of just calling it a "linear model," we make explicit that $y_i = \beta_0 + \beta_1 \log(x_i) + u_i$ is linear (in parameters), while $y_i = 1^{\beta_0} x_i^{\beta_1} + u_i$ is not. This assumption is just there to define the class of estimators we consider (linear estimators).

Assumption SLR.2 (Random Sampling). Our sample of N observations, $\{(y_i, x_i), i = 1, 2, \ldots, N\}$ must be a *random sample* from the population. The probability of including an observation must be equal for all, and it must not depend on who we sampled first.

This assumption is actually very easy to violate. Think, for example, about a situation where you sample only from a certain part of the population. This could be the case when you want to survey students, but only collect responses at lunchtime in the cafeteria. That way, you are systematically missing working students. Or you select part of the sample based on another part, for example, by randomly selecting only $N/2$ students and asking them to ask their best friends to also participate in the survey. Under this procedure, half of the sample is not randomly selected. We need this assumption to describe the population model using individual observations:

$$\mathrm{E}(y_i \mid x_1, \ldots, x_N) = \mathrm{E}(y_i \mid x_i) = \mathrm{E}(y \mid x). \tag{1.56}$$

Of course, we will have to deal with non-random samples very often. Econometric techniques to deal with this problem exist, we are just not going to go into them at this stage.

Assumption SLR.3 (Variation in x). To estimate our model, we need *variation in x*. This means that the x values are not allowed to be all exactly the same.

This is a very basic assumption: If this is not fulfilled, we cannot identify the parameter. This is the problem that we run into when we need to divide by the variance of x when computing the estimate: If that variance is zero, we cannot compute an estimate. When sampling from a population, this assumption is usually easily met – unless the sample is small and variation in the population is minimal in the first place. For an example of when this condition is not met, consider a regression of exam scores on studying time, but every student in the sample studied for exactly ten hours. There is no way we can infer anything from this. The opposite case, where there is no variation in the dependent variable, is unproblematic but boring: If everyone gets 100 points on the exam, but studied for different amounts of time, then we can conclude that studying had no effect on the exam score.

Assumption SLR.4 (Exogenous Errors). The expected value of the error term u must be 0 for every value of x:

$$\mathrm{E}(u_i \mid x_i) = 0. \tag{1.57}$$

Note that this assumption also implies[4] the two *moment conditions*, $\mathrm{E}(u_i) = 0$, and $\mathrm{E}(u_i x_i) = 0$.

We need this assumption because we work with expectations of the format $\mathrm{E}(\cdot \mid x_i)$ in a lot of proofs and derivations. This means that we "fix" the x values and imagine drawing many random samples for *exactly those x_i values*, but different u_i and therefore y_i. We say that x is fixed in repeated samples. Of course, this is unrealistic, especially when we talk about observational data. We are not going to get multiple samples with the same x_i. But this assumption essentially allows us to treat x_i values *as if they were* fixed.

This assumption is actually very easy to violate. This may be unexpected, since it looks so similar to $\mathrm{E}(u_i) = 0$, which we said was trivial. But consider the following example to see why $\mathrm{E}(u_i \mid x_i) = 0$ is far less trivial.

[4] Part 1: Apply the law of iterated expectations: $\mathrm{E}(u_i) = \mathrm{E}(\mathrm{E}(u_i \mid x_i)) = \mathrm{E}(0) = 0$. Part 2: Again apply LIE: $\mathrm{E}(u_i x_i) = \mathrm{E}(\mathrm{E}(u_i x_i \mid x_i)) = \mathrm{E}(\mathrm{E}(u_i \mid x_i) x_i) = \mathrm{E}(0 \cdot x_i) = 0$, since $\mathrm{E}(x_i \mid x_i) = x_i$.

In an experimental study, we randomly select a number of fields. Then, we randomly choose half of them to apply fertilizer. Finally, we record crop yields. In this experiment, the intervention (fertilizer use, x_i) is guaranteed to be independent of unobserved factors. The assumption that $\mathrm{E}(u_i \mid x_i) = 0$ is thus plausible.

Next, we conduct an observational study. We again randomly select a number of fields. Then, we ask farmers whether they applied fertilizer. We record fertilizer use and yields. Here, the intervention may not be independent of unobserved factors. What if fertilizer is used on less fertile fields to compensate? Or on very fertile fields to boost yields even more? Without additional information, there is no way to tell whether the treatment is independent of unobservable factors. If we believe that $\mathrm{E}(u_i \mid x_i) = 0$ is plausible, we thus need to make an *argument* for it.

1.2.7 Expected Value of the OLS Estimator

If Assumptions SLR.1 through SLR.4 hold, we can now prove that the OLS estimator is *unbiased*. We know that we call an estimator unbiased if its expected value equals the true value of the parameter in the population model. So we want to prove that

$$\mathrm{E}\left(\hat{\beta}_j\right) = \beta_j, \qquad j = 0, 1. \tag{1.58}$$

For the proof, we start with the expression for the OLS estimator that we have derived earlier:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})x_i} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})y_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i}. \tag{1.59}$$

As a first step, we split y_i into its components $(\beta_0 + \beta_1 x_i + u_i)$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{N}(x_i - \bar{x})x_i} \tag{1.60}$$

$$\hat{\beta}_1 = \frac{\beta_0 \sum_{i=1}^{N}(x_i - \bar{x}) + \beta_1 \sum_{i=1}^{N}(x_i - \bar{x})x_i + \sum_{i=1}^{N}(x_i - \bar{x})u_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i} \tag{1.61}$$

Now, because $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$ and $\frac{\sum_{i=1}^{N}(x_i - \bar{x})x_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i} = 1$, we can further simplify this to:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(x_i - \bar{x})u_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i}. \tag{1.62}$$

We have decomposed the estimator $\hat{\beta}_1$ into the true parameter, β_1, and the *sampling error*, a linear combination of the error terms: $\frac{\sum_{i=1}^{N}(x_i - \bar{x})u_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i}$. Our goal is now to show that this sampling error has an expected value of zero: If that is the case, then our proof of unbiasedness of the estimator is complete.

The expected value of $\hat{\beta}_1$ is

$$\mathrm{E}(\hat{\beta}_1 | x_1, \ldots, x_N) = \mathrm{E}\left(\beta_1 + \frac{\sum_{i=1}^{N}(x_i - \bar{x})u_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i} \middle| x_1, \ldots, x_N \right). \tag{1.63}$$

Since the true parameter β_1 is not a random variable, we can take it out of the expectation:

$$\mathrm{E}(\hat{\beta}_1|x_1,\ldots,x_N) = \beta_1 + \mathrm{E}\left(\left.\frac{\sum_{i=1}^N (x_i-\bar{x})u_i}{\sum_{i=1}^N (x_i-\bar{x})x_i}\right| x_1,\ldots,x_N\right) \tag{1.64}$$

Now, because $\mathrm{E}(x_i \mid x_i) = x_i$, we can take all x_i outside as well:

$$\mathrm{E}(\hat{\beta}_1|x_1,\ldots,x_N) = \beta_1 + \frac{\sum_{i=1}^N (x_i-\bar{x})\mathrm{E}(u_i|x_1,\ldots,x_N)}{\sum_{i=1}^N (x_i-\bar{x})x_i}. \tag{1.65}$$

This is where our assumptions come into play. First, Assumption SLR.2 allows us to simplify the expectations like this:

$$\mathrm{E}(\hat{\beta}_1|x_1,\ldots,x_N) = \beta_1 + \frac{\sum_{i=1}^N (x_i-\bar{x})\mathrm{E}(u_i|x_i)}{\sum_{i=1}^N (x_i-\bar{x})x_i}. \tag{1.66}$$

Now we are getting very close: $\mathrm{E}(u_i \mid x_i)$ is in the numerator, and Assumption SLR.4 says that it equals zero, so we get

$$\mathrm{E}(\hat{\beta}_1|x_1,\ldots,x_N) = \beta_1. \tag{1.67}$$

By the law of iterated expectations, $\mathrm{E}(\hat{\beta}_1) = \mathrm{E}(\mathrm{E}(\hat{\beta}_1 \mid x_1,\ldots,x_N))$, and thus:

$$\mathrm{E}(\hat{\beta}_1) = \beta_1. \tag{1.68}$$

This concludes the proof: The OLS estimator for β_1 is indeed unbiased. □

What is missing is a proof of the unbiasedness of $\hat{\beta}_0$. We start by writing $\hat{\beta}_0$ as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{1.69}$$

Now, because we already know that $\mathrm{E}(\hat{\beta}_1 \mid x_1,\ldots,x_N) = \beta_1$:

$$\begin{aligned}\mathrm{E}(\hat{\beta}_0 \mid x_1,\ldots,x_N) &= \mathrm{E}(\bar{y} \mid x_1,\ldots,x_N) - \mathrm{E}(\hat{\beta}_1\bar{x} \mid x_1,\ldots,x_N) \\ &= \mathrm{E}(\bar{y} \mid x_1,\ldots,x_N) - \mathrm{E}(\hat{\beta}_1 \mid x_1,\ldots,x_N)\bar{x} \\ &= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \\ &= \beta_0.\end{aligned} \tag{1.70}$$

We have proved that $\hat{\beta}_0$ is unbiased as well. □

1.2.8 Variance of the OLS Estimator

We now know that the OLS estimator is *unbiased*. But we said before that it is also the *best* linear unbiased estimator, meaning that it has the lowest possible variance. We are also going to prove that. To do this, we need to introduce an additional assumption:

Assumption SLR.5 (Homoskedasticity). The variance of the error term u_i is the same for all values of x_i:

$$\mathrm{Var}(u_i \mid x_i) = \mathrm{Var}(u_i) = \sigma^2. \tag{1.71}$$

The variance of the error term is a measure of the variation that is caused by unobserved factors. Now, we are assuming that this variance is constant for all x_i values, and that it

equals some value, which we have called σ^2. We did not need this assumption to prove that the OLS estimator is unbiased. But we do need it to show that it has the lowest possible variance. This assumption is extremely easy to violate. Think, for example, of a cross-sectional dataset of people's education and wages. People with higher education may have greater variation in what they earn. Later in the course, there is an entire chapter on how we can deal with situations like that one.

If Assumptions SLR.1 through SLR.5 are all fulfilled, then we can prove that the OLS estimator has the *lowest possible variance* among all linear unbiased estimators. Then, we say it is the *best* linear unbiased estimator. This property is also called *efficiency*. We are going to structure the proof as follows: First, we will show that the variance of the OLS estimator is

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}, \qquad \text{Var}(\hat{\beta}_0 \mid x_i) = \frac{\sigma^2 N^{-1} \sum_{i=1}^N x_i^2}{\sum_{i=1}^N (x_i - \bar{x})^2}, \tag{1.72}$$

and then, we will show that there cannot be any other linear unbiased estimator with a smaller variance.

We can start with the decomposition from before:

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \text{Var}\left(\beta_1 + \frac{\sum_{i=1}^N (x_i - \bar{x}) u_i}{\sum_{i=1}^N (x_i - \bar{x}) x_i} \middle| x_i \right) \tag{1.73}$$

To simplify our notation, we now define $w_i := \frac{x_i - \bar{x}}{\sum_{i=1}^N (x_i - \bar{x}) x_i}$, so this becomes

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \text{Var}\left(\beta_1 + \sum_{i=1}^N \frac{x_i - \bar{x}}{(x_i - \bar{x}) x_i} u_i \middle| x_i \right) = \text{Var}\left(\beta_1 + \sum_{i=1}^N w_i u_i \middle| x_i \right). \tag{1.74}$$

These *weights* w_i depend only on x_i and are thus fixed conditional on x_i. Additionally, we can now apply Assumption SLR.5:

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \sigma^2 \sum_{i=1}^N w_i^2 \tag{1.75}$$

We can now expand w_i again. Since $w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^N (x_i - \bar{x}) x_i}$, we also have

$$\sum_{i=1}^N w_i^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{\left(\sum_{i=1}^N (x_i - \bar{x}) x_i \right)^2} = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{\left(\sum_{i=1}^N (x_i - \bar{x})^2 \right)^2} = \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2}. \tag{1.76}$$

Thus,

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}. \tag{1.77}$$

This concludes the first part of the proof.

What remains to do is to show whether this variance is the smallest possible for any linear unbiased estimator. Let now $\tilde{\beta}_1$ be any other linear estimator. The difference between $\tilde{\beta}_1$ and the OLS estimator $\hat{\beta}_1$ is that instead of the OLS weights w_i, the other estimator uses arbitrary weights a_i:

$$\tilde{\beta}_1 = \sum_{i=1}^N a_i y_i = \sum_{i=1}^N a_i (\beta_0 + \beta_1 x_i + u_i). \tag{1.78}$$

Since these weights a_i, just like w_i, are based on the x_i values, we can apply SLR.4 to write the expectation as:

$$\mathrm{E}\left(\tilde{\beta}_1 \middle| x_i\right) = \beta_0 \sum_{i=1}^{N} a_i + \beta_1 \sum_{i=1}^{N} a_i x_i. \tag{1.79}$$

Because this is, by assumption, an unbiased estimator, this expression must equal β_1, which implies both that $\sum_{i=1}^{N} a_i = 0$ and that $\sum_{i=1}^{N} a_i x_i = 1$. We continue by expressing the weights of $\tilde{\beta}_1$ as the OLS weights, plus some difference:

$$a_i = w_i + d_i. \tag{1.80}$$

This allows us to rewrite the estimator, using the same decomposition as earlier for the OLS estimator, as:

$$\tilde{\beta}_1 = \beta_1 + \sum_{i=1}^{N} (w_i + d_i) u_i. \tag{1.81}$$

The variance of $\tilde{\beta}_1$ is thus:

$$\mathrm{Var}\left(\tilde{\beta}_1 \middle| x_i\right) = \sigma^2 \sum_{i=1}^{N} (w_i + d_i)^2 = \sigma^2 \sum_{i=1}^{N} \left(w_i^2 + 2 w_i d_i + d_i^2\right). \tag{1.82}$$

Because $\sum_{i=1}^{N} a_i = \sum_{i=1}^{N} (w_i + d_i) = 0$ and $\sum_{i=1}^{N} w_i = 0$, we also have

$$\sum_{i=1}^{N} d_i = 0. \tag{1.83}$$

Also, because $\sum_{i=1}^{N} (w_i + d_i) x_i = \sum_{i=1}^{N} w_i x_i + \sum_{i=1}^{N} d_i x_i = 1$,

$$\sum_{i=1}^{N} d_i x_i = 0. \tag{1.84}$$

Now, since $\sum_{i=1}^{N} d_i = 0$ and $\sum_{i=1}^{N} d_i x_i = 0$, the term $\sigma^2 \sum_{i=1}^{N} (w_i + d_i)^2$ becomes

$$\sum_{i=1}^{N} w_i d_i = \frac{\sum_{i=1}^{N} (x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} d_i = \frac{1}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sum_{i=1}^{N} x_i d_i - \frac{\bar{x}}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sum_{i=1}^{N} d_i = 0, \tag{1.85}$$

which means that the variance of $\tilde{\beta}_1$ is

$$\mathrm{Var}\left(\tilde{\beta}_1 \middle| x_i\right) \quad = \quad \sigma^2 \sum_{i=1}^{N} w_i^2 + \sigma^2 \sum_{i=1}^{N} d_i^2. \tag{1.86}$$

The difference between the variance of this estimator and the variance of the OLS estimator is the right-hand term, $\sigma^2 \sum_{i=1}^{N} d_i^2$. Since this term can never be negative, the variance of $\tilde{\beta}_1$ must always be greater than or equal to that of the OLS estimator $\hat{\beta}_1$. Thus, the OLS estimator has the lowest possible variance of all linear unbiased estimators. □

To find the variance of $\hat{\beta}_0$, we start at the expression we had for the estimator:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{1.87}$$

We begin by substituting $\beta_0 + \beta_1\bar{x} + \bar{u}$ for $\bar{y}$:

$$\hat{\beta}_0 = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}. \tag{1.88}$$

Next, we take the conditional variance:

$$\text{Var}(\hat{\beta}_0 \mid x_i) = \text{Var}\left((\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \mid x_i\right). \tag{1.89}$$

Since $\bar{x}$ is fixed conditional on x_i, we can treat it as constant. Also, we know that $\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$ and $\text{Var}(\bar{u} \mid x_i) = \frac{\sigma^2}{N}$, so

$$\begin{aligned}\text{Var}(\hat{\beta}_0 \mid x_i) &= \bar{x}^2\text{Var}(\hat{\beta}_1 \mid x_i) + \text{Var}(\bar{u} \mid x_i) + 2\bar{x}\text{Cov}(-\hat{\beta}_1, \bar{u} \mid x_i) \\ &= \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2} + \frac{\sigma^2}{N} + 2\bar{x}\text{Cov}(-\hat{\beta}_1, \bar{u} \mid x_i).\end{aligned} \tag{1.90}$$

The covariance term vanishes:

$$\begin{aligned}\text{Cov}(\hat{\beta}_1, \bar{u} \mid x_i) &= \text{Cov}\left(\frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2}, \frac{1}{N}\sum u_j\right) \\ &= \frac{1}{N\sum(x_i - \bar{x})^2}\sum_{i=1}^N (x_i - \bar{x})\text{Var}(u_i) \\ &= \frac{\sigma^2}{N\sum(x_i - \bar{x})^2}\underbrace{\sum_{i=1}^N (x_i - \bar{x})}_{=0} \\ &= 0.\end{aligned} \tag{1.91}$$

Thus, without the covariance term, we have:

$$\begin{aligned}\text{Var}(\hat{\beta}_0 \mid x_i) &= \bar{x}^2 \cdot \frac{\sigma^2}{\sum(x_i - \bar{x})^2} + \frac{\sigma^2}{N} \\ &= \sigma^2\left(\frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} + \frac{1}{N}\right) \\ &= \sigma^2 \cdot \frac{N\bar{x}^2 + \sum(x_i - \bar{x})^2}{N\sum(x_i - \bar{x})^2}.\end{aligned} \tag{1.92}$$

Finally, we use $\sum(x_i - \bar{x})^2 = \sum x_i^2 - N\bar{x}^2$, and thus $N\bar{x}^2 + \sum(x_i - \bar{x})^2 = \sum x_i^2$:

$$\text{Var}(\hat{\beta}_0 \mid x_i) = \frac{\sigma^2\sum_{i=1}^N x_i^2}{N\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{\sigma^2 N^{-1}\sum_{i=1}^N x_i^2}{\sum_{i=1}^N (x_i - \bar{x})^2}. \tag{1.93}$$

This is the variance of $\hat{\beta}_0$.

If we now go back to the variance of $\hat{\beta}_1$, one problem remains: We know that the variance is

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}, \tag{1.94}$$

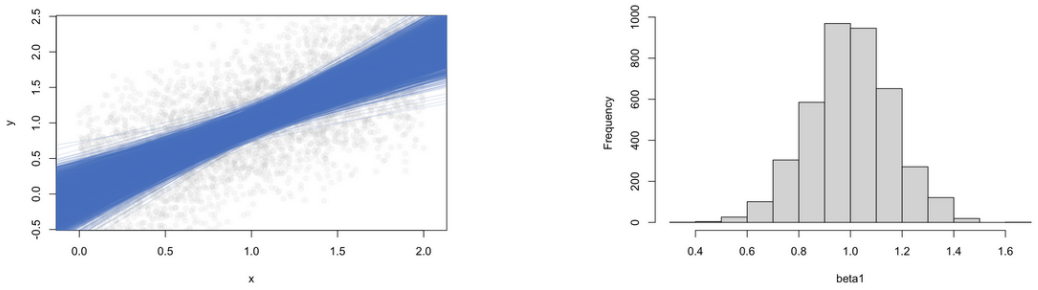


Figure 1.8: Estimating $\hat{\beta}_1$ 4000 times.

but we do not know σ^2, since it is the variance of the unobserved error term. Fortunately, under Assumptions SLR.1 through SLR.5, there exists an unbiased estimator for the variance: The residual sum of squares, divided by $N-2$;

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-2}. \tag{1.95}$$

If we take the square root of this estimator, we get

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}. \tag{1.96}$$

We call this value the *standard error of regression*. It is a consistent (albeit not unbiased[5]) estimator for σ. We can use it to compute the standard error of $\hat{\beta}_1$, which is an estimator of the standard deviation of $\hat{\beta}_1$:

$$\text{se}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2}} \tag{1.97}$$

Similarly, we can compute the standard error of $\hat{\beta}_0$. Knowing the standard errors of the estimators allows us to measure how "precisely" the coefficients are estimated.

Figure 1.8 shows the results of a simulation. 4000 samples were drawn for a population, and the coefficient β_1 was estimated 4000 times. The true value of the parameter was 1. We can see that the draws average about a value of 1, but that they stray quite a bit. Since this is simulated data, we know both the standard deviation of $\hat{\beta}_1$ as well as its standard error (the estimator for the standard deviation): The standard deviation is 0.161, and the standard error is 0.1638.

1.2.9 Regressions with Only One Parameter

Up to now, we have only considered regression models with two parameters, that is, models of the form

$$y_i = \beta_0 + \beta_1 x_i + u_i. \tag{1.98}$$

What happens if we drop one of the parameters and instead estimate $y_i = \beta_1 x_i + u_i$ or $y_i = \beta_0 + u_i$?

Let us first consider the *model without an intercept*,

[5] Unbiasedness means that the expected value of an estimator equals the true parameter, while consistency means that the estimator converges in probability to the true parameter as the sample size increases.

$$y_i = \beta_1 x_i + u_i. \tag{1.99}$$

We can derive the OLS estimator for this case by minimizing the loss function

$$\min_{\beta_1} S(\beta_1) = \sum_{i=1}^{N} (y_i - \beta_1 x_i)^2. \tag{1.100}$$

The (single) first-order condition is

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{N} x_i (y_i - \beta_1 x_i) \stackrel{!}{=} 0. \tag{1.101}$$

Rearranging yields

$$\sum_{i=1}^{N} x_i y_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \tag{1.102}$$

$$\beta_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \tag{1.103}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}. \tag{1.104}$$

Estimating this model is equivalent to imposing the restriction that $\beta_0 = 0$ on the full simple linear regression model. In the case where the true parameter β_0 is actually zero, this still yields an unbiased estimator. Figure 1.9 depicts a situation like this. When, however, the true parameter β_0 is not equal to zero, then our situation looks more like the one in Figure 1.10: Imposing the no-intercept restriction yields biased estimates for β_1.

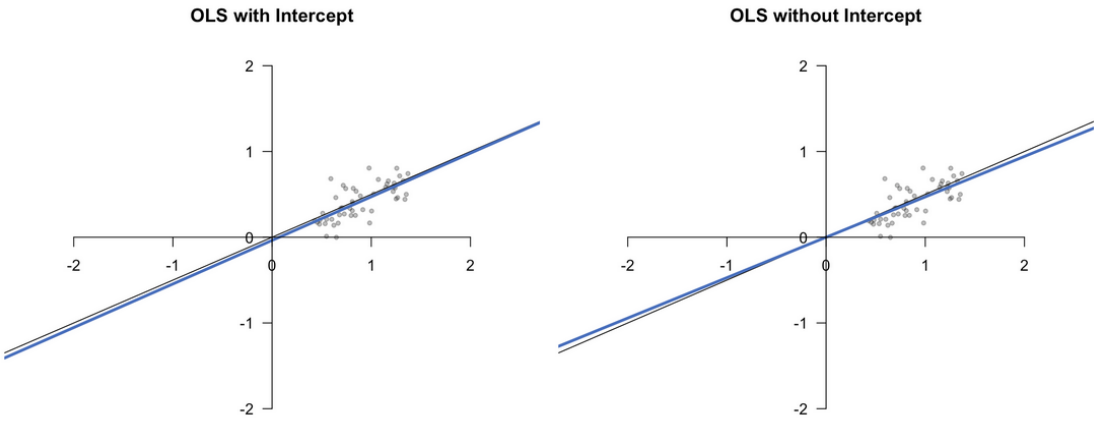


Figure 1.9: OLS with and without intercept, in the case where the true $\beta_0 = 0$.

So, when should we actually estimate a model with no intercept? The answer is: almost never. If we are completely certain that the true intercept is zero, then estimating a model with an intercept would amount to imposing unnecessary structure. In this case, estimating a model with no intercept is preferable. However, it is hard to imagine a situation like this actually occurring, and even if we are somewhat sure that the intercept should be zero, we

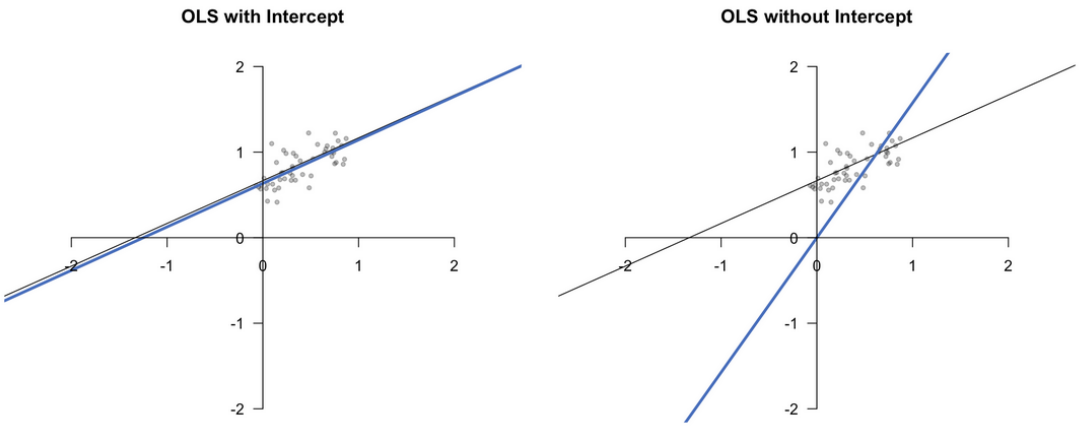


Figure 1.10: OLS with and without intercept, in the case where the true $\beta_0 \neq 0$.

can never be certain. So, except for rare cases where we have overwhelming theoretical justification to run a model without an intercept, we should never do this.

Proving that this estimator is biased when there is a non-zero intercept is actually very simple. We start by doing the same we did in the unbiasedness proof, substituting the true data-generating process (which includes the intercept) in the expression for the no-intercept estimator:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^N x_i(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^N x_i^2} \\ &= \frac{\beta_0 \sum_{i=1}^N x_i + \beta_1 \sum_{i=1}^N x_i^2 + \sum_{i=1}^N x_i u_i}{\sum_{i=1}^N x_i^2} \\ &= \beta_1 + \beta_0 \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N x_i^2} + \frac{\sum_{i=1}^N x_i u_i}{\sum_{i=1}^N x_i^2}.\end{aligned} \tag{1.105}$$

Taking the expectation yields

$$E(\hat{\beta}_1) = \beta_1 + \beta_0 \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N x_i^2}. \tag{1.106}$$

The rightmost term is the bias. We can see that the bias is only zero if the true $\beta_0 = 0$ (i.e., the true DGP had no intercept at all), or the regressor x is mean-zero. We can also see that the bias increases with the size of the true β_0 as well as the mean of x, and it decreases when the variance of x is higher.

Now, what happens if instead of dropping β_0 from the model, we estimate a regression with only the intercept? We get the following model:

$$y_i = \beta_0 + u_i. \tag{1.107}$$

It should come as no surprise that the estimator for β_0 will equal the sample mean of the outcome. But we can still formally derive that. We start again by minimizing the loss function

$$\min_{\beta_0} S(\beta_0) = \sum_{i=1}^{N} (y_i - \beta_0)^2, \tag{1.108}$$

which yields the following first-order condition:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{N} (y_i - \beta_0) \stackrel{!}{=} 0 \tag{1.109}$$

Solving:

$$\begin{aligned} \sum_{i=1}^{N} (y_i - \beta_0) &= 0 \qquad (1.110) \\ \sum_{i=1}^{N} y_i - N\beta_0 &= 0 \\ N\beta_0 &= \sum_{i=1}^{N} y_i \\ \hat{\beta}_0 &= \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}. \end{aligned}$$

1.2.10 Binary Explanatory Variables

So far, we have only considered cases where the explanatory variable is continuous, such as class size, years of education, and so on. In other words, we have discussed how to treat explanatory variables that have a *quantitative interpretation*. But we can easily include *qualitative information* in our model, by using variables that encode characteristics as discrete information.

Suppose, for instance, that we want to analyze the gender pay gap. We are therefore interested whether a given individual i is a woman or not. We can start by defining a variable woman_i like this:

$$\text{woman}_i = \begin{cases} 1 & \text{if } i \text{ is a woman,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.111}$$

In econometrics, we call this type of variable a *binary variable* or *dummy variable*. We can also use it to indicate treatment: Say we are interested in the effects of a job training program. Then we can define a variable

$$\text{participation}_i = \begin{cases} 1 & \text{if } i \text{ participated in the job training program,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.112}$$

How do we interpret the results of our models if the explanatory variable is binary? Consider a model of the form

$$y_i = \beta_0 + \beta_1 x_i + u_i, \tag{1.113}$$

where x is a dummy variable. Because Assumptions SLR.1 through SLR.5 still hold,

$$\mathrm{E}(y_i \mid x_i = 1) = \beta_0 + \beta_1, \text{and} \tag{1.114}$$
$$\mathrm{E}(y_i \mid x_i = 0) = \beta_0. \tag{1.115}$$

This means that β_1 can be interpreted as the expected difference in the outcome between the two groups (the one where $x = 0$ and the one where $x = 1$), and that β_0 is the expected value in the group where $x = 0$. This implies that $\beta_0 + \beta_1$ is the expected value of y for the group where $x = 1$.

You may ask yourself, what if we have categorical information that has more levels than just "yes" and "no"? Maybe you want to incorporate information on whether someone is a woman, a man, or neither. Or you want to evaluate the effects of multiple different training programs. Actually, this is very simple: We just use separate dummy variables for each level. But this requires multiple regression techniques, which we have not discussed yet, so we will defer any discussion of this to the following chapters.

1.2.11 Introduction to Causal Inference

We are going to conclude this chapter with a short introduction to one of the reasons why we are actually doing all of this: We want to answer *causal questions*. One of the most common types of causal question is that we have some kind of *treatment*, often also called an *intervention*, that we want to evaluate.

Knowing *dummy variables* gives us a chance to encode the notion of a treatment in our econometric framework. Essentially, we define a variable treatment_i, where

$$\text{treatment}_i = \begin{cases} 1 & \text{if } i \text{ is treated,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.116}$$

By assigning these values to every individual based on their treatment status, and subsequently including this variable as an explanatory variable in our regression, we can divide the sample into a *treatment group* and a *control group*.

For every individual i, there are now *two possible outcome states*: $y_i(1)$ is the outcome for individual i if it received the treatment, and $y_i(0)$ is i's outcome if it did not receive the treatment. The treatment effect is then simply the difference,

$$\text{causal effect}_i = y_i(1) - y_i(0). \tag{1.117}$$

The problem with this is obvious: We can only ever observe one of the two states. We cannot travel to an alternative reality where i was (not) treated, record that outcome, and come back to compare it to the other outcome. This other, unobserved state is called the *counterfactual outcome*; and the fact that we cannot observe it is called the *fundamental problem of causal inference*.

Looking again at causal effect_i, we notice that it has an i subscript, meaning that it may vary across individuals. While it is infeasible to retrieve this individual causal effect (for reasons discussed above), we can compute the *average treatment effect* (ATE):

$$\text{ATE} = \mathrm{E}(\text{causal effect}_i) = \mathrm{E}(y_i(1) - y_i(0)) = \mathrm{E}(y_i(1)) - \mathrm{E}(y_i(0)). \tag{1.118}$$

If Assumptions SLR.1 through SLR.4 hold, then the OLS estimator for β_1 in the regression

$$y_i = \beta_0 + \beta_1 \text{treatment}_i + u_i \tag{1.119}$$

is an unbiased estimator of the average treatment effect.

This brings us back to something we have discussed before: Assumption SLR.4, which in this context means that the errors are independent of whether somebody is in the treatment group or in the control group, is only guaranteed if the assignment to the treatment group is random. In a randomized controlled trial, this is the case. In an observational study, this is not necessarily the case. For cases where random treatment assignment is not possible, simple linear regression methods cannot yield valid statements about treatment effects. In these cases, we need more advanced methods. The first step in that direction are *multiple linear regression* methods, which we will discuss in the following chapter.

1.3 Multiple Linear Regression

1.3.1 Introduction

In the previous section, we have discussed the basics of *bivariate regression* models, that is, models with two variables, a dependent variable y and an independent variable x. Expectedly, the applications for models like these are limited. In reality, dependent variables are affected by multiple independent variables at the same time, and ideally, we want to model this explicitly. In addition, being able to consider multiple explanatory variables also allows us to include multiple dummy variables at the same time. We have already mentioned that this allows us to encode qualitative information with multiple categories.

For all of this, we need *multivariate regression* models. As the name says, these include more than two variables: one dependent variable y and multiple independent variables. We denote those $x_1, x_2, \ldots, x_K$, which implies that the number of explanatory variables in our model is K. Multivariate regression is similar to bivariate regression, but we still need to explicitly generalize most of what we found in the previous section. Therefore, the contents of this section closely mirror what we discussed before, but now we consider more than two variables.

We will start by discussing the simplest possible extension of the bivariate model: a model with two explanatory variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i. \tag{1.120}$$

Note that the observations of the explanatory variables now have two subscripts. The first, $i = 1, 2, \ldots, N$, still denotes the individual unit of observation. Depending on our data, this can be individuals, firms, countries, or other units. The second subscript, $k = 1, 2, \ldots, K$, indexes the individual explanatory variables. The order in which we write the two subscripts is important, since we will later write the model in matrix form.

Interpretation of the parameters in a multivariate model is somewhat less straightforward than in a bivariate model. Previously, we needed to be careful about the *conditional expectation* interpretation of our effects, but other than that, it was relatively simple. Now, the parameter

$$\beta_1 = \frac{\partial \mathrm{E}(y_i \mid x_{i1}, x_{i2})}{\partial x_{i1}} \tag{1.121}$$

measures the *expected difference* of the variable y when we *change* x by one unit, *holding all other observed variables fixed*. This interpretation is sometimes referred to as a *ceteris paribus* interpretation. As you know, "ceteris paribus" is Latin and essentially means to keep all other things constant. However, this is somewhat incomplete: We can only hold those variables constant that are *observed* and *included* in the model.

As always, it makes sense to view this in the context of an example. Consider the following:

$$\text{wage}_i = \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{experience}_i + u_i. \tag{1.122}$$

In this model, the parameter

$$\beta_1 = \frac{\partial \mathrm{E}(\text{wage}_i \mid \text{education}_i, \text{experience}_i)}{\partial \text{education}_i} \tag{1.123}$$

measures the expected change in the wage, when education increases by one unit, and the individual in question gains no experience at the same time. There are two parts in this sentence where we need to get the precise wording correct. One, we are still modeling an expectation, so we need to talk about the "expected change," a change that will occur "on average," or similar. Two, we need to be explicit about that this is the change we expect "when we hold all other (observed) factors constant," although a simple "ceteris paribus" will do as well.

More variables means more fun, so what about this?

$$\text{wage}_i = \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{experience}_i + \beta_3 \text{age}_i + \beta_4 \text{career years}_i + \beta_5 \text{union}_i + u_i \quad (1.124)$$

Suppose union_i is a dummy variable indicating union membership. Why did we not add the opposite case, non-union_i, as well? The reason is that these two variables would be directly, mechanically, inversely related. When your union membership indicator is 1, your non-membership indicator is 0, and vice versa. It would not make any sense to try to split the observed effect in a part that occurs because i is a member in a union and another part that occurs because i is not *not* a member in a union. For a dummy variable with a "yes-no" interpretation, this makes trivial sense, but it also applies to dummy variables with multiple levels. We must always exclude one level as reference category. We will later see the mathematical reason for this.

Also, experience_i and career years_i are closely correlated. It is very difficult to gain experience without accumulating additional career years, and it is probably also difficult to add years to your career without gaining any experience (although it should certainly be doable). This makes a "ceteris paribus" explanation increasingly difficult (and, frankly, nonsensical). We will also see the reasons for this a bit later on in this section of the course.

So let us start right away with trying to derive the OLS estimators for the multivariate case. Setting up the loss function is easy. We end up with

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K\right) = \underset{\left(\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_K\right)}{\arg\min} \sum_{i=1}^{N} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_K x_{iK}\right)^2. \quad (1.125)$$

Finding partial first derivatives and setting them all to zero is equally easy, so we are going to do just that:

$$-2 \sum_{i=1}^{N} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_K x_{iK}\right) = 0 \quad (1.126)$$

$$-2 \sum_{i=1}^{N} x_{i1} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_K x_{iK}\right) = 0 \quad (1.127)$$

$$\vdots$$

$$-2 \sum_{i=1}^{N} x_{iK} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_K x_{iK}\right) = 0 \quad (1.128)$$

The problem of this approach should be obvious by now. We end up with a system of equations, consisting of $K+1$ linear equations. Since there are also $K+1$ variables in the system, it is solvable. However, we cannot solve it, and thus find the OLS estimators $\hat{\beta}_k$ without using matrix algebra.

But there is one thing we *can* do with these first-order conditions: interpret them as moment conditions.

- The condition for the constant, $\hat{\beta}_0$, tells us that the sample mean of the OLS residuals must be zero. The corresponding population moment condition is

$$\mathrm{E}(u_i) = 0. \tag{1.129}$$

- The condition for the slope parameters, $\hat{\beta}_k$, tells us that the sample covariance between the residuals and any regressor x_{ik} must be zero. This implies the following moment conditions of the population:

$$\mathrm{Cov}(x_{ik}, u_i) = \mathrm{E}(x_{ik}u_i) = 0. \tag{1.130}$$

As we did in the bivariate case, we have two ways to derive the OLS estimators: We can either consider the first-order conditions from the optimization problem, or, equivalently, we can derive the estimators from the moment conditions.

1.3.2 Vector and Matrix Notation

When we deal with multiple variables, the summation notation we have previously used becomes increasingly tedious. It is also very restrictive, which we have seen when we tried to derive the OLS estimators. Therefore, we are now going to switch to write our models in terms of *vectors* and *matrices*. Consider a similar model as before, slightly simplified:

$$\text{wage}_i = \beta_0 \times 1 + \beta_1 \text{education}_i + \beta_2 \text{experience}_i + \beta_3 \text{age}_i + u_i. \tag{1.131}$$

Note that there is an unnecessary (or is it?) 1 that we multiply with the constant parameter. This is actually only a pedagogical trick to direct your attention to how we can stack our *explanatory* variables in a vector:

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ \text{education}_i \\ \text{experience}_i \\ \text{age}_i \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \tag{1.132}$$

Including the 1 yields a vector of four explanatory variables, one of which is just a constant. But this way, we can easily multiply it with our *vector of parameters*, which also has four elements. For this, of course, we need to *transpose* one of the two, since we know that we cannot multiply them otherwise. We choose to transpose the $\boldsymbol{x}$ vector, for reasons that will become clear to you very soon.

$$\boldsymbol{x}_i' = (1, \text{education}_i, \text{experience}_i, \text{age}_i). \tag{1.133}$$

Now look at how beautifully we can multiply them:

$$\boldsymbol{x}_i'\boldsymbol{\beta} = 1 \times \beta_0 + \text{education}_i\beta_1 + \text{experience}_i\beta_2 + \text{age}_i\beta_3. \tag{1.134}$$

Using this, we end up with something that can be characterized as a hybrid form of notation for our model:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \tag{1.135}$$

Using *vector notation* makes our model more compact, and means that we are now able to derive the OLS estimators. But the subscript i is still there, and that means that sums are still going to pop up, for example, in the optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\tilde{\boldsymbol{\beta}}} \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}\right)^2. \tag{1.136}$$

Let us solve this problem:

$$\begin{aligned}
S(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}\right)^2 \qquad (1.137)\\
S(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left(y_i^2 - 2y_i \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{\beta}' \boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta}\right)\\
\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} \left(-2y_i \boldsymbol{x}_i + 2\boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta}\right)\\
\sum_{i=1}^{N} \left(-2y_i \boldsymbol{x}_i + 2\boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta}\right) &= \mathbf{0}\\
\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta} &= \sum_{i=1}^{N} y_i \boldsymbol{x}_i\\
\left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i'\right) \boldsymbol{\beta} &= \sum_{i=1}^{N} \boldsymbol{x}_i y_i\\
\hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{x}_i y_i\right) \qquad (1.138)
\end{aligned}$$

Looks nice, but we can simplify this even more.

The vector notation we had before still described the regression model for a single observation i. We are now going to look for a way to "stack" the observations so that we can describe the model for *all observations* using just one equation. Let us start with the outcomes y_i:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}. \tag{1.139}$$

You can see that this is very easy. We are going to mirror this for the error term:

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}. \tag{1.140}$$

What about $\boldsymbol{x}_i$? Remember that we said that transposing the explanatory variables vector would make sense later on. That is actually now: By stacking $\boldsymbol{x}_i'$, we get the following matrix:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1K} \\ 1 & x_{21} & \dots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{NK} \end{pmatrix}. \tag{1.141}$$

This allows us to, finally, write our model in *matrix notation*:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{1.142}$$

Using this notation, the *sum of squared residuals* is given as

$$\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \tag{1.143}$$

which means that the optimization problem is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\tilde{\boldsymbol{\beta}}} \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}. \tag{1.144}$$

Deriving the OLS estimator from this requires us to differentiate matrices. This is not as convenient as using the method of moments, but it is possible:

$$\begin{aligned} \boldsymbol{u}'\boldsymbol{u} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{1.145}$$

This allows us to differentiate with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \boldsymbol{u}'\boldsymbol{u}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \stackrel{!}{=} 0, \tag{1.146}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1.147}$$

The *method of moments* is simpler. We start with our *moment condition* (in matrix notation, there is only one):

$$\mathrm{E}(\boldsymbol{X}'\boldsymbol{u}) = \boldsymbol{0}. \tag{1.148}$$

If $\boldsymbol{X}$ contains a column of ones, this also implies that $\mathrm{E}(\boldsymbol{u}) = 0$. As usual, we start by replacing the population moments by their sample analogues. Thus, we get

$$\boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{0}. \tag{1.149}$$

With only basic knowledge of matrix algebra, we can now derive our estimator very easily:

$$\begin{aligned} \boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= 0 \\ \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= 0 \\ \boldsymbol{X}'\boldsymbol{y} &= \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} &= \hat{\boldsymbol{\beta}} \end{aligned} \tag{1.150, 1.151}$$

No matter which approach we used, we received the same estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1.152}$$

This is the same estimator as we got before. We now know it in three different notational forms: In *matrix notation* as above, in *vector notation* as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{x}_i y_i\right), \tag{1.153}$$

and in *summation notation* (for the bivariate case) as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{1.154}$$

There is one bit about vector and matrix notation that is useful to know. In the bivariate case, we discussed the variance of the error term on multiple occasions. But now, the error term is a vector of random variables, all of which have mean zero, and a variance of σ^2. Denoting the expected value of the vector is straightforward:

$$\mathrm{E}(\boldsymbol{u}) = \mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0. \end{pmatrix} \tag{1.155}$$

If we think about the moment definitions we briefly discussed in the previous section, it makes sense that the variance of a vector must be a matrix:

$$\mathrm{Var}(\boldsymbol{u}) = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}') \tag{1.156}$$

What we receive is also called a *variance-covariance matrix*. It is a matrix that contains the variances of the individual elements of the vector on the main diagonal, while the off-diagonal elements are the covariances between the individual elements:

$$\mathrm{Var}(\boldsymbol{u}) = \begin{pmatrix} \mathrm{Cov}(u_1, u_1) & \dots & \mathrm{Cov}(u_1, u_N) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(u_N, u_1) & \dots & \mathrm{Cov}(u_N, u_N) \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(u_1) & \dots & \mathrm{Cov}(u_1, u_N) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(u_N, u_1) & \dots & \mathrm{Var}(u_N) \end{pmatrix} \tag{1.157}$$

If the elements of the vector are assumed to be independent, the off-diagonal elements are thus zero. Our most standard assumption for the variance of the error term, which read $\mathrm{Var}(u_i) = \sigma^2$ in the bivariate case, is therefore now given as

$$\mathrm{Var}(\boldsymbol{u}) = \sigma^2 \boldsymbol{I} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}. \tag{1.158}$$

1.3.3 Multivariate vs. Bivariate Models

Multivariate regression is very similar to bivariate regression. The core idea, like before, is that we split the outcomes we observe into an *explained part*, $\hat{y}$, and an unexplained part, $\hat{u}$, the residuals:

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}}_{\hat{y}_i} + \hat{u}_i. \tag{1.159}$$

Just like before, the sample mean of the residuals is zero. This implies that the sample mean of the predicted values equals the sample mean of the observed values, $\bar{y}$. Additionally, in simple linear regression, the point $(\bar{x}, \bar{y})$ lies on the regression line. In multiple linear regression, the point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K, \bar{y})$ lies on the $K+1$-dimensional equivalent of a regression line, which is called a regression *hyperplane*. This is increasingly difficult to picture for more than three dimensions, but in three dimensions (i.e., for $K = 2$,), this would be a plane, such as the one shown in Figure 1.11.

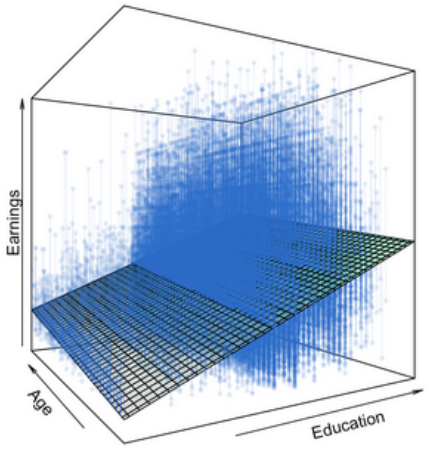


Figure 1.11: A graphical representation of a regression of the wage on education and age in three-dimensional space.

If we estimate a simple linear regression model, and then add additional variables to the same model, the estimates for the coefficient that occurs in both models will not generally match. Consider the models

$$y_i = \beta_0^* + \beta_1^* x_{i1} + u_i \tag{1.160}$$

and

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i. \tag{1.161}$$

The estimates for $\hat{\beta}_1^*$ and $\hat{\beta}_1$ will only be the same in two special cases: One, if $\text{Cov}(x_{i1}, x_{ik}) = 0$ for all $k \neq 1$. This is very rare, as virtually all variables we encounter in real life are usually at least somewhat correlated. Two, if $\beta_k = 0$ for all $k \notin \{0, 1\}$, that is, if all explanatory variables other than the first one are irrelevant.

To see why this is the case, consider the following two-regressor model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, \tag{1.162}$$

and the model without x_2,

$$y_i = \beta_0^* + \beta_1^* x_{i1} + u_i. \tag{1.163}$$

Assume that we can regress x_2 on x_1 like this:

$$x_{i2} = \delta_0 + \delta x_{i1} + v_i. \tag{1.164}$$

We can now substitute:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2(\delta_0 + \delta x_{i1} + v_i) + u_i \\ &= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta)x_{i1} + (u_i + \beta_2 v_i). \end{aligned} \tag{1.165}$$

We can see that

$$\beta_1^* = \beta_1 + \beta_2\delta, \tag{1.166}$$

a relationship which also holds for the estimators,

$$\hat{\beta}_1^* = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}. \tag{1.167}$$

Taking the expectation yields

$$\mathrm{E}\left(\hat{\beta}_1^*\right) = \mathrm{E}(\hat{\beta}_1) + \mathrm{E}(\hat{\beta}_2)\delta = \beta_1 + \beta_2\delta = \beta_1^*, \tag{1.168}$$

which shows that $\hat{\beta}_1^*$ only equals $\hat{\beta}_1$ if either $\hat{\beta}_2$ or $\hat{\delta}$ is zero.

If none of these conditions holds, our estimates will be biased:

- If the longer model, the one with two regressors, is the "correct" model, but we instead estimate the shorter model, then we estimate $\beta_1^* = \beta_1 + \beta_2\delta$ instead of β_1. The resulting bias is called *omitted variable bias*.
- If the shorter model, the model with one regressor, is the "correct" model, but we instead estimate the longer model, then we estimate $\beta_1 = \beta_1^* - \beta_2\delta$ instead of β_1^*. What we receive will be biased as a result of this *overspecification*.

The problem with this is that we never know whether one of the models is "correct" or "true" and the other is not. We may be inclined to think that the larger model is automatically better since it includes more information, but that is not necessarily the case. (Later in the course, we will extensively discuss why.) Therefore, we need to consider which model better fits our assumptions and argue for why we think it is unlikely that there is bias resulting from a misspecification.

In the bivariate case, we discussed a very simple measure for *goodness of fit*, the *coefficient of determination*. Just like we did then, we can now divide the variation we observe in y into variation originating from variation in the x variables, and into an unexplained part:

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}\hat{u}_i^2 \tag{1.169}$$

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR} \tag{1.170}$$

And as before, we can compute the coefficient of determination from these decomposed measures:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}. \tag{1.171}$$

This measure, of course, has the same interpretation – it quantifies the part of the variation in the outcome that is explained by the model – and suffers from all of the same problems that we have already discussed in the context of bivariate models. In the multivariate case, it has one additional important shortcoming: When we add variables to our model, the R^2 will always, mechanically, increase. So, it was a bad measure for comparing models in the first place, but when those models are of different sizes, it becomes completely useless because it will always favor the longer model.

1.3.4 Practical Example

Let us look at a practical example of multivariate regression in R. For this, we will use a dataset from the `AER` package that contains data on hourly wages and some explanatory variables. We start by reading in the dataset and selecting four variables of interest:

```
# Load packages
library(AER) # Contains our dataset
library(dplyr) # Contains select()

# Load data
data("CPSSW8") # Dataset with hourly wages

# Keep only variables we are interested in
CPSSW8 <- CPSSW8 |>
  select(earnings, gender, age, education)
```

With the following command, we can display a summary of all variables in the dataset:

```
# Summary of our variables
summary(CPSSW8)
```

```
    earnings          gender           age          education
 Min.   : 2.003   male   :34348   Min.   :21.00   Min.   : 6.00
 1st Qu.:11.058   female:27047   1st Qu.:33.00   1st Qu.:12.00
 Median :16.250                  Median :41.00   Median :13.00
 Mean   :18.435                  Mean   :41.23   Mean   :13.64
 3rd Qu.:23.558                  3rd Qu.:49.00   3rd Qu.:16.00
 Max.   :72.115                  Max.   :64.00   Max.   :20.00
```

The interpretation of this output is relatively straightforward. For all numeric variables, we get the minimum, quartiles, mean, and maximum; and for categorical variables, we get a tally of the different levels the dataset contains. Here, there is one dummy variable, `gender`, which in this dataset contains two levels, totaling 34,348 observations of men, and 27,047 observations of women. The summary of the `age` variable is what we would suspect of a dataset on the working population, and the values of `education` and `earnings` tell us that the former is likely measured in years and the latter in dollars per hour. Of course, since we have read the documentation of the data before working with it, we already know this, but it never hurts to check.

So, let the fun begin. We are actually going to run a regression:

```
lm(log(earnings) ~ education, data = CPSSW8) |>
  summary()
```

```
Call:
lm(formula = log(earnings) ~ education, data = CPSSW8)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57775 -0.32030  0.02187  0.34346  1.74119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.4848226  0.0114553   129.6   <2e-16 ***
education   0.0940194  0.0008262   113.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5038 on 61393 degrees of freedom
Multiple R-squared:  0.1742,    Adjusted R-squared:  0.1742
F-statistic: 1.295e+04 on 1 and 61393 DF,  p-value: < 2.2e-16
```

Remember what we have learned about *interpretation of the coefficients*: We can see that the slope coefficient for education is 0.094. Since we have a log-level model, this means that we expect people with one additional year of education to have a roughly 9.4 percent higher wage. The intercept is 1.485, which means that people with no education have an average log wage of 1.485. If we exponentiate this, we get 4.41 US Dollars, but we have to be careful about the interpretation of this figure: It is *not* the average wage of people with zero education, since $E(\log(\cdot)) \neq \log(E(\cdot))$.

The next step is adding more variables to the regression.

```
lm(log(earnings) ~ education+gender+age, data = CPSSW8) |>
  summary()
```

```
Call:
lm(formula = log(earnings) ~ education + gender + age, data = CPSSW8
    )

Residuals:
     Min       1Q   Median       3Q      Max
-2.79472 -0.28807  0.02562  0.32439  1.63195

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2206890  0.0130145   93.80   <2e-16 ***
education     0.0941281  0.0007922  118.82   <2e-16 ***
genderfemale -0.2338747  0.0039207  -59.65   <2e-16 ***
age           0.0088690  0.0001839   48.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4816 on 61391 degrees of freedom
Multiple R-squared:  0.2452,    Adjusted R-squared:  0.2452
F-statistic:  6649 on 3 and 61391 DF,  p-value: < 2.2e-16
```

The slope coefficient for `education` is 0.0941. We can interpret this coefficient as follows: For people with one additional year of education, we expect a roughly 9.4 percent higher

wage, holding gender and age constant. Although the coefficient is extremely close to the one from the previous model, we can also see that it is not the same. The coefficient for `gender==female` is −0.234. This means that, in this dataset, women have, on average, 23 percent lower wages, after controlling for education and age. The coefficient for `age` is 0.0089, which means that people who are one year older are expected to have 0.89 percent higher wages, after controlling for education and gender. The intercept is 1.22 and relates to a hypothetical observation with no education, an age of zero years, and whose gender is male (the reference level). Other than that, its interpretation is analogous to before.

When you read a paper, a bachelor's thesis, or anything similar, regression results are usually presented in a format that differs slightly from a plain output table. An example of how this would look like for the two specifications we estimated in this section is given in Table 1.6. Using R, tables like this can easily be produced using `stargazer`, `modelsummary`, or similar packages, which can produce LaTeX, HTML, or plain text output.

	Dependent variable:	
	log(earnings)	
	(1)	(2)
education	0.094***	0.094***
	(0.001)	(0.001)
genderfemale		−0.234***
		(0.004)
age		0.009***
		(0.0002)
Constant	1.485***	1.221***
	(0.011)	(0.013)
Observations	61,395	61,395
R^2	0.174	0.245
Adjusted R^2	0.174	0.245
Residual Std. Error	0.504 (df = 61393)	0.482 (df = 61391)
F Statistic	12,950*** (df = 1; 61393)	6,648.669*** (df = 3; 61391)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 1.6: Regression results for the shorter specification (Column 1) and the longer specification (Column 2) from this section, presented in a table in the style you would usually encounter it in a research paper.

1.3.5 The Gauss-Markov Theorem

We know that the OLS estimator is unbiased for bivariate models. But ideally, we would like to prove that for the multivariate case as well. This is what we are going to do in this section. As before, we are going to start with stating the assumptions that are necessary for our proof. These assumptions are generalized versions of SLR.1 to SLR.4, which we already

know:

1. Linearity in Parameters.
2. Random Sampling.
3. Variation in x.
4. Exogenous Errors.

To prove the *Gauss-Markov Theorem* for the multivariate case, we will need the following assumptions, which we will call MLR.1 to MLR.4:

1. Linearity in Parameters.
2. Random Sampling.
3. *No Perfect Multicollinearity.*
4. Exogenous errors.

Assumption MLR.1 (Linearity in Parameters). The population regression function (PRF) must be linear in its parameters:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i. \tag{1.172}$$

Transformations, for example logarithmic ones, are still unproblematic, since the population regression function remains a linear combination of the *parameters*. Note also that the above model can be written in matrix notation as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{1.173}$$

Assumption MLR.2 (Random Sampling). Our sample of N observations, $\{(y_i, x_{i1}, \ldots, x_{iK}), i = 1, 2, \ldots, N\}$ must be a *random sample* from the population. The probability of including an observation must be equal for all, and it must not depend on who we sampled first.

If this assumption holds, we can again treat observations and eror terms as independent of one another.

Assumption MLR.3 (No Perfect Multicollinearity). The regressor matrix $\boldsymbol{X}$ contains no column that is a linear combination of other columns. Equivalently, $\boldsymbol{X}$ has full rank.

This is the first of our assumptions where generalizing yields a meaningful difference between the bivariate and the multivariate case. Previously, we had assumed variation in x values instead. Actually, it is easy to see how "no variation in x" in a bivariate model is a special case of this assumption. Consider the regressor matrix for a bivariate model:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}. \tag{1.174}$$

If there is no variation in x, then the second column is a linear combination of the first (because it is identical). Thus, "variation in x" is a special case of "no perfect multicollinearity." We needed this assumption before because we needed to be able to divide by $\sum_{i=1}^{N}(x_i - \bar{x})^2$. Now, we need it to avoid a similar mathematical issue: When computing our estimator, we need to invert $(\boldsymbol{X}'\boldsymbol{X})$. This is only possible if $\boldsymbol{X}$ has full rank. Otherwise, we cannot compute our estimator.

Assumption MLR.3 is *violated* when a regressor is a (weighted) sum or difference of other regressors. One example would be studying the relationship between a person's income and their parents' income. If we include parental income (x_{i3}) as well as a measure of both parents' separate incomes (x_{i1} and x_{i2}) in the regression, the assumption is violated since $x_{i3} = x_{i1} + x_{i2}$. A similar problem would arise were we to include the difference between both parents' individual incomes. For *dummy variables*, the assumption is violated if *every level* of the categorical variable is included in the regression, and none is omitted as reference category. If we have data on all 27 E.U. member countries and want to add a country dummy, we need to exclude one country indicator as reference in order to not run into a multicollinearity problem.

The assumption is *not violated* if a regressor is a non-linear combination of other regressors. For example, it is unproblematic to include the squared income of each parent alongside the non-squared income. By including both, we can model parabolic relationships. Alternatively, we can include $x_{i3} = x_{i1} \times x_{i2}$. This is called an *interaction effect*. It allows us to model effects that depend on one another, but makes interpretation more difficult, since a simple "ceteris paribus" interpretation is no longer possible. We will discuss squared regressors and interaction terms in more detail later on. At this point, it is sufficient to know that they exist and that they do not violate Assumption MLR.3.

MLR.3 is not violated either if two regressors are strongly (but not perfectly) correlated. The stronger two regressors are correlated, the less precise OLS estimates become (i.e., their variance becomes larger). But they are still BLUE, because we do not require regressors to be uncorrelated for either our unbiasedness or our variance proof. Simple correlation between them does not violate any Gauss-Markov assumption, as long as the correlation is not perfect. Keep in mind that, while perfect correlation is immediately detected and flagged by statistical software (because it cannot compute estimates), nearly perfect correlation is not. Therefore, you as researcher should be aware about any correlation patterns between your variables and decide on a case-by-case basis whether two of them are unreasonably strongly correlated or not.

Assumption MLR.4 (Exogenous Errors). The expected value of the error term u must be zero for each regressor:

$$\mathrm{E}(u_i \mid x_{i1}, \ldots, x_{iK}) = 0. \tag{1.175}$$

The corresponding assumption in matrix notation is

$$\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{X}) = \boldsymbol{0}. \tag{1.176}$$

Note that the latter version of the assumption is slightly stronger than the former, as it includes not just the regressors but also linear combinations between them. As in the bivariate case, this assumption implies that $\mathrm{E}(u_i) = 0$, and that $\mathrm{Cov}(x_{ik}, u_i) = 0$ or equivalently $\mathrm{E}(x_{ik}u_i) = 0$ for all regressors x_{ik}.

Independence of regressors and unobserved factors is easy to achieve in experiments, but much harder to ensure when using observational data. We call the case in which this

assumption is violated *endogeneity*. When $\mathrm{E}(x_{ik}u_i) \neq 0$, we call x_{ik} an *endogenous regressor*. The following are examples for situations that cause regressors to be endogenous.

- We omit a variable that is correlated with some of the regressors and relevant for explaining the dependent variable. This is what we call *omitted variable bias*, and it is a form of endogeneity. An example would be to regress wage on education, without accounting for talent. Note that as we cannot measure talent, we are unable to get rid of the endogeneity using the tools we know.
- The dependent variable itself influences a regressor. This is called *reverse causality*. For example, during the Covid pandemic, more mask wearing reduced infections, but at the same time, the number of infections also influenced how many people wore masks.
- The true relationship is non-linear.

The presence of endogeneity requires more advanced techniques if we still want to get valid results out of our regressions. You can easily fill an entire course only with these methods. This, by the way, is essentially what we will be doing in Econometrics II.

1.3.6 Expected Value of the OLS Estimator

Now it is time to formally prove the first part of the *Gauss-Markov Theorem* for multivariate regression:

Unbiasedness of the OLS estimator. Under Assumptions MLR.1 to MLR.4, we have $\mathrm{E}\left(\hat{\beta}_k\right) = \beta_k, \qquad k = 0, 1, \ldots, K,$, for every value of the parameters β_j. In matrix notation:

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}, \tag{1.177}$$

where $\boldsymbol{\beta}$ has dimension $(K+1) \times 1$.

The OLS estimator is therefore an *unbiased estimator* of the intercept and the slope parameters. We can, just as we did in the previous section, prove this by splitting the estimator into the true coefficient and a sampling error component. We start by decomposing $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \\
&= \underbrace{\boldsymbol{\beta}}_{\text{true parameter}} + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}}_{\text{sampling error}}.
\end{aligned} \tag{1.178}$$

This is equivalent to the step in the proof for the bivariate case in which we decomposed $\hat{\beta}_1$ into

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(x_i - \bar{x})u_i}{\sum_{i=1}^{N}(x_i - \bar{x})x_i}. \tag{1.179}$$

We continue by taking the expectation conditional on the regressors. The rest of the proof is straightforward:

$$\begin{aligned}
\mathrm{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) &= \mathrm{E}\left(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\middle|\boldsymbol{X}\right) \\
&= \boldsymbol{\beta} + \mathrm{E}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\middle|\boldsymbol{X}\right) \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \underbrace{\mathrm{E}(\boldsymbol{u}|\boldsymbol{X})}_{=0 \text{ (MLR.4)}} \\
&= \boldsymbol{\beta},
\end{aligned} \tag{1.180}$$

where we, just as before, make use of assumption MLR.4 in the last step. Because of the law of iterated expectations,

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) = \boldsymbol{\beta} \quad \Rightarrow \quad \mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}. \tag{1.181}$$

Our estimator is unbiased. □

1.3.7 Variance of the OLS Estimator

Now, we are going to add one assumption in order to be able to make statements about our estimator's variance:

Assumption MLR.5 (Homoskedasticity). The variance of the error term u_i is the same for all x_{ik}:

$$\mathrm{Var}(u_i \mid x_{i1}, \ldots, x_{iK}) = \mathrm{Var}(u_i) = \sigma^2, \tag{1.182}$$

or in matrix notation:

$$\mathrm{Var}(\boldsymbol{u} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_N, \tag{1.183}$$

where $\boldsymbol{I}_N$ is the identity matrix with dimension $N \times N$.

The interpretation of this assumption is analogous to the corresponding SLR assumption. We are assuming that this variance is constant for all values of the regressors, and that it equals σ^2. We are also explicitly assuming that the covariance between all individual errors is zero, that is, that they are independent. Again, we did not need this assumption to prove that the OLS estimator is unbiased. And as before, it is very easy to violate.

Of course, the OLS estimator is BLUE in multivariate settings as well. We are, however, going to skip the proof this time – we know the idea behind how it works. But we are quickly going to derive what the variance of the estimator is equal to:

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) &= \mathrm{Var}\Big(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\Big|\boldsymbol{X}\Big) \\
&= \mathrm{Var}\Big(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u})\Big|\boldsymbol{X}\Big) \\
&= \mathrm{Var}\Big(\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\Big|\boldsymbol{X}\Big) \\
&= \mathrm{Var}\Big(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{u}\Big|\boldsymbol{X}\Big) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{u} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{I}\sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned} \tag{1.184}$$

You can clearly see at which point we are applying Assumption MLR.5 by substituting $\sigma^2 \boldsymbol{I}$ for $\text{Var}(\boldsymbol{u} \mid \boldsymbol{X})$.

The variance of an individual coefficient is given by

$$\text{Var}\left(\hat{\beta}_k \middle| \boldsymbol{X}\right) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2} \times \frac{1}{1 - R_k^2}, \tag{1.185}$$

where R_k^2 is the R^2 from a regression of x_k on all other regressors $x_j, j \neq k$. Using this equation, we can see that a larger σ^2 increases individual coefficients' variance, that is, it will decrease the precision of the estimate. If we, however, have a larger sample (and thus a larger N), our estimates will be more precise. In addition, strong variation in a regressor x_k, as well as weak correlation of that regressor with other regressors, make the estimate for the coefficient associated with the regressor more precise.

The problem is again that we do not know σ^2 and need an estimator for it. It can be shown (we omit the proof) that the estimator

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{N - K - 1} \tag{1.186}$$

is an unbiased estimator of the error variance under Assumptions MLR.1 to MLR.5:

$$\text{E}\left(\frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{N - K - 1}\right) = \text{E}\left(\hat{\sigma}^2\right) = \sigma^2 \tag{1.187}$$

We apply the same degrees of freedom correction as in the bivariate case by dividing by $N - K - 1$ instead of dividing by N.

Note that the variance of the estimator $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1)$ is

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \begin{pmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{Var}(\hat{\beta}_1) \end{pmatrix}. \tag{1.188}$$

We have so far only discussed the sample *variance* of an estimator, and not its sample *covariance*. Statistical software usually estimates only the variances of the parameters, that is, only the diagonal of the variance-covariance matrix. We will need the covariances later for certain statistical tests (such as joint hypothesis tests, which we will discuss in the next section).

To finish this subsection, let us summarize the Gauss-Markov Theorem for multiple regression:

Gauss-Markov Theorem. Under Assumptions MLR.1 to MLR.5, the OLS estimator

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{pmatrix} \tag{1.189}$$

is the *best linear unbiased estimator* (BLUE) of the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$.

1.3.8 Frisch-Waugh-Lovell Theorem

It is not immediately easy to understand what the coefficients in a multivariate model actually measure. Now, we are going to try to make this a bit more intuitive. We are going to do this by discussing a simplified representation of the *Frisch-Waugh-Lovell Theorem* (FWL Theorem), named after Frisch and Waugh (1933) as well as Lovell (1963). Consider the following model:

$$y_i = x_{i1}\beta_1 + \boldsymbol{x}_{i2}'\boldsymbol{\beta}_2 + u_i, \qquad \mathrm{E}\left(\begin{pmatrix} x_{i1} \\ \boldsymbol{x}_{i2} \end{pmatrix} u_i\right) = 0. \tag{1.190}$$

On first sight, this looks like a model representation we have not seen before. But it is actually quite simple. We have written down the first explanatory variable, x_1, separately from the others, which we have grouped in the vector $\boldsymbol{x}_{i2}$. This is mathematically equivalent to "full" vector notation, but we can pay special attention to x_1.

To make this more intuitive, you can think of y_i being the wage, x_{i1} being education, and $\boldsymbol{x}_{i2}$ being a vector containing a column of ones as well as data on gender and the age. We typically call explanatory variables in which we are not primarily interested *control variables*. It is also not uncommon to not denote them separately, similar to the notation we chose here. For our example, the equation becomes:

$$\text{wage}_i = \text{education}_i\beta_1 + \boldsymbol{x}_{i2}'\boldsymbol{\beta}_2 + u_i. \tag{1.191}$$

We are now going to do something that seems a little unintuitive at first, but will give us a new perspective at interpreting coefficients. We will come up with an alternative regression, and then we will find out that the parameter associated with education_i is numerically equivalent in two very different regressions. This will show us that we can interpret the parameter in an additional, very useful way.

We start by relating the outcome, wage_i, only to the vector of control variables, that is, we leave out our variable of interest, education_i:

$$\text{wage}_i = \boldsymbol{x}_{i2}'\boldsymbol{\alpha} + \underbrace{\text{wage}_i^{(R)}}_{\text{error}}. \tag{1.192}$$

From this regression, we "keep" the error $\text{wage}_i^{(R)}$. We can interpret this error as a "filtered" version of the outcome variable: It represents the part of the variation that cannot be explained by the control variables.

Next, we are going to do the exact same thing – pay close attention – to our *explanatory variable of interest*. That is, we are regressing the explanatory variable on the controls:

$$\text{education}_i = \boldsymbol{x}_{i2}'\boldsymbol{\alpha} + \underbrace{\text{education}_i^{(R)}}_{\text{error}}. \tag{1.193}$$

From this regression, we again "keep" the error $\text{education}_i^{(R)}$. As before, we can interpret this error as representing the part of the variation in education that cannot be explained by the control variables. So we now have a "version" of the outcome with the effects of our controls *partialed out*, and a "version" of the explanatory variable of interest with the effects of the controls *partialed out*.

The core of the FWL Theorem is now the following. We can obtain *the exact same parameter* β_1 in two different ways: One way is running a regression like we usually would do, regressing wage_i on both education_i and the controls. But we can get the exact same

parameter by regressing $\text{wage}_i^{(R)}$ on $\text{education}_i^{(R)}$ (full stop, no controls). This is what we call the *FWL regression*: We regress a version of the outcome where all controls are partialed out onto a version of our explanatory variable of interest where all controls are partialed out.

Implementing this in a sample is straightforward, we just have to follow this procedure:

1. Regress y_i on $\boldsymbol{x}_{i2}'$ and obtain the residuals $\hat{y}_i^{(R)}$.
2. Regress x_{i1} on $\boldsymbol{x}_{i2}'$ and obtain the residuals $\hat{x}_{i1}^{(R)}$.
3. Regress $\hat{y}_i^{(R)}$ on $\hat{x}_{i1}^{(R)}$ and obtain the OLS estimator $\hat{\beta}_1$. This estimator is equal to the estimator from the original regression.

The primary use case for this, at least at this point, is that it enables an intuitive interpretation of multivariate regression. We can illustrate this situation with a causal graph like the one in Figure 1.12, where the bubbles represent variables and arrows represent relations between them. We assume that education has an influence on wage, but there is also correlation between the control variables in $\boldsymbol{x}_{i2}'$ and both education and wage.

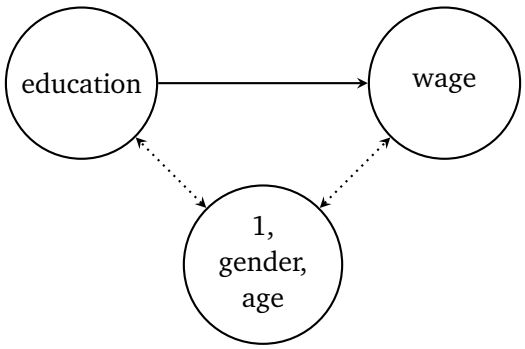


Figure 1.12: Causal graph for the FWL example.

The graph helps with understanding what we do here. We can interpret the error $\text{wage}_i^{(R)}$ as the variation in wage_i that is not explained by $\boldsymbol{x}_{i2}'$. Likewise, the error $\text{education}_i^{(R)}$ represents the variation in education_i that is not explained by the controls. Thus, we are "filtering out" the effects represented by the dashed lines in the graph. We have now two ways to cleanly get our effect of interest, the solid line: Running a bivariate regression between the "filtered" variables, or running the full multivariate regression with controls. This approach shows why we interpret coefficients in multivariate regression models as *ceteris paribus* effects.

We can also check whether this actually works. Let us again use the earnings dataset. Then, we manually go through the entire procedure we have outlined above. We run the auxiliary regressions, save the residuals, and then regress the residuals from the one auxiliary regression onto the residuals from the other auxiliary regression. In addition to that, we estimate the full model so that we can later compare the coefficients.

```
# Load packages
library(AER) # Contains our dataset
library(dplyr) # Contains select()

# Load data
```

```r
data("CPSSW8") # Dataset with hourly wages

# Keep only relevant variables
CPSSW8 <- CPSSW8 |>
  select(earnings, gender, age, education)

CPSSW8$female <- CPSSW8$gender=="female"

# Residuals for y_i
log_earning_residuals <- lm(log(earnings) ~ female + age,
    data = CPSSW8) |>
  residuals()

# Residuals for x_{i1}
education_residuals <- lm(education ~ female + age, data =
    CPSSW8) |>
  residuals()

# Estimate FWL model
partialled_out_model <- lm(log_earning_residuals ~ 0 +
    education_residuals, data = CPSSW8)

# Estimate original model
full_model <- lm(log(earnings) ~ education + female + age,
    data = CPSSW8)
```

Next, we compare the model output of the FWL regression to that of the full regression.

```r
# Output FWL model
summary(partialled_out_model)
```

```
Call:
lm(formula = log_earning_residuals ~ 0 + education_residuals,
    data = CPSSW8)

Residuals:
     Min       1Q   Median       3Q      Max
-2.79472 -0.28807  0.02562  0.32439  1.63195

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
education_residuals 0.0941281  0.0007922   118.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4816 on 61394 degrees of freedom
Multiple R-squared:  0.187,     Adjusted R-squared:  0.187
F-statistic: 1.412e+04 on 1 and 61394 DF,  p-value: < 2.2e-16
```

```r
# Output original model
summary(full_model)
```

```
Call:
lm(formula = log(earnings) ~ education + female + age, data = CPSSW8
    )

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2.79472 -0.28807  0.02562  0.32439  1.63195

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2206890  0.0130145   93.80   <2e-16 ***
education    0.0941281  0.0007922  118.82   <2e-16 ***
femaleTRUE  -0.2338747  0.0039207  -59.65   <2e-16 ***
age          0.0088690  0.0001839   48.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4816 on 61391 degrees of freedom
Multiple R-squared:  0.2452,    Adjusted R-squared:  0.2452
F-statistic:  6649 on 3 and 61391 DF,  p-value: < 2.2e-16
```

We can see that the coefficient associated with `education` is exactly the same in both regressions, 0.0941281.

1.3.9 How Many Variables?

Knowing multiple regression opens the possibility of including (almost) infinitely many regressors in our models. But is it actually always beneficial to include additional variables? Expectedly, the answer is no. But this answer spurs two new questions: How many variables are too many? And how many variables are not enough?

There is no general "rule of thumb," and no universally valid answer to these questions. No serious researcher would even try to come up with a general rule, since whether we include a given variable or not depends entirely on our context and the model we are currently designing. But we can discuss what happens when we omit relevant variables, and what happens when we include unnecessary ones.

We have already tacitly used the term *omitted variable bias* (OVB) a number of times. This kind of bias occurs if we omit relevant variables. In this case, the effect that the omitted variable has on our outcome gets incorrectly attributed to the variables we have included in our model. That this bias exists is fairly easy to prove, which is why we are going to do it.

Assume the true model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \tag{1.194}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are two groups of regressors. What happens if we now estimate

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{1.195}$$

that is, we omit $\boldsymbol{Z}$? Let us try the approach from our earlier proof of unbiasedness. But note that while the expression for the estimator, $\hat{\boldsymbol{\beta}}$ depends only on $\boldsymbol{X}$ and not on $\boldsymbol{Z}$, $\boldsymbol{Z}$ *is* a part of $\boldsymbol{y}$, the observed values:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}
\end{aligned} \tag{1.196}$$

Taking the expectation of this expression, we see that the estimator is no longer unbiased:

$$\begin{aligned} \mathrm{E}\left(\hat{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right) &= \mathrm{E}\left(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \middle| \boldsymbol{X}\right) \\ &= \boldsymbol{\beta} + \mathrm{E}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma} \middle| \boldsymbol{X}\right) + \mathrm{E}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \middle| \boldsymbol{X}\right) \\ &= \boldsymbol{\beta} + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\mathrm{E}\left(\boldsymbol{X}'\boldsymbol{Z} \middle| \boldsymbol{X}\right)\boldsymbol{\gamma}}_{\text{Bias}} + \mathbf{0} \end{aligned} \tag{1.197}$$

	$\mathrm{E}(\boldsymbol{X}'\boldsymbol{Z} \mid \boldsymbol{X})$ positive	$\mathrm{E}(\boldsymbol{X}'\boldsymbol{Z} \mid \boldsymbol{X})$ negative
$\boldsymbol{\gamma}$ positive	Positive bias	Negative bias
$\boldsymbol{\gamma}$ negative	Negative bias	Positive bias

Table 1.7: The direction of omitted variable bias.

Table 1.7 contains a summary of the direction in which the estimates will be biased. For the sake of simplicity, and being able to speak of "positive" and "negative," you can think of $\boldsymbol{X}$ and $\boldsymbol{Z}$ as only containing one variable each; but the general intuition holds for larger matrices as well. $\mathrm{E}(\boldsymbol{X}'\boldsymbol{Z} \mid \boldsymbol{X})$ essentially tells us whether the variables in $\boldsymbol{X}$ are correlated with the variables in $\boldsymbol{Z}$, and $\boldsymbol{\gamma}$ is the regression parameter associated with $\boldsymbol{Z}$. Only if at least one of them is zero, there is no bias: Either $\boldsymbol{X}$ and $\boldsymbol{Z}$ are uncorrelated, or $\boldsymbol{Z}$ was not relevant for explaining $\boldsymbol{y}$ in the first place.

The opposite case, where we unnecessarily include *irrelevant variables*, can also be a problem. This can be for a multitude of reasons: If we have too many variables and they are strongly correlated, our estimates become less precise. If we have more parameters than observations, that is, $N < K$, then Assumption MLR.3 is violated and we cannot estimate our model. Additionally, unnecessary variables can induce endogeneity, a violation of Assumption MLR.4.

To see why the latter can be the case, consider the following example. Suppose we want to estimate the effect of fertilizer on agricultural yields and conduct an experiment where we correctly randomize fertilizer usage. If we now estimate the model

$$\text{yield}_i = \beta_0 + \beta_1 \text{fertilizer}_i + u_i, \tag{1.198}$$

the assumption that $\mathrm{E}(u_i \mid \text{fertilizer}_i) = 0$ is justified, since fertilizer use was randomized. If, however, we estimate instead

$$\text{yield}_i = \beta_0 + \beta_1 \text{fertilizer}_i + \beta_2 \text{weeds}_i + u_i, \tag{1.199}$$

the assumption that $\mathrm{E}(u_i \mid \text{fertilizer}_i, \text{weeds}_i) = 0$ is far less trivial. In fact, it is probably no longer satisfied, because weeds were not randomized and are likely correlated with unobserved factors. Nonetheless, R^2 as a model selection criterion would have chosen the biased model since it contains more variables. This is one of many examples why econometrics is about critically thinking about your research design at least as much as it is about employing statistical methods.

1.4 Testing and Inference

1.4.1 Introduction

We know by now that our OLS estimator $\hat{\boldsymbol{\beta}}$ is a *random variable*. We have discussed its expectation and variance extensively, and have proved that it is both unbiased and efficient. However, being *moments*, expectation and variance only "summarize" the sampling distribution of the OLS estimator. But ideally, we would like to know what the exact distribution looks like, and not only its moments. This is what we will be doing in this section.

But why do we need more information about the distribution of the estimator? Initially, we mentioned that we want to use data to test our hypotheses. At that time, we focused on the words "data" and "hypothesis" – but there is a third important word in this statement: *test*. We already agreed that we need a falsifiable hypothesis, ideally derived from economic theory. But we have not defined what we mean by empirically "testing" it.

We can think of a few approaches how we can falsify (or not falsify) hypotheses. Suppose we run a linear regression and want to know whether the parameter β_1 is equal to zero, meaning the corresponding variable x_1 has no effect on y. The simplest approach would be to run the regression and check whether the estimate differs from zero, that is, $|\hat{\beta}_1| > 0$. This is, of course, fundamentally a bad idea and it is easy to see why. We know that there is some uncertainty around our estimate. If it is very close to zero, there is no way to discern whether it is actually different from zero or whether the true parameter is zero and the difference is only caused by random uncertainty. Also, given any randomness, the probability of the estimate being *exactly* zero is zero.

We can come up with a better idea for a *hypothesis test*. We start by *assuming* that the true β_1 equals zero. Then, we look at our estimate and ask ourselves how likely it is to get such an estimate if the hypothesis that β_1 is zero is actually true. If this probability is very small, this gives us a sense of confidence in that the true parameter is different from zero.

A bit more formally speaking, a hypothesis test consists of the following steps:

1. We begin by formulating a so-called *null hypothesis*:

$$H_0 : \beta_1 = 0. \tag{1.200}$$

   This is the hypothesis we are going to try to *reject*, in this case, it represents the notion that the variable exerts no influence on the outcome. From our null hypothesis, we can also derive an *alternative hypothesis*:

$$H_A : \beta_1 \neq 0 \tag{1.201}$$

2. We assume that the null hypothesis is true, and calculate the *probability* of obtaining the estimate $\hat{\beta}_1$ we actually got *under the null hypothesis*.

3. If that probability is sufficiently low, we *reject* the null hypothesis.

You may have seen that the *alternative hypothesis*, $H_A : \beta_1 \neq 0$, more closely resembles what we would like to find. So why do we not test this hypothesis directly? There are multiple reasons for this. One is that classical statistical tests only allow us to assume that something *is* a certain value, and not that it *is not* a certain value. This is why we can assume that $\beta_1 = 0$, but we cannot assume that $\beta_1 \neq 0$. Intuitively, suppose we assume that $\beta_1 = 0$. In this case, we will be more likely to obtain $\hat{\beta}_1 = 1$ as estimate compared to, say,

$\hat{\beta}_1 = 5$. If we only assumed β_1 to be different from zero, but did not specify a certain value, such thinking would be outright impossible because probabilities to get specific estimates would be very different under the assumption that $\beta_1 = 0.3$ compared to the assumption that $\beta_1 = -10^6$.

Another reason is that we can *never confirm a hypothesis*; we can *only reject* it. Suppose we want to find out whether a given explanatory variable x_1 has an effect on y. Our null hypothesis is that $\beta_1 = 0$, that is, that the variable has no effect. If we find sufficiently strong estimates, we can with some confidence reject that null hypothesis. But whatever we observe, while it can be very clear that it is unlikely to come from a world where $\beta_1 = 0$, it is impossible to settle on a definitive hypothesis and accept it. If $\hat{\beta}_1 = 3$, it may be unlikely that $\beta_1 = 0$, and maybe it is sufficiently unlikely for us to reject that null hypothesis. But if we want to *accept* a different hypothesis instead, which one do we select? $\beta_1 = 3$? $\beta_1 = 2.5$? $\beta_1 = \pi$? There is no way to choose. All of them could be true.

In the following, we will revert to discussing the *sampling distribution* of the OLS estimator since we will need it to calculate the probabilities we have talked about. After we know how to find the distribution, we will come back to learning about hypothesis tests.

1.4.2 Small Samples

Under Assumptions MLR.1 through MLR.5, we were able to make statements about *moments*: the expected value and the variance of the OLS estimator. But this is not enough to make statements about the estimator's distribution. To see why, consider Figure 1.13. It depicts four very different distributions, a standard normal distribution, a uniform distribution from $-\sqrt{3}$ to $\sqrt{3}$, a mixture of two normal distributions, $N(-0.8, 0.36)$ and $N(0.8, 0.36)$, and a standard exponential distribution. The problem with these four is that all of them have the same expected value, 0, and the same variance, 1. This shows that if we want to make statements about how likely it is that the value of the estimator is in a certain interval, knowing only the moments is not sufficient; since even under the Gauss-Markov assumptions, the distribution of $\hat{\beta}_1$ can take very different shapes.

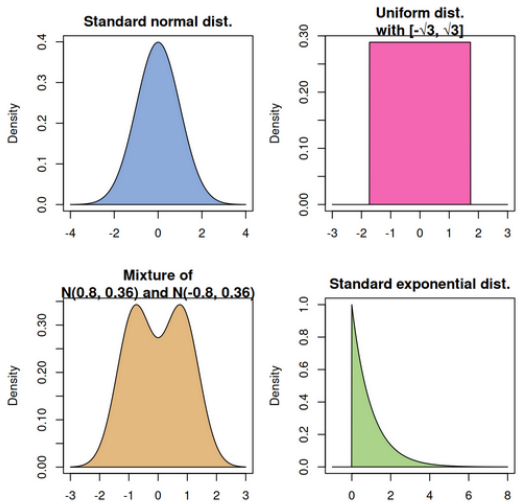


Figure 1.13: Four very different probability distributions. All of them have an expected value of 0 and a variance of 1.

We therefore need an additional assumption.

Assumption MLR.6 (Normality). The error term of the population is independent of the explanatory variables $x_1, \ldots, x_K$ and is normally distributed with mean zero and variance σ^2:

$$u_i \sim \mathrm{N}(0, \sigma^2) \tag{1.202}$$

This assumption implies Assumptions MLR.4 (exogenous errors) and MLR.5 (homoskedasticity). We will still be talking about "Assumptions MLR.1 through MLR.6" to make it clear that MLR.6 is something that we assume "on top" of the other assumptions. Together, we refer to Assumptions MLR.1 to MLR.6 collectively as the *classical linear model* (CLM) assumptions. Under the CLM assumptions, OLS is not only BLUE, but BUE, meaning that it is the *best unbiased estimator*, as opposed to being only the best among the class of *linear* estimators.

We can summarize the CLM assumptions as follows:

$$y_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right). \tag{1.203}$$

This means that the outcome, conditional on the explanatory variables, is normally distributed with the predicted value $\hat{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ as mean and a variance of σ^2. Figure 1.14 offers an illustration of this for the bivariate case (meaning the subscripts are i, not k). For each observation of the explanatory variable x, we *expect* the observed value of y to be on the regression line (which is the set of predicted values), but the value we actually observe will be normally distributed around that value.

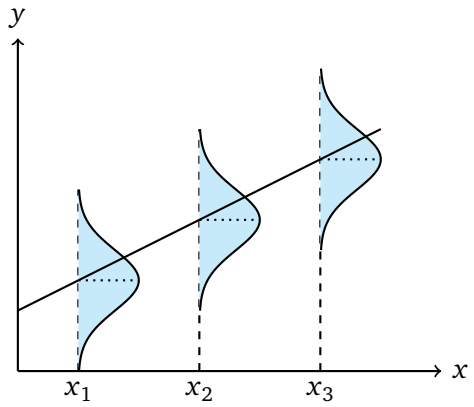


Figure 1.14: For each x_i, the outcome is normally distributed with mean $\boldsymbol{x}_i'\boldsymbol{\beta}$ and variance σ^2.

Of course, since MLR.4 and MLR.5 were already very restrictive, MLR.6 is an *extremely strong assumption*. One argument that is sometimes used to justify it is that the error term is a *sum* of unobserved factors that affect y_i. As such, we can apply the *central limit theorem* and u_i is approximately normally distributed:

Central Limit Theorem. Let $\{X_1, X_2, \ldots, X_N\}$ be a sequence of independently and identically distributed random variables with mean μ and variance σ^2. Then, the distribution function of the standardized random variable

$$Z_N = \frac{\bar{X}_N - \mu}{\sigma / \sqrt{N}} \tag{1.204}$$

(a standardized version of the mean), where $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$, converges in distribution to the distribution function of the standard normal distribution. Intuitively, as N increases, the distribution of the sample mean of the X_i converges to a normal distribution.

As we can see from this, the CLT argument only works if the factors constituting u_i do not follow wildly different distributions themselves. But they very likely do. In addition, nothing guarantees that all of these factors enter the error term additively. For both of these reasons, the assumption of normality of the errors is unreasonably strong. We will continue with it anyway, since we will later see that it is not as important as long as our sample is large enough.

Distribution of the OLS Estimator. Under the CLM assumptions MLR.1 through MLR.6, the OLS estimator, given the sample values of the independent variables, is normally distributed:

$$\hat{\beta}_k \sim \mathrm{N}(\beta_k, \mathrm{Var}(\hat{\beta}_k)), \tag{1.205}$$

where $\mathrm{Var}(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{ik}-\bar{x}_k)^2} \times \frac{1}{1-R_k^2}$, and R_k^2 is the R^2 from a regression of x_k on all other regressors $x_j, j \neq k$.

This means that $\hat{\beta}_k$ is normally distributed. In addition, any linear combination of the estimators $\hat{\beta}_k$ is also normally distributed; and the joint distribution of any subset of them is a multivariate normal distribution. The standardized coefficient

$$\hat{\beta}_k^{(\text{standardized})} = \frac{\hat{\beta}_k - \beta_k}{\mathrm{sd}(\hat{\beta}_k)} \tag{1.206}$$

follows a standard normal distribution.

1.4.3 t-Test

Since we now know the distribution of the OLS estimators, we can start to discuss how we are going to *test hypotheses* about them. First, we will talk about how to test *hypotheses about one parameter* of the population model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i. \tag{1.207}$$

We have already seen that the standardized estimator follows a standard normal distribution:

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{sd}(\hat{\beta}_k)} \sim \mathrm{N}(0,1). \tag{1.208}$$

The problem with this expression, of course, is that we do not know the standard deviation of the estimator and have to replace it by the estimator's *standard error*, which amounts to replacing σ^2 by $\hat{\sigma}^2$. If we standardize the estimator using this standard error instead, it follows a slightly different distribution, a t-distribution with $N-K-1$ degrees of freedom:

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{se}(\hat{\beta}_k)} \sim t_{N-K-1}. \tag{1.209}$$

You can think of the t-distribution as being very similar to the standard normal distribution. The only difference is that it has fatter tails. As the degrees of freedom increase, the t-distribution's tails get thinner and it converges to a standard normal distribution. You can see a comparison in Figure 1.15.

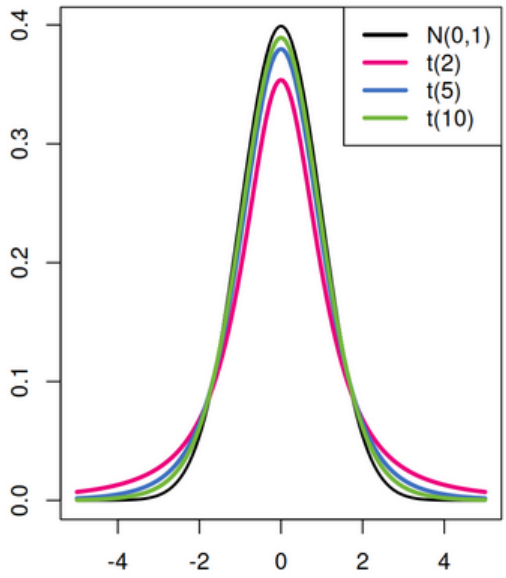


Figure 1.15: t-distributions with different degrees of freedom, compared to a standard normal distribution.

We can now continue by specifying a *null hypothesis*:

$$H_0 : \beta_k = 0, \tag{1.210}$$

or, in words: After accounting for all $x_j, j \neq k$, x_k has no effect on y. In the next step, we are again going to compute what we previously called a standardized coefficient, but now, we are calling it our *test statistic*:

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)}. \tag{1.211}$$

This particular test statistic is called a *t-statistic*. Under the null hypothesis, β_k is assumed to be zero, and so the t-statistic simplifies to

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k}{\text{se}(\hat{\beta}_k)}. \tag{1.212}$$

What is now very convenient for us is that we know the distribution of this test statistic: The t-statistic is t-distributed (hence the name) with $N - K - 1$ degrees of freedom. If our null hypothesis looks like the one above, we are testing it against a *two-sided alternative*. This means that we will reject the null hypothesis both when the estimate is too small and when it is too large. In Figure 1.16, you can see a t-distribution with 25 degrees of freedom. If the null hypothesis is true, then the t-statistics of the estimates we get should be distributed as shown in the figure. The idea of the t-test is now: If the estimate we actually received is *so extreme* that its t-statistic is in the small shaded areas left and right of the two colored

lines, then we consider it unlikely that the null hypothesis is true. Since we then reject the null hypothesis, we call the shaded areas *rejection regions*.

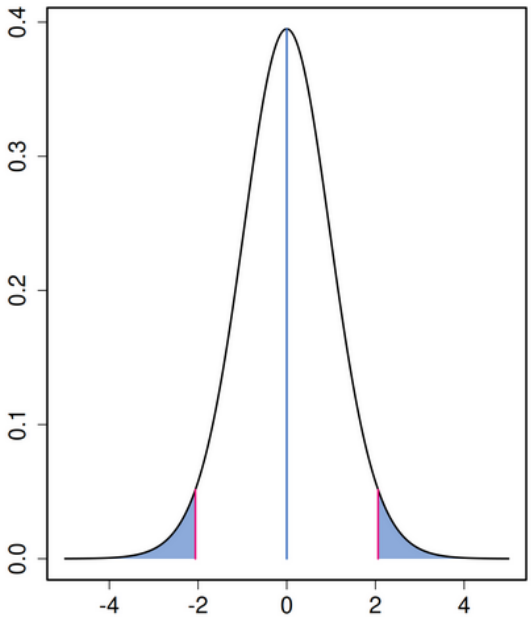


Figure 1.16: A two-sided t-test.

Where the rejection regions lie depends on the shape of the distribution, which in the case of the t-distribution is only determined by its degrees of freedom. In Figure 1.16, both rejection regions jointly cover an area of 0.05. This represents the probability of us rejecting the null hypothesis even though it was true, and this probability we can set to an arbitrary number. We call this number the *significance level*. Most commonly, it is set to 0.05; although, for example, R outputs significance levels of 0.1, 0.05, 0.01, and 0.001 by default.

Based on the significance level we have set, we can determine the thresholds the test statistic must exceed in order for us to reject the null hypothesis. We call these thresholds the *critical values*. For a t-distribution with 25 degrees of freedom, like the one pictured, the critical values are -2.06 and 2.06. This means that we reject the null hypothesis if the absolute value of the t-statistic exceeds 2.06. Since the shape of the distribution changes with its degrees of freedom, the critical values also depend on the degrees of freedom. But since it usually changes only a little, a critical value of 2 is an often-used approximation for eyeballing significance at the 0.05 level. As the t-distribution converges to the normal distribution when its degrees of freedom approach infinity, the critical values for a significance level of 0.05 will approach -1.96 and 1.96.

As mentioned previously, choosing a significance level of 0.05 means that we will falsely reject the null hypothesis in exactly 5 percent of cases. This is intuitive: Under the null hypothesis, there is a 5 percent chance that we will receive an estimate that is so extreme that its t-statistic exceeds the critical values. If that happens, we reject the null even though our estimates actually arose from a situation where the null hypothesis held. We call this situation a *type 1 error*, or alternatively, a *false positive*. We can set the probability for this error ourselves by altering the significance level. The probability of a *type 2 error*, also called a *false negative*, is much harder to assess. This refers to a situation where we fail to reject the null hypothesis even though it is not true.

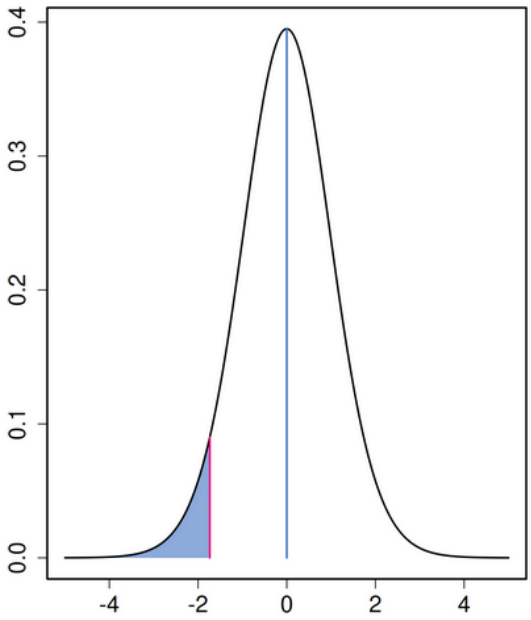


Figure 1.17: A one-sided t-test.

The two-sided t-test with $H_0 : \beta_k = 0$ is not the only t-test we can conduct. Figure 1.17, for example, shows a one-sided t-test. The null and alternative hypotheses in this case are

$$H_0 : \beta_k \geq 0, \qquad H_A : \beta_k < 0, \tag{1.213}$$

that is, we are testing the null hypothesis that the parameter is larger than or equal to zero. The alternative hypothesis is that it is smaller than zero. In a one-sided test, the entire rejection region is on the same side, which means that the critical value for the same significance level and degrees of freedom will be different (Its absolute value will be lower, since "more" is cut off on one of the sides). In the same way, we can run one-sided and two-sided tests with null hypotheses relating to numbers other than zero, such as

$$H_0 : \beta_k \geq (-1), \qquad H_A : \beta_k < (-1) \tag{1.214}$$

or

$$H_0 : \beta_k = 3, \qquad H_A : \beta_k \neq 3. \tag{1.215}$$

Note that since the t-statistic is a standardized measure, its distribution will still be centered around zero.

Suppose we conduct a test as before, with a significance level of $\alpha = 0.05$ and 25 degrees of freedom, and obtain a t-statistic of $t = 2.5$. In Figure 1.18, the areas shaded in pink represent the probability for the t-statistic being even "more extreme" (meaning, having a greater absolute value) than 2.5. The sum of the areas of these two regions, and thus the probability to get a more extreme t-statistic, is 0.019. We call this value the *p-value*, and it is very useful for interpretation. At the critical value, the p-value will always precisely equal the significance level. If we move beyond it, the p-value will decrease. This gives us a useful tool to assess significance without knowing the critical values for every possible distribution: If the p-value is smaller than the significance level, we reject the null hypothesis.

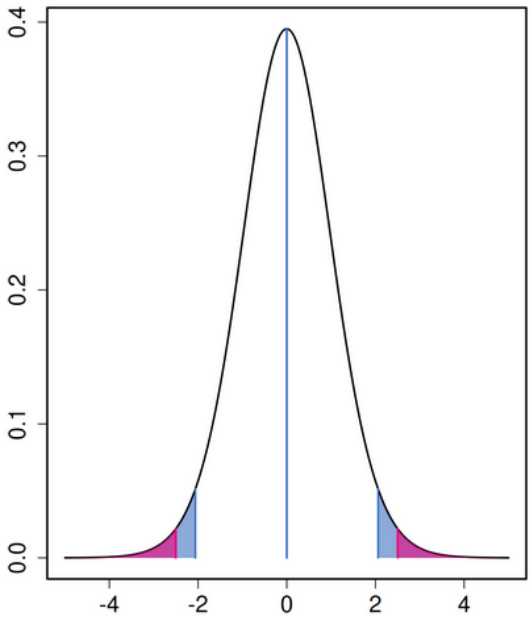


Figure 1.18: A two-sided t-test. Regions where the t-statistic has an absolute value greater than 2.5 are marked in pink.

Regarding all of this, we have to be very precise with which words we use to describe our results and interpret them. There are a lot of very closely related notions, some of which are supported by this testing procedure, while others are not. Suppose the p-value associated with $\hat{\beta}_2$ is 0.03 and the significance level is 0.05. Then, the following statements are supported by the test:

- x_2 is statistically significant at a significance level of 0.05.
- $\hat{\beta}_2$ is statistically significantly different from zero at a 0.05 significance level.
- We reject the null hypothesis at a 0.05 significance level.
- At a significance level of 0.03, the test would be indifferent between rejection and non-rejection.

However, the following statements are not supported:

- We accept the alternative hypothesis.
- The probability that the null hypothesis is true is 3 percent.

... at a 0.95 significance level.

- We are 97 percent confident that x_2 has an effect on y.

When a variable is *statistically significant*, we have evidence for it affecting the outcome. Importantly however, this is not a statement about how meaningful that influence is. That concept is sometimes called *economic significance* or *practical significance*. The basic idea behind it is that not every statistically significant variable is also an important factor of influence on y, so the magnitude of the effect should also be taken into consideration. A

variable that is both statistically significant and has a meaningfully large associated coefficient may then be interpreted as "statistically and economically significant." Of course, what "meaningfully large" translates to depends entirely on the variables, the question, and how the data is scaled.

Under the CLM assumptions, we can also calculate a *confidence interval* for a population parameter β_k. The interpretation of, for example, a 95 percent confidence interval is the following: If we repeatedly draw samples and compute the confidence interval, the interval will contain the true parameter in 95 percent of cases. We *cannot* interpret it as the parameter falling within the interval 95 percent of the time. This is because the population parameter is fixed, while the confidence interval changes. The 95 percent confidence interval for a parameter β_k is given as

$$\left[\hat{\beta}_k - c \times \text{se}(\hat{\beta}_k), \quad \hat{\beta}_k + c \times \text{se}(\hat{\beta}_k)\right], \tag{1.216}$$

where c is the 97.5th percentile of a t-distribution with $N-K-1$ degrees of freedom.

1.4.4 F-Test

Using the t-test, we can only ever impose and test a single restriction at a time, for example

$$\beta_1 = 0. \tag{1.217}$$

But what if we want to test *multiple restrictions* jointly? For example, we may be interested whether a particular group of independent variables collectively has no effect on y:

$$\beta_1 = 0, \beta_2 = 0, \beta_3 = 0. \tag{1.218}$$

We cannot do this with a t-test. To test restrictions like these, we need a different test, called the *F-test*.

In the above example, the null and alternative hypotheses are:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0; \qquad H_A : H_0 \text{ is not true.} \tag{1.219}$$

We are testing three *exclusion restrictions*, that is, multiple hypotheses, simultaneously. Since we cannot rely on t-statistics, which can only test one restriction at a time, we need a different test statistic, and we need to know its distribution.

To find it, we start by writing down our full, unrestricted model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u. \tag{1.220}$$

Next, we apply all of our restrictions and write down the restricted model:

$$y = \beta_0 + \beta_4 x_4 + \beta_5 x_5 + u. \tag{1.221}$$

One straightforward approach is now to evaluate by how much the sum of squared residuals increases when we remove the three additional variables. Since the sum of squared residuals *always* increases when we remove variables from a model, we need a test statistic that evaluates how large the relative increase in the SSR is when we apply our restrictions.

The *F-statistic* is one such test statistic. It is given as,

$$F = \frac{(\text{SSR}_r - \text{SSR}_{ur})/q}{\text{SSR}_{ur}/(N-K-1)}, \tag{1.222}$$

where q is the number of restrictions imposed, SSR_r denotes the SSR from the restricted model, and SSR_{ur} denotes the SSR from the unrestricted model. Under the CLM assumptions, it follows an *F-distribution* with q degrees of freedom in the numerator and $(N-K-1)$ degrees of freedom in the denominator. We can alternatively describe the F-statistic as a function of the R^2 of the restricted and unrestricted models instead of the sums of squared residuals:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(N-K-1)}. \tag{1.223}$$

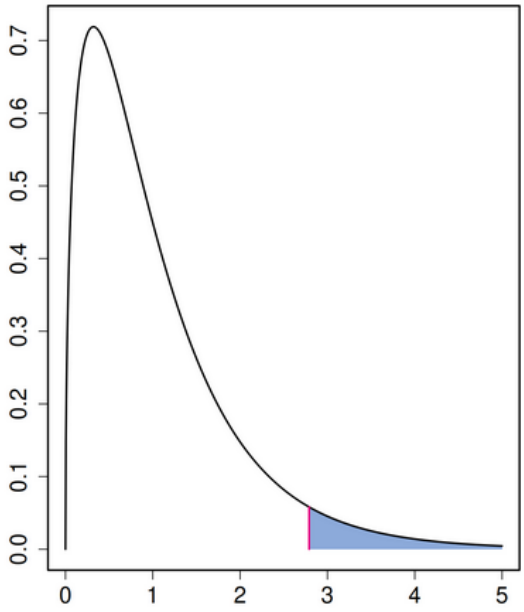


Figure 1.19: The distribution of an F-statistic with 3 and 50 degrees of freedom.

Figure 1.19 shows an F-test of 3 restrictions, meaning that the numerator degrees of freedom are 3. The denominator degrees of freedom are 50. We can see that the value of the F-statistic is never negative, which makes sense given the explanation as a measure of the increase in the sum of squared residuals, which we said was always positive. In the case depicted here, the critical value is 2.798. The F-test is always one-sided, which means that we reject the null hypothesis if we obtain an F-statistic that is greater than this critical value. If we can reject the null hypothesis, we say that the variables on which we imposed the restrictions are jointly significant. If we fail to reject the null, we call them jointly insignificant. Conducting an F-test for only one restriction will always give the same result as the corresponding t-test; however, several individually insignificant variables can still be jointly significant.

Most often, we encounter the F-test as *global F-test*, meaning that we test the joint significance of all explanatory variables in the model. This test is also automatically computed by most statistical software immediately upon running a regression. The F-statistic for this case can be written as

$$F = \frac{R^2/K}{(1-R^2)/(N-K-1)} \tag{1.224}$$

since the "restricted model" here is simply no model and thus has an R^2 of zero. For both this global F-statistic as well as any other F-statistic, statistical software usually provides a p-value. It has the same interpretation as a p-value associated with a t-statistic, which greatly simplifies interpretation.

1.4.5 Interpretation of Regression Tables

t-statistics and p-values are a large part of regression tables which we always had to disregard so far. So let us run a regression again, and see how much we can now interpret. This time, we are using the `mlb1` baseball dataset from the Wooldridge (2020) textbook.

```
# Load packages
library(wooldridge) # Contains the dataset
library(dplyr)
library(car) # For F-test later

# Load data
data("mlb1") # Baseball data

# Keep only the variables we're interested in
mlb1 <- mlb1 |>
  select(salary, years, gamesyr, bavg, hrunsyr, rbisyr)
```

We start by estimating a regression of the logarithmized salary of baseball players on their career length, their games per year, their batting average, the number of home runs per year, and the number of runs batted in per year.

```
lm(log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr,
   data = mlb1) |>
  summary()
```

```
Call:
lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +
    rbisyr, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max
-3.02508 -0.45034 -0.04013  0.47014  2.68924

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.119e+01  2.888e-01  38.752  < 2e-16 ***
years       6.886e-02  1.211e-02   5.684 2.79e-08 ***
gamesyr     1.255e-02  2.647e-03   4.742 3.09e-06 ***
bavg        9.786e-04  1.104e-03   0.887    0.376
hrunsyr     1.443e-02  1.606e-02   0.899    0.369
rbisyr      1.077e-02  7.175e-03   1.500    0.134
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7266 on 347 degrees of freedom
Multiple R-squared:  0.6278,    Adjusted R-squared:  0.6224
F-statistic: 117.1 on 5 and 347 DF,  p-value: < 2.2e-16
```

Next to the estimate for each coefficient, there is its standard error. Next to the standard error, R automatically computes the t-statistic (which is just the ratio of the estimate and the

standard error) and also outputs the associated p-value. Significant explanatory variables are marked using a series of symbols for different significance levels. A variable with one star next to it is significant at the 0.05 level, with more stars indicating significance even when the level is set lower. In the last line of the output, we can see the F-statistic for joint significance of all explanatory variables as well as its associated p-value.

Next, we estimate a smaller (restricted) model, where we impose the restriction that `bavg`, `hrunsyr`, and `rbisyr` are zero, and compare the coefficients as well as the individual t-statistics.

```
lm(log(salary) ~ years + gamesyr, data = mlb1) |>
  summary()
```

```
Call:
lm(formula = log(salary) ~ years + gamesyr, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66858 -0.46412 -0.01177  0.49219  2.68829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.223804   0.108312 103.625  < 2e-16 ***
years        0.071318   0.012505   5.703  2.5e-08 ***
gamesyr      0.020174   0.001343  15.023  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7527 on 350 degrees of freedom
Multiple R-squared:  0.5971,    Adjusted R-squared:  0.5948
F-statistic: 259.3 on 2 and 350 DF,  p-value: < 2.2e-16
```

But what if we actually *test* the three restrictions we imposed? For that, we can save the larger model again and then conduct an F-test using the `linearHypothesis()` function.

```
model <- lm(log(salary) ~ years + gamesyr + bavg + hrunsyr +
   rbisyr, data = mlb1)

linearHypothesis(model, c("bavg = 0", "hrunsyr = 0", "rbisyr
   = 0"))
```

```
Linear hypothesis test:
bavg = 0
hrunsyr = 0
rbisyr = 0

Model 1: restricted model
Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    350 198.31
2    347 183.19  3    15.125 9.5503 4.474e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that even though none of the three variables which we imposed the restrictions on was individually significant, they were *jointly significant*.

1.4.6 Large Samples

All of the properties of the OLS estimator we have discussed so far apply to *finite samples*, no matter how large or small N is. This includes the unbiasedness of the estimators, the Gauss-Markov theorem, and everything else we mentioned. But it only includes everything we discussed in this section *if we assume MLR.6* (normality of errors) – and we have already said that MLR.6 is an unreasonably strong assumption.

Fortunately, this does not mean that everything we said about testing is useless. This is because in addition to its *finite sample properties*, OLS has certain *large sample properties*. These are properties that arise when N approaches infinity. If we deal with a particularly small sample, these may not hold. Actually, they may not hold for any sample size N – but as N grows, we get at least closer to a situation where they hold. This is useful because some finite sample properties hold in large samples even when we relax certain assumptions – for instance (you may have guessed it) MLR.6.

Without assuming normality of errors, the t-statistic does not necessarily follow a t-distribution, and the F-statistic does not necessarily follow a F-distribution. This means that we cannot test hypotheses about parameters the way we did before. Conveniently, using the central limit theorem, the following can be shown for large samples:

Asymptotic Normality. Under Assumptions MLR.1 through MLR.5, the t-statistic is asymptotically normally distributed:

$$\frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)} \xrightarrow{\text{d}} \text{N}(0,1) \qquad \text{or} \qquad \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)} \xrightarrow{\text{d}} \text{t}_{N-K-1}. \tag{1.225}$$

Since the t-distribution converges in distribution to a standard normal distribution as its degrees of freedom increase, we can use either the left or the right representation. This means that we can use the t-statistic just the way we used it *with* assuming MLR.6, given that our sample size is large enough. The asymptotic normality of OLS estimators also implies that the F-statistic is asymptotically F-distributed in larger samples.

In large samples, there is also an alternative to the F-test: the *Lagrange multiplier test* (LM test). It rarely leads to different results than the F statistic, but the testing procedure has notable parallels to tests we will discuss in later sections of the course, so we are going to briefly discuss it. The basic idea of this test is to check whether the additional explanatory variables in the full model can explain the residuals from the restricted model. If they can, we reject the null hypothesis of joint insignificance of the restricted variables. We obtain the LM statistic by following these steps:

1. Estimate only the restricted model.
2. Take the residuals from the regression from Step 1, and regress them on all K independent variables from the full model.
3. Compute the LM statistic
$$\text{LM} = NR^2, \tag{1.226}$$
where R^2 is the R^2 from the regression in Step 2.

Under Assumptions MLR.1 through MLR.5, the LM statistic is asymptotically χ^2_q-distributed (its small sample distribution is, however, unknown).

1.5 More on Multiple Regression

1.5.1 Large Samples

In this section, we will get to know some additional properties of multiple linear regression. We will begin by talking about *large sample properties*. This is similar to how we ended the previous section, but then we were talking about tests, and now we are talking about more general properties.

We already know that the OLS estimator is *unbiased*. Now, we will learn that it is also *consistent*. Unbiasedness and consistency are somewhat similar properties that are often confused, so it is important to always distinguish them properly: An estimator is *unbiased* if its expected value equals the true parameter. An estimator is *consistent* if the estimates *converge in probability* to the true parameter as N increases. It is evident from this definition that consistency is fundamentally a large sample property, while unbiasedness holds in every sample, no matter how small. Under Assumptions MLR.1 to MLR.4, the OLS estimator is both unbiased and consistent.

The following is a quick sketch of a proof for the consistency of the OLS estimator:

$$\begin{aligned}
\operatorname{plim} \hat{\boldsymbol{\beta}} &= \operatorname{plim} ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= \operatorname{plim} ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u})) \\
&= \operatorname{plim} ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) \\
&= \operatorname{plim} \boldsymbol{\beta} + \operatorname{plim} ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) \\
&= \boldsymbol{\beta} + \operatorname{plim} (\boldsymbol{X}'\boldsymbol{X})^{-1} \operatorname{plim} \boldsymbol{X}'\boldsymbol{u} \\
&= \boldsymbol{\beta} + \operatorname{plim} (N^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1} \operatorname{plim} (N^{-1}\boldsymbol{X}'\boldsymbol{u}),
\end{aligned} \tag{1.227}$$

where the *probability limit* $\operatorname{plim} X_n = X$ means that X_n converges in probability to X as $N \to \infty$. In the final step, we multiply the second term once by N^{-1} and once by $(N^{-1})^{-1}$ so that we can directly apply the law of large numbers. Since we know that $\operatorname{plim} (N^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}$ is invertible, the only thing we need to show in order to show that the OLS estimator is consistent is that $\operatorname{plim} (N^{-1}\boldsymbol{X}'\boldsymbol{u}) = \boldsymbol{0}$. This is the case because, when N approaches infinity, the sample covariance $\boldsymbol{X}'\boldsymbol{u}$ converges to the population covariance. This then means that we can directly apply MLR.4, in which we have assumed that all x_k are uncorrelated with the error term. We therefore get

$$\operatorname{plim}(N^{-1}\boldsymbol{X}'\boldsymbol{u}) = \operatorname{plim} N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' u_i = \boldsymbol{0}. \tag{1.228}$$

Although we have used MLR.4 when outlining the consistency proof, we could actually have gone with a weaker assumption. Let us explicitly state this weaker assumption:

Assumption MLR.4' (Zero Mean and Zero Correlation.) The error term has expected value zero and is uncorrelated with any explanatory variable:

$$\mathrm{E}(u) = 0, \qquad \operatorname{Cov}(x_k, u) = 0 \quad \text{for } k = 1, \ldots, K. \tag{1.229}$$

The stronger Assumption MLR.4 implies this weaker Assumption MLR.4'. Under Assumption MLR.4', a nonlinear function of a regressor, such as x_1^2, is allowed to be correlated

with the error term, which would not be the case under Assumption MLR.4. Expectedly, if MLR.4 does not hold, but MLR.4' does, the OLS estimator is biased, but consistent.

Let us take a closer look at the expression for the inconsistency of the OLS estimator:

$$\text{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \text{plim}\,(N^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}\text{plim}\,(N^{-1}\boldsymbol{X}'\boldsymbol{u}). \tag{1.230}$$

For one element of $\hat{\boldsymbol{\beta}}$, for example $\hat{\beta}_1$, we can alternatively write:

$$\text{plim}\,\hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x_1, u)}{\text{Var}(x_1)}. \tag{1.231}$$

This tells us that the direction of the inconsistency depends on the covariance between x_1 and u. If $\text{Cov}(x_1, u) > 0$, the inconsistency will be positive, and if $\text{Cov}(x_1, u) < 0$, the inconsistency will be negative. This knowledge, however, is of limited practical use, since we never observe u and thus cannot say anything about its covariance with x_1.

1.5.2 Scaling, Transforming, Interacting

We already know since the very beginning that scaling a variable changes certain coefficients in certain ways. Scaling the outcome affects all coefficients, whereas scaling one input scales only the associated coefficient. Have a look at the following two equations to remind yourself of this fact:

$$y^* = 10\beta_0 + 10\beta_1 x_1 + 10\beta_2 x_2 + 10u, \qquad y^* = 10 \times y \tag{1.232}$$

$$y = \beta_0 + \frac{\beta_1}{10} x_1^* + \beta_2 x_2 + u, \qquad x_1^* = 10 \times x_1 \tag{1.233}$$

But now that we know more about hypothesis testing, we should briefly consider what happens to test statistics when their associated variables are scaled. After all, it would be very unfortunate if scaling our variable would suddenly render it insignificant. Fortunately, this would not make any sense, since test statistics are standardized, and thus all t-statistics, F-statistics, and p-values remain the same after we scale a variable. Confidence intervals scale in the same way as the associated variable.

Now, let us revisit logarithmic transformations. Remember Table 1.8, which also occurred earlier. As we know, $\%\Delta\hat{y} \approx 100 \times \Delta\log(y)$ is an approximation that works especially well for small variables. If changes are larger, we can calculate the exact change:

$$\%\Delta\hat{y} = 100 \times (\exp(\hat{\beta}_k \Delta x_k) - 1). \tag{1.234}$$

However, this exact value differs based on the size of the change. For example, it will be different for $\Delta x_k = 1$ and $\Delta x_k = -1$. Since the approximation lies between these two values, it can be a useful benchmark for interpretation even when the percentage change is large.

There are several good reasons for log-transforming variables. The most straightforward is that an economic model, or our intuition, tells us that the relationship we want to model is an elasticity or a semi-elasticity. These relationships are adequately captured only by log-transformed models. An additional reason is that, if all $y_i > 0$, models with a log-transformed outcome often come closer to fulfilling the CLM assumptions than models with plain y as the dependent. Taking the logarithm can mitigate issues with heteroskedasticity. Also, if a variable has very extreme values, taking the logarithm can reduce the influence of outliers.

Model	**Dep. Variable**	**Indep. Variable**	**Interpretation**
Level-Level	y	x	$+1$ in $x \Leftrightarrow +\beta_1$ in y
Level-Log	y	$\log(x)$	$+1\%$ in $x \Leftrightarrow +\beta_1/100$ in y
Log-Level	$\log(y)$	x	$+1$ in $x \Leftrightarrow +\beta_1 \times 100\%$ in y
Log-Log	$\log(y)$	$\log(x)$	$+1\%$ in $x \Leftrightarrow +\beta_1\%$ in y

Table 1.8: Interpretations of Level-Level, Level-Log, Log-Level, and Log-Log models.

All of this sounds like logarithms are a convenient one-size-fits-all solution for a whole bunch of commonly encountered problems. But they are not. Whenever we log-transform a variable, we need a reason to do so. There are also good reasons *not* to logarithmize a variable. Again, the most straightforward is us not wanting to model an elasticity or semi-elasticity. If we want to model a linear relationship, we need an untransformed model. Another reason is that logarithms can not only mitigate, but also *create* extreme values: If a lot of our data is very close to zero, their logarithms will be negative values with huge magnitudes.

We have briefly mentioned *quadratic functions* before, so let us now talk about them in more detail. They allow us to model nonlinear relationships of the following kind:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u. \tag{1.235}$$

Importantly, since a quadratic function is not a linear function, including both x_1 and x_1^2 does not cause issues with MLR.3. Where it does cause an issue, however, is interpretation: It makes no sense to interpret β_1 without considering β_2. We cannot induce a change in x while, at the same time, holding x^2 constant. β_1 therefore cannot be interpreted as a partial effect. Consider the following equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2. \tag{1.236}$$

Here, we can approximate

$$\frac{\Delta \hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \tag{1.237}$$

This is one way to interpret partial effects in a model that contains quadratic terms in addition to linear ones. If we are interested in a specific effect starting from a particular initial value, we can just plug into the equation and do not need an approximation. Often, we will also compute average partial effects, which we will discuss very soon.

But first, we are going to talk about *interaction terms*. We can use interaction terms in situations where the effect of one variable depends on the value of another variable. Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \; x_1 \times x_2 + u. \tag{1.238}$$

Here, $x_1 \times x_2$ represents an *interaction* between x_1 and x_2. Similarly to before, parameters do not directly represent partial effects. For example, the partial effect of x_1 is

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2. \tag{1.239}$$

This means that the effect of x_1 on y depends on x_2: If β_3 is positive, the effect of x_1 on y is stronger when x_2 is high; and if β_3 is negative, the effect of x_1 on y is stronger where x_2 is low.

In both of these cases, quadratic terms and interactions (as well in all other situations where coefficients do not represent partial effects), the *average partial effect* (APE) is a useful summary metric. Suppose we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + u. \tag{1.240}$$

We can calculate average partial effects by estimating the model, plugging in the estimates we get, computing the partial effect for each observation (based on actual data we insert), and then averaging those individual partial effects to get one average partial effect. Alternatively, the *partial effect at the average* (PEA) is sometimes calculated by plugging the sample mean of all x variables into the model and then computing a partial effect. However, calculating averages for dummy variables and nonlinearly transformed variables can be problematic.

We can see how average partial effects look like in a real regression. To do so, we load a dataset from the `wooldridge` package that contains data on exam results and attendance, and run a regression:

```
library(wooldridge) # Contains the dataset
library(dplyr) # Contains useful functions
library(margins) # for APE

data("attend") # Data on exam results and attendance

model_1 <- lm(stndfnl ~ atndrte + priGPA + ACT + I(priGPA^2)
   + I(ACT^2) + priGPA * atndrte, data = attend)
summary(model_1)
```

```
Call:
lm(formula = stndfnl ~ atndrte + priGPA + ACT + I(priGPA^2) +
    I(ACT^2) + priGPA * atndrte, data = attend)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1698 -0.5316 -0.0177  0.5737  2.3344

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.050293   1.360319   1.507 0.132225
atndrte        -0.006713   0.010232  -0.656 0.512005
priGPA         -1.628540   0.481003  -3.386 0.000751 ***
ACT            -0.128039   0.098492  -1.300 0.194047
I(priGPA^2)     0.295905   0.101049   2.928 0.003523 **
I(ACT^2)        0.004533   0.002176   2.083 0.037634 *
atndrte:priGPA  0.005586   0.004317   1.294 0.196173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8729 on 673 degrees of freedom
Multiple R-squared:  0.2287,    Adjusted R-squared:  0.2218
F-statistic: 33.25 on 6 and 673 DF,  p-value: < 2.2e-16
```

We can see a lot of coefficients, but the interpretation of none of them is straightforward. To be able to interpret the influence that the different variables have on the final exam score, we need to compute, for instance, average partial effects:

```
library(margins)

margins(model_1)

summary(margins(model_1))
```

```
Average marginal effects

lm(formula = stndfnl ~ atndrte + priGPA + ACT + I(priGPA^2) +     I(
    ACT^2) + priGPA * atndrte, data = attend)

  atndrte priGPA     ACT
 0.007737 0.3588 0.07606

  factor     AME      SE       z       p   lower   upper
     ACT  0.0761  0.0112  6.7914  0.0000  0.0541  0.0980
 atndrte  0.0077  0.0026  2.9384  0.0033  0.0026  0.0129
  priGPA  0.3588  0.0778  4.6121  0.0000  0.2063  0.5112
```

Now, we can see that the ACT score (a college entrance exam), attendance, and the GPA a student previously had all have a positive influence on the final exam score. This influence is statistically significant at the 0.05 significance level, since p-values are smaller than 0.05. The average effect of the attendance rate increasing by one percentage point, for instance, is an increase in the standardized final exam score by about 0.008. A change from an attendance rate of 50 percent to 100 percent thus would yield to an average improvement of the standardized final score by about 0.4 standard deviations. The other variables and their effects can be interpreted in a similar fashion.

1.5.3 Goodness of Fit

We already know that the *coefficient of determination* R^2, given by

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}, \tag{1.241}$$

is only of limited use when we want to compare models. This becomes especially apparent in the multivariate case. When we add more variables to a model, the explained sum of squares will *always* increase (technically, it can also stay the same), and so the R^2 can only increase as well. Since a larger model is not always better, using the R^2 to compare models of different size is going to lead to a bunch of plainly wrong decisions. Also, a low R^2 is not necessarily a "bad" sign or even something we always care about. Remember that a low coefficient of determination only means that the explained variation is rather small, compared to the entire variation in the outcome. This may imply imprecise estimates, but does not have to. For instance, in a randomized experiment, we only need one explanatory variable to estimate its effect precisely. Since there still is a bunch of variation that we do not explain, the R^2 will be low even though we have conducted the cleanest research we possibly can.

One way to deal with this is to compute the *adjusted* R^2. It effectively deducts a penalty for additional variables:

$$R^2_{\text{adj.}} = 1 - \frac{\text{SSR}/(N-K-1)}{\text{SST}/(N-1)} = 1 - \left(1-R^2\right) \times \frac{N-1}{N-K-1}. \tag{1.242}$$

The penalty term $\frac{N-1}{N-K-1}$ becomes larger as K increases, and so larger models that would otherwise have the same R^2 have a lower $R^2_{\text{adj.}}$. When N is small and K is large, $R^2_{\text{adj.}}$ can be substantially lower than R^2. In extreme cases, $R^2_{\text{adj.}}$ can even be negative.

The adjusted R^2 allows us to do something that we were not able to do before: *compare non-nested models*. The F-test allowed us to compare *nested* models, that is, we could use it in situations where one model was a special case of another. Now, we can compare models with different numbers of variables that are non-nested, which we can neither do with the regular R^2 nor with an F-test. We can also compare models with different functional forms of an explanatory variable. However, we cannot use the adjusted R^2 to compare models with different dependent variables, and that includes models with different transformations of the dependent variable. In other words, whether to use y or $\log(y)$ is still left only to our judgment.

Let us try that out using the baseball dataset we already know:

```
data("mlb1")

model_2 <- lm(log(salary) ~ years + gamesyr + hrunsyr, data =
    mlb1)
model_3 <- lm(log(salary) ~ years + gamesyr + bavg + rbisyr,
   data = mlb1)

summary(model_2)

summary(model_3)
```

```
Call:
lm(formula = log(salary) ~ years + gamesyr + hrunsyr, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max
-3.07403 -0.46113 -0.02591  0.45723  2.64506

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.344333   0.107645 105.387  < 2e-16 ***
years        0.068106   0.012123   5.618 3.96e-08 ***
gamesyr      0.016228   0.001525  10.639  < 2e-16 ***
hrunsyr      0.035863   0.007249   4.948 1.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7287 on 349 degrees of freedom
Multiple R-squared:  0.6235,    Adjusted R-squared:  0.6202
F-statistic: 192.6 on 3 and 349 DF,  p-value: < 2.2e-16



Call:
lm(formula = log(salary) ~ years + gamesyr + bavg + rbisyr, data =
    mlb1)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2.97116 -0.45464 -0.05178  0.46468  2.67529

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.128e+01  2.737e-01  41.197  < 2e-16 ***
years       6.973e-02  1.207e-02   5.776 1.70e-08 ***
gamesyr     1.116e-02  2.145e-03   5.202 3.37e-07 ***
bavg        7.398e-04  1.071e-03   0.691     0.49
rbisyr      1.652e-02  3.229e-03   5.117 5.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7264 on 348 degrees of freedom
Multiple R-squared:  0.6269,    Adjusted R-squared:  0.6226
F-statistic: 146.2 on 4 and 348 DF,  p-value: < 2.2e-1
```

We can see that the adjusted R^2 is lower than the regular R^2 for both models (which makes sense, since the penalty term is always larger than one). However, in this case, the penalty is not enough – the larger model is still better.

1.5.4 Dummy Variables

We have learned that we can use *dummy variables* to encode qualitative information in our models. In the following model, x_1 is a dummy variable:

$$y = \beta_0 + \beta_1 x_1 + \cdots + u, \qquad x_1 \in \{0, 1\} \tag{1.243}$$

In simple linear regression, interpretation was quite straightforward: β_0 was the mean in the group where $x_1 = 0$, and $\beta_0 + \beta_1$ was the mean in the group where $x_1 = 1$. In multiple linear regression, interpretation works analogously:

$$\mathrm{E}(y \mid x_1 = 1) = \beta_0 + \beta_1 + \cdots, \qquad \mathrm{E}(y \mid x_1 = 0) = \beta_0 + \cdots. \tag{1.244}$$

One thing we have already hinted at, but not discussed in detail, is that we can use *multiple dummy variables* to encode qualitative information that has more than two levels. Each category becomes its own dummy variable, except for one category, which is set as reference category. This is to avoid multicollinearity with the intercept. To see this, consider the following example: Suppose we want to use a car's color as a regressor. In the population, there are black, red, and blue cars, and so we construct

$$\begin{aligned}
\text{black}_i &= \begin{cases} 1 & \text{if } i \text{ is black,} \\ 0 & \text{otherwise} \end{cases}, \\
\text{red}_i &= \begin{cases} 1 & \text{if } i \text{ is red,} \\ 0 & \text{otherwise} \end{cases}, \\
\text{blue}_i &= \begin{cases} 1 & \text{if } i \text{ is blue,} \\ 0 & \text{otherwise} \end{cases}.
\end{aligned} \tag{1.245}$$

Assume we then estimate the following model:

$$y = \beta_0 + \beta_1 \,\text{black} + \beta_2 \,\text{red} + \beta_3 \,\text{blue} + u. \tag{1.246}$$

The *regressor matrix* in this model looks like this:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{1.247}$$

You can immediately see that the fourth column, x_3, is a linear combination of the other columns: $x_3 = 1 - x_1 - x_2$.

This is what we call the *dummy trap*. If we include both an intercept and every category of our dummy variable, we have *perfect multicollinearity*, and Assumption MLR.3 is violated. To avoid this, we can do one of two things. Either we estimate a model without an intercept. Since the dummy variables essentially serve as separate intercepts for all categories of the dummy variable, omitting the intercept does not cause any bias. Or we can apply the more common solution of omitting one of the category dummies while retaining the intercept. This way, the effect of the reference category is absorbed by the intercept.

Let us apply this solution to our example and define blue_i as the benchmark. We can then estimate:

$$y = \beta_0 + \beta_1 \,\text{black} + \beta_2 \,\text{red} + u. \tag{1.248}$$

Now, we have to be careful with how we interpret the parameters. β_0 is now the expected outcome for a blue car, since blue_i is represented by the intercept. $\beta_0 + \beta_1$ is the expected outcome for a black car, which means that β_1 is the expected difference between black and blue cars. Similarly, $\beta_0 + \beta_2$ is the expected outcome for a red car, and β_2 is the expected difference between red and blue cars. If we additionally include, for example, a numerical variable, then these values can be interpreted *group-specific intercepts*: For instance, $\beta_0 + \beta_1$ is the expected outcome for a black car where the additional numerical variable is equal to zero.

Since we now also know about *interactions*, we can do one more thing: Model *different slopes* for different groups. Assume we have one numerical variable in addition to the color categories in our model, x_3:

$$y = \beta_0 + \beta_1 \,\text{black} + \beta_2 \,\text{red} + \beta_3 x_3 + \beta_4 \,\text{black} \times x_3 + \beta_5 \,\text{red} \times x_3 + u. \tag{1.249}$$

The interactions between our dummy variables and the numerical variable allow for differing slope patterns between categories. Interpretation of these coefficients works analogously to how it worked for intercepts: The interpretation of β_0, β_1 and β_2 stays the same. β_3 is the slope coefficient associated with x_3 for the reference category, that is, blue cars. $\beta_3 + \beta_4$ is the slope coefficient associated with x_3 for black cars, and $\beta_3 + \beta_5$ is the slope coefficient associated with x_3 for red cars.

To see how such an interaction looks like, we can estimate a simple model. We use the `wage1` dataset to estimate a simple gender pay gap regression again, with a female dummy as the gender variable. This time, however, we also let education explain the wage, and allow for differing slopes with regard to education for female and non-female observations:

```
1 data("wage1")
2 model_4 <- lm(wage ~ female + educ + female * educ, data =
    wage1)
3 summary(model_4)
```

```
Call:
lm(formula = wage ~ female + educ + female * educ, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1611 -1.8028 -0.6367  1.0054 15.5258

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20050    0.84356   0.238    0.812
female      -1.19852    1.32504  -0.905    0.366
educ         0.53948    0.06422   8.400 4.24e-16 ***
female:educ -0.08600    0.10364  -0.830    0.407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.186 on 522 degrees of freedom
Multiple R-squared:  0.2598,    Adjusted R-squared:  0.2555
F-statistic: 61.07 on 3 and 522 DF,  p-value: < 2.2e-16
```

Because both `female` and `educ` appear twice, we cannot interpret significance straightforwardly. Both variables relating to gender are insignificant, but they might be jointly significant, which we can test using an F-test. So we will stick to naively interpreting the coefficients: One additional year of education leads to an increase of about 54 cents in an individual's hourly wage, but if that individual is a woman, this slope effect is only $0.539 - 0.086 = 0.453$ dollars.

We can also use a *dummy variable as dependent variable*. For example, we can examine which factors influence whether someone passes their econometrics course (study time, motivation, etc.). We can model this the following way:

$$y = \beta_0 + \beta_1 x_1 + \cdots + u, \qquad y_i = \begin{cases} 1 & \text{if } i \text{ passes the econometrics course,} \\ 0 & \text{otherwise} \end{cases}. \tag{1.250}$$

We call a model like this a *linear probability model* (LPM), since what we are modeling as outcome can be interpreted as a probability: $\hat{y}_i = 0.82$ would then mean that i has an 82-percent chance of passing econometrics. Accordingly, we interpret β_k as changes in this probability in percentage points. This is the simplest of many possible ways to model binary dependent variables, and as such, it has advantages (mostly its simplicity), but it also has disadvantages. The most straightforward disadvantage of the linear probability model is that predicted outcomes can fall outside the $[0, 1]$ range, and probabilities outside of that range have no interpretation. Because of this, there are a bunch of alternative, but more complicated models, which we will learn about in Econometrics II. That said, the linear probability model is still used frequently, mostly because it is easy to interpret and estimate.

To conclude this section, we can estimate a linear probability model. The `mroz` dataset contains observations on the labor force participation of married women and some relevant associated demographics. We are going to model female labor force participation, where `inlf` is 1 if a married woman is in the labor force and 0 otherwise, as depending on the non-wife income, the wife's education, experience, and age, as well as the presence of kids under and over six years of age.

```
1 data("mroz")
2 model_5 <- lm(inlf ~ nwifeinc + educ + exper + I(exper^2) +
    age + kidslt6 + kidsge6, data = mroz)
```

```
summary(model_5)
```

```
Call:
lm(formula = inlf ~ nwifeinc + educ + exper + I(exper^2) + age +
    kidslt6 + kidsge6, data = mroz)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93432 -0.37526  0.08833  0.34404  0.99417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5855192  0.1541780   3.798 0.000158 ***
nwifeinc    -0.0034052  0.0014485  -2.351 0.018991 *
educ         0.0379953  0.0073760   5.151 3.32e-07 ***
exper        0.0394924  0.0056727   6.962 7.38e-12 ***
I(exper^2)  -0.0005963  0.0001848  -3.227 0.001306 **
age         -0.0160908  0.0024847  -6.476 1.71e-10 ***
kidslt6     -0.2618105  0.0335058  -7.814 1.89e-14 ***
kidsge6      0.0130122  0.0131960   0.986 0.324415
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4271 on 745 degrees of freedom
Multiple R-squared:  0.2642,    Adjusted R-squared:  0.2573
F-statistic: 38.22 on 7 and 745 DF,  p-value: < 2.2e-16
```

You can use this model to practice interpretation of quadratic terms. We are only going to interpret the coefficient associated with `kidslt6`, the dummy variable indicating whether there are kids that are younger than six in the household. The effect is -0.26, which can be interpreted straightforwardly: Having a kid that is younger than six decreases the expected probability that the wife participates in the labor force by 26 percent.

1.6 Heteroskedasticity

1.6.1 What Is Heteroskedasticity?

When we introduced Assumption MLR.5 (Homoskedasticity),

$$\text{Var}(u_i \mid x_{i1}, \dots, x_{iK}) = \text{Var}(u_i) = \sigma^2, \tag{1.251}$$

we also discussed that this assumption is frequently violated. The assumption is violated when the *variance* of the error is related to some of the explanatory variables. This is very often the case: For example, people who have more years of education likely have more variance in their incomes. Some will have very high incomes, while others will have incomes that are more similar to the incomes of people with less education. Similarly, people with higher incomes may have greater variance in how much CO_2 emissions they cause. All of these cases are instances of *heteroskedasticity*.

Heteroskedasticity occurs when certain individuals, or groups of individuals, have more or less unexplained variation than the rest of the sample. More formally, we can say that the variance of the error term depends on i:

$$\text{Var}(u_i \mid x_{i1}, \dots, x_{iK}) = \text{E}(u_i^2 \mid x_{i1}, \dots, x_{iK}) = \sigma_i^2 \qquad \neq \sigma^2, \tag{1.252}$$

or in matrix notation,

$$\text{Var}(\boldsymbol{u} \mid \boldsymbol{X}) = \text{E}(\boldsymbol{u}\boldsymbol{u}' \mid \boldsymbol{X}) = \text{diag}(\sigma_1^2, \dots, \sigma_N^2) \qquad \neq \sigma^2 \boldsymbol{I}. \tag{1.253}$$

Since neither the proof for unbiasedness nor the proof sketch for consistency relied on Assumption MLR.5, the OLS estimator is still unbiased and consistent with a heteroskedastic error term. However, the formula we used to calculate $\text{Var}(\hat{\boldsymbol{\beta}})$ and $\text{s.e.}(\hat{\boldsymbol{\beta}})$ is no longer valid, and OLS is also no longer *efficient*.

There are a lot of problems that result from us not being able to compute the variance of our estimator. One, standard errors are no longer accurate, so all of our t-statistics, F-statistics, and p-values are now misleading. Two, the estimator being *inefficient* means that there must now be a *better* linear unbiased estimator. Unfortunately, no solution exists to fix the heteroskedasticity problem altogether.

More fortunately, however, there are some strategies that we can apply to mitigate the effects. The simplest one is to just ignore the problem. This may be justified when our samples is very large, as the efficiency problem becomes smaller in large samples. Other than that, we can keep the inefficient OLS estimator, but replace the standard errors with ones that are robust to heteroskedasticity. That way, we can compute accurate t-statistics and F-statistics. Alternatively, we can use a different estimator, one that is efficient even when there is heteroskedasticity. In addition to these two strategies, we are also going to discuss testing procedures we can use to find out whether the errors are actually heteroskedastic.

1.6.2 Robust Standard Errors

Originally, we had assumed that

$$\text{Var}(\boldsymbol{u} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_N. \tag{1.254}$$

Under that assumption, the variance of the OLS estimator was

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{1.255}$$

Now, we make a less restrictive assumption:

$$\text{Var}(\boldsymbol{u} \mid \boldsymbol{X}) = \text{E}(\boldsymbol{u}\boldsymbol{u}' \mid \boldsymbol{X}) = \text{diag}(\sigma_1^2, \dots, \sigma_N^2) =: \boldsymbol{\Omega} \tag{1.256}$$

Under this assumption, the variance of the OLS estimator is:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) &= \text{Var}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \mid \boldsymbol{X}\right) \\
&= \text{Var}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \mid \boldsymbol{X}\right) \\
&= \text{Var}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \mid \boldsymbol{X}\right) \\
&= \text{Var}\left(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \mid \boldsymbol{X}\right) \\
&= \text{Var}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \mid \boldsymbol{X}\right) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{Var}(\boldsymbol{u} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned} \tag{1.257}$$

Upon trying to compute the variance using this formula, we encounter a familiar problem: $\boldsymbol{\Omega}$ is not observed. Fortunately,

$$\text{diag}(\hat{u}_1^2, \dots, \hat{u}_N^2) \tag{1.258}$$

is a consistent estimator for $\boldsymbol{\Omega}$. Using this estimator, we can construct the following *consistent estimator for the variance of* $\hat{\boldsymbol{\beta}}$:

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}(\hat{u}_1^2, \dots, \hat{u}_N^2)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{1.259}$$

This estimator is sometimes called the *sandwich estimator*.

Standard errors that are computed with this estimator for the variance are called *heteroskedasticity-robust standard errors*. The t-statistics and F-statistics we can compute using these standard errors are similarly referred to as *robust* test statistics. Robust standard errors are valid both under heteroskedasticity and under homoskedasticity, while non-robust standard errors are only valid under homoskedasticity. However, there is a small tradeoff: t-statistics that are computed with robust standard errors are only approximately t-distributed if the sample is large enough. In small samples, the distribution can differ substantially. t-statistics computed with non-robust standard errors are exactly t-distributed even in small samples, but *only* as long as the errors are homoskedastic.

We can easily use robust standard errors in R. To do this, we use the `CASchools` dataset again:

```
# Load packages
library(AER) # Contains our dataset
library(dplyr)
library(sandwich) # Robust standard errors
library(lmtest) # Robust tests

# Load data
data("CASchools")

# Compute variables with mutate()
CASchools <- CASchools |>
  mutate(student_teacher_ratio = students / teachers,
         test_score = (read + math)/2)
```

```
14
15 model_1 <- lm(test_score ~ student_teacher_ratio + income + I
       (income^2) + lunch + english, data = CASchools)
16 summary(model_1)
```

```
Call:
lm(formula = test_score ~ student_teacher_ratio + income + I(income
    ^2) +
    lunch + english, data = CASchools)

Residuals:
     Min       1Q   Median       3Q      Max
-30.5291  -4.8919  -0.1249   4.9673  28.9267

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           677.244735   5.945707 113.905  < 2e-16 ***
student_teacher_ratio  -0.548224   0.229643  -2.387   0.0174 *
income                  0.519290   0.267289   1.943   0.0527 .
I(income^2)             0.002939   0.004794   0.613   0.5401
lunch                  -0.404505   0.030473 -13.274  < 2e-16 ***
english                -0.192395   0.031561  -6.096 2.49e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.454 on 414 degrees of freedom
Multiple R-squared:  0.8055,    Adjusted R-squared:  0.8031
F-statistic: 342.9 on 5 and 414 DF,  p-value: < 2.2e-16
```

```
17 robust_se <- vcovHC(model_1, type = "HC1")
18 coeftest(model_1, vcov. = robust_se)
```

```
t test of coefficients:

                        Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)           677.2447352   6.7115437 100.9074 < 2.2e-16 ***
student_teacher_ratio  -0.5482237   0.2568858  -2.1341   0.03342 *
income                  0.5192895   0.2757895   1.8829   0.06041 .
I(income^2)             0.0029392   0.0048905   0.6010   0.54817
lunch                  -0.4045051   0.0335072 -12.0722 < 2.2e-16 ***
english                -0.1923953   0.0334659  -5.7490 1.745e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.6.3 Tests for Heteroskedasticity

When we decide whether or not we should apply one of our strategies to deal with heteroskedasticity, it would be of great help to know whether there actually is heteroskedasticity. Fortunately, there are ways to test whether specific forms of heteroskedasticity are present. This is not ideal, as we usually do not know what form of heteroskedasticity we are dealing with, but it is better than nothing.

The simplest approach is the *Breusch-Pagan test*, from Breusch and Pagan (1979). It is a type of *LM test* (now you know why we discussed this testing procedure earlier on). When we conduct a Breusch-Pagan test, we check whether σ_i^2 depends linearly on the regressors:

$$\sigma_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_K x_{iK} + \text{error}. \tag{1.260}$$

The null hypothesis of this test is that the variance does not depend on any of the regressors, that is,

$$H_0 : \delta_1 = \cdots = \delta_K = 0. \tag{1.261}$$

In other words, rejecting the null hypothesis means that we have reason to believe that errors are heteroskedastic. In large samples, under the null hypothesis, the LM statistic of this test is asymptotically χ^2-distributed with K degrees of freedom.

Of course, when we want to actually conduct a Breusch-Pagan test, we run into the problem that σ_i^2 is unknown. Thus, the testing procedure is as follows:

1. Estimate the main regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ using OLS and retain the residuals $\hat{u}_i$.
2. Use squared residuals in place of the variance and estimate the following auxiliary regression:
$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_K x_{iK} + \text{error}, \tag{1.262}$$
and retain the R^2 of this regression.
3. The statistic NR^2 is the approximate LM statistic and is χ_K^2 distributed in large samples.

There is also a variant of the Breusch-Pagan test that allows for more flexibility in the type of heteroskedasticity: The *White test* by White (1980). The difference to the Breusch-Pagan test is that the White test includes all possible squared terms and interactions of the regressors. To conduct a White test, we can follow this procedure:

1. Estimate the main regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ using OLS and retain the residuals $\hat{u}_i$.
2. Estimate this auxiliary regression:
$$\begin{aligned}\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_K x_{iK} + \delta_{K+1} x_{i1}^2 + \cdots + \delta_{2K} x_{iK}^2 + \\ \delta_{2K+1} x_{i1} x_{i2} + \cdots + \delta_{(K(K+3))/2} x_{i,K-1} x_{iK} + \text{error},\end{aligned} \tag{1.263}$$
and retain the R^2 of this regression.
3. The statistic NR^2 is the approximate LM statistic and is χ_K^2 distributed in large samples.

One very apparent problem with this is that the number of regressors in the auxiliary regression grows explosively (specifically, there are $(K(K+3))/2$ regressors) as we add more variables. If K is large and N is small, this may quickly grow to a problem. Fortunately, an alternative version of the White test exists that avoids this issue. It uses the following auxiliary regression instead:

$$\hat{u}_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + \text{error}, \tag{1.264}$$

that is, we regress $\hat{u}_i^2$ on the fitted values from the first step. Since $\hat{y}_i$ is a linear function of the explanatory variables, $\hat{y}_i^2$ is a specific function of the squares and cross-products of the explanatory variables.

In R, we can implement the two testing procedures like this:

```
# Breusch-Pagan Test
cat("Breusch-Pagan Test:")
bptest(model_1)

# White Test (alternative form)
cat("White Test:")
bptest(model_1, ~ fitted(model_1) + I(fitted(model_1)^2),
   data = CASchools)
```

```
Breusch-Pagan Test:

  studentized Breusch-Pagan test

data:  model_1
BP = 4.7983, df = 5, p-value = 0.441

White Test:

  studentized Breusch-Pagan test

data:  model_1
BP = 0.13015, df = 2, p-value = 0.937
```

1.6.4 Weighted Least Squares

There is one strategy left that we have named, but not talked about: Finding a different estimator, one that is efficient even under heteroskedasticity. Now, we are going to introduce one approach to find such an estimator. Suppose we want to estimate the following regression and know that OLS is inefficient:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i. \tag{1.265}$$

If we know the error variances σ_i^2, we can construct an efficient estimator. To do this, we divide the entire regression by $\sigma_i = \sqrt{\sigma_i^2}$:

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_{i1}}{\sigma_i} + \cdots + \beta_K \frac{x_{iK}}{\sigma_i} + \frac{u_i}{\sigma_i}. \tag{1.266}$$

The idea behind why we are doing this is actually very simple: We are scaling the variance in a way that it is exactly equal for all i. If $\text{Var}(u_i) = \sigma_i^2$, then $\text{Var}(u_i/\sigma_i) = 1$. This means that Assumption MLR.5 is satisfied by construction.

The estimator we get from this equation is called *weighted least squares* (WLS), since we weight observations with higher variance less than those with lower variance. In matrix notation, the above regression is

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta}_{\text{WLS}} + \tilde{\boldsymbol{u}}, \tag{1.267}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{X}$, $\tilde{\boldsymbol{u}} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{u}$, and $\boldsymbol{\Omega} = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$, and the WLS estimator is given as

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}. \tag{1.268}$$

The WLS estimator is a special case of the *generalized least squares* (GLS) estimator. The difference is that the GLS estimator can be used with any variance-covariance matrix $\boldsymbol{\Omega}$, while WLS specifically refers to the diagonal one we used above.

The variance of the WLS estimator is:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{WLS}} \mid \boldsymbol{X}) = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}. \tag{1.269}$$

We can estimate this variance using $\hat{\boldsymbol{\Omega}}$. This allows us to obtain standard errors, which we can then use for our testing procedures. The variance of the WLS estimator is *lower* than that of the OLS estimator (for which we omit the proof). It is given by

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{1.270}$$

One problem remains: We cannot estimate this. From the beginning, we assumed that we know σ_i^2, which we never do. What we can do, however, is to estimate σ_i^2. For example, we can assume that

$$\sigma_i^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_i + \cdots + \delta_K x_K), \tag{1.271}$$

where we use the exponential function to avoid negative values. Then, we take logarithms and plug in $\hat{u}_i^2$ for σ_i^2:

$$\log(\hat{u}_i^2) = \alpha_0 + \delta_1 x_i + \cdots + \delta_K x_K + \text{error}, \qquad \alpha_0 = \log(\sigma^2) + \delta_0. \tag{1.272}$$

Next, we denote the fitted values from this regression $\hat{g}_i$ and use $\hat{\sigma}_i = \sqrt{\exp(\hat{g}_i)}$ as weights. The resulting estimator is called *feasible generalized least squares* (fGLS).

We can use the following procedure if we want to use fGLS:

1. Regress y on $x_1, \ldots, x_K$ using OLS and retain the residuals $\hat{u}$.
2. Compute $\log(\hat{u}^2)$ using these residuals.
3. Regress $\log(\hat{u}^2)$ on $x_1, \ldots, x_K$ using OLS and retain the fitted values $\hat{g}$.
4. To obtain variance estimates, compute $\hat{\sigma}_i^2 = \exp(\hat{g}_i)$.
5. Finally, regress y on $x_1, \ldots, x_K$ using WLS, with weights $1/\sqrt{\hat{\sigma}_i^2}$.

The one remaining problem with this is that we do not know the "true" functional form of the heteroskedasticity, and we have only applied *one* possible form. WLS, however, is only guaranteed to be efficient if this form is correctly specified. Fortunately, if this is not the case, fGLS is still more efficient than OLS, provided the sample is large enough. Additionally, fGLS is consistent, although it is not unbiased.