Econometrics Notes

In progress.

Anthony Tay

2025-09-01

Table of contents

Preface

These notes were written to accompany the econometrics courses that I teach at the School of Economics, Singapore Management University (SMU):

- ECON207 Intermediate Econometrics (BSc Econ)
- ECON682 Econometric Analysis (Econometrics core for MSc Econ / MSc Fin. Econ.)
- ECON6001 Time Series Econometrics (MSc Economics - Quantitative Economics Track)

The specific chapters you will use are listed in your course outline.

What is Econometrics?

Econometrics draws on statistics, economic theory, and mathematics to develop tools for estimating economic relationships, for the purposes of decision making, prediction and forecasting, inferring causal effects, evaluating the efficacy of policy interventions and initiatives, testing the validity of economic theories and their underlying assumptions, and answering a multitude of questions that are ultimately empirical in nature. Examples include the following:

- Pricing decisions by firms require knowledge of the price sensitivity of demand for their products. These are provided by estimates of the products' price elasticities of demand.
- Monetary authorities / central banks build empirical forecasting models of the economy to help anticipate outcomes such as high inflation or economic recessions and predict the outcome of potential policy responses.
- House prices that are very much higher than that predicted by an empirical model linking house prices to economic fundamentals may indicate imbalances in the economy that require policy intervention.
- There is a long list of public initiatives undertaken by authorities to encourage certain behaviors in people and firms, or to improve economic, health, educational and other outcomes in populations. To what extent do they work?
- Many theories in various fields such as industrial organization, economic growth, economic geography among others, assume constant returns to scale in production. Are such assumptions in line with empirical evidence, or would they fall when tested against data.
- Estimates of the economic effect of climate change must factor in adaptation by industries. While we can expect industries to at least try to adapt, is there evidence that they are able to do so effectively and quickly enough?

Such applications present many challenges. The challenge in forecasting applications is to find predictors that have stable relationships with the variable being forecast, and to determine and estimate the form of these relationships. In some cases there are many potential predictors, each limited in predictive ability on its own, but perhaps powerful in totality. The challenge is to estimate usable forecasting relationships with those predictors.

Causal inference – empirically teasing out causal relations from correlative ones – must deal with confounding effects. For example, the causal link from years of education to earnings is tangled up with the effects of individual characteristics such as ability, work experience at the time of sampling, family background, among others things, all of which drive both earnings and

the decision to pursue more years of education. Any attempt to interpret a correlation between years of education and earnings as a causal effect must somehow control for these factors. The ideal situation is if we could hold everything fixed apart from the candidate "causal" variable x and observe what happens to the 'explained' variable y when we change x, but of course this is impossible. What are the alternatives? In some applications, one might be able to employ a randomized controlled trial (RCT) wherein subjects are randomly assigned into a treatment group and a non-treatment group. The randomization breaks the link between the confounding characteristics and the treatment, and enables one to interpret the correlation between treatment and outcome as evidence of causation. In most cases, however, researchers have to depend on observational data, where information regarding a sample drawn from a population is observed without any intervention from the researcher. In these cases, clever methods must be devised to tease out causality from correlation.

Econometric methods must also take into consideration the data formats found in economic data – whether data is made up of a sample from a population taken at some point in time (we call this "cross-sectional data"), or several cross-sections resampled over multiple periods ("pooled cross-sections") or the same cross-sectional sample re-observed over multiple periods ("panel or logistic data"), or observations of variables taken over multiple time periods ("time-series data"), and so on. In some applications, the researcher has to take special steps to counter the complications that arise because of the format of the data. In other examples, the specific features of a data format can be exploited to assist in empirical causal inference. Other data related issues include measurement error, and the fact that we often are only able to employ data that are, at best, proxies of the actual variables we would like to study.

Econometricians have always relied on computers to implement their formulas. This reliance has further increased as computer-based statistical methods – where algorithms have replaced formulas – have become more important over the past few decades. The econometrician now must add computing skills, in addition to economic theory, mathematics and statistics, to her list of competencies.

Mathematical Prerequisites

I assume that the reader is able to do simple differentiation and integration, and understands the basics of optimization theory. We will use a considerable amount of matrix algebra. These notes contain a full length chapter on this topic. We will also use probability theory and statistics extensively. These notes contain a quick review of both topics. For a fuller coverage of the mathematical prerequisites for econometrics (and economics and data science) at the upper undergraduate / masters level, the reader may wish to consult Tay, Preve, and Baydur (2025).

Software

The computations in these notes were done in `R`. Data are available from your course webpages, and I assume these are stored in a 'data' folder in your working directory. There are many introductions to R on the web. I will proceed on the assumption that you have studied some of these, and that you have a working installation on your computer. I recommend running R within the RStudio Integrated Development Environment (IDE). The chapter "Introduction to R" contains brief instructions on installing R and RStudio, and a quick primer on using R.

Chapter 1
A Brief Introduction to R

1.1 Getting Set Up

First download and install the R software from The R Project for Statistical Computing website. Then download and install RStudio from the RStudio website (go to Products, RStudio under the Open Source tab, and download the Open Source Edition of RStudio Desktop). RStudio is an "integrated development environment" (IDE) comprising a set of programs that help you to develop and run R code. You do not need RStudio to run R, but we will do so.

When you first run RStudio, you will see an RStudio desktop open up with three or four windows. There will be one with tabs such as `Files`, `Plots`, `Packages`, `Help`, and `Viewer`, another window with tabs marked `Console`, `Terminal`, `Jobs`, and a third window with `Environments`, `History`, `Connections`, and other tabs. If you go to the menu and select `File>New File>R Script`, a fourth window will open up with an `Untitled1` tab. This is the Editor window, and the `Untitled1` tab is a blank **R Script** file.

You will issue commands in the `Console` tab, or write your instructions in a R script file and execute them from the Editor window. If you ask R to display the results of calculations, these will show up in the `Console` tab. Graphics produced will show up in the `Plots` tab. Objects created will be listed in the `Environment` tab. Executing commands from an R script is best practice; you can save the commands, modify them, correct errors and redo computations easily. Use the console for testing commands, to make inquiries of objects, and for other one-off actions.

Activity: Go to your console tab, type `x <- 4.5` and press enter. Type `x` and enter.

```
x <- 4.5
x
```

```
[1] 4.5
```

You have just used the **assignment operator** `<-` to create an object named `x` containing the number 4.5.[1] The object `x` now appears in your `Environment` tab. You can also use the `=` operator, but we almost always use `<-` for assignment and `=` when giving values to parameters in functions. The second line prints the value of `x` to your console.

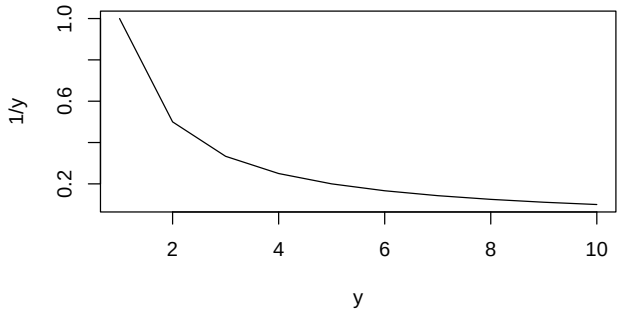
If you make a calculation, say `2+2`, and don't assign the result to a name, the result of that calculation is displayed in the console window, but is thereafter inaccessible, lost in computer memory until overwritten by R.

Activity: Open a new R Script in the Editor window (or use the 'untitled' script if it is already open), and type in the following lines.

```
# Any line or part of line following a '#' is ignored by R
# We use this feature to add comments to code
```

[1] Technically, R stores the value 4.5 as binary code somewhere in your computer, creates a name `x`, and a pointer linking the name to the memory location where the value is stored. But at this stage, you can think of it as having created an object named `x` with the value 4.5.

```
y <- seq(1,10)          # Create an integer sequence from 1 to 10
plot(y,1/y,type='l')    # Create a line plot
```



On your Editor window, click on the `Source` button (alternatively, select all lines and hit `Ctrl+Enter`). This will cause all of the lines in the R script to run one after the other. A line plot will appear on your `Plots` tab. If you select `Export>Copy to Clipboard`, the plot appears in a pop-up window. If you click `Copy Plot`, the plot is placed on your clipboard and you can then paste the plot into, say, a Word document. New plots are placed over current plots. Use the arrows in the plot window to go back and forth between plots. Press the red circle with a white X to erase the current plot. Press the broom icon to erase all plots. To save your script, click on the floppy disk icon and save the file with an appropriate name. The saved file will have the ".R" extension.

In R, you store your data in **vectors**, **matrices**, **lists**, **data frames** (and variants of it), and **time series** objects (and variants of it). These are different **data structures**, i.e., different ways that R can organize your data.

You will work with different kinds of **data types**, including **integer** (whole numbers, without decimals), **double** (or **floating-point** for numbers with decimals), **character** (for text data), **logical** (or **boolean**, to indicate `TRUE` or `FALSE`), and **complex** (for complex numbers). Numbers such as `1` and `2` can be stored either as integers or doubles. To force a whole number to be an integer, we append an `L` after the number, e.g., `1L`. Integer and double are also collectively known as **numeric** data types. There is also a **factor** data type for categorical data.

All actions in R are carried out using **functions** such as `seq()` and `plot()`, and **operators** such as `<-` and `:` (operators are actually also functions). Functions in R are sets of instructions designed to perform certain tasks. Pre-written R functions are organized into **libraries**. Every installation of R comes with some libraries pre-installed (including the `base`, `datasets`, `graphics` and `stats` libraries) and which are automatically loaded every time you start R. There are many other libraries written by independent programmers that provide additional functionality and that are not pre-installed in R. To access these functions, you have first to install the package into your R installation. You can then load the package into any R session that requires the functions in that package. Finally, you can write your own functions.[2]

[2]The packages that we will use include `tidyverse` (Wickham et al. (2019)) for data management and plotting, `readxl` (Wickham and Bryan (2023)) for importing data, `car` (Fox and Weisberg (2019)) and `sandwich` (Zeileis, Köll, and Graham (2020)) for econometrics related algorithms, and `patchwork` (Pedersen (2023)), `latex2exp` (Meschiari (2023)), `gridExtra` (Auguie (2015)) and `plot3D` (Karline (2015)) for additional plotting functionality.

To find out more about any given R function, enter `? function_name` into the console and the relevant documentation will come up, e.g.,

```
? seq  # Enter this and see what happens
? `:`  # With this sort of operators, you have to use surround them by backticks
```

As we mentioned earlier, objects that you create go into your "environment". To see what is in your environment, use `ls()`. To remove all objects in your environment, use `rm(list=ls())`. Try the following line-by-line.

```
ls()            # ls ~ list objects
rm(list=ls())   # rm ~ remove. Environment is now clear.
```

Most of the data that you work with will be imported from an external files (`.csv`, `.xlsx`, etc.) and stored data frames. We will import some later. You will also want to "create" data within `R`. For instance, you may want to create a sequence of integers, or a value that will be used as a constant in your work, or generate a sequence of random numbers and so on.

1.2 Data Types

Activity: Run the following commands and queries as suggested, line by line.

```
## The following illustrate some of the major data types in R
a1 <- 1L            # Integer
a2 <- 4             # Integer or Double?
a3 <- 2.3           # Double
a4 <- TRUE          # Logical
a5 <- "Two"         # Character, making up a "string"
a6 <- "12"          # Another string
a7 <- 2+1i          # Complex

## Use typeof() to query the data type of the object
typeof(a1)         # Try with the others objects you created

## The is.integer(), is.double(), is.numeric(), is.character(), is.complex()
## functions make more specific queries as to the object's data type. An example
## is shown. Try each of the query function on each of the objects above.
is.integer(a2)
## In some cases, you can "coerce" R to change data types using functions like
## as.integer(), as.double(), as.numeric(), as.logical(), as.character(), as.complex().
## An example is shown below. Try these functions of each of the objects created so far.
as.integer(a4)

## Sometimes R will do the coercion for you
3 + TRUE
```

There are **special values** in R:

- `NA` stands for "Not Available" or "Missing". It is by default a logical datatype, but can be converted to other data types.
- `NULL` is an empty object, with no datatype.
- `Inf` stands for "Infinity" and comes about when you do operations like $1/0$. It is by default a `numeric` datatype (specifically `double`, but coerce-able to `complex`).
- `NaN` stands for "Not a Number" and comes about when you do operations like `Inf-Inf`. It is, ironically, of `numeric` datatype by default (specifically `double`, coerce-able to `complex`).

The following are examples of how these values can arise, or how they may be used.

```
1/0       # This will give you Inf
Inf - Inf # Gives NaN. Inf - Inf is NOT equal to zero
0/0       # Also gives NaN. 0/0 is NOT equal to one. Please.
a <- NA   # Basically saying the data that's suppose to be there is missing
```

`NULL`, `NA`, `Inf` and `NaN` are reserved words. You cannot use them as names of objects. Other reserved words include: `if`, `else`, `while`, `repeat`, `for`, `next`, `in`, `function`, `break`, `TRUE`, `FALSE`.

Computers can only store real numbers up to some degree of accuracy:

Activity: The `sqrt()` function returns the square root of a number. Execute the following code. Do the results surprise you?

```
sqrt(2)
sqrt(2)*sqrt(2)
sqrt(2)*sqrt(2) == 2  # use == to make equality comparisons
sqrt(2)*sqrt(2) - 2
```

The last of these commands may return a result such as `e-16`, which stands for $\times 10^{-16}$. Computers, of course, cannot store irrational numbers to infinite accuracy, and this can lead to surprising results when making comparisons, or inaccurate results when performing complicated tasks that involve a very large number of calculations. For now, just bear this in mind. The degree of accuracy is generally not going to be an issue for us (except when making comparisons).

1.2.1 Arithmetic and Logical Operators

The **arithmetic operators** include: **Addition** (`+`), **Subtraction** (`-`), **Multiplication** (`*`), **Division** (`/`), and **Exponent** (`^`). The usual **operator precedence** apply: operations in parentheses are evaluated first, followed by `^`, followed by `(*,/)`, followed by `(+/-)`. Ties between multiplication and division, and between addition and subtraction, are broken by evaluating from left to right. Always use parentheses when in doubt.

Activity: Enter $8/2 * (2+2)$. Do you agree with the result?

```
 8 / 2 * (2+2)
```

You may recognize this from an internet meme, asking what is $8 \div 2(2+2)$ and the answer depends on whether you treat $2(2+2)$ as a single entity. If I say "4 divided by $2n$" do I mean "4 divided by $(2n)$" or "4 divided by 2, times n". I mean the former. In `R`, you cannot write `8/2(2+2)`, you have to write `8/2*(2+2)` which means 8 divided by 2 times 4.

The **relational operators** are:

- Less than `<`
- Greater than `>`
- Less than or equal to `<=`
- Greater than or equal to `>=`
- Equal to `==`
- Not equal to `!=`

Comparisons using relational operators result in the logical outcomes `TRUE` or `FALSE`.

```
2 != 3
```

```
[1] TRUE
```

There are usually a number of different ways to make a comparison, e.g.,

```
2 != 3
!(2 == 3)
```

The `!` is the logical operator "not", or negation. The **logical operators** are:

- Logical Negation: `!`
- Logical And: `&`, `&&`
- Logical Or: `|`, `||`

We will use `&` and `|` for now, and explain `&&` and `||` later.

Let A and B be two statements, each of which are either true or false. If A is true, !A is false. If A is false, then !A is true. The statement A & B is true only if both are true. If one or both statements are false, then A & B is false. The statement A | B is true if one or both are true. If both statements are false, then A | B is false.

In mathematics and computer programming, "or" is always non-exclusive. A or B is true means either (i) A is true, (ii) B is true, or (iii) both are true. Also note that 'and' takes precedence over 'or', so the statement "A or B and C" means "A or (B and C)". Question: will R evaluate the following as true or false?

```
# Is the following TRUE or FALSE
(1 < 2) | (2 < 3) & (4 < 2)
# What about this?
(1 < 2) | (4 < 2) & (6 < 5)
```

The order of precedence is, from highest to lowest: not, and, or, implies, equivalent to. Use parentheses to ensure the order is as you want it.

1.3 Data Structures

1.3.1 Vectors

The **vector** datatype is the most basic data structure in `R`. It is an ordered set of data items. Even single values are stored as a vector.

Activity: Earlier we created the data objects `x` and `y`. Enter the following commands one at a time, and study the outcome.

```
is.vector(x)  # query if object is data type
length(x)     # how many items are in it?
typeof(x)     # what data type does it contain?
## Repeat the above with the object "y"
```

Activity: The following commands all produce vectors. Run the following lines one at a time. After each line output the variable to your console, and study them

```
b01 <- 5
b02 <- 1:26                       # `:` ~ colon operator, gives integers from:to
b03 <- seq(from=2, to=15, by=2)   # the seq() function is more flexible
b04 <- c(37, 42, 29, pi)          # c() ~ "combine" things in a vector or list
b05 <- c("Q1", "Q2", "Q3", "Q4")  # A piece of character data is a "string"
b06 <- rep(1, times=5)            # rep() ~ "replicate". Can simply say rep(1,5)
b07 <- rep(b05, times=4)          # what happens here?
b08 <- rep(1980:1983, each=4)     # and here?
b09 <- c(1<2, 2==4, 4>=3, 1+1==2) # gives logical values!
b10 <- c(1+1i, 0+1i, 2+3i)        # complex numbers!!
b11 <- letters                    # built-in vector, like "pi" in a04
b12 <- LETTERS                    # built-in vector
b13 <- month.abb                  # built-in vector
b14 <- month.name                 # built-in vector
```

Use `is.vector()` to verify that all the objects you just created are vectors. Use `typeof()` to check their data types. Use `is.integer()`, `is.double()`, `is.numeric()`, `is.character()`, `is.logical()`, `is.complex()` to query the data type of the elements of an object. For example:

```
is.vector(b01)   # Should return TRUE
typeof(b05)      # Should return 'character'
is.double(b02)   # Should return FALSE. R has opted to store these as integer.
is.integer(b02)
is.logical(b09)
```

A few things to remember about `R` vectors:

- In matrix algebra, we have row vectors and column vector:

$$\text{a row vector: } \begin{bmatrix} 1 & 2 & 4 & 8 \end{bmatrix} \qquad \text{a column vector: } \begin{bmatrix} 1 \\ 2 \\ 4 \\ 8 \end{bmatrix}.$$

  In R, vectors have no shape: they are simply ordered collections of data items, one item following another, but not organized into a row or a column. Vectors have length (here meaning "number of items") but no dimension.
- There are no scalars. The object `b01` is just a vector (of length 1).
- Each vector can only hold data of a single datatype. You cannot mix datatypes in a vector

You access elements of a vector using the "extract and replace" operator `[`.

Activity: What do the following do?

```
b12[2]           # indexing from R starts with 1. This returns the 2nd item in b.
b12[c(1,1,3)]    # returns the 1st, 1st, and 3rd items
b12[22:26]       # returns 22nd to 26th items
b12[-(1:3)]      # negative indices remove items. Cannot mix with positive indices

head(b12,5)      #
head(b12,-5)     # Frequently helpful if accessing the start or end of vectors
tail(b12,5)      # Check them out!
tail(b12,-5)     #

c01 <- 1:4          # A new vector
c02 <- c01[c(2,4)]  # Copies 2nd and 4th elements of c01 into c02
c02                 # Check it out.
c01[2] <- 20        # What does this do?
c01                 # 2nd element of c01 has been changed,
c02                 # but c02 is not changed. It is its own object.
```

Activity: What happens in the next activity is a bit tough to figure out. First find out about the `%%` operator. Then try to figure out what the following lines do? Remember `b02` is 1, 2, …, 26 and `b11` is `a`, `b`, …, `z`. The point of this activity is to show that you can extract from a vector using logical values.

```
i <- !(b02 %% 2)      # First check out b02 %% 2, then check out !(b02 %% 2)
evenletters <- b11[i] # Then see what 'evenletters' is
```

You can also gives names to the positions of elements in a vector, and access the elements by their position name.

Activity: Try the following.

```
names(b02) <- b12
b02
b02[c("A","C","D","C")]
```

Since a vector can hold data of one type only, if you attempt to mix data types in a vector, R will try to coerce the data types "upwards" – logicals become integers or higher, integers become doubles or higher, doubles becomes complex or higher, complex becomes character. In the first vector in the following example, we try to mix a logical, double, and complex values. The result is a complex vector. In the second case, we mix a logical with an integer and a character. The result is a character vector.

```
c1 <- c(F, 4.5, 1+1i)
c2 <- c(T, 1, "r")
```

1.3.2 Factor Datatype

The **factor** datatype is used for **categorical** variables. The following vectors contain the names of a sample of people, their ages, the region of Singapore they live in[3], and birth month.

[3]For purposes of urban planning, Singapore's Urban Redevelopment Authority (URA) divides the country into five regions: Central, East, North, North-East and West. These are further subdivided into 55 planning areas.

```
name <- c("Abe", "Ben", "Claire", "Daniel", "Edwin",
          "Fred", "Gina", "Harry", "Ivy", "Judy")
age <- c(16, 24, 16, 23, 25, 40, 33, 31, 31, 60)
region <- c("West", "North-East", "West", "Central", "East",
            "North-East", "West", "West", "East", "North")
bmonth <- c("Apr", "Jun", "Oct", "Jan", "Apr",
            "Sep", "Jun", "Jul", "Aug", "Apr")
```

Both `name`, `region`, and `bmonth` are currently character vectors. We can convert `region` into factor datatype.

```
region <- factor(region)
region
```

```
 [1] West       North-East West       Central    East       North-East
 [7] West       West       East       North
Levels: Central East North North-East West
```

We'll convert `bmonth` into an ordered factor data type:

```
bmonth <- factor(bmonth, levels=month.abb, ordered=TRUE)  # remember what month.abb is?
bmonth
```

```
 [1] Apr Jun Oct Jan Apr Sep Jun Jul Aug Apr
12 Levels: Jan < Feb < Mar < Apr < May < Jun < Jul < Aug < Sep < ... < Dec
```

1.3.3 Data Frames

Most of the time, you will store your data for analysis in a data structure called a **data frame**, or one of its variants. You can think of this as a rectangular "spreadsheet" of data, each column containing data on some variable, with different data types allowed across columns.

```
customers <- data.frame(Name=name, Age=age, Region=region, BMonth = bmonth)
customers
```

```
     Name Age     Region BMonth
1     Abe  16       West    Apr
2     Ben  24 North-East    Jun
3  Claire  16       West    Oct
4  Daniel  23    Central    Jan
5   Edwin  25       East    Apr
6    Fred  40 North-East    Sep
7    Gina  33       West    Jun
8   Harry  31       West    Jul
9     Ivy  31       East    Aug
10   Judy  60      North    Apr
```

You can access the contents of this data frame in various ways, illustrated below.

```
customers[1:3,]  # All columns of the first three rows
```

```
    Name Age     Region BMonth
1    Abe  16       West    Apr
2    Ben  24 North-East    Jun
```

```
3 Claire  16       West    Oct
```

```
customers[age==16,c("Name", "BMonth")]  # Name and Birth month of customers aged 16
```

```
    Name BMonth
1    Abe    Apr
3 Claire    Oct
```

```
customers$Name[6:10] # Names of all customers 6 to 10
```

```
[1] "Fred"  "Gina"  "Harry" "Ivy"   "Judy"
```

1.3.4 Matrices, Lists

Other useful data structures include **matrices** and **lists**. A matrix is a vector given a "dimension attribute." The following code creates a matrix with two rows from a vector.

```
mat1 <- matrix(c(1,2,3,4,5,6), nrow=2)
mat1
```

```
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```
attributes(mat1)
```

```
$dim
[1] 2 3
```

Notice that the matrix is filled up by columns. This is the default. To fill by rows, use the `byrow==TRUE` option

Lists are like vectors, except that you can have different data types *and* even different data structures in a list (including other lists). You access items in a list using `[[..]]`. In the following code, we create a list of six items, from previously defined objects.

```
mylist <- list(first=b01, second=b02, third=b03, fourth=b04, fifth=b05, sixth=mat1)
mylist
```

```
$first
[1] 5

$second
 A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

$third
[1]  2  4  6  8 10 12 14

$fourth
[1] 37.000000 42.000000 29.000000  3.141593

$fifth
[1] "Q1" "Q2" "Q3" "Q4"

$sixth
     [,1] [,2] [,3]
```

```
[1,]    1    3    5
[2,]    2    4    6
```

We gave names to the items in the list when creating the list. This is optional. The following are some examples of how items in a list can be accessed.

```
mylist[[3]]
```

```
[1]  2  4  6  8 10 12 14
```

```
mylist[["second"]]
```

```
 A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
```

```
mylist[[6]][1,1:2]
```

```
[1] 1 3
```

In the last example, `mylist[[3]]` returns a matrix, and `mylist[[3]][1,1:2]` returns the (1,1)th and (1,2)th items of this matrix.

1.3.5 Time Series

Another data structure is **time-series**, for holding data ordered in time. The following example converts a numerical vector of random numbers into a "quarterly" time series.

```
set.seed(13)    # for replicability, use own choice of integer
u <- runif(8)   # generates a vector of 8 numbers from a U(0,1) distribution
u.ts <- ts(u, start=c(2010,1), frequency = 4)
u.ts
```

```
           Qtr1       Qtr2       Qtr3       Qtr4
2010 0.71032245 0.24613730 0.38963444 0.09138367
2011 0.96206454 0.01093333 0.57429518 0.76439799
```

The `ts()` function converts a vector to time series. The `frequency=4` indicates that the data are quarterly (4 observations per year), and the `start` option then gives the starting period.

You can use the `class()` function to query an object as to its data structure.

```
class(name)       # For vectors, this function returns the datatype.
class(age)        #    E.g., class(age) returns "numeric" instead of "vector".
class(region)     #    You should read that as "age is 'a numeric vector'".
class(bmonth)
class(customers)
class(mat1)
class(mylist)
class(ts)
```

```
[1] "character"
[1] "numeric"
[1] "factor"
[1] "ordered" "factor"
[1] "data.frame"
[1] "matrix" "array"
[1] "list"
```

```
[1] "function"
```

The `class()` function returns the "class" attribute which identifies the data structure of the object. You should see what you get when you apply the `attribute()` function to the objects listed above, for example:

```
attributes(region)
```

```
$levels
[1] "Central"    "East"       "North"      "North-East" "West"

$class
[1] "factor"
```

Notice that `class(m)` returns `"matrix" "array"`. An R array is a data structure with more than two dimensions. Matrices are 2-dimensional arrays.

1.4 Importing Data

Most of the time, we will read in our data from an external file.

Example 1.1. I assume you have the data set **Anscombe.xlsx** (available on course website) stored in a 'data' sub-folder of your working directory. We will use the function `read_excel()` from the package `readxl` to read in the data. If the package has not yet been installed, install it with the command

```
install.packages("readxl")   # don't forget the quotes
```

You only have to do install a package once (unless you want to update it). Thereafter, just load the package with `library()` whenever you want to use the functions in this package.

```
library(readxl)   # No quotes!
```

Now we read in the data:

```
df2 <- read_excel("data\\Anscombe.xlsx")
df2
```

```
# A tibble: 11 x 8
      x1    y1    x2    y2    x3    y3    x4    y4
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1    10  8.04    10  9.14    10  7.46     8  6.58
 2     8  6.95     8  8.14     8  6.77     8  5.76
 3    13  7.58    13  8.74    13 12.7      8  7.71
 4     9  8.81     9  8.77     9  7.11     8  8.84
 5    11  8.33    11  9.26    11  7.81     8  8.47
 6    14  9.96    14  8.1     14  8.84     8  7.04
 7     6  7.24     6  6.13     6  6.08     8  5.25
 8     4  4.26     4  3.1      4  5.39    19 12.5
 9    12 10.8     12  9.13    12  8.15     8  5.56
10     7  4.82     7  7.26     7  6.42     8  7.91
11     5  5.68     5  4.74     5  5.73     8  6.89
```

Investigating a large data frame by simply printing it out to screen is not feasible. You can use `head()` and `tail()` to print only the first few or last few observations. Alternatively, you

can use `str()` to give you a summary of the data frame (str = structure).

```
head(df2,3)
```

```
# A tibble: 3 x 8
     x1    y1    x2    y2    x3    y3    x4    y4
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    10  8.04    10  9.14    10  7.46     8  6.58
2     8  6.95     8  8.14     8  6.77     8  5.76
3    13  7.58    13  8.74    13 12.7      8  7.71
```

```
tail(df2,3)
```

```
# A tibble: 3 x 8
     x1    y1    x2    y2    x3    y3    x4    y4
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    12 10.8     12  9.13    12  8.15     8  5.56
2     7  4.82     7  7.26     7  6.42     8  7.91
3     5  5.68     5  4.74     5  5.73     8  6.89
```

```
str(df2)
```

```
tibble [11 x 8] (S3: tbl_df/tbl/data.frame)
 $ x1: num [1:11] 10 8 13 9 11 14 6 4 12 7 ...
 $ y1: num [1:11] 8.04 6.95 7.58 8.81 8.33 ...
 $ x2: num [1:11] 10 8 13 9 11 14 6 4 12 7 ...
 $ y2: num [1:11] 9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
 $ x3: num [1:11] 10 8 13 9 11 14 6 4 12 7 ...
 $ y3: num [1:11] 7.46 6.77 12.74 7.11 7.81 ...
 $ x4: num [1:11] 8 8 8 8 8 8 8 19 8 8 ...
 $ y4: num [1:11] 6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
```

The `read_excel()` function reads data into a modified data frame called a `tibble`. This modification is part of the larger "tidyverse" initiative. For the moment, we can treat the two data structures (tibble vs data frame) as essentially the same thing. We will use the tidyverse suite of libraries for data wrangling, and for graphics.

1.5 Plotting Data

`R` comes with a very good base graphics package pre-installed (and automatically loaded whenever you start an R session). We used the `plot()` function from this package earlier. There is another package called `ggplot2` that contains many functions for producing very good graphics (`gg` = Grammar of Graphics). We will use both in these notes, but for now we use the latter.

You can install the `ggplot2` package separately, but we will instead install the `tidyverse` package which includes several libraries, `ggplot2` being one of them.

```
install.packages("tidyverse")   # don't forget the quotes
```

Once the tidyverse package is installed, you can load it into your R session if you need to use it. Remember you don't need to re-install libraries once you have done so (unless you are updating the package). However, you do need to load the package every time you start an R session, should you be planning to use the functions in that package in the session.

```
library(tidyverse)    # no quotes!
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

In addition to `ggplot2`, there are other libraries that are helpful for constructing plots. One such package used in this book is `patchwork`. We assume you have already installed this package. In the example below, we use these `ggplot` and `patchwork` to plot the data that we just imported into R.

```
library(patchwork)
p1 <- df2 %>% ggplot() + geom_point(aes(x=x1, y=y1), size=1) + theme_classic()
p2 <- df2 %>% ggplot() + geom_point(aes(x=x2, y=y2), size=1) + theme_classic()
p3 <- df2 %>% ggplot() + geom_point(aes(x=x3, y=y3), size=1) + theme_classic()
p4 <- df2 %>% ggplot() + geom_point(aes(x=x4, y=y4), size=1) + theme_classic()
(p1 | p2) / (p3 | p4) # this is from patchwork package
```

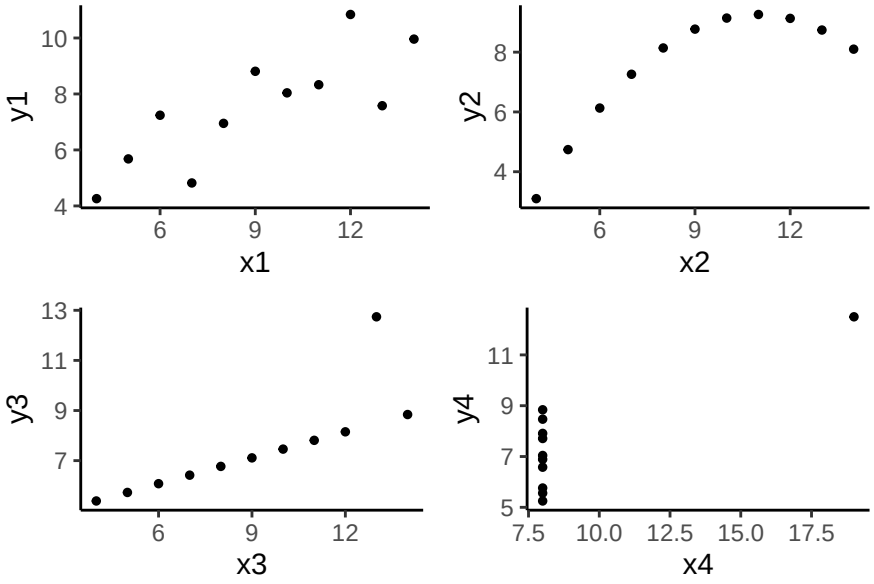


Figure 1.1: Anscombe Quartet.

In the code above, we created four separate figures, named `p1`, `p2`, `p3`, `p4` and used the `patchwork` library to create a composite figure comprising the four plots. When creating the individual scatterplots, we used the **pipe operator `%>%`** to "send" the dataframe / tibble to the ggplot function, and then 'added' a scatterplot with the `geom_point()` function. The `aes` option (which stands for aesthetics) is used to indicate the x-variable, y-variable, color-variable, and so on. The `theme_classic()` function is used to create a certain "look" for the plots.

The pipe operator `%>%` is helpful when doing several things to a data frame in sequence, and can help create very readable code. This operator is not part of base R, but is provided by the package `magrittr` which is included in the `dplyr` package which is included in the `tidyverse` package.

1.6 More on the R Environment

In your Environment tab, look for the menu button marked "Global Environment" and click on the little black triangle on the right of it. You will see a large list of "environments", most of which are libraries that were loaded in your R session, either automatically or by yourself using the `library()` command. The "Global Environment", which contains all the variables that you created in your session, is always first. The libraries are ordered as they were loaded (latest on top). To see all the functions in a loaded package, say the package `ggplot2`, you can use the command `ls("package:ggplot2")`. Just entering `ls()` will list the the contents of the Global Environment.

One issue that you should pay attention to is 'masking'. When we loaded the `tidyverse` package we saw two warnings: that **dplyr::filter() masks stats::filter()** and **dplyr::lag() masks stats::lag()**. Both `dplyr` and `stats` libraries have a `lag()` function. Because the `dplyr` package was loaded on top of the `stats` package, the `dplyr` version 'masks' the `stats` version, and calling `lag()` will call the `dplyr` version. However, the two versions behave differently: the `stats` version requires the input to be a time series object, whereas the input to the `dplyr` version *cannot* be a time series object. Worse, `lag(x,1)` in one means something quite different from `lag(x,1)` in the other. We illustrate this issue in the next example. To be explicit about which version you wish to use, indicate the package using `::`, as in `stats::lag()`.

Example 1.2. In this example, we create a vector 1:10, and convert it into a **time series object** from 2019Q1 to 2021Q2. We then apply the `dplyr` version to the vector, and the `stats` version to the time series.

```
x <- 1:10
x
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```
lag(x,1)  # dplyr version is used
```

```
 [1] NA  1  2  3  4  5  6  7  8  9
```

```
x.ts <- ts(1:10, start=c(2019,1), frequency=4)
x.ts
```

```
     Qtr1 Qtr2 Qtr3 Qtr4
2019    1    2    3    4
2020    5    6    7    8
2021    9   10
```

```
stats::lag(x.ts,1)
```

```
     Qtr1 Qtr2 Qtr3 Qtr4
2018                   1
2019    2    3    4    5
2020    6    7    8    9
2021   10
```

We see that the `dplyr` version lags the data whereas the `stats` version creates a "leading" series. To use the `stats` version to lag, we have to say `stats::lag(x.ts,-1)`.

1.7 User-Defined Functions, Conditional Statements, Loops

You can define your own functions.

Example 1.3. A one-line function to calculate the area of a circle.

```
area_circle <- function(r){pi*r^2}
area_circle(2)
```

```
[1] 12.56637
```

Example 1.4. A more complicated function

```
circle_summary <- function(r=1){
  if (!is.numeric(r)){
    stop("Error: Input is not numeric.")
  } else if (r<=0 | is.nan(r) | is.infinite(r)) {
    print("Error: Please input a positive finite value for the radius.")
    return(NULL)
  } else {
    result = list("radius" = r, "area"=pi*r^2, "circumference"=2*pi*r)
    return(result)
  }
}
```

When the set of instructions is executed, a function object named `circle_summary` appears in your environment. Thereafter we can call it whenever we want to use it.

```
A1 = circle_summary(); A1   # radius defaults to 1
```

```
$radius
[1] 1

$area
[1] 3.141593

$circumference
[1] 6.283185
```

```
A2 = circle_summary(2); A2;
```

```
$radius
[1] 2

$area
[1] 12.56637

$circumference
[1] 12.56637
```

```
A3 = circle_summary(-1); A3
```

```
[1] "Error: Please input a positive finite value for the radius."
```

```
NULL
```

```
A4 = circle_summary("two"); A4
```

```
Error in circle_summary("two"): Error: Input is not numeric.
```

The `circle_summary()` function requires one input `r`, which has the default value of 1. The function also contains "if-else" statements that carry out the following conditional actions:

- check if you put in a non-numeric value;
- if you did, print a error message and stop;
- If you did not input a non-numeric, check if it is negative or `NaN` or `Inf`;
- If so, print a different error message and return `NULL` (but don't stop the program);
- If the numeric value is not negative and not `NaN` and not `Inf`, then return a list comprising the radius, area and circumference of the circle.

Blocks of code are bound with "{…}". The way we placed the braces is somewhat conventional. Indentations and writing long commands over several lines also help with readability.

Every function call has its "own namespace":

Example 1.5. In the following example, the assignment of the value 3 to `x` inside the function does not change the value of `x` outside of the function.

```
an_example_function <- function(x){
  cat("x =", x, "was passed into the function.\n")
  x <- 3;
  cat("The function changes the value to: x = ", x, ".\n", sep="")
}
x <- 1
cat("The declared value of x: x = ", x, ".\n", sep="")
an_example_function(x)
cat("The value outside the function remains unchanged: x = ", x, ".\n")
```

```
The declared value of x: x = 1.
x = 1 was passed into the function.
The function changes the value to: x = 3.
The value outside the function remains unchanged: x =  1 .
```

We use the function `cat()` to print to screen (cat == "concatenate and print"). The special code "\n" refers to a line break. The function automatically adds a space between entries. To tell the function not to add the space, set the option `sep=""`.

Another essential programming technique is the "for-loop". The following code, which contains a loop and a nested loop, illustrates how they work.

Example 1.6. Can you figure out what is going on in the program below?

```
for (A in c(TRUE,FALSE)){
  for (B in c(TRUE,FALSE)){
    cat("A is",A,"and B is",B,"then A & B is",A & B,"\n")
  }
}
```

```
A is TRUE and B is TRUE then A & B is TRUE
A is TRUE and B is FALSE then A & B is FALSE
A is FALSE and B is TRUE then A & B is FALSE
A is FALSE and B is FALSE then A & B is FALSE
```

Finally, we illustrate the "while" loop:

```
x <- 0
while (x<10){
  cat("x = ", x, ", x < 10 is ", x<10, " so we enter the loop,\n", sep="")
  x=x+2   # this means replace current value of x with current value + 2
}
cat("x = ", x, ", x < 10 is ", x<10, " so we skip the loop.\n", sep="")
```

```
x = 0, x < 10 is TRUE so we enter the loop,
x = 2, x < 10 is TRUE so we enter the loop,
x = 4, x < 10 is TRUE so we enter the loop,
x = 6, x < 10 is TRUE so we enter the loop,
x = 8, x < 10 is TRUE so we enter the loop,
x = 10, x < 10 is FALSE so we skip the loop.
```

Can you see the danger of inadvertently entering an infinite loop? If you condition a `while` loop on a condition that will always be satisfied, the loop will so on forever. You will have to break the loop using `Esc` or `Ctrl-C`.

Chapter 2
Probability and Statistics Review

This chapter uses the following R libraries.

```
library(tidyverse)
library(patchwork)
library(latex2exp)
```

Other libraries may be introduced later in the chapter.

2.1 Introduction

This chapter contains a quick review of probability and statistics. The central question in **statistics** is: how to learn about a **population** given a **sample** of observations from that population.

Example 2.1. (a) The population of interest may be the set of all "US non-institutional working civilians aged 16 or above in 2019". Perhaps we wish to learn what the average hourly earnings was in this population in 2018.
(b) The population of interest may be the set of all Singapore households in 2020. Perhaps we wish to learn how many dogs SG households owned on average in 2020.
(c) Consider a machine producing a certain product. The population may be all the goods that the machine can potentially produce in its lifetime. We might be interested in the defect rate of the machine.
(d) The population of interest may be the set of outcomes of an infinite number of potential tosses of a coin. We might be interested in whether the coin is fair (if the probability of obtaining heads is 0.5).

The populations in (a) and (b) are finite (though fairly large) **tangible** populations, whereas (c) and (d) are **intangible**, effectively infinite populations.

For Example 2.1(a) suppose you select, soon after 2019, a small number (1000? 5000?) of individuals from this population and ask each person sampled how much they earned per hour on average in 2018. Let X_i be individual i's response. You calculate

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \tag{2.1}$$

where n is your **sample size**. You use the **sample average** $\overline{X}$ as an **estimate** of the **population average**. Will your sample average be a good estimate of the population average?

Statistics uses probability theory (the concepts of random variables, probability distributions, expectations and variances, etc.) to come up with estimation rules such as (2.1) and to determine the properties of such rules. We will review some probability concepts before return to the statistical problem of estimation and hypothesis testing.

2.2 Random Variables

Let X represent the numerical outcome of an action where (i) there is a range of possible outcomes, (ii) there is randomness in terms of which outcome is obtained each time the action is taken. We call X a **random variable**.

Example 2.2. Each of the following describes a random variable, denoted X.
(a) You randomly select a person from the population in Example 2.1(a) and let X be this persons average hourly earnings in 2019.
(b) You randomly select a household from the population in Example 2.1(b) and let X be the number of dogs in this household.
(c) You take a product produced by the machine in Example 2.1(c) and observe if it is defective, $X = 1$ if defective and $X = 0$ if not defective.
(d) You toss the coin in Example 2.1(d) and note the outcome *heads* or *tails*. You set $X = 1$ if *heads* and $X = 0$ if *tails*.

Although materially quite different, you can see that the problem of learning about the defect rate of a machine, and the probability of heads in a toss of a coin are mathematically identical.

2.3 Probability Distributions

Random outcome is not the same as *arbitrary* outcome. For instance, in Example 2.2(a) you are more likely to receive an answer like "\$20 per hour" than an answer like "\$2000 per hour". The value of X is random, but follows some "rule" in the sense that some values or range of values are more likely than other values or ranges of values. This "rule" is summarized in the **probability distribution function (pdf)** of the random variable.

Example 2.3. Suppose X takes possible values 0 or 1 and that outcome $X = 1$ occurs with probability p, and outcome $X = 0$ occurs with probability $1-p$, where $0 \leq p \leq 1$. Probabilities are defined so they are never negative, and such that total probabilities always add to 1. Then the probability distribution of X is $f_X(x) = \Pr(X = x)$ where $x = 0, 1$ and

$$f_X(x) = p^x(1-p)^{1-x}\,,\ x = 0, 1\,. \tag{2.2}$$

We say that X has a **Bernoulli distribution** with parameter p, or $X \sim \text{Bernoulli}(p)$. This pdf is useful for modelling the populations in Example 2.2(c) and (d), where in (c) p is the defect rate of the machine (hopefully very small), whereas in (d) p is the probability of obtaining heads (which we might expect to be 0.5 or close to it).

Example 2.4. The random variable X has the Poisson distribution with parameter $\lambda > 0$ if

$$f_X(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}\,,\ x = 0, 1, 2, \ldots\,. \tag{2.3}$$

We write $X \sim \text{Poisson}(\lambda)$. Fig. 2.1 shows the Poisson probability distribution functions for $\lambda = 0.4$. This distribution might be a good description of the population of households in Example 2.2(b) where each bar in Fig. 2.1 represents the population proportion of households with x number of dogs.

```
x = 0:8
lambda = 0.4
fpois = dpois(x,lambda)
df <- data.frame(x=as.factor(x), fx=fpois)
ggplot(df, aes(x=x, y=fx)) + geom_col(width=0.2) +
  ylab("Pr(X=x)") + theme_minimal()
```

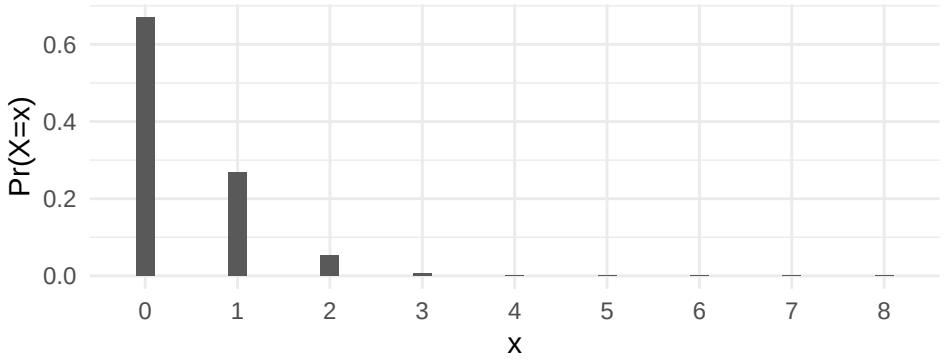


Figure 2.1: The Poisson distribution with $\lambda = 0.4$

Activity: Plot the Poisson pdf for different values of λ.

The Bernoulli and Poisson distributions have discrete ranges (distinct and separate possible values). If X has a distribution with discrete range, it is a **discrete random variable**. Note that the Poisson theoretically has an infinite range, though in practice the probability of a Poisson random variable taking a large integer is usually very small unless λ is very large.

If X takes possible values in a continuum such as the intervals $(-\infty, \infty)$, $(0, \infty)$ or $(0, 1)$, then it is a **continuous random variable**. For continuous random variables, the probability distribution function does not give $\Pr(X = x)$. Instead, the integral of the pdf from $x = a$ to $x = b$ gives the probability that an outcome of X falls between a and b,

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)\, dx\,.$$

Some refer to discrete distributions as **probability mass functions** and continuous distributions as **probability density functions**. I will use **probability distribution function** (pdf) for both.

Example 2.5. A random variable X has the **normal distribution**, denoted $X \sim \text{Normal}(\mu, \sigma^2)$ or $X \sim N(\mu, \sigma^2)$, if its pdf is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \; x \in \mathbb{R}. \tag{2.4}$$

The range of a normal random variable is the entire real line. The pdf of the normal distribution has the familiar symmetric bell-shape, centered at μ. The parameter σ^2 controls how "spread out" the pdf is. The normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the **standard normal distribution**. The normal distribution has a special place in probability theory for reasons that will soon become clear. The normal distribution is also called the Gaussian distribution.

Fig. 2.2 displays the normal pdf with $\mu = 1$ and $\sigma^2 = 2$, with some regions shaded.

```
mu <- 1
sigma <- sqrt(2)  ## note, sigma not sigma^2
x <- seq(mu-4*sigma, mu+4*sigma, length.out = 1000)
fnorm <- dnorm(x, mean=mu, sd = sigma)
df <- data.frame(x=x, fx=fnorm)
shade1 <- subset(df, x <= -1)
shade2 <- subset(df, x >= 2 & x <= 3)
ggplot(df, aes(x=x, y=fx)) +
  geom_line(color="black") + geom_area(data = shade1, aes(x=x, y=fx), fill="grey", alpha=0.3) +
  geom_area(data = shade2, aes(x=x, y=fx), fill="grey", alpha=0.3) +
  geom_vline(xintercept = c(-1, 2, 3), linetype="dashed", color = "black")+
  scale_x_continuous(breaks=seq(floor(min(x)), ceiling(max(x)), by=1)) +
  ylab("f(x)") +  theme_minimal()
```

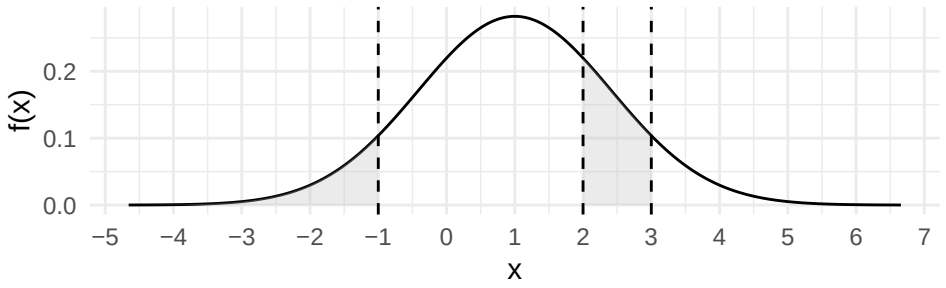


Figure 2.2: The normal distribution with $\mu = 1$ and $\sigma^2 = 2$

The R command `pnorm(x=a, mean, sd)` calculates the **cumulative distribution function (cdf)** of the normal distribution, defined as

$$F_X(x) = \Pr(X \le x) = \int_{-\infty}^{x} f_X(u)\, du$$

for the Normal(μ, σ) distribution. We can calculate the probability $\Pr(a \le x \le b)$ as

$$\Pr(a \le x \le b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(u)\, du - \int_{-\infty}^{a} f_X(u)\, du\,.$$

We have

```
  cat("For X ~ N(1, 2):\n")
  cat("Pr(X <= -1) =", pnorm(-1, mean=1, sd=sqrt(2)), ", ")
  cat("Pr(2 <= X <= 3) =", pnorm(3, mean=mu, sd=sigma)-pnorm(2, mean=1, sd=sqrt(2)))
```

```
For X ~ N(1, 2):
Pr(X <= -1) = 0.0786496 , Pr(2 <= X <= 3) = 0.1611005
```

Given $\Pr(X \le x) = \int_{-\infty}^{x} f_X(u)\, du = p$, the command `qnorm(p, mean, sd)` finds x. That is, the `qnorm(p, mean, sd)` function finds the p-th quantile of the normal distribution.

```
  cat("For X ~ N(1, 2):\n")
  cat("If Pr(X <= x) = 0.0786496, then x =", qnorm(0.0786496, mean=1, sd=sqrt(2)))
```

```
For X ~ N(1, 2):
If Pr(X <= x) = 0.0786496, then x = -1
```

Activity: For the standard normal, find c such that (i) $\Pr(X \leq c) = 0.025$, (ii) $\Pr(X \geq c) = 0.025$. Find (iii) $\Pr(X \leq -2)$ for the Normal$(0, 4)$ distribution.

Is the normal distribution a good description of hourly earnings in the population in Example 2.2(a)? We will come to this in a moment, after discussing the idea of sampling.

If you randomly select n people from the population in Example 2.2(a) and let X_i be the average hourly earnings of individual i, $i = 1, \ldots, n$, then each X_i is a random variable. The set $\{X_1, \ldots, X_n\}$ is your **sample**. If you select these n people in such a way that each member of the population has an equal chance of getting selected, and n is large enough, then you will get a sample that is **representative** of the population.[1] The characteristics of the sample should be similar to the characteristics of the population in terms of relative numbers of males and females, proportion of the various races, and so on. The distribution of observations in the sample should also be similar to the population distribution.

The dataset `earnings2019.csv` contains a random sample of almost 5000 U.S. non-institutional working civilians who had worked in 2018 (these individuals are part of the 2019 wave of the University of Michigan Panel Survey of Income Dynamics). The survey collected information regarding the surveyed individuals on variables including average hourly earnings ($earn$) in the previous year, number of years of schooling, age, and race. Fig. 2.3 shows **histogram density estimates** of the distributions of $earn$ and $\ln earn$. The horizontal axes in both (a) and (b) are divided into bins, and the frequency with which $earn_i$ and $\ln earn_i$ falls into each bin is noted. The rectangles are then scaled so that their areas sum to one.

```
dat1 <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
p1 <- ggplot(dat1, aes(x = earn)) +
  geom_histogram(aes(y = after_stat(density)), fill = "lightblue", color="black") +
  xlab("average earning per hour") + theme_bw()
p2 <- ggplot(dat1, aes(x = log(earn))) +
  geom_histogram(aes(y = after_stat(density)), fill = "lightblue", color="black") +
  xlab("log(average earning per hour)") + theme_bw()
p1 | p2
```

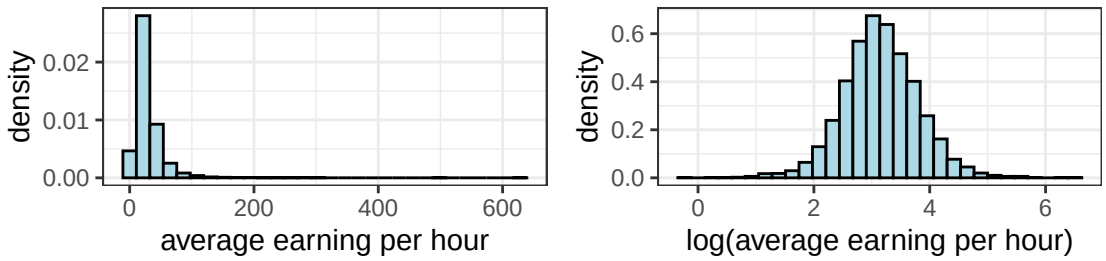


Figure 2.3: Histogram density estimate of $earn$ and $\ln earn$.

If our sample is representative of the population, then Fig. 2.3(a) strongly suggests that the normal distribution is not a reasonable distribution with which to model population earnings. Besides, average hourly earnings take only non-negative values whereas a normal distribution has range $(-\infty, \infty)$. However, Fig. 2.3(b) does suggests that the normal distribution may be a reasonable model for the population of $\ln earn$.

[1] How to actually do this is the subject of the field called "survey sampling".

2.4 Expectations

The **mean** or **expected value** of a random variable X is defined as

$$E(X) = \begin{cases} \sum_x x f_X(x) = \sum_x x \Pr(X = x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{+\infty} x f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases} \tag{2.5}$$

The symbol $\sum_x$ means "sum over the possible values of X". The expected value of a random variable is therefore the weighted sum of its possible values, weighted by their corresponding probabilities.

- If $X \sim \text{Bernoulli}(p)$, then $E(X) = 1 \cdot p + 0 \cdot (1-p) = p$.
- If $X \sim \text{Poisson}(\lambda)$, then

$$E(X) = \sum_{x=0}^{\infty} x \Pr(X = x) = \sum_{x=0}^{\infty} \frac{x e^{-\lambda} \lambda^x}{x!} = \lambda\,.$$

- If $X \sim \text{Normal}(\mu, \sigma^2)$, then

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\, dx = \mu\,.$$

For detailed proofs of these results, and other omitted proofs in this chapter, see Tay, Preve, and Baydur (2025).

2.4.1 Properties of Expectations

If X is a random variable, then $g(X)$ is also a random variable, with expectation:

$$E(g(X)) = \begin{cases} \sum_X g(x) f_X(x) = \sum_X g(x) \Pr(X = x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

For example, if $X \sim \text{Bernoulli}(p)$, we have $E(X^2) = 1^2 p + 0^2(1-p) = p$. It can also be shown that if $X \sim \text{Poisson}(\lambda)$, then $E(X^2) = \lambda + \lambda^2$, and if $X \sim \text{Normal}(\mu, \sigma^2)$, then $E(X^2) = \sigma^2 + \mu^2$.

It is straightforward to show using the properties of summation and integration that

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X)) \tag{2.6}$$

where a and b are constants. It follows that $E(a) = a$ for constants a, and

$$E(a + bX) = a + bE(X)\,. \tag{2.7}$$

We will show later than if X and Y are any two random variables, then

$$E(aX + bY) = aE(X) + bE(Y)\,. \tag{2.8}$$

2.4.2 Variance

The **variance** of a random variable X is its expected squared deviation from mean, i.e.,

$$Var(X) = E((X - E(X))^2) = \begin{cases} \sum_X (x - E(X))^2 f_X(x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases} \tag{2.9}$$

The variance is a measure of the spread of the probabilities about the mean. It is sometimes referred to as the "second central moment". The square root of the variance of X is the **standard deviation** of X, and can be viewed as a measure of how far any given draw might be from the mean. Note that the unit of measurement of the standard deviation follows that of the variable itself. For instance, if X is measured in dollars, then the standard deviation is also measured in dollars, whereas the variance is measured in "squared dollars".

There is another expression for the variance that is often easier to use:

$$Var(X) = E((X - E(X))^2) = E(X^2 - 2XE(X) + E(X)^2) = E(X^2) - E(X)^2\,. \tag{2.10}$$

Using this expression for the variance, it is straightforward to show that:

- If $X \sim \text{Bernoulli}(p)$, then $Var(X) = p(1-p)$.
- If $X \sim \text{Poisson}(\lambda)$, then $Var(X) = \lambda$.
- If $X \sim \text{Normal}(\mu, \sigma^2)$, then $Var(X) = \sigma^2$.

It follows from (2.10) that

$$Var(aX + b) = a^2\, Var(X)\,. \tag{2.11}$$

We will show later that for any two random variables X and Y, we have

$$Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) + 2ab\, Cov(X, Y) \tag{2.12}$$

where $Cov(X, Y)$ is the covariance between the two random variables. We just note, for the moment, that if two random variables X and Y are **independent** (meaning that knowing the outcome of one tells you nothing about the other), then $Cov(X, Y) = 0$, and

$$Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) \qquad \textit{NB: Only if } Cov(X, Y) = 0.$$

We will discuss independence and covariance / correlation later on in this chapter.

2.5 Jensen's Inequality

Since $Var(X) \geq 0$, (2.10) implies that $E(X^2) \geq E(X)^2$. This is a special case of **Jensen's inequality**, which says that for any random variable X, we have

$$E(g(X)) \geq g(E(X)) \quad \text{for any convex function } \; g\,. \tag{2.13}$$

The inequality is reversed if g is concave. Property (2.7) says that (2.13) holds with equality if the transformation is linear, i.e., if $g(X) = a + bX$.

2.6 Distributions related to the normal distribution

We mention a few more distributions before discussing estimation and hypothesis testing. These distributions are all somehow related to the normal distribution.

2.6.1 Log-normal distribution

A random variable X has the **log-normal distribution** with parameters μ and σ^2 if $\ln X \sim \text{Normal}(\mu, \sigma^2)$. Its pdf is

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, \; x \in (0, \infty)\,. \tag{2.14}$$

We write $X \sim \text{Log-normal}(\mu, \sigma^2)$. We plot the pdf of a log-normal distribution in Fig. 2.4.

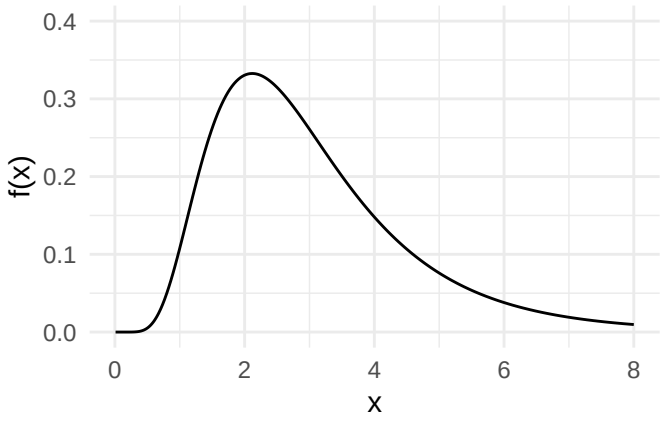


Figure 2.4: The Log-normal pdf with $\mu = 1$, $\sigma^2 = 1/4$.

If $X \sim \text{Log-normal}(\mu, \sigma^2)$, then

$$E(X) = \exp(\mu)\exp\left(\frac{\sigma^2}{2}\right) \quad \text{and} \quad Var(X) = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)\,.$$

Another useful property of the log-normal distribution is that if $X \sim \text{Log-normal}(\mu, \sigma^2)$, then

$$Median(X) = \exp(\mu)\,.$$

Since the histogram for the $\ln(earn)$ data in our dataset `earnings2019.csv` appears normal, it seems like the log-normal distribution would be an appropriate model for the population distribution of *earn*.

2.6.2 Chi-square distribution

If $X \sim \text{Normal}(0, 1)$, then X^2 has the "**chi-square distribution** with one degree of freedom", denoted $\chi^2(1)$. If $X_1, X_2, \ldots, X_k$ are independent standard normal random variables, then $\sum_{i=1}^{k} X_i^2$ has a chi-square distribution with k degrees of freedom, denoted $\chi^2(k)$. If $X \sim \chi^2(k)$, then $E(X) = k$ and $Var(X) = 2k$. Fig. 2.5 displays the chi-square distribution for a few different values of the "degree of freedom" parameter.

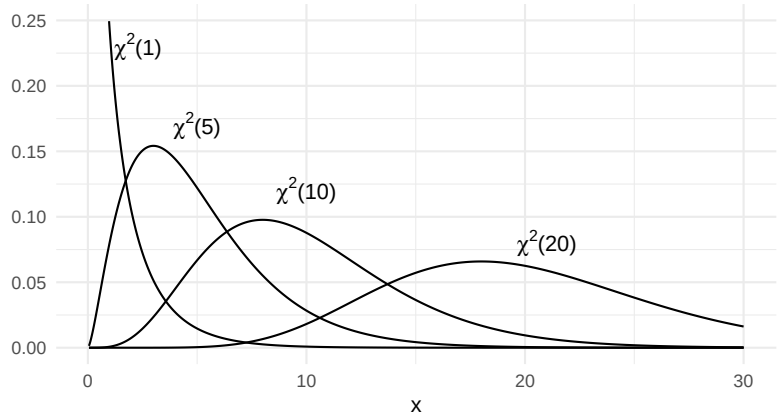


Figure 2.5: The χ^2 distribution.

2.6.3 Student-t distribution

If X and W are independent variables with $X \sim \text{Normal}(0, 1)$ and $W \sim \chi^2(v)$, then

$$\frac{X}{\sqrt{W/v}} \sim \text{t}(v)$$

where $\text{t}(v)$ denotes the **student-t distribution** with v degrees of freedom. A student-t random variate has zero mean, and variance $\frac{v}{v-2}$ (the variance does not exist unless $v > 2$). The student-t pdf is similar to that of the standard normal pdf in that it is symmetrically bell-shaped and centered about zero. However, it has fatter tails than a normal distribution. This means that a student-t random variable has greater probability of extreme realizations than a comparable normal variate. The student-t pdf has the property that it converges to the standard normal pdf as $v \to \infty$. Fig. 2.6 shows the student-t pdf with degree-of-freedom parameter $v = 1$, 5, 10, and 20, and also the pdf of the standard normal. The $\text{t}(1)$ and $\text{t}(5)$ distributions are indicated, with the $\text{t}(10)$ and $\text{t}(20)$ distributions "between" the $\text{t}(5)$ and the standard normal pdf. The following table compares the tail probabilities of the normal and the student-t distribution.

```
norm_vs_t_tail <- cbind(
                    pnorm(c(-2.57, -1.96, -1.64)),
                    pt(c(-2.57, -1.96, -1.64), 1),
                    pt(c(-2.57, -1.96, -1.64), 5),
                    pt(c(-2.57, -1.96, -1.64), 10),
                    pt(c(-2.57, -1.96, -1.64), 20),
                    pt(c(-2.57, -1.96, -1.64), 30))
colnames(norm_vs_t_tail) <- c("N(0,1)", "t(1)", "t(5)", "t(10)", "t(20)", "t(30)")
rownames(norm_vs_t_tail) <- c("P(X<-2.57)", "P(X<-1.96)", "P(X<-1.64)")
round(norm_vs_t_tail, 4)
```

```
           N(0,1)   t(1)   t(5)  t(10)  t(20)  t(30)
P(X<-2.57) 0.0051 0.1181 0.0250 0.0139 0.0091 0.0077
P(X<-1.96) 0.0250 0.1502 0.0536 0.0392 0.0320 0.0297
P(X<-1.64) 0.0505 0.1743 0.0810 0.0660 0.0583 0.0557
```

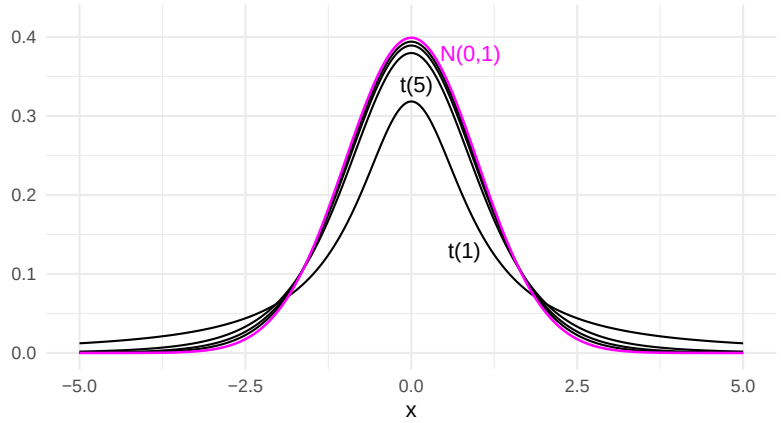


Figure 2.6: The student t distribution.

2.6.4 F distribution

If W_1 and W_2 are independent chi-square random variables with degrees of freedom v_1 and v_2 respectively, then

$$\frac{W_1/v_1}{W_2/v_2} \sim F(v_1, v_2)$$

where $F(v_1, v_2)$ denotes the "**F distribution** with v_1 and v_2 degrees of freedom". If $X \sim F(v_1, v_2)$, then

$$E(X) = \frac{v_2}{v_2 - 2} \quad \text{and} \quad Var(X) = 2\left(\frac{v_2}{v_2 - 2}\right)^2 \frac{v_1 + v_2 - 2}{v_1(v_2 - 4)}.$$

The F-distribution is also related to the student-t and chi-squared distributions in that

- If $X \sim t(v)$, then $X^2 \sim F(1, v)$,
- If $X \sim F(v_1, v_2)$, then the pdf of $v_1 X$ converges to the $\chi^2(v_1)$ pdf as $v_2 \to \infty$.

Fig. 2.7 shows the $F(3, 10)$ pdf.

Figure 2.7: The F distribution.

2.7 Estimation

In all four examples in Example 2.1, the aspect of the population that we want to learn about is the population average, or **population mean**. This corresponds to the expected value $E(X)$ of the distribution that we use to model the population, where X is the random variable representing the outcome of a single draw. Our goal then is to estimate $E(X)$.

We can begin by assuming that the population follows a certain distribution. For example, we can use the log-normal distribution as a model for the population in Example 2.1(a), or the Poisson distribution for the population in Example 2.1(b), or the binomial distribution for the populations in Example 2.1(c) and (d). For the moment, however, we shall keep things general, and merely assume that the population is characterized by the parameters $E(X) = \mu$ and $Var(X) = \sigma^2$, without assuming a specific distribution for the population.[2]

We assume that you have a representative i.i.d. random sample $\{X_i\}_{i=1}^n$ from the population. The abbreviation "i.i.d." means **identically and independently distributed**. Since we assume the sample is representative of the population, each X_i will be identically distributed, with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1, \ldots, n$. The "independently distributed" part means that having drawn one particular individual does not alter the probability of another individual being drawn. We will talk about independent variables in the chapter, and later in the course we will see examples of non-independent samples. For now, we will simply assume an independent sample, which implies that $Var(X_i + X_j) = Var(X_i) + Var(X_j)$ for all $i \neq j$.

Since we are trying to estimate the population mean, it seems sensible to use the **sample mean** as the estimator for the population mean:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}\,. \tag{2.15}$$

You should think of (2.15) as a random variable. Every time you draw n observations from the population, you will get a different sample, and therefore a different sample mean. Put briefly, since the X_i's are random, then so is their average. Being a random variable, the sample mean will have an expected value, a variance, and a distribution.

An estimator $\hat{\theta}$ for some parameter θ is unbiased if $E(\hat{\theta}) = \theta$. It turns out that the sample mean is an unbiased estimator of the population mean:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu\,. \tag{2.16}$$

That is, the distribution of your estimator is centered about the population mean. You can interpret this to mean that by following this estimation rule, you will not systematically over- or under-estimate your target parameter. Another way to understand unbiasedness is to imagine the following: suppose a number of statisticians were to each draw n (different) samples from the population, and each calculate the sample mean for their own sample. Each statistician will

[2]Note that the mean and variance need not actually be separate parameters. In the case of the normal, they are indeed separate, but in the Poisson, both the mean and variance are equal to λ. In the binomial we have $E(X) = p$ and $Var(X) = p(1-p)$. In the log-normal, there are two parameters typically called μ and σ^2 but these do not represent the mean and variance of X. In what follows, I will use μ and σ^2 to represent the mean $E(X)$ and variance $Var(X)$ of X, no matter what the actual distribution of X is.

have a different sample mean. Some will overestimate the true value of $E(X)$ whereas others will underestimate it. What unbiasedness says is that *on average*, the statisticians will be correct; their sample means will center about the true value of $E(X)$.

How big of an error can we expect to make using this estimator? We can answer this question by looking at the variance of the estimator. The variance of the estimator measures the spread of the estimator's distribution around its mean, so it is a measure of how "precise" the estimator is. Assuming we have an i.i.d. sample, we have

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}\,. \tag{2.17}$$

The expression (2.17) tells you that the larger your sample size the smaller would your estimator variance be. A larger n means a smaller variance, so you get a more precise estimator. In your sample, some sample observations will be above the population mean, some below. When you take the average, the negative and positive errors cancel, so your overall error becomes smaller.

If you want to get a numerical estimate of the variance of the sample mean, you will have to estimate σ^2. How do we do that? Since

$$Var(X) = E((X - E(X))^2)$$

one obvious suggestion is to use the estimator

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\,. \tag{2.18}$$

Then the variance of the sample mean can be estimated as

$$\widetilde{Var}(\overline{X}) = \frac{\widetilde{\sigma^2}}{n}\,.$$

However, (2.18) turns out to be a biased estimator for σ^2. We can show this using the fact that

$$E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$\text{and} \quad E(\overline{X}^2) = Var(\overline{X}) + E(\overline{X})^2 = \sigma^2/n + \mu^2\,.$$

We have

$$E(\widetilde{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

The estimator (2.18) therefore systematically under-estimates the variance of the sample observations. If your sample size n is large, the bias may be negligible for all intents and purposes in which case there shouldn't be any problem using (2.18). Nonetheless, it is easy to derive an unbiased estimator for σ^2, namely

$$\widehat{\sigma^2} = \frac{n}{n-1}\widetilde{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{2.19}$$

The intuition for why the divisor in (2.19) has to be $n-1$ instead of n is that the deviations from sample mean always sum to zero. This means that there are only $n-1$ 'free' deviations from sample mean. For example, given $\sum_{i=1}^{n}(X_i - \overline{X})$ and the first $n-1$ deviations $(X_i - \overline{X})$, $i = 1, 2, ..., n-1$, you can determine the nth deviation as $(X_i - \overline{X}) = -\sum_{i=1}^{n-1}(X_i - \overline{X})$. One "degree of freedom" was lost because we had to use the observations to compute the sample mean in order to compute the deviations from sample mean.

Once you have obtained $\widehat{\sigma^2}$, you can replace σ^2 in (2.17) with it. We often report the **standard error** for the sample mean, defined as

$$\text{s.e.}(\overline{X}) = \sqrt{\frac{\widehat{\sigma^2}}{n}}\,.$$

Example 2.6. For our average hourly earnings example, we have

```
cat("Estimate of 2019 average hourly earnings\n")
X = dat1$earn
n = length(X)
Xbar = sum(X)/n
s2hat = sum((X-Xbar)^2)/(n-1)
Xse = sqrt(s2hat/n)
cat("sample mean ($):", round(Xbar,2), "   ")
cat("standard error ($):", round(Xse,2))
```

```
Estimate of 2019 average hourly earnings
sample mean ($): 29.23    standard error ($): 0.37
```

Using R built-in functions:

```
cat("Estimate of 2019 average hourly earnings\n")
cat("sample mean ($):", round(mean(X),2), "   ")
s2hat = var(X)
cat("standard error ($):", round(sqrt(s2hat/n),2))
```

```
Estimate of 2019 average hourly earnings
sample mean ($): 29.23    standard error ($): 0.37
```

Example 2.7. Coins can be weighted so that one side shows more frequently than the other in random tosses of the coin. Let $X = 1$ if heads shows, and $X = 0$ if tails shows and let p be the probability of obtaining heads. Then $X \sim \text{Binomial}(p)$. Suppose you randomly toss the coin n times and record the outcomes, giving you an i.i.d. sample $\{X_i\}_{i=1}^{n}$ such that $E(X_i) = p$ and $Var(X_i) = p(1-p)$ for all $i = 1, \ldots, n$. Then p can be estimated using the sample mean $\hat{p} = \overline{X}$ which is equal to the proportion of 1s in the sample. The variance of the estimator is

$$Var(\hat{p}) = p(1-p)/n$$

which can be estimated by

$$\widehat{Var}(\hat{p}) = \frac{\widehat{\sigma^2}}{n} \quad \text{where} \quad \widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\,.$$

We can take the square root of $\widehat{Var}(\hat{p})$ to get the standard error of the estimator.

Example 2.7 is an example of estimating the population mean of an infinite conceptual population. It is nonetheless mathematically identical to the next example, where we are estimating the population mean of a finite tangible population.

Example 2.8. Suppose you are interested in estimating the proportion p of smokers in a large population of size N. If X is a random draw from this population (with "smoker" = 1, "non-smoker"=0), then $X \sim \text{Binomial}(p)$. Suppose you randomly sample n people and ask if they smoke. Suppose you have done the appropriate randomization in selecting your sample, so that you can consider your sample $\{X_i\}_{i=1}^n$ to be a representative i.i.d. draw from the population.

As in the coin toss example, $E(X) = p$ so an unbiased estimator for p is the sample mean, which is the proportion of smokers in your sample. You can estimate the standard error of the estimator as in the previous example.

Note that the maximum value of the variance $Var(\hat{p}) = p(1-p)/n$ occurs at $p = 0.5$, i.e., for fixed n, the largest value of the standard error is

$$\sqrt{\frac{0.5(1-0.5)}{n}} = \sqrt{\frac{0.25}{n}}\,.$$

How many samples do you need to collect to ensure that the standard error is less than 0.01, or 1%. We have

$$\sqrt{\frac{0.25}{n}} < 0.01 \Rightarrow \frac{0.25}{n} < 0.0001 \Rightarrow n > 2500\,.$$

This is regardless of the size of the population.

2.8 Hypothesis Testing

To test if the population mean is equal to some specific numerical value μ_0, we check if the sample mean is "improbably far" from μ_0 when $\mu = \mu_0$ is assumed to be true. If it is, we construe this as evidence that the **null hypothesis** $H_0 : E(X) = \mu_0$ is false, and reject it in favor of the **alternative hypothesis** $H_A : E(X) \neq \mu_0$. But how far is "improbably far"? To provide an answer to this question, we need to derive the distribution of the sample mean when $\mu = \mu_0$, and to do so we need to know the distribution of X. If all you know is that $E(X) = \mu$ and $Var(X) = \sigma^2$, then you do not have enough information to derive the distribution of the sample mean.

In the case of the coin toss example, the structure of the problem does give us enough information to derive the finite sample distribution of the sample mean. Suppose $n = 2$. Then the possible values of the sample mean are 0, 1/2 and 1, corresponding to sample outcomes $(0, 0)$, $(0, 1)$ or $(1, 0)$, and $(1, 1)$ respectively. The corresponding probabilities are $(1-p)^2$, $2p(1-p)$, and p^2. For $n = 3$, the possible outcomes for the sample mean are:

- $\overline{X} = 0$, corresponding to outcome $(0, 0, 0)$, which occurs with probability $(1-p)^3$;
- $\overline{X} = 1/3$, corresponding to outcomes with 1 head out of 3 tosses. There are $\binom{3}{1} = 3$ such outcomes, so the probability is $3p(1-p)^2$.
- $\overline{X} = 2/3$, corresponding to outcomes with 2 heads out of 3 tosses. There are $\binom{3}{2} = 3$ such outcomes, so the probability is $3p^2(1-p)$.
- $\overline{X} = 1$, corresponding to outcome $(1, 1, 1)$ which occurs with probability p^3.

For a sample of size n, the possible values of the sample mean are i/n, $i = 0, 1, ..., n$, each corresponding to a set of $\binom{n}{i}$ outcomes comprising i heads out of n tosses, so the probability of obtaining a sample mean of i/n is

$$\Pr\left(\overline{X} = \frac{i}{n}\right) = \binom{n}{i} p^i (1-p)^{n-i}, \ i = 0, 1, 2, ..., n. \tag{2.20}$$

We can use (2.20) to help us decide if we should reject the hypothesis that the coin is fair.

Example 2.9. Suppose we have a sample of 20 coin tosses, and suppose that the coin is in fact fair, i.e., $p = 0.5$. The following is the probability distribution function $\Pr(\overline{X} = i/n)$, $i = 0, 1, ..., n$, of the sample mean, calculated using (2.20) with $p = 0.5$, and displayed in Fig. 2.8.

```
p <- 0.5
n <- 20
i <- 0:n       # i integers from 0 to 20
phat <- 0:n/n  # possible values of sample means
Pr_phat <- choose(n,i)*p^i*(1-p)^(n-i)
dim(Pr_phat) <- c(1,n+1) # make into row vector for presentation
colnames(Pr_phat) = paste0("p_hat=",i/n)
rownames(Pr_phat) = "Prob:"
noquote(format(Pr_phat, scientific=T,digits=6)) # another way to print to screen
```

```
      p_hat=0     p_hat=0.05  p_hat=0.1   p_hat=0.15  p_hat=0.2   p_hat=0.25
Prob: 9.53674e-07 1.90735e-05 1.81198e-04 1.08719e-03 4.62055e-03 1.47858e-02
      p_hat=0.3   p_hat=0.35  p_hat=0.4   p_hat=0.45  p_hat=0.5   p_hat=0.55
Prob: 3.69644e-02 7.39288e-02 1.20134e-01 1.60179e-01 1.76197e-01 1.60179e-01
      p_hat=0.6   p_hat=0.65  p_hat=0.7   p_hat=0.75  p_hat=0.8   p_hat=0.85
Prob: 1.20134e-01 7.39288e-02 3.69644e-02 1.47858e-02 4.62055e-03 1.08719e-03
      p_hat=0.9   p_hat=0.95  p_hat=1
Prob: 1.81198e-04 1.90735e-05 9.53674e-07
```

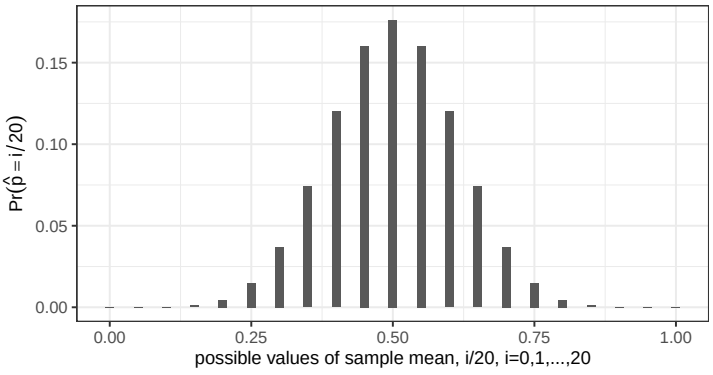


Figure 2.8: Distribution of sample mean, n=20, p=0.5.

Notice that there are non-zero probabilities on every possible outcome of the sample mean. This means that any reasonable decision rule that we use to reject or not reject the null hypothesis will carry a non-zero probability of rejection even when the null hypothesis is true (we call this

a "Type 1 error"). For example, suppose we use the rule "Reject $p = 0.5$ in favor of the alternative $p \neq 0.5$ if the frequency of heads $\hat{p}$ is less than 0.3 or greater than 0.7", which seems not unreasonable. We can calculate from the table above that in using this rule, there is a probability of approximately 0.0414 that we reject the null even though p is in fact equal to 0.5.

```
round(sum(Pr_phat[i/n<0.3])+sum(Pr_phat[i/n>0.7]),4)
```

```
[1] 0.0414
```

We can reduce the probability of Type 1 error by allowing for a larger range for $\hat{p}$ (perhaps reject if $\hat{p} < 0.05$ or $\hat{p} > 0.95$), but then the test loses power to reject a false hypothesis (i.e., the probability of failing to reject a wrong hypothesis — a "Type 2 error" — increases). In practice, researchers usually opt for rules such that the probability of an incorrect rejection of a true hypothesis is around 0.01, or 0.05, or 0.10.

What about the average hourly earnings example? Since we did not assume a specific distribution for the population, we cannot derive the distribution of the sample mean, which we need for hypothesis testing. All we know is that $E(\overline{X}) = \mu$ and $Var(X) = \sigma^2$. Had we assumed that X is normally distributed, then the sample mean $\overline{X}$ would also be normally distributed

$$\overline{X} \sim \text{Normal}\left(\mu\,, \frac{\sigma^2}{n}\right) \quad \text{or} \quad \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Normal}(0,1)\,.$$

Then, replacing σ^2 with $\widehat{\sigma^2}$ from (2.19), it can be shown that

$$t\text{-stat} = \frac{\overline{X} - \mu}{\sqrt{\widehat{\sigma^2}/n}} \sim \text{t}(n-1)\,. \tag{2.21}$$

We can then use this to do hypothesis testing. The t-stat measures the number of standard errors the sample mean is away from μ. Suppose you want to test $H_0 : \mu = 30$ vs $H_a : \mu \neq 30$. If the null hypothesis that $\mu = 30$ is true, then the t-stat with μ substituted for 30 will have the t-distribution. The idea then is to reject the hypothesis if the sample mean is too far away from the hypothesized $\mu = 30$. We use the following decision rule

$$\text{Reject } H_0 \;\text{ if }\; |t\text{-stat}| > c_{\alpha/2}$$

where $c_{\alpha/2}$ is the $\alpha/2$-percentile of the $\text{t}(n-1)$ distribution. This sets the probability of rejecting the null hypothesis when it is true at α. The typical values of α are 0.01, 0.05, 0.10.

This is illustrated in Fig. 2.9(a) which plots the t-distribution. For a 0.05-significance test (where the probability of rejecting a true hypothesis is controlled at 0.05), we would choose $c_{0.05/2}$ such that the probability in the shaded region is 0.05. For moderate values of n this is around 2, that is, we reject the hypothesis if the parameter estimate is more than twice the standard deviation away from the hypothesized value of the parameter. For large values of n, the t-distribution is practically the same as the normal distribution, and $c_{\alpha/2}$ is 1.96 for $\alpha = 0.05$.

An easier way is to report the p-value, defined as the area of the shaded region illustrated in Fig. 2.9(b), where t-stat‘ is the value of your t-statistic. If the p-value is greater than α, we do not reject the hypothesis at the α-significance level. If it is smaller than α, we reject the hypothesis.

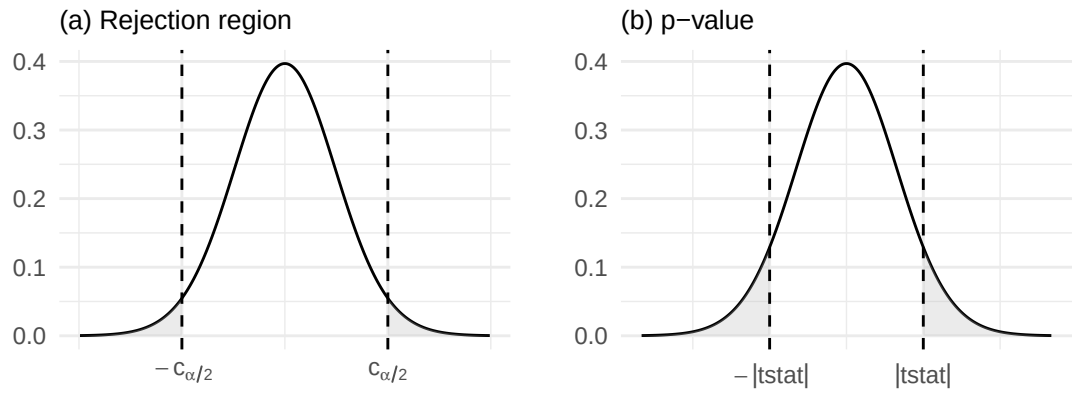


Figure 2.9: Hypothesis testing with the t-test

Example 2.10. For our average hourly earnings example, to test $H_0 : \mu = 30$ vs $H_A : \mu \neq 30$, we have

```
X=dat1$earn
n=length(X)
tstat = (mean(X)-30)/sqrt(var(X)/n)
pval = 2*pt(-abs(tstat), n-1)
cat("t-stat:", round(tstat, 3), "   p-val:", round(pval, 3))
```

```
t-stat: -2.087    p-val: 0.037
```

We reject the hypothesis at 0.05 significance level, but not at 0.01 significance level.

The big question mark with the test in Example 2.10 is that the assumption that the population is normally distributed is not appropriate. It might be an appropriate assumption for $\ln earn$ but not for $earn$. If $earn$ is not normally distributed, the t-stat does not have a t-distribution, and so the conclusion of the t-test might be suspect. It turns out that the test we just did is probably fine given our sample size. We explain in Section 2.9.2 the reason for this conclusion.

Notice that the sample mean remains unbiased. We did not need to make any assumptions about the specific form of the distribution in our proof that the sample mean is unbiased for the population mean. The only question is with regard to the conclusion of the hypothesis test.

2.9 Asymptotic Analysis

Asymptotic analysis refers to results that apply "in the limit", as the sample size goes to infinity. It serves to approximate the finite sample properties of estimators if the sample size is reasonably large, and is especially helpful when the finite sample properties are unknown. We continue to focus on the sample mean, which we now denote as $\overline{X}_n$ to indicate the sample size used to calculate it.

2.9.1 Consistency and the Law of Large Numbers

We have mentioned the desirability of larger sample sizes. For the general problem of estimating the population mean of a random variable X with mean μ and variance σ^2 using the sample mean, we have $Var(\overline{X}_n) = \sigma^2/n \to 0$ as $n \to \infty$. Since $\overline{X}_n$ is unbiased, and its variance collapses

to zero as the sample size goes to infinity, the estimator converges to the population mean as the sample size grows larger and larger.

The convergence of $\overline{X}_n$ to μ is not quite the same as the convergence of, say, the deterministic sequence $1/n$ to zero. In the latter case, I know that if n is large enough, then $1/n$ will *definitely* be within a certain distance of 0. For instance, if $n > 1000$, then $1/n < 0.001$ *for sure*. In the case of $\overline{X}_n$, which is a sequence of *random variables*, we cannot make such a definite claim.

The convergence concept we use for random variables is "convergence in probability". A sequence of random variables Y_n is said to **converge in probability** to some value c as $n \to \infty$ if for any $\epsilon > 0$ (no matter how small),

$$\Pr(\,|Y_n - c| > \epsilon\,) \to 0 \;\text{ as }\; n \to \infty\,.$$

This allows for some probability that the distance between Y_n and c exceeds ϵ at any sample size n, but as n increases towards infinity, this probability becomes vanishingly small. We write $\operatorname{plim} Y_n = c$ or $Y_n \xrightarrow{p} c$. It can be shown that if $E(Y_n)$ converges to c and $Var(Y_n)$ converges to zero, then Y_n converges in probability to c. The sample mean, therefore, converges in probability to the population mean.

In the context of parameter estimation, we say that an estimator is **consistent** if it converges in probability to the true value of the parameter it is estimating. The sample mean $\overline{X}_n$ is a consistent estimator for μ under quite general conditions. This result is known as a **law of large numbers**. There are several laws of large numbers, each describing a set of conditions which, if met, guarantee the consistency of the sample mean. We state one such law below:

Theorem 2.1 (Khinchine's Weak Law of Large Numbers, WLLN)**.** *If* $\{X_i\}_{i=1}^n$ *are i.i.d. with* $E(X_i) = \mu < \infty$ *for all* i*, then*

$$\overline{X}_n \xrightarrow{p} \mu\,.$$

There are other kinds of convergence concepts for sequences of random variables, but for the moment we consider only convergence in probability. The theorem above is referred to as a *weak* law of large numbers because the convergence concept used is convergence in probability, and there are "stronger" forms of probabilistic convergence.

The following result is used frequently:

Proposition 2.1. *If* $g(.)$ *is a continuous function, then*

$$Y_n \xrightarrow{p} c \Rightarrow g(Y_n) \xrightarrow{p} g(c). \tag{2.22}$$

That is, if $\operatorname{plim} Y_n$ *exists, and* $g(.)$ *is continuous, then* $\operatorname{plim} g(Y_n) = g(\operatorname{plim} Y_n)$.

For example, if Y_n converges in probability to c, then $Y_n^2 \xrightarrow{p} c^2$. Result (2.22) extends to continuous functions of multiple variables. This implies that if $Y_n \xrightarrow{p} c_y$ and $Z_n \xrightarrow{p} c_z$, then

- $Y_n + Z_n \xrightarrow{p} c_y + c_z$,
- $Y_n Z_n \xrightarrow{p} c_y c_z$,
- $Y_n / Z_n \xrightarrow{p} c_y / c_z$, as long as c_z is not zero.

Example 2.11. Suppose $\{X_i\}_{i=1}^n$ is an i.i.d. sample, with $E(X_i) = \mu < \infty$ and $var(X_i) = \sigma^2 < \infty$ for all i. We show that the biased estimator

$$\widetilde{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$$

is consistent for the population variance σ^2. Since $\{X_i\}_{i=1}^n$ are i.i.d., so are $\{X_i^2\}_{i=1}^n$. Furthermore, $E(X_i^2) = \sigma^2 + \mu^2 < \infty$, so

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{p} \sigma^2 + \mu^2 .$$

Since $\overline{X}_n \xrightarrow{p} \mu$ and "power of two" is a continuous function, we have $\overline{X}_n^2 \xrightarrow{p} \mu^2$. Therefore $\widetilde{\sigma^2}$ converges in probability to $\sigma^2 + \mu^2 - \mu^2 = \sigma^2$. This example shows that consistent estimators can be biased in finite samples.

Example 2.12. Since $\widehat{\sigma_n^2} = \frac{n}{n-1}\widetilde{\sigma_n^2}$, and because $\frac{n}{n-1} \to 1$ and $\widetilde{\sigma^2} \xrightarrow{p} \sigma^2$, we have $\widehat{\sigma_n^2} \xrightarrow{p} \sigma^2$.

Example 2.13. Since both $\widehat{\sigma_n^2}$ and $\widetilde{\sigma_n^2}$ are consistent estimators for σ^2, both $(\widehat{\sigma_n^2})^{1/2}$ and $(\widetilde{\sigma_n^2})^{1/2}$ are consistent estimators for σ.

It should be noted that unbiasedness, unlike consistency, generally does not carry over from estimators to non-linear functions of estimators. For instance, we saw earlier that $E(\overline{X}_n^2) \geq \mu^2$ even though $E(\overline{X}_n) = \mu$.

It may seem that unbiasedness is a more relevant way to judge an estimator than consistency since we never have infinite sample sizes, but consistency is still useful as it ensures that as sample size grows, our estimates become more reliable. Furthermore, in more complex applications it can be difficult or impossible to find unbiased estimators, but it is often reasonably straightforward to find consistent ones. We have also seen that it is easy to find consistent estimators of continuous functions of parameters once we have consistent estimators for the parameters.

In Example 2.9 we derived the distribution of the sample mean in the coin toss example, and calculated this distribution for a fair coin with sample size 20. We repeat this exercise, this time for a coin with $p = 0.25$, for sample sizes of 5, 10, 20, 100, 200 and 400. We present the probability distribution functions graphically in Fig. 2.10. The convergence in probability of the sample mean to the true value of p can be seen in these graphs.

2.9.2 Asymptotic Normality

The distribution of the sample mean in the example above, with $p = 0.25$, is unsurprisingly skewed in small samples because of the low probability of heads relative to tails. However, the shape of the distribution appears to quickly becomes quite symmetric as sample size grows, and appears to converge to a familiar bell-shaped distribution. Of course, in the limit the distribution collapses to a degenerate one with all of the probability at $p = 0.25$. This is because the variance of the sample mean, $Var(\hat{p}) = p(1-p)/n$ goes to zero as $n \to \infty$. Suppose, however, that we scale the sample mean (after subtracting p) by $\sqrt{n}$, i.e., suppose we look at the distribution of

$$\sqrt{n}(\hat{p} - p). \tag{2.23}$$

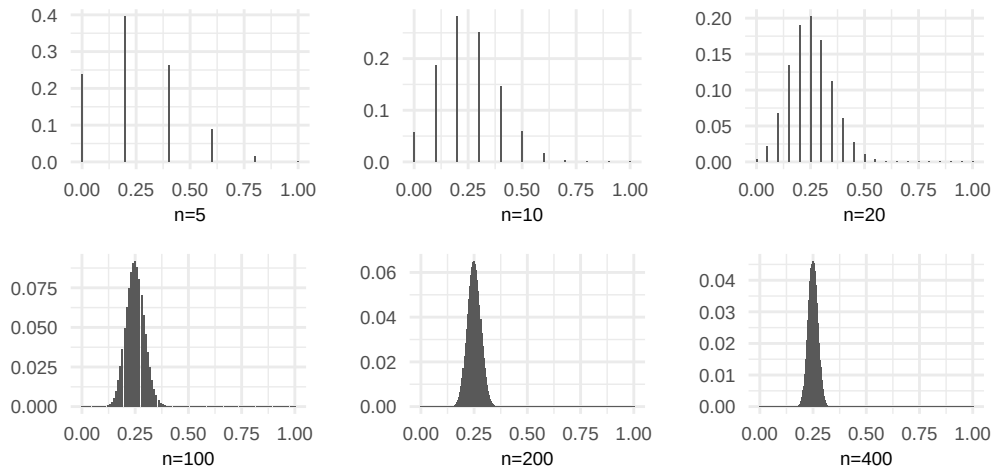


Figure 2.10: Consistency of sample mean to p=0.25.

This random variable now has mean 0 and a non-collapsing variance $np(1-p)/n = p(1-p)$. We can then talk about the shape of (2.23) as $n \to \infty$ without the distribution collapsing to a single point. The plots in Fig. 2.11 show the same distributions as in Fig. 2.10, but after centering and scaling as in (2.23). The distributions appear to take the shape of a normal distribution as sample size increases.

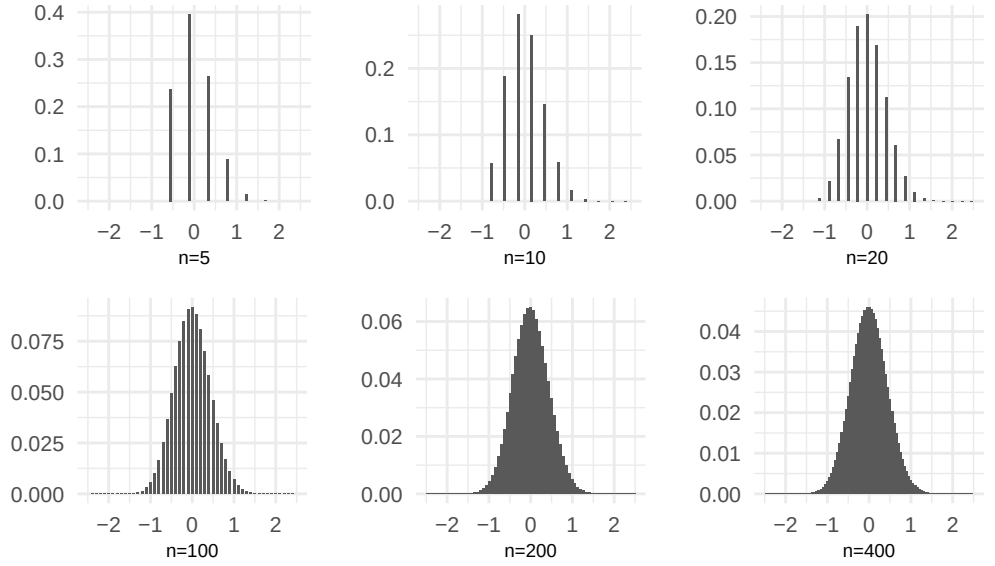


Figure 2.11: Convergence of distribution to normal (pdf view).

2.9.3 The Central Limit Theorem

The convergence of the cdf of the (centered and scaled) sample mean in the coin toss example to a normal cdf is an instance of the **Central Limit Theorem** (CLT), a key result in probability theory. As with the law of large numbers, there are many CLTs, each listing out a set of conditions under which convergence to normality is guaranteed. We state one such CLT:

Theorem 2.2 (Lindeberg-Levy CLT)**.** *If* $\{X_i\}_{i=1}^n$ *are i.i.d. with* $E(X_i) = \mu < \infty$ *and* $Var(X_i) = \sigma^2 < \infty$ *for all* i*, then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$$

where $\xrightarrow{d}$ *means* ***convergence in distribution****.*[3]

Our plots of the distribution of $\sqrt{n}(\hat{p}_n - p)$ in the coin toss example suggests convergence in distribution to Normal$(0, p(1-p))$. The sample $\{X_i\}$ in the coin toss example does meet the requirements of the Lindeberg-Levy CLT, so in fact $\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \text{Normal}(0, p(1-p))$

Sometimes we want to indicate that a sequence of random variables Y_n converges in distribution to the cdf of some random variable Y. To so do, we write $Y_n \xrightarrow{d} Y$.

Proposition 2.2 (Properties of convergence in distribution)**.**

(a) If $g(.)$ *is a continuous function and* $Y_n \xrightarrow{d} Y$*, then* $g(Y_n) \xrightarrow{d} g(Y)$*.*

(b) If $Y_n \xrightarrow{p} Y$*, then* $Y_n \xrightarrow{d} Y$*.*

(c) If $a_n \xrightarrow{p} a$ *and* $Y_n \xrightarrow{d} Y$*, then* $a_n Y_n \xrightarrow{d} aY$*.*

Example 2.14. If $Y_n \xrightarrow{d} Y \sim \text{Normal}(0,1)$, then $Y_n^2 \xrightarrow{d} Y^2 \sim \chi^2(1)$, since the square of a standard normal is $\chi^2(1)$.

Example 2.15. If $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$ and s_n^2 is any consistent estimator of σ^2, then $1/s_n = (1/s_n^2)^{1/2}$ converges in probability to $1/\sigma$, and therefore

$$t = \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} \xrightarrow{d} \text{Normal}(0,1). \tag{2.24}$$

If $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$, we would be justified, in large enough samples, to say that the distribution of $\sqrt{n}(\overline{X}_n - \mu)$ is approximately Normal$(0, \sigma^2)$, or that $\overline{X}_n$ is approximately Normal$(\mu, \sigma^2/n)$. This last statement is sometimes written $\overline{X}_n \overset{a}{\sim} \text{Normal}(\mu, \sigma^2/n)$, where the "$a$" stands for "approximately" (some take "a" to stand for "asymptotically").

Result (2.24) is useful for hypotheses testing when one is unable or unwilling to make an assumption regarding the distribution of the population. Suppose $\{X_i\}_{i=1}^n$ is an i.i.d. sample with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. The sample mean $\overline{X}_n$ is a consistent estimator for μ and $\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ is a consistent estimator for σ^2. To test the null hypothesis $H_0 : \mu = \mu_0$, we need to compute the distribution of the sample mean, but you cannot do this unless you know the distribution of each X_i. Result (2.24), however, tells us that if our sample size is large enough, then under the null hypothesis,

$$t = \frac{\overline{X}_n - \mu_0}{\sqrt{\widehat{\sigma^2}/n}} \overset{a}{\sim} \text{Normal}(0,1). \tag{2.25}$$

It suggests that we use the decision rule "reject the null if $|t| > c_\alpha$" where c_α is that value such that $\Pr(|t| > c_\alpha) = \alpha$, where α is the chosen "level of significance" of the test, i.e., the

[3]Technically, this means that the **cumulative distribution function** (cdf) of the random variable on the left converges pointwise to the cdf of the distribution indicated on the right.

probability of rejecting the null when it is true, and where the value c_α is found from the standard normal distribution. For 0.01, 0.05, 0.10 levels of significance, the appropriate values of c_α are approximately

```
round(qnorm(c(0.995, 0.975, 0.95)),3)
```

```
[1] 2.576 1.960 1.645
```

respectively. The 0.05 level of significance test, in particular, says to reject $H_0 : \mu = \mu_0$ if the absolute distance from the sample mean to the hypothesized value μ_0 is more than 1.96 (or approximately 2) standard errors.

A test based on the statistic in (2.25) and rejection values (or 'critical values') based on the Normal$(0, 1)$ distribution, would be an approximate test in the sense that the true significance level may not be exactly α, as intended. Nonetheless, it is a way forward in a situation where an exact test is unavailable, such as in the case of Example 2.10.

2.9.4 Working with Log-Transformed Variables

Returning to the average hourly earnings example, suppose we had computed the sample mean of $\ln earn_i$ instead of the sample mean of $earn_i$. Then, in an effort to obtain an estimate of the population mean of $earn$, we take exponents of the sample mean of $\ln earn_i$. For the data in `earn2019.csv`, we have

```
X = dat1$earn
Xbar = mean(X)
lnX = log(dat1$earn)
lnXbar = mean(lnX)
cat("sample mean of earn:", round(Xbar,2), "\n")
cat("sample mean of log earn:", round(lnXbar,2), "\n")
cat("exponent of sample mean of log earn:", round(exp(lnXbar),2), "\n")
```

```
sample mean of earn: 29.23
sample mean of log earn: 3.15
exponent of sample mean of log earn: 23.36
```

What are we getting? To make arguments simpler, we consider consistency instead of unbiasedness, and assume that $earn \sim$ Log-normal(μ, σ^2). Then we have $E(\ln earn) = \mu$, and from the properties of the log-normal distribution, we have

$$E(earn) = \exp \mu \exp(\sigma^2/2) > \exp \mu = Median(earn)$$

where the inequality arises because $\exp(\sigma^2/2) > 1$. We know the sample mean of $\ln earn_i$ is consistent for μ, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} \ln earn_i \xrightarrow{p} \mu \,,$$

therefore the exponent of the sample mean of $\ln earn_i$ is consistent for the the population median of $earn$:

$$\exp \left(\frac{1}{n} \sum_{i=1}^{n} \ln earn_i \right) \xrightarrow{p} \exp \mu \,.$$

However, the exponent of the sample mean of $\ln earn_i$ consistently underestimates the population mean of $earn$.

If we want to get an estimate of $E(earn)$ from the sample mean of $\ln earn_i$, we have to apply a correction. Since $\widehat{\sigma^2} \xrightarrow{p} \sigma^2$, a consistent estimator for $E(earn)$ is

$$\widehat{E(earn)} = \exp \hat{\mu} \exp \widehat{\sigma^2} \,.$$

For our data, we have

```
s2hat=var(lnX)
cat("exponent of sample mean of log earn, with correction :", round(exp(lnXbar)*exp(s2hat/2),2))
```

```
exponent of sample mean of log earn, with correction : 28.91
```

which is quite close to the sample mean of *earn*. One might argue, nonetheless, that as the distribution of *earn* is so skewed, the median might be a better measure of location than the mean.

2.10 Chapter 2 Exercises

Exercise 2.1. Let $\{X_i\}_{i=1}^n$ be an iid sample from a Bernoulli distribution with parameter p, i.e.,

$$E(X_i) = p \quad \text{and} \quad Var(X_i) = p(1-p) \quad \text{for all} \quad i = 1, \dots, n.$$

An unbiased estimator for p is the sample mean $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$. Since $Var(\hat{p}) = p(1-p)/n$, consider estimating $Var(\hat{p})$ using

$$\widehat{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \,.$$

Show that this is equivalent to using

$$\widetilde{Var}(\hat{p}) = \frac{\widetilde{\sigma^2}}{n}$$

where

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \,.$$

Exercise 2.2. Given a sample $\{X_i\}_{i=1}^n$, consider a weighted average

$$\widetilde{X} = \sum_{i=1}^n w_i X_i = w_1 X_1 + w_2 X_2 + \dots w_n X_n$$

as an estimator for the population mean. The sample mean $\overline{X}$ is a special case of $\widetilde{X}$ with $w_i = 1/n$ for all $i = 1, \dots, n$. Show that $\widetilde{X}$ is an unbiased estimator for the population mean as long as $\sum_{i=1}^n w_i = 1$. Assuming that $\sum_{i=1}^n w_i = 1$, show that

$$Var(\widetilde{X}) \geq Var(\overline{X}) \,.$$

Exercise 2.3. A measure of the quality of an estimator $\hat{\theta}$ for a parameter θ is the **mean squared estimation error**

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

Show that

$$MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$$

where $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Exercise 2.4. It can be shown that if Y_i, $i = 1, 2, ..., n$ are iid draws from a normal distribution with mean μ and variance σ^2, then the variance of the unbiased variance estimator

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

is

$$Var(\widehat{\sigma^2}) = \frac{2\sigma^4}{n-1}.$$

For this question, you may take the above fact as given. Because $\widehat{\sigma^2}$ is unbiased, its MSE also has the same expression.

(a) Show that the biased estimator $\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ has a smaller variance than $\widehat{\sigma^2}$.

(b) Show that $MSE(\widetilde{\sigma^2}) = \frac{2n-1}{n^2}\sigma^4$.

(c) Show that $MSE(\widetilde{\sigma^2}) < MSE(\widehat{\sigma^2})$.

This is an example where a biased estimator might be prefered to an unbiased one. Here the MSE of the unbiased estimator is lowered by reducing variance at the expense of a little bit of bias.

Exercise 2.5. The quantile function `qnorm(p, mean, sd)`, when evaluated at probability p, returns the value of c for which $\Pr(X \leq c) = p$, when $X \sim \text{Normal}(\text{mean}, \text{sd}^2)$. The default values of `mean` and `sd` are 0 and 1 respectively. For example:

```
qnorm(0.5)  # c such that Pr(X<=c)=0.5 when X~N(0,1)
```

```
[1] 0
```

The corresponding functions for the student-t, chi-sq, and F distributions are `qt(p, df)`, `qchisq(p,df)`, and `qf(p,df1, df2)` respectively.

(a) For $p = 0.01, 0.05, 0.10$, plot the lines c and $-c$ against v such that

i. $\Pr(|X| > c) = p$ when $X \sim \text{t}(v)$ for $v = 10, 20, ..., 200$.

ii. $\Pr(|X| > c) = p$ when $X \sim \text{Normal}(0, 1)$. (This will be two constant lines.)

The code to do (a) is given below.

```
v <- seq(10, 200, 10)
c1 <- qt(0.995, v); c2 <- qt(0.975, v); c3 <- qt(0.95, v)
n1 <- qnorm(0.995); n2 <- qnorm(0.975); n3 <- qnorm(0.95)
plt_df <- data.frame(c1=c1, n1=n1, c2=c2, n2=n2, c3=c3, n3=n3, v=v)
p1 <- ggplot(data=plt_df) +
  geom_line(aes(y= c1, x=v), linetype='dashed', color='cadetblue', linewidth=1) +
  geom_line(aes(y=-c1, x=v), linetype='dashed', color='cadetblue', linewidth=1) +
  geom_line(aes(y= n1, x=v), linetype='solid', color ='cadetblue', linewidth=1) +
  geom_line(aes(y=-n1, x=v), linetype='solid', color ='cadetblue', linewidth=1) +
  geom_line(aes(y= c2, x=v), linetype='dashed', color='chocolate', linewidth=1) +
  geom_line(aes(y=-c2, x=v), linetype='dashed', color='chocolate', linewidth=1) +
  geom_line(aes(y= n2, x=v), linetype='solid', color ='chocolate', linewidth=1) +
  geom_line(aes(y=-n2, x=v), linetype='solid', color ='chocolate', linewidth=1) +
  geom_line(aes(y= c3, x=v), linetype='dashed', color='darkorchid', linewidth=1) +
  geom_line(aes(y=-c3, x=v), linetype='dashed', color='darkorchid', linewidth=1) +
  geom_line(aes(y= n3, x=v), linetype='solid', color ='darkorchid', linewidth=1) +
  geom_line(aes(y=-n3, x=v), linetype='solid', color ='darkorchid', linewidth=1) +
  theme_bw() + ylim(-3.5,3.5) + xlim(1, 200)
p1
```

(b) For $p = 0.01, 0.05, 0.10$, plot the lines c against v such that $\Pr(X > c) = p$ when $X \sim \chi^2(v)$ for $v = 10, 20, \ldots, 200$.

(c) For $p = 0.01, 0.05, 0.10$, plot the lines c against v_2 such that $\Pr(X > c) = p$ when $X \sim F(v_1, v_2)$ for $v_2 = 10, 20, \ldots, 200$. Do this for $v_1 = 2, 3, 4, 5$, one figure per value of v_1.

2.11 Appendix: The Summation Notation

The uppercase sigma "Σ" is used to denote summation. For a set of numbers $\{x_1, x_2, \ldots, x_n\}$, define

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \,.$$

Example 2.16. The sample average (also sample mean, arithmetic mean) of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \,.$$

Example 2.17. Write the sum $4 + 8 + 12 + 16 + 20 + 24$ in summation notation. Ans: $\sum_{i=1}^{6} 4i$.

Example 2.18. In economics and finance, the **present value** of a future amount of money is the amount today that, if invested at a certain rate, returns that future sum. Suppose the following payments are to be made: a_1 at the end of the first period, a_2 at the end of the second period, and so on, for n periods. At a fixed interest rate of r per period, the present value of the payments is

$$\frac{a_1}{1+r} + \frac{a_2}{(1+r)^2} + \cdots + \frac{a_n}{(1+r)^n} = \sum_{i=1}^{n} \frac{a_i}{(1+r)^i} \,.$$

Example 2.19. $\sum_{i=1}^{n} c = nc$.

Expressions using the summation notation are not unique.

Example 2.20. Write $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11}$ in summation notation.

$$\text{Ans: } \sum_{i=1}^{6} (-1)^{i-1} \frac{1}{2i-1}.$$

$$\text{Alternative Ans: } \sum_{i=0}^{5} (-1)^{i} \frac{1}{2i+1} \,.$$

2.11.0.1 Rules for summation notation

The summation notation greatly simplifies notation but this is only helpful if you know how to manipulate expressions written using it. There are only two rules to learn:

i. $$\sum_{i=1}^{n} (a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$$

ii. $$\sum_{i=1}^{n} (ca_i) = c \sum_{i=1}^{n} a_i \quad \text{where } c \text{ is some constant.}$$

Example 2.21. Given any set of numbers $\{x_1, x_2, \ldots, x_n\}$, we have

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0\,.$$

That is, the sum of deviations of any set of numbers from its sample average is always zero.

Proof: $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = n\overline{x} - n\overline{x} = 0.$

Example 2.22. Given n pairs of numbers (x_i, y_i), $i = 1, 2, \ldots, n$, we have

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n}(y_i - \overline{y})x_i\,. \tag{2.26}$$

Proof: For the first equality in (2.26), we have

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n}(x_i - \overline{x})y_i - \sum_{i=1}^{n}(x_i - \overline{x})\overline{y} \\ &= \sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} = \sum_{i=1}^{n}(x_i - \overline{x})y_i.\end{aligned}$$

The second equality in (2.26) can be shown in similar fashion.

Chapter 3
Conditional Expectations / Linear Regression Overview

We will use the following packages in these notes:

```
library(tidyverse)
library(patchwork)
library(latex2exp)
```

3.1 Joint and Conditional Probabilities

We model the joint behavior of two random variable using a **joint probability distribution function**. We will use a simple example with two discrete random variables to illustrate the main ideas.

3.1.1 Joint and Marginal Distributions

Suppose X and Y are discrete random variables with range $x = 1, 2, 3, 4, 5$ and $y = 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0$. Their joint pdf $f_{X,Y}(x, y)$ gives you the probability of events of the form $X = x$ and $Y = y$, i.e.,

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y)\,.$$

Suppose the joint pdf of X and Y are as given below:

	6	0	0	0	0	$\frac{1}{20}$
	5.5	0	0	0	$\frac{1}{20}$	$\frac{2}{20}$
	5	0	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
y	4.5	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0
	4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0	0
	3.5	$\frac{2}{20}$	$\frac{1}{20}$	0	0	0
	3	$\frac{1}{20}$	0	0	0	0
		1	2	3	4	5
				x		

(3.1)

so we have, for example,

$$\Pr(X = 1, Y = 3) = f_{X,Y}(1, 3) = \frac{1}{20}, \Pr(X = 3, Y = 3) = 0, \Pr(X \geq 4, Y \geq 5) = \frac{7}{20}\,.$$

In a large sample of observations $\{X_i, Y_i\}_{i=1}^n$ from this population, you should find that approximately 1/20th of the observations will comprise the pair $(X_i, Y_i) = (1, 3)$, approximately 35% of the sample will have $X_i \geq 4$ and $Y_i \geq 5$, and will be no observations with $X_i = 3$ and $Y_i = 3$.

What is the probability of observing $X = 1$ (regardless of the value of the accompanying Y value)? To find $\Pr(X = 1)$, add up all the probabilities of events where $X = 1$, i.e.,

$$\begin{aligned}\Pr(X = 1) &= \Pr(Y = 3, X = 1) + \Pr(Y = 3.5, X = 1) + \cdots + \Pr(Y = 6, X = 1) \\ &= \frac{1}{20} + \frac{2}{20} + \frac{1}{20} = 0.2\,.\end{aligned}$$

You can repeat this calculation for $\Pr(X = 2)$, $\Pr(X = 3)$, $\Pr(X = 4)$, $\Pr(X = 5)$. You should find that:

x	1	2	3	4	5
$\Pr(X = x)$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$

This is the "**marginal**" (or "**unconditional**") pdf of X. In this example X is uniformly distributed over the values $X = 1, 2, ..., 5$. If you look at a large sample of observations $\{X_i, Y_i\}_{i=1}^n$ from this population, you will find that 1/5th of the pairs (X_i, Y_i) in your sample will have $X_i = 1$, 1/5th will have $X_i = 2$ and so on.

Similar calculations will give you the marginal distribution of Y.

	6	0	0	0	0	$\frac{1}{20}$		6	$\frac{1}{20}$	
	5.5	0	0	0	$\frac{1}{20}$	$\frac{2}{20}$		5.5	$\frac{3}{20}$	
	5	0	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$		5	$\frac{4}{20}$	
y	4.5	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0		4.5	$\frac{4}{20}$	
	4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0	0	$\rightarrow$	4	$\frac{4}{20}$	$\rightarrow$ $E(Y) = 4.5$
	3.5	$\frac{2}{20}$	$\frac{1}{20}$	0	0	0		3.5	$\frac{3}{20}$	$Var(Y) = 0.625$
	3	$\frac{1}{20}$	0	0	0	0		3	$\frac{1}{20}$	
		1	2	3	4	5		y	$\Pr(Y = y)$	
				x						

$$\downarrow$$

x	1	2	3	4	5
$\Pr(X = x)$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$

$\rightarrow$ $E(X) = 3$, $Var(X) = 2$

The marginal distribution of Y in our example is somewhat "bell-shaped". You can calculate the (unconditional) means and variances of X and Y from their marginal pdfs using the usual formulas.

In general, the joint pdf of two random variables is the function $f_{X,Y}(x, y)$. For discrete random variables, the marginal pdf of X is computed as $f_X(x) = \sum_y f_{X,Y}(x, y)$ where $\sum_y$ indicates summation over the possible values of Y. Likewise, the marginal pdf of Y is computed as $f_Y(y) = \sum_x f_{X,Y}(x, y)$. For continuous random variables, the marginals are computed as integrals:

$$f_X(x) = \int_Y f_{X,Y}(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_X f_{X,Y}(x, y)\, dx\,.$$

We can extend the joint pdf concept to multiple variables, e.g., $f_{X,Y,Z}(x, y, z)$, and so on.

3.1.2 Covariance and Correlation

It seems clear from (3.1) that there is a positive relationship between X and Y. One way to describe the relationship between the random variables is to calculate the **covariance** between X and Y, defined as

$$\sigma_{X,Y} = Cov(X, Y) = E((X - E(X))(Y - E(Y))).$$

In our example, we have

$$\begin{aligned}
Cov(X,Y) = {} & (5-3)(6.0-4.5)\tfrac{1}{20}+ \\
& (4-3)(5.5-4.5)\tfrac{1}{20} + (5-3)(5.5-4.5)\tfrac{2}{20}+ \\
& (3-3)(5.0-4.5)\tfrac{1}{20} + (4-3)(5.0-4.5)\tfrac{2}{20} + (5-3)(5.0-4.5)\tfrac{1}{20}+ \\
& (2-3)(4.5-4.5)\tfrac{1}{20} + (3-3)(4.5-4.5)\tfrac{2}{20} + (4-3)(4.5-4.5)\tfrac{1}{20}+ \\
& (1-3)(4.0-4.5)\tfrac{1}{20} + (2-3)(4.0-4.5)\tfrac{2}{20} + (3-3)(4.0-4.5)\tfrac{1}{20}+ \\
& (1-3)(3.5-4.5)\tfrac{2}{20} + (2-3)(3.5-4.5)\tfrac{1}{20}+ \\
& (1-3)(3.0-4.5)\tfrac{1}{20} \\
= {} & 1
\end{aligned}$$

One problem with the covariance measure is that it is not invariant to scale. For instance, suppose X is currently measured in thousands of dollars. If we re-scale to dollars by multiplying X by 1000, then the covariance becomes

$$\begin{aligned}
Cov(1000X, Y) &= E((Y - E(Y))(1000X - E(1000X))) \\
&= 1000E((Y - E(Y))(X - E(X))) \\
&= 1000\,Cov(X, Y).
\end{aligned}$$

For this reason, the correlation coefficient

$$\rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}},$$

which is invariant to scale and always lies between -1 and 1, is more informative.

Given a sample $\{X_i, Y_i\}_{i=1}^n$ from a joint pdf $f_{X,Y}(x, y)$, we can estimate the covariance using the **sample covariance**

$$\hat{\sigma}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of X_i and Y_i respectively. To estimate the correlation coefficient, we can divide the sample covariance by the sample standard deviations, i.e.,

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}} \ .$$

The following properties of means, variances and covariances are easy to show: if a and b are constants, we have

i. $E(aX + bY) = aE(X) + bE(Y)$,

ii. $Var(aX + bY) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X, Y)$,

iii. $Cov(X, Y) = E(XY) - E(X)E(Y)$,

iv. $Cov(X, X) = Var(X)$.

From ii, we see that the variance of a sum is the sum of the variances only if the variables are uncorrelated. From iii we see that $Cov(X, Y) = E(XY)$ if either X or Y has mean zero.

3.1.3 Conditional Distributions

Another way of describing the relationship between random variables is via conditional distributions. These describe the behavior of one variable for various values of the other. For instance, if we observe $X = 1$ but do not observe the Y realization, what can we predict about the behavior of Y? For the joint pdf in (3.1), we know that only three values of Y are possible when $X = 1$, with $Y = 1$ and $Y = 4$ equally likely, and $Y = 3.5$ twice as likely as either of these. Other values of Y have probability zero. To describe the behavior of Y when $X = 1$ as a probability distribution function, we have to make the total probabilities when $X = 1$ sum to one, so we divide each of these probabilities by their sum (i.e., by $\Pr(X = 1)$) to obtain the conditional probabilities:

$$\Pr(Y = 3 \mid X = 1) = \frac{\frac{1}{20}}{\frac{4}{20}} = \frac{1}{4}\,,\ \Pr(Y = 3.5 \mid X = 1) = \frac{\frac{2}{20}}{\frac{4}{20}} = \frac{1}{2}\,,\ \Pr(Y = 4 \mid X = 1) = \frac{1}{4}\,,$$

$$\Pr(Y = 4.5 \mid X = 1) = \Pr(Y = 5.5 \mid X = 1) = \Pr(Y = 6.5 \mid X = 1) = 0\,.$$

This collection of probabilities make up the **conditional pdf** of Y given $X = 1$. Making these calculations for each value of X gives

$$\begin{array}{cc|ccccc} & & \multicolumn{5}{c}{\Pr(Y \mid X)} \\ & 6 & 0 & 0 & 0 & 0 & \frac{1}{4} \\ & 5.5 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} \\ & 5 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ Y & 4.5 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ & 4 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ & 3.5 & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 \\ & 3 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \hline & & 1 & 2 & 3 & 4 & 5 \\ & & & & X & & \end{array} \tag{3.2}$$

Each column of (3.2) represents a complete pdf, so we have a collection of five pdfs, one for each possible value of X.

For any given value of $X = x$, we can use the corresponding conditional pdf to compute the conditional mean of Y given $X = x$, and the conditional variance of Y given $X = x$. For $X = 1$, we have:

$$E(Y \mid X = 1) = 3(\tfrac{1}{4}) + 3.5(\tfrac{1}{2}) + 4(\tfrac{1}{4}) + 4.5(0) + 5(0) + 5.5(0) + 6(0) = 3.5$$
$$Var(Y \mid X = 1) = (3 - 3.5)^2(\tfrac{1}{4}) + (3.5 - 3.5)^2(\tfrac{1}{2}) + (4 - 3.5)^2(\tfrac{1}{4}) + 0 + 0 + 0 + 0 = 0.125$$

Repeating these calculations for each value of X we get:

X	1	2	3	4	5
$E(Y \mid X)$	3.5	4	4.5	5	5.5
$Var(Y \mid X)$	0.125	0.125	0.125	0.125	0.125

(3.3)

Notice that $E(Y \mid X)$ is a function of X. In our example, the conditional mean of Y given X increases with X. In particular, we have $E(Y \mid X) = 3 + 0.5X$, $X = 1, 2, 3, 4, 5$. The conditional variance in this example turns out to be constant: $Var(Y \mid X) = 0.125$ for all X. In general it may vary with X.

In this example, knowledge of the value of X gives us information that we can use to refine our view of the behavior of Y. For instance, if we know that X is small (relative to its mean), then we know that the mean of Y will also tend to be small (relative to its unconditional mean). If we know that the X is large, then we also know that the Y outcome will be large. If we do not observe X, then our view regarding the mean value of Y will have to cover all possible values of X, which is what the unconditional mean of Y does. The fact that X gives us information about Y is also reflected in the reduction in variance from $Var(Y) = 0.625$ to $Var(Y \mid X) = 0.125$. The inequality $Var(Y \mid X) \leq Var(Y)$ does not hold in general, but it does hold when $Var(Y \mid X)$ is constant.

For general continuous random variables, the conditional distributions can be computed as

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \text{ and } f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

when $f_Y(y) \neq 0$ and $f_X(x) \neq 0$. Another way of writing this is

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y).$$

This decomposition of joint pdfs can be extended to more than two variables, e.g., we have

$$f_{X,Y,Z}(x, y, z) = f_{Z|Y,X}(z \mid y, x) f_{Y|X}(y \mid x) f_X(x)\,.$$

3.1.4 Manipulating Conditional Moments

Manipulation of conditional expectations and variances follows one simple principle: whatever is being conditioned on can be treated as "fixed" (i.e., like a constant) as far as that expectation or variance is concerned.

Example 3.1.

- $E(aXY \mid X) = aXE(Y \mid X)\,,\ Var(aXY \mid X) = a^2X^2\,Var(Y \mid X)$,
- $E(aX \mid X) = aX$ (contrast with $E(aX) = aE(X)$, a constant),
- $Var(aX \mid X) = 0$ (contrast with $Var(aX) = a^2\,Var(X)$),
- If $Y = \beta_0 + \beta_1 X + \epsilon$ with $E(\epsilon \mid X) = 0$ and $Var(\epsilon \mid X) = \sigma^2$, then

$$E(Y \mid X) = \beta_0 + \beta_1 X \text{ and } Var(Y \mid X) = \sigma^2. \tag{3.4}$$

In linear regression analysis, we often begin with an assumption that the conditional expectation takes some form, such as (3.4), the objective being to estimate the coefficients β_0 and β_1.

3.1.5 The Law of Iterated Expectations

Recall that for the joint pdf in (3.1), we have $E(Y) = 4.5$, and

X	1	2	3	4	5
$E(Y \mid X)$	3.5	4	4.5	5	5.5
$\Pr(X)$	0.2	0.2	0.2	0.2	0.2

While $E(Y)$ is a single number, $E(Y \mid X)$ is a random variable when considered over all possible values of X. In our example, $E(Y \mid X)$ is a (uniformly distributed) random variable with possible values 3.5, 4.0, 4.5, 5.0, and 5.5. If we calculate the mean of this random variable, we get

$$E(E(Y \mid X)) = 3.5(0.2) + 4(0.2) + 4.5(0.2) + 5(0.2) + 5.5(0.2) = 4.5 = E(Y).$$

This equality is not a coincidence, but an example of the Law of Iterated Expectations:

$$E_X(E_{Y|X}(Y \mid X)) = E_Y(Y). \tag{3.5}$$

We add the subscript to the expectation notation in (3.5) to be clear as to the probabilities over which the expectations are taken, e.g., $E_{Y|X}$ indicates that the expectation is taken over the conditional probabilities of Y given X, whereas E_Y and E_X indicate that the expectations are taken under the marginal distributions of Y and X respectively. We often drop the subscripts for cleaner exposition.

The Law of Iterated Expectations says (roughly speaking) that we can get the 'overall' average of Y by taking the Y average for each possible value of X, and then taking the average of those averages. More generally, we have

$$E_{X,Y}(g(X,Y)) = E_X(E_{Y|X}(g(X,Y)) \mid X)\,.$$

If $g(X,Y) = Y$, we get the Law of Iterated Expectations as stated in (3.5).

We demonstrate two results implied by the Law of Iterated Expectations:

i. If $E(Y \mid X) = c$, then $E(Y) = c$,

ii. If $E(Y \mid X) = c$, then $Cov(X,Y) = 0$.

Result (i) says that if the expected value of Y is c for every possible value of X, then the "overall" mean must be that same constant, and (ii) says that $E(Y \mid X) = c$ is a sufficient condition for $Cov(X,Y) = 0$.

The derivation of these results is straightforward. For (i), if $E(Y \mid X) = c$, then

$$E(Y) = E(E(Y \mid X)) = E(c) = c.$$

For (ii), we note that

$$E(YX) = E(E(YX \mid X)) = E(XE(Y \mid X)) = E(cX) = cE(X)\,.$$

Therefore

$$Cov(X,Y) = E(XY) - E(X)E(Y) = cE(X) - cE(X) = 0.$$

Although constant conditional mean implies zero covariance, the converse does not necessarily hold. For instance, suppose X is zero mean and has a symmetric distribution (which together implies that $E(X^3) = 0$). Suppose $Y = X^2$. Then $E(Y \mid X) = X^2$ but

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(XE(Y \mid X)) - 0E(X) = E(X^3) = 0.$$

Two remarks:

- the Law of Iterated Expectations can be extended to more than two variables. For example, for random variables W, X and Y, we have

$$E(X \mid Y) = E(E(X \mid W, Y) \mid Y).$$

- we also have a **Law of Iterated Variance**, or **Law of Total Variance**:

$$Var(Y) = E(\,Var(Y \mid X)) + \,Var(E(Y \mid X)). \tag{3.6}$$

3.1.6 Independent Random Variables

Two random variables are said to be **independent** if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{3.7}$$

For discrete random variables, this means that

$$\Pr(Y = i, X = j) = \Pr(Y = i) \Pr(X = j)$$

for all possible values of Y and X. Independence of X and Y implies

$$f_{Y|X}(y \mid x) = f_Y(y) \text{ and } f_{X|Y}(x \mid y) = f_X(x).$$

Knowledge of the realization for one variable does not add any information regarding the probabilistic behavior of the other.

Independence implies $E(Y \mid X) = E(Y)$, $Var(Y \mid X) = Var(Y)$, and so on. Independence of X and Y implies zero covariance between the two random variables: if X and Y are independent, then

$$E(XY) = E(XE(Y \mid X)) = E(X)E(Y),$$

$$\text{therefore} \quad Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

However, zero covariance does not imply independence. Exercise 3.6 at the end of this section presents an example where X and Y have zero covariance, but where the variance of Y increases in X.

Bivariate *normal* random variables (see appendix) are an exception: if two random variables have a bivariate normal distribution and are uncorrelated, then they are also independent.

3.2 Chapter 3 Exercises A

Exercise 3.1. Starting from the definition $Cov(X,Y) = E((X - E(X))(Y - E(Y)))$ and using the properties of expectations, show that $Cov(X,Y) = E(XY) - E(X)E(Y)$.

Exercise 3.2. Show for example (3.1) that the correlation coefficient of X and Y is 0.8944.

Exercise 3.3. Show that

$$Cov(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2 + b_3Y_3) = \sum_{i=1}^{2}\sum_{j=1}^{3} a_ib_jCov(X_i, Y_j).$$

Exercise 3.4. Explain why the correlation coefficient always lies between -1 and 1, inclusive.

Hint: For arbitrary α, we have $Var(X - \alpha Y) \geq 0$. Expand $Var(X - \alpha Y)$ and let $\alpha = Cov(X,Y)/Var(Y)$

Exercise 3.5. For the joint pdf (3.1), find the conditional distribution of Y given $X \geq 3$, and the corresponding conditional mean and variance.

Exercise 3.6. Suppose Y and X have the following joint distribution function:

	10	0	0	0	0	0.1
	9	0	0	0	0.1	0
	8	0	0	0.1	0	0
	7	0	0.1	0	0	0
	6	0.1	0	0	0	0
Y	5	0.1	0	0	0	0
	4	0	0.1	0	0	0
	3	0	0	0.1	0	0
	2	0	0	0	0.1	0
	1	0	0	0	0	0.1
		1	2	3	4	5
				X		

i. Find the marginal distribution of X and of Y.

ii. Find the conditional distribution, conditional mean, and conditional variance of Y given X, and of X given Y.

iii. Find $Cov(X,Y)$.

iv. In what way is the conditional distribution of Y related to X?

Exercise 3.7. Show that if $E(Y \mid X) = a + bX$, then

$$b = \frac{Cov(X,Y)}{Var(X)} \quad \text{and} \quad a = E(Y) - bE(X)\ .$$

If you know that $E(Y \mid X) = 3 + 0.5X$ and $Var(X) = 2$, what is $Cov(X,Y)$?

Exercise 3.8. Prove (3.6). Use this relationship to show that

i. $Var(Y) = E(\,Var(Y \mid X))$ if $E(Y \mid X)$ is constant.

ii. $Var(Y \mid X) \leq Var(Y)$ if $Var(Y \mid X)$ is constant.

Exercise 3.9. Suppose Y and X have the following joint pdf:

	5	0.01	0.04	0.03	0.01	0.01
	4	0.02	0.08	0.06	0.02	0.02
Y	3	0.04	0.16	0.12	0.04	0.04
	2	0.02	0.08	0.06	0.02	0.02
	1	0.01	0.04	0.03	0.01	0.01
		1	2	3	4	5
				X		

Are the variables independent? Are they identically distributed (i.e., do they have the same marginal distributions?) Change the probabilities in the joint pdf of X and Y so that the two variables are independently and identically distributed (but not uniformly distributed).

3.3 Overview of Linear Regression

Suppose our population of interest is the set of all US non-institutional working civilians aged 16 or over in 2019. What you wish to learn about this population is the relationship between earnings and years of schooling: how much does each additional year of schooling contribute to average hourly earnings. You have a representative random (iid) sample from this population, stored in the file `earnings2019.csv`. Your sample size is $n = 4946$

```
options(width=100)
dat1 <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
dat1 <- dat1 %>%
  mutate(                                    # race variable is
    race_white = if_else(race=="White", 1, 0),   # "White", "Black", "Other"
    race_black = if_else(race=="Black", 1, 0),   # Convert to three dummy var,
    race_other = if_else(race=="Other", 1, 0)    # one for each race
  ) %>%                                      # then
  select(-race)                              # remove race variable
dat1 %>% summary()                           # and produce summary
```

```
      age            height           educ            feduc            meduc            tenure
 Min.   :19.00   Min.   :40.00   Min.   : 7.00   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
 1st Qu.:33.00   1st Qu.:64.00   1st Qu.:12.00   1st Qu.: 4.000   1st Qu.: 4.000   1st Qu.: 3.000
 Median :40.00   Median :67.00   Median :14.00   Median : 4.000   Median : 4.000   Median : 6.000
 Mean   :41.99   Mean   :67.45   Mean   :14.31   Mean   : 5.425   Mean   : 5.523   Mean   : 9.177
 3rd Qu.:51.00   3rd Qu.:70.00   3rd Qu.:16.00   3rd Qu.: 7.000   3rd Qu.: 7.000   3rd Qu.:13.000
 Max.   :82.00   Max.   :83.00   Max.   :17.00   Max.   :26.000   Max.   :26.000   Max.   :54.000
      wexp             male             earn            totalwork      race_white
 Min.   : 1.000   Min.   :0.0000   Min.   :  0.7428   Min.   :1000   Min.   :0.0000
 1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.: 15.5048   1st Qu.:1936   1st Qu.:0.0000
 Median : 7.000   Median :0.0000   Median : 22.9995   Median :2080   Median :1.0000
 Mean   : 9.251   Mean   :0.4646   Mean   : 29.2315   Mean   :2182   Mean   :0.5623
 3rd Qu.:13.000   3rd Qu.:1.0000   3rd Qu.: 35.0235   3rd Qu.:2428   3rd Qu.:1.0000
 Max.   :51.000   Max.   :1.0000   Max.   :628.9308   Max.   :5824   Max.   :1.0000
   race_black       race_other
 Min.   :0.000   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000
 Median :0.000   Median :0.0000
 Mean   :0.311   Mean   :0.1268
 3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :1.000   Max.   :1.0000
```

The following are scatter plots of `earn` against `educ`, and `log(earn)` against `educ`.

```
p1 <- ggplot(dat1, aes(y=earn, x=educ)) + geom_point(size=0.5) +
  scale_x_continuous(breaks=7:17) + theme_classic()
p2 <- ggplot(dat1, aes(y=log(earn), x=educ)) + geom_point(size=0.5) +
  scale_x_continuous(breaks=7:17) + theme_classic()
p1 | p2
```

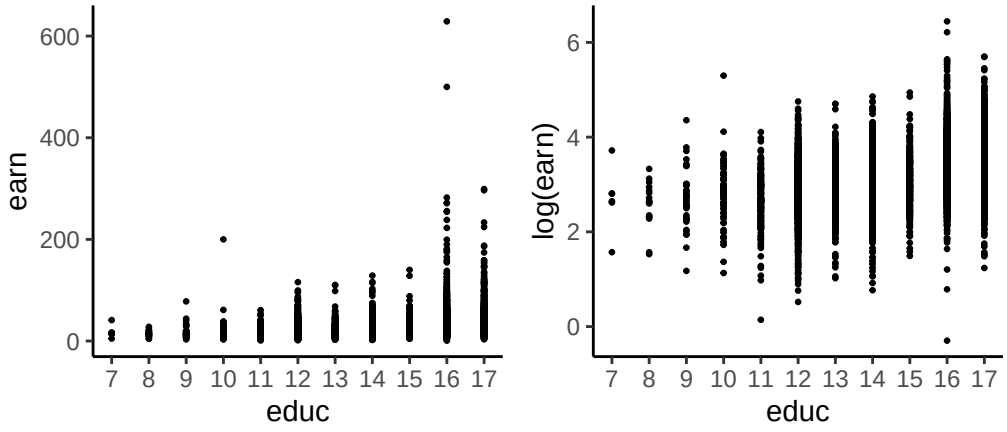


Figure 3.1: Scatterplots of earn and log(earn) against educ.

To be more specific, suppose we are interested in estimating the conditional expectation of `ln(earn)` given `educ`, which we assume to be linear, i.e., we want to estimate

$$E(\ln earn \mid educ) = \beta_0 + \beta_1 educ\,. \tag{3.8}$$

This means estimating the parameters β_0 and β_1.

Why do we specify the conditional expectation in **log-linear form** as in (3.8) rather than the fully linear form

$$E(earn \mid educ) = \beta_0 + \beta_1 educ? \tag{3.9}$$

For one thing, we can see from Fig. 3.1 that the log-linear form seems to better describe the data than the fully linear specification. Secondly, (3.9) implies that expected earnings go up by β_1 dollars for each additional year of `educ`, regardless of the level of schooling, i.e., increasing schooling from 7 to 8 leads to the same dollar increase in earnings as increasing schooling from 15 to 16 years. This seems unrealistic. On the other hand, the log-linear form says that expected earnings go up by $100\beta_1$ percent for each additional year of `educ`, which seems more realistic.

If we define

$$\epsilon = \ln earn - \beta_0 - \beta_1 educ\,, \tag{3.10}$$

then we can write

$$\ln earn = \beta_0 + \beta_1 educ + \epsilon\,. \tag{3.11}$$

We will refer to ϵ as the **noise term**, which has the property

$$\begin{aligned} E(\epsilon \mid educ) &= E(\ln earn - \beta_0 - \beta_1 educ \mid educ) \\ &= E(\ln earn \mid educ) - \beta_0 - \beta_1 educ = 0 \end{aligned} \tag{3.12}$$

as long as our specification of the conditional expectation (3.8) is correct. We can write our model of the population as

$$\ln earn = \beta_0 + \beta_1 educ + \epsilon\,,\ E(\epsilon \mid educ) = 0\,. \tag{3.13}$$

Equation (3.12) also implies $E(\epsilon) = 0$.

Let $\{Y_i, X_i\}_{i=1}^n$ represent your sample, where Y_i represents $\ln earn_i$, X_i represents $educ_i$, and $n = 4946$. Since your sample is representative of the population, we can write

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i\,,\ E(\epsilon_i \mid X_i) = 0 \ \text{ for } \ i = 1, \ldots, n\,. \tag{3.14}$$

If your sample is iid, we can extend (3.14) to

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ E(\epsilon_i \mid X_1, \ldots, X_n) &= 0 \\ E(\epsilon_i \epsilon_j \mid X_1, \ldots, X_n) &= 0 \ \text{ for } \ i, j = 1, \ldots, n\,,\ i \neq j\,. \end{aligned} \tag{3.15}$$

This is our **simple linear regression** model. Given estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1, we define the following terms:

- Fitted values: $\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$,
- Residuals: $\hat{\epsilon}_i = Y_i - \widehat{Y_i} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.

for all $i = 1, \ldots, n$. Of course, we have $Y_i = \widehat{Y_i} + \hat{\epsilon}_i$.

There are a number of ways to obtain estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Some approaches lead to the same estimators whereas others produce different estimators. The **method of moments approach** proceeds as follows: Since your sample is representative of the population and your population satisfies $E(\epsilon \mid X) = 0$, which implies

$$E(\epsilon) = 0 \ \text{ and } \ E(\epsilon X) = 0\,, \tag{3.16}$$

we can choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sample counterparts of (3.16) also hold, i.e., we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^{mm} &= 0 \\ \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^{mm} X_i &= 0 \end{aligned}$$

where $\hat{\epsilon}_i^{mm} = (Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_i)$.

That is

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0^{mm}-\hat{\beta}_1^{mm}X_i)&=0\\ \sum_{i=1}^{n}(Y_i-\hat{\beta}_0^{mm}-\hat{\beta}_1^{mm}X_i)X_i&=0\end{aligned} \tag{3.17}$$

This is a system of two equations which can be solved for the $\hat{\beta}_0^{mm}$ and $\hat{\beta}_1^{mm}$. We label the estimators with mm to indicate that they were obtained from the method of moments approach.

Another approach is to say that we want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the estimated regression line fits the data well in the sense of making the residual sum of squares as small as possible, i.e., choose

$$\hat{\beta}_0^{ols},\hat{\beta}_1^{ols}=\underset{\widehat{\beta}_0\,,\widehat{\beta}_1}{\arg\min}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2\,. \tag{3.18}$$

This is the **ordinary least squares (OLS)** approach. Elementary optimization theory says that $\hat{\beta}_0^{ols},\hat{\beta}_1^{ols}$ must satisfy the necessary first-order conditions

$$\begin{aligned}(1)\qquad \left.\frac{\partial RSS}{\partial\hat{\beta}_0}\right|_{\hat{\beta}_0^{ols},\,\hat{\beta}_1^{ols}}&=-2\sum_{i=1}^{n}(Y_i-\hat{\beta}_0^{ols}-\hat{\beta}_1^{ols}X_i)=0\\ (2)\qquad \left.\frac{\partial RSS}{\partial\hat{\beta}_1}\right|_{\hat{\beta}_0^{ols},\,\hat{\beta}_1^{ols}}&=-2\sum_{i=1}^{n}(Y_i-\hat{\beta}_0^{ols}-\hat{\beta}_1^{ols}X_i)X_i=0\end{aligned} \tag{3.19}$$

where RSS refers to the **residual sum of squares** $\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2$. Notice that the conditions (3.19) are the same as the method of moments conditions (3.17) so both approaches lead to the same estimator.

Yet another approach is the **least absolute deviation (LAD)** approach:

$$\hat{\beta}_0^{lad},\hat{\beta}_1^{lad}=\underset{\widehat{\beta}_0\,,\widehat{\beta}_1}{\arg\min}\sum_{i=1}^{n}\mid Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\mid. \tag{3.20}$$

This approach leads to a different set of estimators.

We shall focus on the OLS / MM approach. We will call it the "OLS" approach and label the estimators, fitted values and residuals with an OLS superscript.

3.3.1 OLS Formulas for the Simple Linear Regression Model

Solving (3.19) gives

$$\begin{aligned}\hat{\beta}_0^{ols}&=\overline{Y}-\hat{\beta}_1^{ols}\overline{X}\\ \hat{\beta}_1^{ols}&=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})X_i}{\sum_{i=1}^{n}(X_i-\overline{X})X_i}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\end{aligned} \tag{3.21}$$

The details are as follows:

$$(1)\ \Rightarrow\ \sum_{i=1}^{n}Y_i-n\hat{\beta}_0^{ols}-\hat{\beta}_1^{ols}\sum_{i=1}^{n}X_i=0\ \Rightarrow\ \overline{Y}-\hat{\beta}_0^{ols}-\hat{\beta}_1^{ols}\overline{X}=0\ \Rightarrow\ \hat{\beta}_0^{ols}=\overline{Y}-\hat{\beta}_1^{ols}\overline{X}$$

Substituting $\hat{\beta}_0^{ols}$ into (2) then gives

$$\sum_{i=1}^{n}(Y_i - (\overline{Y} - \hat{\beta}_1^{ols}\overline{X}) - \hat{\beta}_1^{ols}X_i)X_i = 0$$
$$\sum_{i=1}^{n}[(Y_i - \overline{Y}) - \hat{\beta}_1^{ols}(X_i - \overline{X})]X_i = 0$$
$$\sum_{i=1}^{n}(Y_i - \overline{Y})X_i - \hat{\beta}_1^{ols}\sum_{i=1}^{n}(X_i - \overline{X})X_i = 0 \quad \Rightarrow \quad \hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})X_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i}\,.$$

Since $\sum_{i=1}^{n}(Y_i - \overline{Y})X_i = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i$, we can also write $\hat{\beta}_1^{ols}$ as

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{3.22}$$

or even

$$\hat{\beta}_1^{ols} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{sample covariance}(X_i, Y_i)}{\text{sample variance}(X_i)}\,. \tag{3.23}$$

Ultimately, $\hat{\beta}_1^{ols}$ measures the sample covariance between X_i and Y_i, scaled by the sample variance of X_i.

From (3.23) it should be clear that we need

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 \neq 0\,,$$

i.e., we need the sample variance of X_i to be greater than zero, in order to get an OLS estimator for β_1. This makes sense. The parameter β_1 measures how Y_i varies with X_i in your data. To make this measurement you must have variation in your regressor X_i.

Notice also that we can write

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n}\left(\frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)Y_i = \sum_{i=1}^{n} w_i Y_i\,.$$

where $w_i = (X_i - \overline{X})/\sum_{i=1}^{n}(X_i - \overline{X})^2$. Any estimator that can be written as a weighted sum of Y_i is called a **linear estimator**.

Once you have $\hat{\beta}_1^{ols}$, you can obtain $\hat{\beta}_0^{ols}$ from the first equation in (3.21). The estimated model (the **sample regression line**) is

$$\hat{Y} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols}X\,. \tag{3.24}$$

The OLS fitted values are:

$$\hat{Y}_i^{ols} = \hat{\beta}_0^{ols} + \hat{\beta}_1^{ols}X_i\,,\ i = 1, \ldots, n\,. \tag{3.25}$$

The OLS residuals are

$$\hat{\epsilon}_i^{ols} = Y_i - \hat{Y}_i^{ols} = Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols}X_i\,,\ i = 1, \ldots, n\,. \tag{3.26}$$

3.4 Properties of the OLS Estimators

We shall focus on $\hat{\beta}_1$. Similar remarks hold for $\hat{\beta}_0$ but the details for $\hat{\beta}_0$ are best left for when we discuss the general multiple linear regression case.

3.4.1 Unbiasedness

First rewrite $\hat{\beta}_1^{ols}$ as

$$\begin{aligned}
\hat{\beta}_1^{ols} &= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\
&= \frac{\beta_0 \sum_{i=1}^n (X_i - \overline{X}) + \beta_1 \sum_{i=1}^n (X_i - \overline{X})X_i + \sum_{i=1}^n (X_i - \overline{X})\epsilon_i)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \\
&= \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\sum_{i=1}^n (X_i - \overline{X})X_i} \,.
\end{aligned} \tag{3.27}$$

Then

$$E(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) \;=\; \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})E(\epsilon_i \mid X_1, \ldots, X_n)}{\sum_{i=1}^n (X_i - \overline{X})X_i} \;=\; \beta_1 \,.$$

It follows that

$$E(\hat{\beta}_1^{ols}) = \beta_1 \,.$$

That is, $\hat{\beta}_1^{ols}$ is unbiased for β_1 . By following the OLS approach, you will not systematically over- or under-estimate β_1. Notice, however, that we made use of the assumption $E(\epsilon_i \mid X_1, \ldots, X_n) = 0$ which in turn depends on the assumption that we specified the conditional expectation correctly. There are other reasons that might cause $E(\epsilon_i \mid X_1, \ldots, X_n) = 0$ to fail to hold. We shall come to them in due course.

3.4.2 Consistency

The OLS estimator $\hat{\beta}_1^{ols}$ is also consistent for β_1. We present two rough arguments.

Rough argument 1:

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} \,. \tag{3.28}$$

Appealing to LLN:

- Numerator converges in probability to population $Cov(X, Y)$.
- Denominator converges in probability to population $Var(X)$.

$$\hat{\beta}_1^{ols} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} \xrightarrow{\;p\;} \frac{Cov(X, Y)}{Var(X)} = \beta_1 \,. \tag{3.29}$$

Rough argument 2:

$$\hat{\beta}_1^{ols} = \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}\,. \tag{3.30}$$

- Numerator in second term converges in probability to population $Cov(X, \epsilon)$.
- Denominator in second term converges in probability to population $Var(X)$.

If population $Cov(X, \epsilon) = 0$ and population $Var(X) \neq 0$, then

$$\hat{\beta}_1^{ols} \;=\; \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} \;\xrightarrow{p}\; \beta_1 + \frac{Cov(X, \epsilon)}{Var(X)} = \beta_1\,. \tag{3.31}$$

3.4.3 Standard Errors

Standard errors should be calculated for all estimators. For $\hat{\beta}_1^{ols}$, we have

$$\begin{aligned} \hat{\beta}_1^{ols} &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \overline{X})\epsilon_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ Var(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) &= \frac{\sum_{i=1}^n (X_i - \overline{X})^2\, Var(\epsilon_i \mid X_1, \ldots, X_n)}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2} \end{aligned} \tag{3.32}$$

If we are willing to assume, in addition to all the assumptions we have made so far, that

$$Var(\epsilon_i \mid X_1, \ldots, X_n) = \sigma^2 \tag{3.33}$$

then (3.32) simplifies to

$$\begin{aligned} Var(\hat{\beta}_1^{ols} \mid X_1, \ldots, X_n) &= \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \overbrace{Var(\epsilon_i \mid X_1, \ldots, X_n)}^{\sigma^2}}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2} \\ &= \frac{\sigma^2 \sum_{i=1}^n (X_i - \overline{X})^2}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\,. \end{aligned} \tag{3.34}$$

In order to get a numerical estimate for the variance, we have to get an estimate for σ^2. Nonetheless, (3.34) is informative. It says that the OLS estimator $\hat{\beta}_1$ has a smaller variance if (i) σ^2 is small (the data is less noisy), (ii) n is larger (since the denominator is a sum of n non-negative terms) and (iii) if there is more variation in your X_i sample.

The assumption (3.33) is called **homoskedasticity**. It will hold if $Var(\epsilon \mid X)$ is constant in population and you have a representative iid sample from the population. It says that the "noisiness" of the data does not depend on any of the X observations. If $Var(\epsilon_i \mid X_1, \ldots, X_n)$ is not constant, then we say there is **heteroskedasticity**. Fig. 3.2 shows data from the data set `heterosk.csv`. In (a), for a regression of Z on X, the error terms appear homoskedastic whereas in (b), for a regression of Y on X, the error terms appear to be heteroskedastic, in

particular, the variance appears to be increasing in X.

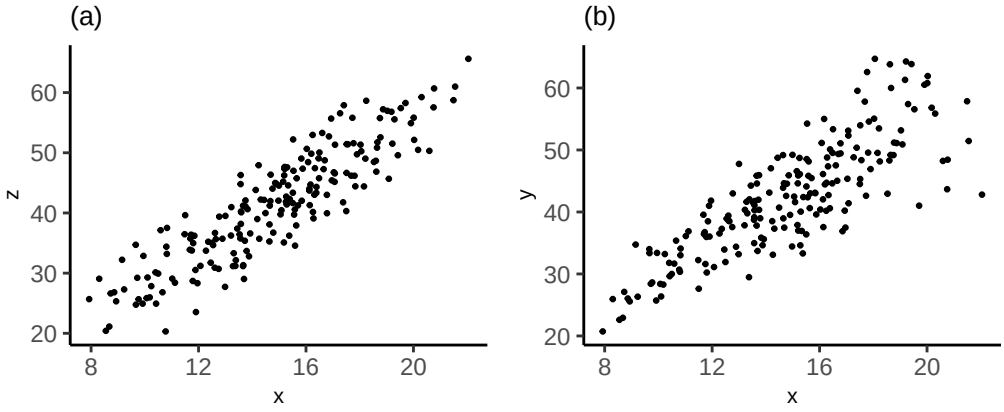


Figure 3.2: (a) Homoskedastic errors, (b) Heteroskedastic errors,

The formula (3.34) is valid only in the homoskedasticity case. In that case, it can be shown that an unbiased estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\epsilon}_{i,ols}^2 \,.$$

We define the standard error of $\hat{\beta}_1$ as

$$\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \,.$$

Furthermore, if $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, we have

$$\text{t-stat} = \frac{\hat{\beta}_1^{ols} - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \quad \sim \quad \text{t}(n-2) \,.$$

If not, we can appeal to the asymptotic result

$$t = \frac{\hat{\beta}_1^{ols} - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \quad \overset{a}{\sim} \quad \text{Normal}(0,1) \,.$$

The t-stat can be used to test hypotheses on β_1.

Example 3.2. For our returns to education example, using data in `earn2019.csv`, we have

```
slr <- function(y, x){
  n <- length(y)
  ybar <- mean(y)
  xbar <- mean(x)
  beta1hat <- sum((x-xbar)*y)/sum((x-xbar)*x)
  beta0hat <- ybar - beta1hat*xbar
  yhat <- beta0hat + beta1hat*x
  ehat <- y - yhat
  rss <- sum(ehat^2)
```

```
  beta1hat_se <- sqrt((rss/(n-2))/sum((x-xbar)^2))
  beta1hat_t <- beta1hat / beta1hat_se
  cat("beta0hat:", round(beta0hat,3), "\n")
  cat("beta1hat:", round(beta1hat,3),
      "  s.e.:", round(beta1hat_se,3),
      "  t-stat:", round(beta1hat_t,3),
      "  p-val:", round(2*pt(-abs(beta1hat_t), n-2),3))
  results <- list(beta0hat=beta0hat, beta1hat=beta1hat,
                  beta1hat_se=beta1hat_se, ehat=ehat, yhat=yhat)
}

cat("Regression of ln earn on educ, with intercept\n\n")
mdl1 <- slr(y=log(dat1$earn), x=dat1$educ)
```

```
Regression of ln earn on educ, with intercept

beta0hat: 1.32
beta1hat: 0.128   s.e.: 0.004   t-stat: 32.17   p-val: 0
```

We left out the standard error for $\hat{\beta}_0$ since we haven't yet developed the formula for it (we will do so in a later chapter). The t-statistic for $\hat{\beta}_1$ is calculated as $\hat{\beta}_1/\text{s.e.}(\hat{\beta}_1)$ and is meant for testing the hypothesis $\beta_1 = 0$. The associate p-value for this test is also given.

Instead of writing your own code, you can use R's built-in `lm()` function to estimate the regression:

```
mdl1a <- lm(log(earn)~educ, dat1)
cat("Dependent variable: ", as.character(formula(mdl1a))[2], "\n")
mdl1a %>% summary() %>% coef()
cat("n =", nobs(mdl1a))
```

```
Dependent variable:  log(earn)
             Estimate  Std. Error  t value      Pr(>|t|)
(Intercept) 1.3199894 0.057541299 22.93986 9.401431e-111
educ        0.1279965 0.003978753 32.17000 2.442343e-206
n = 4946
```

The data, the sample regression line and the histogram of the residuals are shown in Fig. 3.3.

```
regdat <- data.frame(educ=dat1$educ, earn=dat1$earn, fitted=mdl1$yhat, resid=mdl1$ehat)
p1 <- ggplot(data=regdat) +
        geom_point(aes(y=log(earn), x=educ), size=0.5) +
        geom_point(aes(y=fitted, x=educ), size=1.5, color='blue') +
        geom_abline(intercept=mdl1$beta0hat, slope=mdl1$beta1hat, color='red') +
        theme_classic()
p2 <- ggplot(regdat, aes(x = resid)) +
  geom_histogram(aes(y = after_stat(density)), fill = "lightblue", color="black", bins=30) +
  xlab("residuals") + theme_bw()
p1 | p2
```

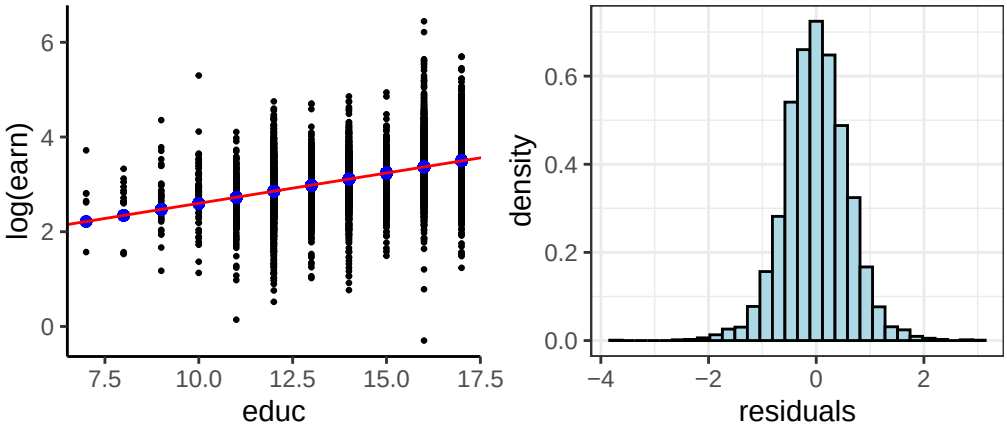


Figure 3.3: Regression of ln(earn) on educ.

What if you are unsure if your errors are homoskedastic? When deriving the variance of $\hat{\beta}_1$, we obtained (3.32), repeated below:

$$Var(\hat{\beta}_1^{ols} \mid X_1, \dots, X_n) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \, Var(\epsilon_i \mid X_1, \dots, X_n)}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2} \, .$$

If $Var(\epsilon_i \mid X_1, \dots, X_n)$ is not a constant, we cannot bring it out of the summation in the numerator, as we did in the derivation of the variance formula in the homoskedasticity case. Without specifying $Var(\epsilon_i \mid X_1, \dots, X_n)$ any further, we cannot proceed with the derivation. However, it can be shown that under quite general conditions, replacing $Var(\epsilon_i \mid X_1, \dots, X_n)$ with the squared residuals produces a consistent estimator for the variance of $\hat{\beta}_1$ even under heteroskedasticity, i.e.,

$$\widehat{Var}_{HC}(\hat{\beta}_1^{ols}) \;=\; \frac{\sum_{i=1}^n (X_i - \overline{X})^2 \, \hat{\epsilon}_{i,ols}^2}{\left(\sum_{i=1}^n (X_i - \overline{X})^2\right)^2} \;\xrightarrow{p}\; Var(\hat{\beta}_1^{ols}) \, .$$

This variance estimator is called the **heteroskedasticity-robust / heteroskedasticy-consistent / White standard errors**. Consistency holds despite the fact that $E(\hat{\epsilon}_{i,ols}^2)$ is in general a biased and inconsistent estimator for $Var(\epsilon_i)$. Furthermore, $\widehat{Var}_{HC}(\hat{\beta}_1^{ols})$ remains consistent even under homoskedasticity. Some researchers use this as their default variance estimator.

```
slr_hc0 <- function(y, x){
  n <- length(y)
  ybar <- mean(y)
  xbar <- mean(x)
  beta1hat <- sum((x-xbar)*y)/sum((x-xbar)*x)
  beta0hat <- ybar - beta1hat*xbar
  yhat <- beta0hat + beta1hat*x
  ehat <- y - yhat
  hc0 <- sum((x-xbar)^2*ehat^2)/sum((x-xbar)^2)^2
  beta1hat_se <- sqrt(hc0)
  beta1hat_t <- beta1hat / beta1hat_se
```

```
  cat("beta0hat:", round(beta0hat,3), "\n")
  cat("beta1hat:", round(beta1hat,3),
      "  s.e. (het. robust):", round(beta1hat_se,4),
      "  t-stat:", round(beta1hat_t,3),
      "  p-val:", round(2*pt(-abs(beta1hat_t), n-2),4))
  results <- list(beta0hat=beta0hat, beta1hat=beta1hat,
                  beta1hat_se=beta1hat_se, ehat=ehat, yhat=yhat)
}
mdl2 <- slr_hc0(y=log(dat1$earn), x=dat1$educ)
```

```
beta0hat: 1.32
beta1hat: 0.128   s.e. (het. robust): 0.0041   t-stat: 31.584   p-val: 0
```

Alternatively, you can use the `coeftest()` function from the `lmtest` package together with the `vcovHC` function from the `sandwich` package in the following way. The reason for the `sandwich` package name will be explained in a later chapter.

```
library(lmtest)     # lmtest::coeftest to calculate t-test
library(sandwich)   # using robust s.e. calculated with sandwich::vcovHC
coeftest(mdl1a, vcov=vcovHC, type="HC0")  # Robust standard errors
#NB: mdl1a was estimated earlier using the lm() function
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.3199894  0.0581594  22.696 < 2.2e-16 ***
educ        0.1279965  0.0040525  31.584 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The OLS coefficient estimates are the same, of course. We are only re-calculating the standard errors. As it turns out, the standard errors are very similar, suggesting that heteroskedasticity is not much of an issue here.

3.4.4 Using the Estimated Regression Model

3.4.4.1 Prediction

We can use our estimated model

$$\widehat{\ln earn} = \underset{(0.0582)}{1.32} + \underset{(0.0041)}{0.128}\, educ$$

for predictive purposes. Imagine going back to the population and sampling one more individual with, say, 12 years of `educ`. What is the model's prediction for this individual's `ln(earn)`? The model predicts that this individual's average hourly earnings `ln(earn)` to be

```
cat("Predicted ln ave. hr. earn for individual with 12 years educ is ")
cat(round(mdl1$beta0hat + mdl1$beta1hat*12,2), "\"log-dollars\"")
```

```
Predicted ln ave. hr. earn for individual with 12 years educ is 2.86 "log-dollars"
```

Of course, the scatterplot in Fig. 3.1(b) suggests that the potential error of this prediction might be quite high, the reason being that there are many factors that affect earnings, and these have not been accounted for in the model. Since the parameter estimates are unbiased, we can

expect the prediction to also be unbiased (to be correct on average) but the *prediction standard error* (which we will consider later in the course) would also be quite high.

The model gives predictions for `ln earn`. To convert this to `earn` you can take exponentials. In view of the comments at the end of the previous chapter, you might view this as a "conditional median prediction" rather than a "conditional expectation prediction".

```
cat("Predicted ave. hr. earn for individual with 12 years educ is ")
cat(round(exp(mdl1$beta0hat + mdl1$beta1hat*12),2), "dollars (cond. median pred.).")
```

```
Predicted ave. hr. earn for individual with 12 years educ is 17.39 dollars (cond. median pred.).
```

3.4.4.2 Measuring direct causal effects?

The model that we just estimated also predicts that an individual with one more year in `educ` will have 12.8% higher average hourly earnings. Can we take this to mean that every additional year of schooling directly "causes" average hourly earnings to go up by 12.8%? One has to be very careful when interpreting regression results from a causal perspective. The problem again is that there are many factors that affect earnings, and your regression is taking into account only the effect of `educ`. All of the other factors' influence on `ln earn` may end up being attributed to `educ`.

For instance, suppose there are two factors X and Z that affect Y, and that the conditional expectation of Y given X, and of Y given X and Z are

$$\begin{aligned} E(Y \mid X) &= \alpha_0 + \alpha_1 X\,, \\ E(Y \mid X, Z) &= \beta_0 + \beta_1 X + \beta_2 Z\,. \end{aligned} \tag{3.35}$$

If your intention is to capture the direct effect of X on Y, then you would want to estimate β_1, not α_1. The parameter β_1 tells you how the expected value of Y differs with X for two individuals in your population with the same Z:

$$E(Y|X = x_0 + 1, Z = z_0) - E(Y|X = x_0, Z = z_0) = \beta_1\,.$$

Since Z is unchanged, the difference in the expected value of Y must be due the difference in X. We say that β_1 measures the effect of X on Y **controlling** for Z.

A simple linear regression of Y on X, however, will give you an unbiased estimator for α_1, not β_1, and it is straightforward to show that if $\beta_2 \neq 0$ and $Cov(X, Z) \neq 0$, then $\alpha_1 \neq \beta_1$. We have

$$\begin{aligned} E_{Y|X}(Y \mid X) &= E_{Z|X}(E_{Y|X,Z}(Y \mid X, Z)) \\ &= E_{Z|X}(\beta_0 + \beta_1 X + \beta_2 Z) \\ &= \beta_0 + \beta_1 X + \beta_2 E_{Z|X}(Z \mid X)\,. \end{aligned} \tag{3.36}$$

Suppose that $E_{Z|X}(Z \mid X) = \delta_0 + \delta_1 X$. Then (3.36) becomes

$$\begin{aligned} E(Y \mid X) &= \beta_0 + \beta_1 X + \beta_2(\delta_0 + \delta_1 X) \\ &= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)X. \end{aligned} \tag{3.37}$$

From Exercise 3.7, we know that if $E_{Z|X}(Z \mid X) = \delta_0 + \delta_1 X$, then

$$\delta_0 = E(Z) - \delta_1 E(X) \text{ and } \delta_1 = \frac{Cov(X,Z)}{Var(X)}. \tag{3.38}$$

Substituting these into (3.37) gives

$$E(Y \mid X) = (\beta_0 + \beta_2\delta_0) + \left(\beta_1 + \beta_2 \frac{Cov(X,Z)}{Var(Z)}\right) X. \tag{3.39}$$

Comparing (3.39) with $E(Y \mid X) = \alpha_0 + \alpha_1 X$, we see that

$$\alpha_1 = \beta_1 + \beta_2 \frac{Cov(X,Z)}{Var(Z)} \tag{3.40}$$

so $\alpha_1 \neq \beta_1$ unless $\beta_2 = 0$ or $Cov(X,Z) = 0$. A simple linear regression of Y on X gives you an unbiased estimator for α_1, but a biased estimator for β_1.

Here is another way of looking at this issue. Suppose we have

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

and define $u = Y - \beta_0 - \beta_1 X - \beta_2 Z$. Then we can write

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u\,,\ E(u \mid X, Z) = 0\,. \tag{3.41}$$

Suppose X and Z are correlated. Then if we write

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3.42}$$

where the β_0 and β_1 in (3.42) are the same as the β_0 and β_1 in (3.41), then

$$\epsilon = \beta_2 Z + u\,.$$

If X and Z are correlated, then ϵ and X must be correlated, so we no longer have $E(\epsilon \mid X) = 0$. This causes the OLS estimator for β_1 in the simple linear regression of Y on X to be biased for β_1.

Example 3.3. The file `multireg_eg.csv` contains observations on variables X, Y and Z. The variable Z takes on integer values 1 to 5 only. Fig. 3.4(a) shows a scatterplot of Y against X. Fig. 3.4(b) shows the same scatterplot, but this time the value of Z corresponding to each (X,Y) pair is indicated.

The scatterplot in Fig. 3.4(a) shows a strong negative relationship between X and Y. Yet Fig. 3.4(b) shows that for any fixed level of Z, the effect of X on Y is actually positive. What is going on here is that the variable Z, which is negatively correlated with X (the observations drift left as Z goes from 1 to 5), also affects Y positively, for any fixed X. Therefore an increase in X, which tends to increase Y, also tends to be accompanied by a fall in Z, which leads to a fall in Y. The latter effect is strong enough that higher values of X become associated with lower values of Y. We refer to Z as a confounding variable.

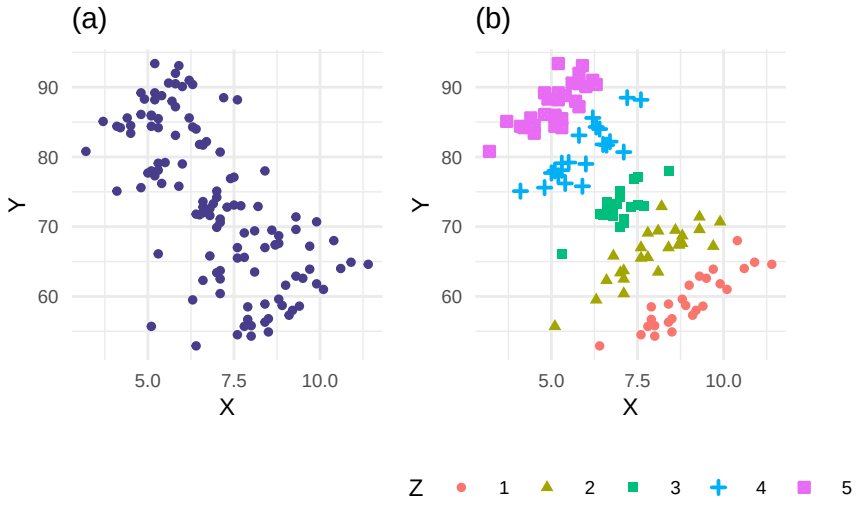


Figure 3.4: Illustrating confounding factors.

Imagine, for instance, that Y represents final exam scores in a certain course, Z is the background preparedness of the students ranging from very poor (1) to excellent (5), and X is study hours per week. Given any level of preparedness, studying clearly increases test scores. Background preparedness is also positively correlated with final exam scores. Because students who have strong background preparedness spend less time studying for the course (they probably allocate study time to the courses that they are less prepared for), when we look at test scores vs study time, it will appear that studying reduces test scores.

In this example, clearly α_1 is negative in

$$E(Y \mid X) = \alpha_0 + \alpha_1 X$$

and a simple linear regression of Y on X produces

```
lm(Y~X, data=df) %>% summary() %>% coef()
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 102.86406  3.2199095 31.94626 6.373574e-60
X            -4.23213  0.4474755 -9.45779 3.857133e-16
```

The separate effects of X and Z on Y are captured in the conditional expectation

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z\,.$$

In this example, β_1 is positive since for any fixed Z, the relationship between Y and X is positive. How might we get estimates of these separate effects? Looking ahead to the next chapter, we see that a *multiple linear regression* of Y on X and Z manages to disentangle the effects

```
lm(Y~X+Z, data=df) %>% summary() %>% coef()
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 21.188337  2.0816539 10.17861 8.212834e-18
X            3.114465  0.2054364 15.16024 2.219609e-29
Z           10.109717  0.2379845 42.48057 5.955647e-73
```

We shall explore multiple linear regression in more detail in the next chapter — why it works in disentangling causal effects, when it works well, when it does not, and what are the trade-offs.

3.5 Chapter 3 Exercise B

Exercise 3.10. Let $\hat{\beta}_0^{ols}$ and $\hat{\beta}_1^{ols}$ be the OLS estimators for β_0 and β_1 in the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

estimated on the sample $\{X_i, Y_i\}_{i=1}^n$. Let $\hat{\epsilon}_i^{ols}$ and $\widehat{Y}_i^{ols}$, $i = 1, \dots, n$, be the OLS residuals and fitted values.

(a) Explain why the OLS residuals have sample mean equal to zero, and that the OLS residuals and regressors X_i are **orthogonal** (meaning that $\sum_{i=1}^n X_i \hat{\epsilon}_i^{ols} = 0$).

(b) Show that $\overline{Y} = \overline{\widehat{Y}}$ where $\overline{\widehat{Y}}$ is the sample mean of the fitted values.

(c) Explain why the OLS residuals and regressors X_i are uncorrelated.

(d) Explain why the OLS residuals and OLS fitted values are uncorrelated.

(e) Show that the estimated regression line always passes through the point $(\overline{X}, \overline{Y})$.

(f) When will $\hat{\beta}_1^{ols} = 0$? What will $\hat{\beta}_0^{ols}$ be in this case?

(g) If $\overline{X} = 0$, show that $\hat{\beta}_1^{ols}$ reduces to

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\,.$$

What will $\hat{\beta}_0^{ols}$ be in this case?

Exercise 3.11. The simple linear regression *without intercept* is the model

$$Y_i = \beta_1 X_i + \epsilon_i \, i = 1, \dots, n\,.$$

(a) Derive the OLS estimator for β_1.
(b) Explain why the OLS residuals now *need not* have sample mean equal to zero.
(c) Explain why the OLS residuals and regressors remain orthogonal, but are no longer necessarily uncorrelated.
(d) For OLS estimation of the linear regression model *with* intercept, we required that there be some variation in your regressors, i.e., we cannot have $X_1 = \cdots = X_n = c$ for some fixed value c. Explain why OLS estimation of the linear regression model without intercept **is still possible** even if there is no variation in X_i.

Exercise 3.12. For the simple linear regression with intercept, estimated by OLS on the dataset $\{X_i, Y_i\}_{i=1}^n$, define the OLS fitted values $\{\widehat{Y}_i\}_{i=1}^n$ and residuals $\{\hat{\epsilon}_i\}_{i=1}^n$ in the usual way, so that

$$Y_i = \widehat{Y}_i + \hat{\epsilon}_i\,,\ i = 1, \dots, n\,.$$

To simplify notation, we drop the "ols" subscripts from the residuals and fitted values.
(a) Show that $\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \widehat{Y}_i^2 + \sum_{i=1}^n \hat{\epsilon}_i^2$.
(b) Show that

$$\underbrace{\sum_{i=1}^n (Y_i - \overline{Y})^2}_{TSS} \;=\; \underbrace{\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2}_{FSS} \;+\; \underbrace{\sum_{i=1}^n \hat{\epsilon}_i^2}_{RSS}\,. \tag{3.43}$$

This equation can be described as "Total Sum of Squares = Fitted Sum of Squares + Residual Sum of Squares", or $TSS = FSS + RSS$.[1] Dividing (3.43) throughout by $n-1$, the equation can also be described as a variance decomposition, specifically:

$$\text{sample var.}(Y_i) = \text{sample var.}(\widehat{Y}_i) + \text{sample var.}(\hat{\epsilon}_i) \tag{3.44}$$

(c) The R^2 is defined as

$$R^2 = 1 - \frac{RSS}{TSS}\,.$$

It is used as a measure of **goodness-of-fit**, since $RSS = 0$ when you have a perfect fit. Where you don't have a perfect fit, $RSS > 0$ and therefore $R^2 < 1$.

i. When will $R^2 = 0$? Can R^2 fall below zero?

ii. What is R^2 for the case where $Y_1 = \cdots = Y_n = c$ for some constant c?

3.6 Appendix: The Bivariate Normal Distribution

Two random variables X and Y have the bivariate normal distribution if their joint pdf have the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left\{-\frac{1}{2}\frac{\tilde{x}^2 - 2\rho_{xy}\tilde{x}\tilde{y} + \tilde{y}^2}{1-\rho_{xy}^2}\right\} \tag{3.45}$$

where $\tilde{x} = \dfrac{x-\mu_x}{\sigma_x}$ and $\tilde{y} = \dfrac{y-\mu_y}{\sigma_y}$. The bivariate normal distribution has five parameters, with the following interpretation:

- μ_x : unconditional mean of X,
- μ_y : unconditional mean of Y,
- σ_x^2 : unconditional variance of X,
- σ_y^2 : unconditional variance of Y, and
- ρ_{xy} : correlation coefficient of X and Y, i.e., $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$ where $\sigma_{xy} = Cov(X,Y)$.

We write

$$(X,Y) \sim \text{Normal}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})\,.$$

Contour plots are helpful for visualizing bivariate normal distributions. We show the contour plots of a bivariate normal distribution with

$$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}) = (1, 0, 1, 2, 0.9).$$

in Fig. 3.5(a). Alternatively, one can look at the 3D plot of the bivariate pdf in Fig. 3.5(b).

[1] Sometimes you will see $\sum_{i=1}^n \hat{\epsilon}_i^2$ labeled as "ESS" or "SSE" standing for "Error Sum of Squares" and $\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2$ labeled as "RSS" or "SSR" for "Regression Sum of Squares". Yet others call $\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2$ the "Explained Sum of Squares" and label it "ESS" or "SSE". You can see the potential for confusion. I prefer to call $\hat{\epsilon}_i$ "residuals" rather than "errors", and although "Fitted Sum of Squares" does not appear to be in common use, it seems appropriate for $\sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2$ and unambiguous. Furthermore, in view of the relationship in (a), the relationship in (b) is more appropriately called "Total Sum of Squares, Centered = Fitted Sum of Squares, Centered + Residual Sum of Squares", but we will stick with TSS = FSS + RSS.

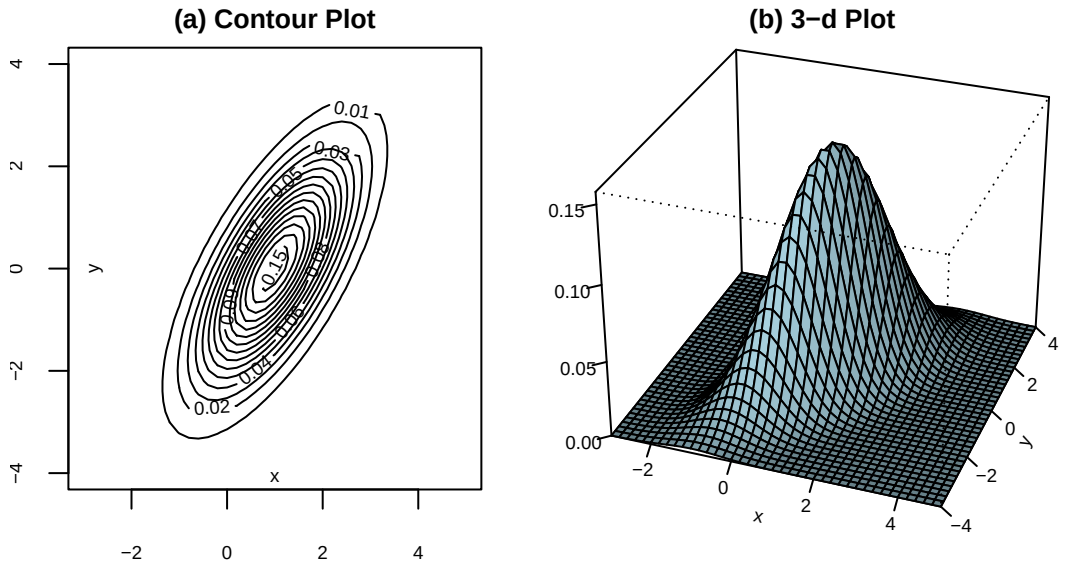


Figure 3.5: Bivariate normal distribution (parameter values given in text).

The marginal and conditional distributions of a bivariate normal random variables are also normal. To see this, we "complete the square" on $\tilde{x}^2 - 2\rho_{xy}\tilde{x}\tilde{y} + \tilde{y}^2$ to get

$$\begin{aligned}
\tilde{x}^2 - 2\rho_{xy}\tilde{x}\tilde{y} + \tilde{y}^2 &= (\tilde{x} - \rho_{xy}\tilde{y})^2 + (1-\rho_{xy}^2)\tilde{y}^2 \\
&= \left[\frac{x-\mu_x}{\sigma_x} - \frac{\sigma_{xy}}{\sigma_x\sigma_y}\frac{y-\mu_y}{\sigma_y}\right]^2 + (1-\rho_{xy}^2)\left(\frac{y-\mu_y}{\sigma_y}\right)^2 \\
&= \frac{1}{\sigma_x^2}\left[x - \mu_x - \frac{\sigma_{xy}}{\sigma_y^2}(y-\mu_y)\right]^2 + (1-\rho_{xy}^2)\left(\frac{y-\mu_y}{\sigma_y}\right)^2 \\
&= \frac{1}{\sigma_x^2}\left[x - (\alpha+\beta y)\right]^2 + (1-\rho_{xy}^2)\left(\frac{y-\mu_y}{\sigma_y}\right)^2
\end{aligned}$$

where $\alpha = \mu_x - \beta\mu_y$ and $\beta = \dfrac{\sigma_{xy}^2}{\sigma_y^2}$. Then the pdf can be written as

$$\begin{aligned}
&f_{X,Y}(x,y) \\
&= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left\{-\frac{1}{2}\frac{1}{1-\rho_{xy}^2}(\tilde{x}^2 - 2\rho_{xy}\tilde{x}\tilde{y} + \tilde{y}^2)\right\} \\
&= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \times \exp\left\{-\frac{1}{2}\frac{1}{1-\rho_{xy}^2}\left[\frac{1}{\sigma_x^2}\left[x-(\alpha+\beta y)\right]^2 + (1-\rho_{xy}^2)\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\} \\
&= \underbrace{\frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2(1-\rho_{xy}^2)}} \exp\left\{-\frac{1}{2}\frac{\left[x-(\alpha+\beta y)\right]^2}{\sigma_x^2(1-\rho_{xy}^2)}\right\}}_{A} \times \underbrace{\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}}_{B}
\end{aligned}$$

If we compare expressions A and B with the expression for the normal pdf, we see that B is a normal pdf $f_Y(y)$ with mean μ_y and variance σ_y, and if we take y as fixed, then A is a (conditional) normal pdf $f_{X|Y}(x \mid y)$ with mean $\alpha + \beta y$ and variance $\sigma_x^2 - \sigma_{xy}^2/\sigma_y^2$. That is, if X and Y have the bivariate normal distribution (3.45), then

- the marginal distribution of Y is Normal(μ_y, σ_y^2),
- the conditional distribution of X given Y is Normal$(\mu_{x|y}, \sigma_{x|y}^2)$ where

$$\mu_{x|y} = \mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y) \quad \text{and} \quad \sigma_{x|y}^2 = \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}.$$

Notice that the condition mean of X given Y is linear in Y. It can be written as

$$\mu_{x|y} = \alpha_x + \beta_x y$$

where $\alpha_x = \mu_x - \beta_x \mu_y$ and $\beta_x = \dfrac{\sigma_{xy}}{\sigma_y^2}$. Similarly,

- the marginal distribution of X is Normal$(\mu_x\, \sigma_x^2)$,
- the conditional distribution of Y given X is Normal$(\mu_{y|x}, \sigma_{y|x}^2)$ where

$$\mu_{y|x} = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x) \quad \text{and} \quad \sigma_{y|x} = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}.$$

The conditional mean can be written as

$$\mu_{y|x} = \alpha_y + \beta_y x$$

where $\alpha_y = \mu_y - \beta_y \mu_x$ and $\beta_y = \dfrac{\sigma_{xy}}{\sigma_x^2}$.

Setting $\rho_{xy} = 0$ causes expression A above to collapse to the marginal distribution of x. Then we have

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

It follows immediately that if X and Y are bivariate normal and uncorrelated, then they are independent random variables. It can also be shown that if X and Y are bivariate normal, then

$$aX + bY \sim \text{Normal}(a\mu_x + b\mu_y\,,\, a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_{xy})\,.$$

We omit the proof of this last result. (Of course, the expressions for the mean and variance of $aX + bY$ hold for all random variables, but the fact that a linear combination of normals is also normal is not immediately obvious).

Chapter 4
Multiple Linear Regression

We use the following packages in this chapter.

```
library(tidyverse); library(patchwork); library(readxl); library(car)
```

4.1 Introduction

We begin with a few examples to motivate the multiple linear regression model, before getting into the details.

Example 4.1. Recall Example 3.3 where we had data on three variables X, Y and Z (in dataset `multireg_eg.csv`) where Y and X were very strongly negatively correlated, but for fixed values of Z, an increase in X is actually associated with an *increase* in Y, as illustrated in Fig. 3.4, recreated below as Fig. 4.1.

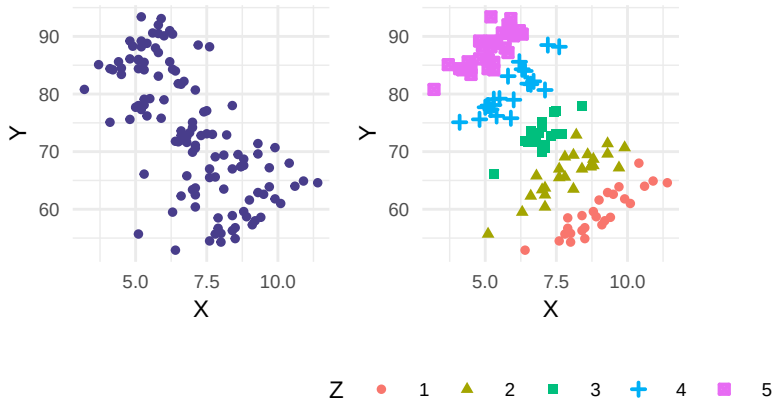


Figure 4.1: Illustrating confounding factors

We showed that the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i \,,$$

estimated using OLS, manages to disentangle the effects of X and Z on Y. We recreate the regressions from Example 3.3 below, this time including the R^2 measure of goodness-of-fit (see Exercise 3.12).

```
dat1 <- read_csv("data\\multireg_eg.csv",col_types = c("n","n","n"))
dat1_lm1_sum <- lm(Y~X, data=dat1) %>% summary()
dat1_lm2_sum <- lm(Y~X+Z, data=dat1) %>% summary()
cat("Dependent variable: Y\n\n")
dat1_lm1_sum %>% coefficients %>% round(4);
cat("R-squared:", round(dat1_lm1_sum$r.squared, 3), "\n\n")
dat1_lm2_sum %>% coefficients %>% round(4);
cat("R-squared:", round(dat1_lm2_sum$r.squared, 3))
```

```
Dependent variable: Y

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.8641     3.2199 31.9463        0
X            -4.2321     0.4475 -9.4578        0
R-squared: 0.431

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.1883     2.0817 10.1786        0
X             3.1145     0.2054 15.1602        0
Z            10.1097     0.2380 42.4806        0
R-squared: 0.965
```

How is it that multiple linear regression model (and the OLS estimation of it) manages to disentangle the effects of X and Z on Y? Besides disentangling the effects of multiple factors, notice that the standard error on X also falls when Z is added to the regression. Is this always the case? We see that the R^2 goes up. Will this always happen?

Example 4.2. Regressing `ln earn` on `height` using data in `earnings2019.csv` produces the following result:

```
dat2<-read_csv("data\\earnings2019.csv",show_col_types=FALSE) %>%
  mutate(white=if_else(race=="White", 1,0),
         black=if_else(race=="Black", 1,0),
         other=if_else(race=="Other", 1,0)) %>% select(-race)
dat2_lm1_sum <- summary(lm(log(earn)~height, data=dat2))
dat2_lm1_sum %>% coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2382     0.1536  8.0617        0
height        0.0284     0.0023 12.4766        0
```

```
cat("R-squared:", round(dat2_lm1_sum$r.squared, 3))
```

```
R-squared: 0.031
```

An increase in height of one inch is associated with 2.84 percent increase in average hourly earnings, and the result is statistically significant. Of course, this must surely be another case of omitted factors. In particular, including `male` changes the estimated relationship between `ln earn` and `height`:

```
dat2_lm2_sum <- summary(lm(log(earn)~height+male, data=dat2))
dat2_lm2_sum %>% coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8140     0.2003  9.0568        0
height        0.0191     0.0031  6.1896        0
male          0.1109     0.0248  4.4668        0
```

```
cat("R-squared:", round(dat2_lm2_sum$r.squared, 3))
```

```
R-squared: 0.034
```

Including `male` (1 if individual is male, 0 if female) reduces the coefficient on `height` to 0.0191, suggesting that at least part of the influence of `height` on `ln earn` can be explained as a *gender gap*, the effect of which in the previous regression was attributed to height, since men are taller than women on average. Again, how does OLS estimation of the multiple linear regression

manage to disentangle the effects of `height` and `male`. The standard error on `height` goes up with `male` is added. Why does this happen when in the previous example, the standard error fell when a confounding factor was included?

Is it possible for the coefficient estimate and standard error on one variable to remain essentially unchanged when an important explanatory factor is added? Consider adding `age` instead of `male` to the `height` equation. We get

```
dat2_lm3_sum <- summary(lm(log(earn)~height+age, data=dat2))
dat2_lm3_sum %>% coefficients %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9072     0.1557  5.8253        0
height        0.0286     0.0023 12.7034        0
age           0.0075     0.0008  9.8971        0
```

```
cat("R-squared:", round(dat2_lm3_sum$r.squared, 3))
```

```
R-squared: 0.049
```

Adding `age`, which is a significant factor, leaves the coefficient on `height` practically unchanged. When does adding a factor change previous results and when does it not?

Example 4.3. Even if the objective is to estimate the conditional expectation of Y on a single factor only, the multiple linear regression model can provide additional flexibility in functional form. In the previous chapter, we specified the conditional expectation of $\ln earn$ given $educ$ as

$$E(\ln educ \mid educ) = \beta_0 + \beta_1 educ\,.$$

For the data in `earnings2019.csv`, one might argue that a non-linear specification might be better, especially at lower levels of `educ`, see Fig. 4.2(a). Using the multiple linear regression framework, we can allow for non-linear specifications of the conditional expectation, such as:

$$E(\ln educ \mid educ) = \beta_0 + \beta_1 educ + \beta_2 educ^2\,.$$

Regressing $\ln earn$ on $educ$ and $educ^2$ does produce a more satisfying fit.

```
ggplot(dat2, aes(y=log(earn),x=educ)) + geom_point(size=0.5) + scale_x_continuous(breaks=7:17) +
  geom_smooth(formula=y~x+I(x^2), se=FALSE, method="lm") + theme_classic()
```

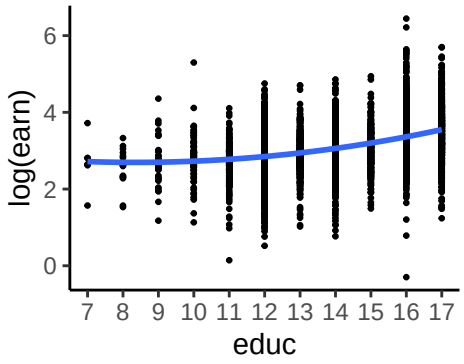


Figure 4.2: ln earn vs educ, with quadratic regression line.

You can also have interaction effects, such as

$$\ln educ = \beta_0 + \beta_1 male + \beta_2 educ + \beta_3 educ.male + \epsilon\,.$$

When $male = 0$ the equation becomes

$$\ln educ = \beta_0 + \beta_2 educ + \epsilon$$

whereas when $male = 1$, we have

$$\ln educ = \beta_0 + \beta_1 + (\beta_2 + \beta_3) educ + \epsilon$$

thereby allowing the returns to education parameter (the coefficient on $educ$) to depend on the sex of the individual. You might be interested in testing whether $\beta_3 = 0$ or $\beta_1 = \beta_3 = 0$. For our dataset, we have:[1]

```
dat2_lm4_sum <- summary(lm(log(earn)~educ*male, data=dat2)) ## Note the formula and the output!
dat2_lm4_sum %>% coefficients %>% round(4)
cat("R-squared:", round(dat2_lm4_sum$r.squared, 3))
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.0809 11.9219   0.0000
educ          0.1438     0.0055 26.0461   0.0000
male          0.5353     0.1125  4.7586   0.0000
educ:male    -0.0183     0.0078 -2.3510   0.0188
R-squared: 0.217
```

It seems that the returns to education is a little bit lower for men than for women, whereas there is a large gender gap even after controlling for years of education. Males with, say, 10 years of education earn $53.53 - 1.83(10) = 35.23\%$ more than women with the same number of years of education. Of course, there are certainly more factors that need to be controlled for.

In this chapter, we focus on ordinary least squares (OLS) estimation of the multiple linear regression with two regressors:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon.$$

We focus on the two regressor case in the first instance in order to build intuition regarding issues such as bias-variance tradeoffs, how the inclusion of an additional variable helps to "control" for the confounding effect of that variable, and basic ideas about joint hypotheses testing. We continue to assume that you are working with cross-sectional data. Be reminded that the variables Y, X and Z may be transformations of the variables of interest. Furthermore, X and Z may be transformations of the *same* variable, e.g., we may have $Z = X^2$.

[1] Here's how formulas work in `R`: (a) Specifying `log(earn) ~ educ:male` estimates a regression of $\ln earn$ on $educ \times male$ only (and an intercept term). (b) Specifying `log(earn) ~ educ*male` estimates a regression of $\ln earn$ on $educ$, $male$ and $educ : male$. (c) To regress $\ln earn$ on $educ$ and $educ : male$ only, use the formula `log(earn) ~ educ + educ:male` or `log(earn) ~ educ + I(educ*male)`. The `I(.)` function is the "as-is" function, i.e., treating what it is applied to as a new variable, e.g., `I(educ*male)` creates the new variable $educ \times male$. (d) Specifying `log(earn) ~ (educ + male + age)^2` estimates a regression of $\ln earn$ on $educ$, $male$, age and pairwise interactions $educ \times male$, $educ \times age$, $male \times age$. Note that squared terms are not included. (e) To remove intercept term, include `-1` in your formula, e.g., `log(earn) ~ educ - 1`.

4.2 OLS Estimation of the Multiple Linear Regression Model

Let $\{Y_i, X_i, Z_i\}_{i=1}^n$ be your sample. For any estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ (whether or not obtained by OLS), define the **fitted values** to be

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \tag{4.1}$$

and the **residuals** to be

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i \tag{4.2}$$

for $i = 1, 2, ..., n$. The OLS method chooses $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ to be those values that minimize the **residual sum of squares** $RSS = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i)^2$, i.e.,

$$\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols} = \text{argmin}_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i)^2. \tag{4.3}$$

The phrase "$\text{argmin}_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2}$" means "the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ that minimize …". The OLS estimators can be found by solving the first order conditions for the minimization problem:

$$\begin{aligned}
\left.\frac{\partial RSS}{\partial \hat{\beta}_0}\right|_{\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}} &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i - \hat{\beta}_2^{ols} Z_i) = 0\,, \\
\left.\frac{\partial RSS}{\partial \hat{\beta}_1}\right|_{\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}} &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i - \hat{\beta}_2^{ols} Z_i) X_i = 0\,, \\
\left.\frac{\partial RSS}{\partial \hat{\beta}_2}\right|_{\hat{\beta}_0^{ols}, \hat{\beta}_1^{ols}, \hat{\beta}_2^{ols}} &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0^{ols} - \hat{\beta}_1^{ols} X_i - \hat{\beta}_2^{ols} Z_i) Z_i = 0\,.
\end{aligned} \tag{4.4}$$

We can also write the first order conditions as[2]

$$\sum_{i=1}^n \hat{\epsilon}_i^{ols} = 0\,, \quad \sum_{i=1}^n \hat{\epsilon}_i^{ols} X_i = 0\,, \quad \text{and} \quad \sum_{i=1}^n \hat{\epsilon}_i^{ols} Z_i = 0\,. \tag{4.5}$$

Instead of solving the three-equation three-unknown system (4.4) directly, we are going to take an alternative but entirely equivalent approach. This alternative approach is indirect, but more illustrative. We focus on the estimation of $\hat{\beta}_1^{ols}$. You can get the solution for $\hat{\beta}_2^{ols}$ by switching X_i with Z_i in the steps shown. After obtaining $\hat{\beta}_1^{ols}$ and $\hat{\beta}_2^{ols}$, you can use the first equation in (4.4) to compute

$$\hat{\beta}_0^{ols} = \overline{Y} - \hat{\beta}_1^{ols} \overline{X} - \hat{\beta}_2^{ols} \overline{Z}\,.$$

[2]Alternatively, we can take the method of moments approach. Defining $\epsilon = Y - \beta_0 - \beta_1 X - \beta_2 Z$, we have $E(\epsilon \mid X, Z) = 0$ as long as $E(Y \mid X, Z)$ is in fact equal to $\beta_0 - \beta_1 X - \beta_2 Z$. The condition $E(\epsilon \mid X, Z) = 0$ in turn implies $E(\epsilon) = 0$, $E(\epsilon X) = 0$ and $E(\epsilon Z) = 0$. First note that we have

$$E(\epsilon \mid X) = E(E(\epsilon \mid X, Z) \mid X) = 0 \;\text{ and }\; E(\epsilon \mid Z) = E(E(\epsilon \mid X, Z) \mid Z) = 0\,.$$

This gives

$$E(\epsilon X) = E(E(\epsilon X \mid X, Z)) = E(X E(\epsilon \mid X, Z)) = E(0) = 0\,.$$

A similar argument shows that $E(\epsilon Z) = 0$. Replacing the population moments $E(\epsilon) = 0$, $E(\epsilon X) = 0$ and $E(\epsilon Z) = 0$ with their sample counterparts produces essentially the same conditions as (4.5).

We begin with the following "auxiliary" regressions:

1. Regress X_i on Z_i, and collect the residuals $r_{i,x|z}$ from this regression, i.e., compute

$$r_{i,x|z} = X_i - \hat{\delta}_0 - \hat{\delta}_1 Z_i \ , \ i = 1, 2, ..., n$$

where $\hat{\delta}_0$ and $\hat{\delta}_1$ are the OLS estimators for the intercept and slope coefficients from a regression of X_i on a constant and Z_i. We can write

$$X_i = \widehat{X}_i + r_{i,x|z} \ \text{ where } \ \widehat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i \ , \ i = 1, \dots, n \, .$$

The fitted values $\widehat{X}_i$ are perfectly correlated with Z_i, and the residuals $r_{i,x|z}$ are perfectly uncorrelated with Z_i. You can think of this regression as "breaking up" X_i into two parts, $\widehat{X}_i$ and $r_{i,x|z}$. The residuals contain the movements in X_i after filtering out that part of X_i that is correlated with Z_i.

2. Regress Y_i on Z_i, and collect the residuals $r_{i,y|z}$ from this regression, i.e., compute

$$r_{i,y|z} = Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 Z_i \ , \ i = 1, 2, ..., n$$

where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are the OLS estimators for the intercept and slope coefficients from a regression of Y_i on a constant and Z_i. As in the first regression, the residuals $r_{i,y|z}$ contain the movements in Y_i after filtering out that part of Y_i that is correlated with Z_i.

The OLS estimator $\hat{\beta}_1^{ols}$ obtained from solving the first order conditions (4.4) turns out to be equal to the OLS estimator of the coefficient on $r_{i,x|z}$ in a regression of $r_{i,y|z}$ on $r_{i,x|z}$ (you can exclude the intercept term here; the sample means of both residuals are zero by construction, so the estimator for the intercept term if included will also be zero). In other words,

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^n r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^n r_{i,x|z}^2}. \tag{4.6}$$

To see this, note that since $\{r_{i,x|z}\}_{i=1}^n$ are OLS residuals from a regression of X_i on an intercept term and Z_i, we have $\sum_{i=1}^n r_{i,x|z} = 0$ and $\sum_{i=1}^n r_{i,x|z} Z_i = 0$. This implies

$$\sum_{i=1}^n r_{i,x|z} \hat{X}_i = \sum_{i=1}^n r_{i,x|z} (\hat{\delta}_0 + \hat{\delta}_1 Z_i) = 0$$

and furthermore,

$$\sum_{i=1}^n r_{i,x|z} X_i = \sum_{i=1}^n r_{i,x|z} (\hat{X}_i + r_{i,x|z}) = \sum_{i=1}^n r_{i,x|z}^2.$$

Now consider the sum $\sum_{i=1}^n r_{i,x|z} r_{i,y|z}$. We have

$$\begin{aligned} \sum_{i=1}^n r_{i,x|z} r_{i,y|z} &= \sum_{i=1}^n r_{i,x|z} (Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 Z_i) = \sum_{i=1}^n r_{i,x|z} Y_i \\ &= \sum_{i=1}^n r_{i,x|z} (\hat{\beta}_0^{ols} + \hat{\beta}_1^{ols} X_i + \hat{\beta}_2^{ols} Z_i + \hat{\epsilon}_i) = \hat{\beta}_1^{ols} \sum_{i=1}^n r_{i,x|z}^2 + \sum_{i=1}^n r_{i,x|z} \hat{\epsilon}_i. \end{aligned} \tag{4.7}$$

Finally, we note that the first order conditions (4.4) imply that

$$\begin{aligned}\sum_{i=1}^{n} r_{i,x|z}\hat{\epsilon}_i &= \sum_{i=1}^{n} \hat{\epsilon}_i(X_i - \hat{\delta}_0 - \hat{\delta}_1 Z_i) \\ &= \sum_{i=1}^{n} \hat{\epsilon}_i X_i - \hat{\delta}_0 \sum_{i=1}^{n} \hat{\epsilon}_i - \hat{\delta}_1 \sum_{i=1}^{n} \hat{\epsilon}_i Z_i = 0.\end{aligned}$$

Therefore

$$\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z} = \hat{\beta}_1^{ols} \sum_{i=1}^{n} r_{i,x|z}^2$$

which gives (4.6).

Note that in order for (4.6) to be feasible, we require $\sum_{i=1}^{n} r_{i,x|z}^2 \neq 0$. This means that X_i and Z_i cannot be *perfectly* correlated. They can be correlated, just not perfectly so. Furthermore, in the auxiliary regression of X_i on Z_i, we require some variation in Z_i, i.e., it cannot be that all the Z_i, $i = 1, 2, ..., n$ have the same value c. Similarly, to derive $\hat{\beta}_2^{ols}$, we require variation in X_i. All this is perfectly intuitive. If there is no variation in X_i in the sample, we cannot measure how Y_i changes with X_i. Similarly for Z_i. If X_i and Z_i are perfectly correlated, we will not be able to tell whether a change in Y_i is due to a change in X_i or in Z_i, since they move in perfect lockstep. We can summarize all of these requirements by saying that we require

$$c_1 + c_2 X_i + c_3 Z_i = 0 \text{ for all } i = 1, 2, ..., n \iff (c_1, c_2, c_3) = (0, 0, 0).$$

The "$\Leftarrow$" implication clearly always holds; that's not the important part. The important part is the "$\Rightarrow$" implication, which does not always hold. In particular, if X_i or Z_i take the same value for all $i = 1, ..., n$, or if X_i and Z_i are perfect correlated, then the "$\Rightarrow$" implication will not hold. If $X_i = c$ for all $i = 1, ..., n$, then set $c_1 = -c, c_2 = 1, c_3 = 0$ and we get $c_1 + c_2 X_i + c_3 Z_i = -c + c + 0 = 0$ for all $i = 1, ..., n$. If $Z_i = c$ for all i, then set $c_1 = -c, c_2 = 0, c_3 = 1$. If $X_i = a + bZ_i$ for all i, (with $b \neq 0$) then set $c_1 = -a, c_2 = 1, c_3 = -b$ which gives $c_1 + c_2 X_i + c_3 Z_i = -a + a + bZ_i - bZ_i = 0$ for all $i = 1, ..., n$.

The arguments presented in this section show the essence of how confounding factors are 'controlled' in multiple regression analysis. Suppose we want to measure how Y_i is affected by X_i. If Z_i is an important determinant of Y_i that is correlated with X_i, but omitted from the regression, then the measurement of the influence of X_i on Y_i will be distorted. In an experiment, we would control for Z_i by literally holding it fixed when collecting X_i and Y_i. In applications in economics, this is impossible. What multiple regression analysis does instead is to strip out all variation in Y_i and X_i that are correlated with Z_i, and then measure the correlation in the remaining variation between Y_i and X_i.

4.3 Algebraic Properties of OLS Estimators

We drop the 'OLS' superscript in our notation of the OLS estimators, residuals, and fitted values from this point, and write $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\epsilon}_i$ and $\hat{Y}_i$ for $\hat{\beta}_0^{ols}$, $\hat{\beta}_1^{ols}$, $\hat{\beta}_2^{ols}$, $\hat{\epsilon}_i^{ols}$ and $\hat{Y}_i^{ols}$ respectively. We will reinstate the 'ols' superscript whenever we need to emphasize that OLS was used, or when comparing OLS estimators to estimators derived in another way.

Many of the algebraic properties carry over from the simple linear regression model. These properties all come about because the OLS estimators satisfy the first order conditions

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0, \quad \sum_{i=1}^{n} X_i \hat{\epsilon}_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} Z_i \hat{\epsilon}_i = 0.$$

1. This implies that the fitted values $\hat{Y}_i$ and the residuals are also uncorrelated.
2. The point $(\overline{X}, \overline{Y}, \overline{Z})$ always lies on the estimate regression function.
3. We continue to have $\overline{Y} = \overline{\hat{Y}}$.
4. The $TSS = FSS + RSS$ equality continues to hold in the multiple regression case

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{\hat{Y}})^2 + \sum_{i=1}^{n} \hat{\epsilon}_i^2. \tag{4.8}$$

5. As in the simple linear regression case, we can use (4.8) to define the goodness-of-fit measure:

$$R^2 = 1 - \frac{RSS}{TSS}. \tag{4.9}$$

   It should be noted that the R^2 will never decrease as we add more variables into the regression. This is because OLS minimizes RSS, and therefore maximizes R^2. For example, the R^2 from the regression $Y = \beta_0 + \beta_1 X + \beta_2 Z + u$ will never be less than the R^2 from the regression $Y = \alpha_0 + \alpha_1 X + \epsilon$, and will generally be greater, unless it so happens that $\hat{\beta}_0 = \hat{\alpha}_0$, $\hat{\beta}_1 = \hat{\alpha}_1$ and $\hat{\beta}_2 = 0$. If $\hat{\beta}_1$, $\hat{\beta}_2$ or $\hat{\beta}_3$ take on any other value, it will be because the RSS can be further minimized, and the R^2 made larger. For this reason, the "Adjusted R^2"

$$\text{Adj.-}R^2 = 1 - \frac{\frac{1}{n-K} RSS}{\frac{1}{n-1} TSS}$$

   is sometimes used, where K is the number of coefficients estimated (including the intercept term; for the 2-regressor case that we are focusing on, $K = 3$). The idea is to use unbiased estimates of the variances of ϵ_i and Y_i in computing the goodness-of-fit. Since both RSS and $n - K$ decrease when additional variables (and parameters) are added into the model, the adjusted R^2 will increase only if RSS falls enough to lower the value of $RSS/(n-k)$. The adjusted R^2 may be used as an alternate measure of goodness-of-fit, but it should not be used as a model selection tool, for reasons we shall come to in a later chapter.
6. In the derivation (4.7) of the OLS estimator $\hat{\beta}_1$, we noted that

$$\begin{aligned} \sum_{i=1}^{n} r_{i,x|z} r_{i,y|z} &= \sum_{i=1}^{n} r_{i,x|z} (Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 Z_i) \\ &= \sum_{i=1}^{n} r_{i,x|z} Y_i. \end{aligned}$$

   This implies that the estimator can also be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} r_{i,x|z} r_{i,y|z}}{\sum_{i=1}^{n} r_{i,x|z}^2} = \frac{\sum_{i=1}^{n} r_{i,x|z} Y_i}{\sum_{i=1}^{n} r_{i,x|z}^2} \tag{4.10}$$

which is the formula for the simple linear regression of Y_i on $r_{i,x|z}$. In other words, you can also get the OLS estimator $\hat{\beta}_1$ by regressing Y_i on $r_{i,x|z}$ without first stripping out the covariance between Y_i and Z_i. Equation (4.10), however, does not fully reflect what happens in a regression of Y_i on two regressors.

7. The expression (4.10) shows that $\hat{\beta}_1$ is a linear estimator, i.e.,

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i$$

where the weights here are $w_i = r_{i,x|z} / \sum_{i=1}^{n} r_{i,x|z}^2$. Note that the weights w_i are made up solely of observations $\{X_i\}_{i=1}^n$ and $\{Z_i\}_{i=1}^n$, since they are the residuals from a regression of X_i on Z_i. Furthermore, the weights have the following properties:

$$\begin{aligned}
&\sum_{i=1}^{n} w_i = 0\,, \\
&\sum_{i=1}^{n} w_i Z_i = \frac{\sum_{i=1}^{n} r_{i,x|z} Z_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = 0\,, \\
&\sum_{i=1}^{n} w_i X_i = \frac{\sum_{i=1}^{n} r_{i,x|z} X_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = 1\,, \\
&\sum_{i=1}^{n} w_i^2 = \frac{\sum_{i=1}^{n} r_{i,x|z}^2}{(\sum_{i=1}^{n} r_{i,x|z}^2)^2} = \frac{1}{\sum_{i=1}^{n} r_{i,x|z}^2}\,.
\end{aligned}$$

8. If the sample correlation between X_i and Z_i is zero, then the coefficient estimate $\hat{\delta}_1$ in the auxiliary regression where we regressed X on Z would be zero, and $\hat{\delta}_0$ would be equal to the sample mean of $\overline{X}$. In other words, we would have $r_{i,x|z} = X_i - \overline{X}$, so

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} r_{i,x|z} Y_i}{\sum_{i=1}^{n} r_{i,x|z}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) Y_i}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

This is, of course, just the OLS estimator for the coefficient on X_i in the *simple* linear regression of Y on X. In other words, if the sample correlation between X_i and Z_i is zero, then including Z_i in the regression would not change the value of the simple linear regression estimator for the coefficient on X_i. We will see shortly that including the additional variable may nonetheless still reduce the estimator variance.

4.4 Statistical Properties of OLS Estimators

The assumption that $E(\epsilon \mid X, Z) = 0$ in population, and that we have a representative i.i.d. sample from the population, leads to the following (which we will also call "assumptions"):

(A1) $E(\epsilon_i \mid \mathrm{x}, \mathrm{z}) = 0$ for all $i = 1, ..., n$,

(A2) $E(\epsilon_i \epsilon_j \mid \mathrm{x}, \mathrm{z}) = 0$ for all $i \neq j,\ i, j = 1, ..., n$ where x denotes $X_1, X_2, ..., X_n$, and z to denotes $Z_1, Z_2, ..., Z_n$.

We will also assume homoskedastic errors

(A3) $E(\epsilon_i^2 \mid \mathrm{x}, \mathrm{z}) = \sigma^2$ for all $i = 1, ..., n$.

Assumption (A1) in particular, leads to unbiased OLS estimators. To see this, first write

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i = \sum_{i=1}^{n} w_i(\beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i) = \beta_1 + \sum_{i=1}^{n} w_i \epsilon_i \,.$$

Taking conditional expectations. noting that the weights w_i comprise only x and z, gives

$$E(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z}) = \beta_1 + \sum_{i=1}^{n} w_i E(\epsilon_i \mid \mathrm{x}, \mathrm{z}) = \beta_1 \,.$$

It follows that the unconditional mean is $E(\hat{\beta}_1) = \beta_1$.

The conditional variance of $\hat{\beta}_1$ under our assumptions is

$$\begin{aligned} Var(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z}) &= Var\left(\beta_1 + \sum_{i=1}^{n} w_i \epsilon_i \,\middle|\, \mathrm{x}, \mathrm{z} \right) \\ &= \sum_{i=1}^{n} w_i^2 \, Var(\epsilon_i \mid \mathrm{x}, \mathrm{z}) \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} r_{i,x|z}^2} \,. \end{aligned} \tag{4.11}$$

Since the R^2 from the regression of X_i on Z_i is

$$R_{x|z}^2 = 1 - \frac{\sum_{i=1}^{n} r_{i,x|z}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \,,$$

we can also write $Var(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z})$ as

$$Var(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z}) = \frac{\sigma^2}{(1 - R_{x|z}^2) \sum_{i=1}^{n} (X_i - \overline{X})^2} \,. \tag{4.12}$$

Expression (4.12) clearly shows the trade-offs involved in adding a second regressor. Suppose the true data generating process is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon \,,\ E(\epsilon \mid X, Z) = 0$$

but you ran the regression

$$Y = \beta_0 + \beta_1 X + u.$$

If X and Z are correlated, then X and u are correlated, and you will get biased estimates of β_1. By estimating the multiple linear regression, you are able to get an unbiased estimate of β_1 by controlling for Z. However, the variance of the OLS estimator for β_1 changes from

$$Var(\hat{\beta}_1 \mid \mathrm{x}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

in the simple linear regression to the expression in (4.12) for the multiple linear regression. Since σ^2 is the variance of ϵ, and σ_u^2 is the variance of a combination of the uncorrelated variables Z

and ϵ, we have $\sigma^2 < \sigma_u^2$. This has the effect of reducing the estimator variance (which is good!). However, since $0 < 1 - R_{x|z}^2 < 1$, the denominator in the variance expression is smaller in the multiple linear regression case than in the simple linear regression case. This is because in the multiple regression, we have stripped out all variation in X that is correlated with Z, resulting in *reduced effective variation* in X, which in turn increases the estimator variance. In general (and especially in causal applications), one would usually consider the trade-off to be in favor of the multiple regression. However, if X and Z are highly correlated ($R_{x|z}^2$ close to 1), then the reduction in effective variation in X may be so severe that the estimator variance becomes very large. This tends to reduce the size of the t-statistic, leading to rejection of statistical significance even in cases where the size of the estimate itself may suggest strong *economic* significance.

To compute a numerical estimate for the conditional variance of $\hat{\beta}_1$, we have to estimate σ^2. An unbiased estimator for σ^2 in the two-regressor case is

$$\widehat{\sigma^2} = \frac{1}{n-3}\sum_{i=1}^{n}\hat{\epsilon}_i^2. \tag{4.13}$$

The RSS is divided by $n-3$ because three "degrees-of-freedom" were used in computing $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ and these were used in the computation of $\hat{\epsilon}_i$. We shall again leave the proof of unbiasedness of $\widehat{\sigma^2}$ for when we deal with the general case. We estimate the conditional variance of $\hat{\beta}_1$ using

$$\widehat{Var}(\hat{\beta}_1 \mid \mathrm{x}, \mathrm{z}) = \frac{\widehat{\sigma^2}}{(1 - R_{x|z}^2)\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{4.14}$$

The standard error of $\hat{\beta}_1$ is the square root of (4.14).

Example 4.4. We illustrate the main points of this chapter using our `multireg_eg.csv` dataset, reproduced yet again below as Fig. 4.3.

```
df <- read_csv("data\\multireg_eg.csv",col_types = c("n","n","n"))
p1 <- ggplot(data=df) + geom_point(aes(x=X, y=Y), color="darkslateblue") + theme_minimal() + theme1
p2 <- ggplot(data=df) + geom_point(aes(x=X, y=Y, shape=as.factor(Z), color=as.factor(Z))) +
      theme_minimal() + theme2 + labs(shape='Z', color='Z')
p1|p2
```

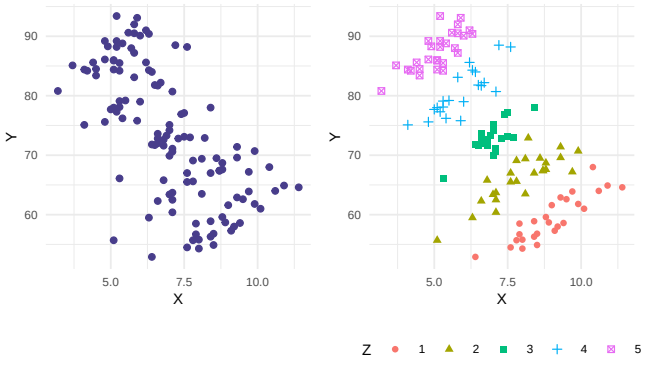


Figure 4.3: The multireg_eg.csv dataset.

We run two regressions below. The first is a simple linear regression of Y on X. The second is a multiple linear regression of Y on X and Z.

```
mdl1 <- lm(Y~X, data=df)
coef(summary(mdl1))
cat("R-squared:", summary(mdl1)$r.squared,"\n\n")
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 102.86406  3.2199095 31.94626 6.373574e-60
X            -4.23213  0.4474755 -9.45779 3.857133e-16
R-squared: 0.4311877
```

```
mdl2 <- lm(Y~X+Z, data=df)
coef(summary(mdl2))
cat("R-squared:", summary(mdl2)$r.squared,"\n\n")
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 21.188337  2.0816539 10.17861 8.212834e-18
X            3.114465  0.2054364 15.16024 2.219609e-29
Z           10.109717  0.2379845 42.48057 5.955647e-73
R-squared: 0.9653668
```

The simple linear regression shows the negative relationship between Y and X when viewed over all outcomes of Z. The multiple regression disentangles the effect of X and Z on Y. In this case, inclusion of Z has also substantially reduced the standard error on the estimate of the coefficient on X. As expected, the R^2 has gone up with the inclusion of Z, in this case by a very large amount.

We replicate below the multi-step approach to obtaining the coefficient estimate on X:

```
mdl1a <- lm(Y~Z, data=df)
r_yz <- residuals(mdl1a)
mdl1b <- lm(X~Z, data=df)
r_xz <- residuals(mdl1b)
df$r_yz <- r_yz
df$r_xz <- r_xz
coef(summary(lm(r_yz~r_xz-1, data=df)))  # "-1" in the formula means exclude the intercept
```

```
     Estimate Std. Error  t value     Pr(>|t|)
r_xz 3.114465  0.2037027 15.28926 7.425184e-30
```

The numerical estimate of the coefficient on `r_xz` is identical to that on `X` in the previous regression. The standard errors are similar, but not the same. We emphasize that the auxiliary regression approach is for illustrative purposes only. The standard errors, t-statistic, etc. should all be taken from the previous (multiple) regression. The plots in Fig. 4.4 illustrate the effect of 'controlling' for `Z`.

```
plot_theme <- theme_minimal() + theme(legend.position = "none", plot.title=element_text(size=10))
p1 <- ggplot(data=df) + geom_point(aes(x=X,y=Y, shape=as.factor(Z), color=as.factor(Z))) +
  xlim(c(2.5,12.5)) + ylim(c(50,100)) + ggtitle('(a)') + plot_theme
p2 <- ggplot(data=df) + geom_point(aes(x=r_xz,y=Y, shape=as.factor(Z), color=as.factor(Z))) +
  xlim(c(-5,5)) + ylim(c(50,100)) + ggtitle('(b)') + plot_theme
p3 <- ggplot(data=df) + geom_point(aes(x=r_xz,y=r_yz, shape=as.factor(Z), color=as.factor(Z))) +
  xlim(c(-5,5)) + ylim(c(-25,25)) + ggtitle('(c)') + plot_theme
p1|p2|p3
```

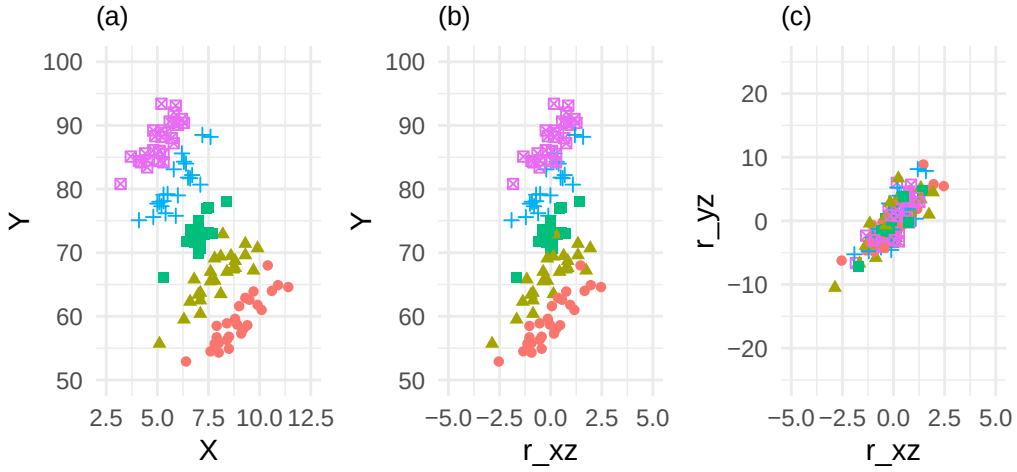


Figure 4.4: Controlling for a confounding variable

The range of the y-axis in the three plots are the same (50 to 100 in the first two, -25 to 25 in the third). Likewise the range of the x-axis is the same across all three plots (2.5 to 12.5 in the first, -5 to 5 in the second and third). When we regress X on Z and take the residuals, we remove the effect of Z on X and also center the residuals around zero (OLS residuals always have sample mean zero). You can see from the second diagram that the variation in X is reduced, which tends to increase the X coefficient estimator variance. You can also see that the positive relationship between Y and X controlling for Z already shows up, albeit with a lot of noise. By also removing the effect of Z on Y (which we do when we include Z in the regression), we reduce the variation of Y, compare (b) and (c). This *reduces* the estimator variance. In this example, the reduction in the variation on Y is very large relative the reduced variation in X, so the overall effect is that the variance on $\hat{\beta}_1$ falls. The slope coefficient in the simple regression of the residuals $r_{i,y|z}$ on $r_{i,x|z}$ in the last panel gives the effect of X on Y, controlling for Z.

4.5 Hypothesis Testing

To test if β_k is equal to some value r_k in population, we can again use the t-statistic as in the simple linear regression case:

$$t = \frac{\hat{\beta}_k - r_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}}. \tag{4.15}$$

If the noise terms are conditionally normally distributed, then the t-statistic has the t-distribution, with degrees-of-freedom $n - K$ where $K = 3$ in the two-regressor case with intercept. If we do not assume normality of the noise terms, then (as long as the necessary CLTs apply) we use instead the approximate test, using the t-statistic as defined above, but using the rejection region derived from the standard normal distribution. You can also test whether a linear combination of the parameters are equal to some value. For example, in the regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

you may wish to test, say, $H_0 : a_1\beta_1 + a_2\beta_2 = r_1$ vs $H_A : a_1\beta_1 + a_2\beta_2 \neq r_1$.

Example 4.5. Suppose the production technology of a firm can be characterized by the Cobb-Douglas production function:

$$Q(L, K) = AL^{\alpha}K^{\beta}$$

where $Q(L, K)$ is the quantity produced using L units of labor and K units of capital. The constants A, α and β are the parameters of the model. If we multiply the amount of labor and capital by c, we get

$$Q(cL, cK) = A(cL)^{\alpha}(cK)^{\beta} = c^{\alpha+\beta}AL^{\alpha}K^{\beta}.$$

The sum $\alpha + \beta$ therefore represents the "returns to scale". If $\alpha + \beta = 1$, then there is constant returns to scale, e.g., doubling the amount of labor and capital ($c = 2$) results in the doubling of total production. If $\alpha + \beta > 1$ then there is increasing returns to scale, and if $\alpha + \beta < 1$, we have decreasing returns to scale. A logarithmic transformation of the production function gives

$$\ln Q = \ln A + \alpha \ln L + \beta \ln K.$$

If we have observations $\{Q_i, L_i, K_i\}_{i=1}^{n}$ of the quantities produced and amount of labor and capital employed by a set of similar firms in an industry, we could estimate the production function for that industry using the regression

$$\ln Q_i = \ln A + \alpha \ln L_i + \beta \ln K_i + \epsilon_i \,.$$

A test for constant returns to scale would be the test $H_0 : \alpha + \beta = 1$ vs $H_A : \alpha + \beta \neq 1$.

The t-statistic for testing a hypothesis such as

$$H_0 : a_1\beta_1 + a_2\beta_2 = r_1 \;\text{ vs }\; H_0 : a_1\beta_1 + a_2\beta_2 \neq r_1$$

is

$$t = \frac{a_1\hat{\beta}_1 + a_2\hat{\beta}_2 - r_1}{\sqrt{\widehat{Var}(a_1\hat{\beta}_1 + a_2\hat{\beta}_2)}}. \tag{4.16}$$

To compute this t-statistic you will need an estimate of the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$, the derivation of which we leave for a later chapter. There are packages, of course, that will help you carry out this hypothesis test. For now, we consider an alternative way of forcing your linear regression to provide you with the relevant t-statistic, which is to re-parameterize your equation. For example, to test the hypothesis $H_0 : \beta_1 + \beta_2 = 1$, we can reparameterize the regression equation as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X - X + X - \beta_2 X + \beta_2 Z + \epsilon$$

$$Y - X = \beta_0 + (\beta_1 + \beta_2 - 1)X + \beta_2(Z - X) + \epsilon$$

By regressing $Y - X$ on X and $Z - X$ (and an intercept), you can test $H_0 : \beta_1 + \beta_2 = 1$ by testing if the coefficient on X is equal to zero.

Example 4.6. Using data in the `earnings2019.xlsx`, we estimate the regression

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon\,.$$

To test the hypothesis $H_0 : \beta_2 = \beta_3$ (equivalent to $H_0 : \beta_2 - \beta_3 = 0$), we re-parameterize the regression equation as

$$\ln earn = \beta_0 + \beta_1 educ + (\beta_2 - \beta_3) black + \beta_3 (female + black) + \beta_4 black.female + \epsilon\,.$$

```
dat2$female <- 1 - dat2$male
mdl_earnings1 <- lm(log(earn) ~ educ + black*female, data=dat2)
mdl_earnings1 %>% summary %>% coefficients %>% round(4)
```

```
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)    1.5369     0.0571  26.9219   0.0000
educ           0.1279     0.0039  32.9577   0.0000
black         -0.2600     0.0270  -9.6468   0.0000
female        -0.2807     0.0196 -14.3238   0.0000
black:female   0.0873     0.0355   2.4601   0.0139
```

```
mdl_earnings2 <- lm(log(earn) ~ educ + black + I(female+black) + female:black, data=dat2)
mdl_earnings2 %>% summary %>% coefficients %>% round(4)
```

```
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)         1.5369     0.0571  26.9219   0.0000
educ                0.1279     0.0039  32.9577   0.0000
black               0.0206     0.0271   0.7607   0.4469
I(female + black)  -0.2807     0.0196 -14.3238   0.0000
black:female        0.0873     0.0355   2.4601   0.0139
```

The hypothesis is not rejected.

In some cases, we may wish to test multiple hypotheses, e.g., in the two-variable regression, we may wish to test

$$H_0 : \beta_1 = 0 \;\text{ and }\; \beta_2 = 0 \;\text{ vs }\; H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

One possibility would be to do individual t-tests for each of the two hypotheses, but we should be aware that two individual 5% tests is not equivalent to a joint 5% test. The following example illustrates this problem.

Example 4.7. We generate 100 observations of three uncorrelated variables X, Y and Z. We regress Y on X and Z, and collect the t-statistics on X and Z. We repeat the experiment 1000 times (with different draws each time, of course, but the same parameters).

```
set.seed(3)
nreps <- 1000
tx <- tz <- rep(NA,nreps)
n <- 100
for (i in 1:nreps){
  X <- rnorm(n, mean=0, sd=2)
  Z <- rnorm(n, mean=0, sd=2)
  Y <- rnorm(n, mean=0, sd=2)
  df_test <- data.frame(X,Y,Z)
```

```
  mdlsim <- lm(Y~X+Z, data=df_test)
  tx[i] <- coef(summary(mdlsim))[2,'t value']
  tz[i] <- coef(summary(mdlsim))[3,'t value']
}
rjt_x <- sum(tx<qt(0.025,n-3) | tx>qt(0.975,n-3))/nreps
cat("Freq. of rejection of Beta_X = 0:", rjt_x, "\n")
```

```
Freq. of rejection of Beta_X = 0: 0.058
```

```
rjt_z <- sum(tz<qt(0.025,n-3) | tz>qt(0.975,n-3))/nreps
cat("Freq. of rejection of Beta_Z = 0:", rjt_z, "\n")
```

```
Freq. of rejection of Beta_Z = 0: 0.054
```

```
rjt_x_or_z <- sum(tz<qt(0.025,197) | tz>qt(0.975,197) |
                  tx<qt(0.025,197) | tx>qt(0.975,197))/nreps
cat("Freq. of rejection of Beta_X = 0 and Beta_Z = 0 using two t-tests:", rjt_x_or_z, "\n")
```

```
Freq. of rejection of Beta_X = 0 and Beta_Z = 0 using two t-tests: 0.112
```

When using a 5% t-test, we reject the (true) hypothesis that $\beta_x = 0$ in about 6% of the experiments, close to 5%. These rejections are regardless of whether the t-test for $\beta_z = 0$ rejects or does not reject. Likewise, the 5% t-test for $\beta_z = 0$ rejects the hypothesis 5.5% of the time, roughly five percent. However, if we say we reject $\beta_x = 0$ and $\beta_z = 0$ if either t-tests rejects the corresponding hypothesis, then the frequency of rejection is much larger, roughly double.

We plot the t-stats below, indicating the critical values for the individual tests. The proportion of points above the upper horizontal line or below the lower one is about 0.05. Similarly, the proportion of points to the left of the left vertical line or to the right of the right one is roughly 0.05. The number of point that meet either of the two sets of criteria is much larger, roughly the sum of the two proportions.

```
df_t <- data.frame(tx,tz)
ggplot(data=df_t) + geom_point(aes(x=tx,y=tz)) +
  geom_hline(yintercept = qt(0.025,n-3), lty='dashed', col='blue') +
  geom_hline(yintercept = qt(0.975,n-3), lty='dashed', col='blue') +
  geom_vline(xintercept = qt(0.025,n-3), lty='dashed', col='blue') +
  geom_vline(xintercept = qt(0.975,n-3), lty='dashed', col='blue') +
  theme_minimal()
```

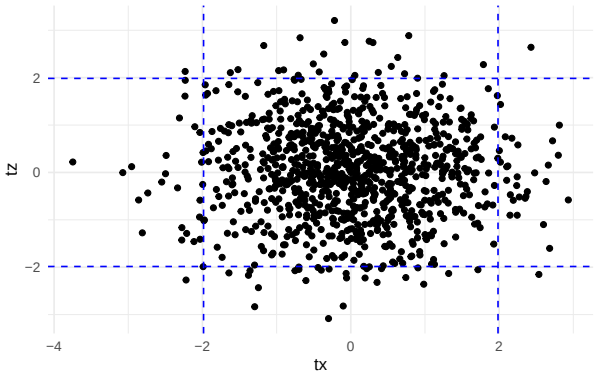


Figure 4.5: Rejection rate when compounding two t-tests

To jointly test multiple hypotheses, we can use the F-test. Suppose in the regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

we wish to jointly test the hypotheses

$$H_0 : \beta_1 = 1 \text{ and } \beta_2 = 0 \text{ vs } H_A : \beta_1 \neq 1 \text{ or } \beta_2 \neq 0 \text{ (or both)}.$$

Suppose we run the regression twice, once unrestricted, and another time with the restrictions in H_0 imposed. The regression with the restrictions imposed is

$$Y = \beta_0 + X + \epsilon$$

so the restricted OLS estimator for β_0 is the sample mean of $Y_i - X_i$, i.e.,

$$\hat{\beta}_0^{rols} = (1/n)\sum_{i=1}^{n}(Y_i - X_i).$$

Calculate the RSS from both equations. The "unrestricted RSS" and "restricted RSS" are

$$RSS_{ur} = \sum_{i=1}^{n}\hat{\epsilon}_i^2 \quad \text{and} \quad RSS_r = \sum_{i=1}^{n}\hat{\epsilon}_{i,\,rols}^2$$

respectively, where $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_i$ and $\hat{\epsilon}_{i,rols} = Y_i - \hat{\beta}_0^{rols} - X_i$. Since OLS minimizes RSS, imposing restrictions never decrease RSS, i.e.,

$$RSS_r \geq RSS_{ur}.$$

It can be shown that if the hypotheses in H_0 are true (and the noise terms are normally distributed), then

$$F = \frac{(RSS_r - RSS_{ur})/J}{RSS_{ur}/(n-K)} \sim F(J, n-K) \tag{4.17}$$

where J is the number of restrictions being tested (in our example, $J = 2$) and K is the number of coefficients to be estimated (including intercept; in our example, $K = 3$).

The F-statistic is always non-negative. The idea is that if the hypotheses in H_0 are true in population, then imposing the restrictions on the regression would not increase the RSS by much (since the unrestricted parameter estimates should be close to their population values), and the F-statistic will be close to zero. On the other hand, if one or more of the hypotheses in H_0 are false in population, then imposing them onto the regression will cause the RSS to increase substantially, and the F-statistic will be large. We take a very large F-statistic, meaning

$$F > \mathrm{F}_{1-\alpha}(J, n-K),$$

as statistical evidence that one or more of the hypothesis is false, where $\mathrm{F}_{1-\alpha}(J, n-K)$ is the $(1-\alpha)$-percentile of the $\mathrm{F}(J, n-K)$ distribution and where α is typically 0.10, 0.05 or 0.01,

Since $R^2 = 1 - RSS/TSS$, we can write the F-statistic in terms of R^2 instead of RSS. You are asked in an exercise to show that the F-statistic can be written as

$$F = \frac{(R^2_{ur} - R^2_r)/J}{(1 - R^2_{ur})/(n - K)}.$$

Imposing restrictions cannot increase R^2, and in general will decrease it. The F-test essentially tests if the R^2 drops significantly when the restrictions are imposed. If the hypotheses being tested are true, then the drop should be slight. If one or more are false, the drop should be substantial, resulting in a large F-statistic.

If you cannot assume that the noise terms are conditionally normally distributed, then you will have to use an asymptotic approximation. It can be shown that

$$JF \xrightarrow{d} \chi^2(J)$$

as $n \to \infty$, where J is the number of hypotheses being jointly tested, and F is the F-statistic (4.17). We refer to this as the "Chi-square Test".

Example 4.8. We continue with Example 4.4, which considered the regression

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon\,.$$

This time we test the hypothesis $H_0 : \beta_2 = \beta_3$ and $\beta_4 = 0$, vs the alternative that at least one of these two restrictions do not hold in population. The restricted regression is

$$\begin{aligned} \ln earn &= \beta_0 + \beta_1 educ + \beta_2 black + \beta_2 female + 0 black.female + \epsilon \\ &= \beta_0 + \beta_1 educ + \beta_2 (black + female) + \epsilon \end{aligned}$$

We carry out the F-test of our hypotheses below:

```
mdl_unres <- lm(log(earn) ~ educ + black*female, data=dat2)
mdl_res <- lm(log(earn) ~ educ + I(black + female), data=dat2)
RSS_unres <- sum(residuals(mdl_unres)^2)
RSS_res   <- sum(residuals(mdl_res)^2)
J = 2
n_minus_K <- mdl_unres$df.residual
Fstat = ((RSS_res - RSS_unres) / J) / (RSS_unres / n_minus_K)
pval = 1-pf(Fstat, J, n_minus_K)
cat("F-stat:", Fstat, "   pvalue:", pval)
```

```
F-stat: 4.546114    pvalue: 0.01065276
```

The joint hypothesis is rejected (at 0.05 significance level, but not 0.01 significance level).

We can use the function `linearHypothesis()` from the `car` package to carry out `F` tests.

```
mdl_unres <- lm(log(earn) ~ educ + black*female, data=dat2)
linearHypothesis(mdl_unres, c("black=female", "black:female=0"))
```

```
Linear hypothesis test:
black - female = 0
black:female = 0
```

```
Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   4943 1606.5
2   4941 1603.5  2    2.9507 4.5461 0.01065 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To get the chi-square version of the test:

```
linearHypothesis(mdl_unres, c("black=female", "black:female=0"), test="Chisq")
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
1   4943 1606.5
2   4941 1603.5  2    2.9507 9.0922    0.01061 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see, the chi-square statistic is just J times the F statistic.

You can use the F-test to test a single hypothesis, e.g., to test $\beta_2 = \beta_3$:

```
linearHypothesis(mdl_unres, c("black=female"))
```

```
Linear hypothesis test:
black - female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   4942 1603.7
2   4941 1603.5  1   0.18778 0.5786 0.4469
```

Notice that the F-statistic for testing this single hypothesis is just twice that of the t-statistic for the same hypothesis in Example 4.4 (we will prove this to be the case in general in a later chapter). The p-value is the same.

Finally we note that the `summary()` output of an `lm object` contains a F-statistic. This is the F-statistic to test the joint hypotheses that all of the coefficients in the regression (not including the intercept) are zero, versus the alternative that at least one of those coefficients are not zero. The output also contains the adjusted R^2 for the estimated regression.

```
mdl_unres %>% summary()
```

```
Call:
lm(formula = log(earn) ~ educ + black * female, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5992 -0.3448  0.0022  0.3515  2.8614

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.536856   0.057086  26.922   <2e-16 ***
educ          0.127858   0.003879  32.958   <2e-16 ***
black        -0.260050   0.026957  -9.647   <2e-16 ***
female       -0.280661   0.019594 -14.324   <2e-16 ***
black:female  0.087299   0.035486   2.460   0.0139 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5697 on 4941 degrees of freedom
Multiple R-squared:  0.2389,    Adjusted R-squared:  0.2383
F-statistic: 387.7 on 4 and 4941 DF,  p-value: < 2.2e-16
```

4.6 Exercises

Exercise 4.1. What is the main conceptual difference between the regressions

$$\text{(A)} \qquad \ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon$$

$$\text{(B)} \qquad \ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \epsilon\,.$$

In particular, what does the inclusion of the interaction term $black.female$ in model A allow us to capture that model B does not.

Exercise 4.2. Suppose your estimated sample regression function is

$$\widehat{log(earn)} = 1.35 + 0.080\, age - 0.0008\, age^2$$

where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are positive and $\hat{\alpha}_2$ is negative.
(a) At what age are wages predicted to start declining with age?
(b) Does the intercept have any reasonable economic interpretation?

Exercise 4.3. Show that the F-statistic in (4.17) can be written as

$$F = \frac{(R^2_{ur} - R^2_r)/J}{(1 - R^2_{ur})/(n - K)}$$

where R_{ur} and R_r are the R^2 from the unrestricted and restricted regressions respectively, J is the number of restrictions being tested, n is the number of observations used in the regression, and K is the number of coefficient parameters in the unrestricted regression model (including intercept). What does this expression simplify to if the hypothesis being tested is that all the slope coefficients (excluding the intercept) are equal to zero?

Exercise 4.4. Prove Equation (4.8).

Exercise 4.5. Each of the following regressions produces a sample regression function whose slope is $\hat{\beta}_1$ when $X_i < \xi$ and $\hat{\beta}_1 + \hat{\alpha}_1$ when $X_i \geq \xi$. Which of them produces an estimated regression function that is continuous at ξ?

i. $Y_i = \beta_0 + \beta_1 X_i + \alpha_1 D_i X_i + \epsilon$ where $D_i = 1$ if $X_i > \xi$, $D_i = 0$ otherwise;

ii. $Y_i = \beta_0 + \alpha_0 D_i + \beta_1 X_i + \alpha_1 D X_i + \epsilon$;

iii. $Y_i = \beta_0 + \beta_1 X_i + \alpha_1 (X_i - \xi)_+ + \epsilon_i$ where

$$(X_i - \xi)_+ = \begin{cases} X_i - \xi & \text{if } X_i > \xi \,, \\ 0 & \text{if } X_i \leq \xi \,. \end{cases}$$

Exercise 4.6. Suppose

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + u$$
$$Z = \delta_1 X + v$$

where u and v are independent zero-mean noise terms. Suppose you have a random sample $\{Y_i, X_i, Z_i\}_{i=1}^n$ from this population and you ran the regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Show that the OLS estimator $\hat{\beta}_1$ will be biased for α_1. What is its expectation?

Exercise 4.7. Suppose

$$Y = \alpha_0 + \alpha_1 Z + u$$
$$X = \delta_1 Z + v$$

where u and v are independent zero-mean noise terms with variances σ_u^2 and σ_v^2 respectively. You may also assume that Z is an independent random variable with mean zero and variance σ_Z^2. Clearly the random variables Y and X are both directly driven by Z. Apart from the common influence of Z, the random variables Y and X are not connected in any way. Suppose you have a random sample $\{Y_i, X_i, Z_i\}_{i=1}^n$ from this population and you ran the regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Show that the OLS estimator $\hat{\beta}_1$ will be biased for α_1. What is its expectation? (*Hint: First find out what $E(Z \mid X)$ is. You may assume that it has the form $E(Z \mid X) = a + bX$; Find out what a and b are. Then find $E(Y \mid X)$.*)

Exercise 4.8. Show for the multiple linear regression model (with intercept term included) that the R^2 is the square of the correlation coefficient of Y_i and $\hat{Y}_i$, i.e.,

$$R^2 = \left[\frac{\sum_{i=1}^N (Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\sum_{i=1}^N (Y_i - \overline{Y})^2} \sqrt{\sum_{i=1}^N (\hat{Y}_i - \overline{\hat{Y}})^2}} \right]^2$$

This is where the name "R^2" comes from.

Exercise 4.9. For this question, we use the data in `earnings2019.csv`, which we earlier read into R as `dat2`.

(a) Run the following code:

```
lm(log(earn)~height, data=dat2) %>% summary %>% coefficients %>% round(4)
lm(log(earn)~height + male, data=dat2) %>% summary %>% coefficients %>% round(4)
```

and explain why the coefficient estimate for `height` goes down when `male` is added to the regression, but its standard error goes up.

(b) Run the following code:

```
lm(log(earn)~educ, data=dat2) %>% summary %>% coefficients %>% round(4)
lm(log(earn)~educ+tenure, data=dat2) %>% summary %>% coefficients %>% round(4)
```

Based on these results alone:

i. Do you think `educ` and `tenure` are strongly correlated or weakly correlated? Positively or negatively?

ii. Do you think that the R^2 from the regression of $\ln earn$ on $tenure$ is near zero, near one, or moderately valued?

State your reasoning.

Exercise 4.10. Suppose $male$ is a dummy variable indicating if an observation is male (=1) or not male (=0), and the variable $female$ is defined as $female = 1 - male$.
(a) Explain why you cannot run the regression

$$\text{(A)} \quad \ln earn = \beta_0 + \beta_1 male + \beta_2\, female + \epsilon$$

but you can run the regression

$$\text{(B)} \quad \ln earn = \beta_1 male + \beta_2\, female + \epsilon\,.$$

(b) In the regression

$$\text{(C)} \quad \ln earn = \alpha_0 + \alpha_1\, male + \epsilon\,,$$

show that the OLS estimator $\hat{\alpha}_0$ is equal to the sample mean of $\ln earn$ for all female individuals in the sample, and $\hat{\alpha}_0 + \hat{\alpha}_1$ is the sample mean of $\ln earn$ for all male individuals in the sample. By suitable re-parameterization of equation B, show that the OLS estimator $\hat{\beta}_1$ is equal to the sample mean of $\ln earn$ for all male individuals in the sample and $\hat{\beta}_2$ is equal to the sample mean of $\ln earn$ for all female individuals in the sample.

Chapter 5
Matrix Algebra

We have so far avoided matrix algebra in our discussion of the linear regression model, one reason being that we can get good intuition from the "summation-notation formulas" for parameter estimators, standard errors, and other statistics. However, we have had to limit ourselves to two regressors, and even then we have had to skip over several results. To go any further, we will need matrix algebra. In this chapter we cover the basics: definitions, notation, and elementary operations, and its applications in solving systems of linear equations. This chapter is an abridged version of Chapter 7 of Tay, Preve, and Baydur (2025).

5.1 Definitions and Notation

A **matrix** is a rectangular collection of numbers. The following is a matrix with m rows and n columns:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

Such a matrix is said to have "dimension" or "order" $m \times n$. The number that appears in the (i,j)th position, i.e., in the ith row and jth column, is called the (i,j)th element/entry/component of the matrix. We count rows from top to bottom, and columns from left to right. If $m = n$, the matrix is a **square matrix**. If $m = 1$ and $n > 1$, we have a **row vector**. If $m > 1$ and $n = 1$, we have a **column vector**. If $m = n = 1$, we have a **scalar**.

The term "vector" is used in many ways in mathematics. Sometimes a vector refers to an ordered list of numbers $(x_1, x_2, \ldots, x_n)$. Such an object has no "shape". It is merely an ordered sequence of n elements. Column and row vectors, on the other hand, are "two-dimensional" objects, in the sense of having a "height" (number of rows) and "width" (number of columns). In the context of matrix algebra, the word "vector" alone usually means a column vector, but not always.

Example 5.1. The matrix A below is a square matrix, b is a column vector and c is a row vector.

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \; c = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix}.$$

Matrices and vectors are often written in bold lettering, or with some sort of mark to distinguish them from scalars and other objects. We will not do so in these notes. The reader will have to rely on context to distinguish scalars from vectors and matrices. Where context is unclear, we will be more explicit.

Some additional notation:

- It is often convenient to indicate an $m \times n$ matrix A by $(a_{ij})_{m\times n}$.
- We can refer to the (i,j)th element of a matrix A by $(A)_{ij}$ or $(A)_{i,j}$.

The utility of these two notational conventions should become clearer as the chapter progresses.

Two matrices of the same dimension are said to be equal if each of their corresponding elements are equal, i.e.,

$$A = B \Leftrightarrow (A)_{ij} = (B)_{ij} \text{ for all } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n.$$

Two matrices of different dimensions cannot be equal.

A **zero matrix** is one whose elements are all zero. It is simply written as 0 although sometimes subscripts are added to indicate its dimension.

The **diagonal** of a $n \times n$ square matrix refers to the (i,i)th elements of the matrix, i.e., to the elements $(A)_{ii}$, $i = 1, 2, \ldots, n$. A **diagonal matrix** is a square matrix with all off-diagonal elements equal to zero, i.e., a square matrix A is diagonal if $(A)_{ij} = 0$ for all $i \neq j$, $i, j = 1, 2, ..., n$. Diagonal matrices are sometimes written $\text{diag}(a_{11}, a_{22}, ..., a_{nn})$.

Example 5.2. The matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \text{diag}(1, 4, 0)$$

is a diagonal matrix. Note that there is nothing in the definition of a diagonal matrix that says its diagonal elements cannot be zero.[1]

An **identity matrix** is a square matrix with all diagonal elements equal to one and all off-diagonal elements equal to zero, i.e.,

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \text{diag}(\underbrace{1, 1, \ldots, 1}_{n \text{ terms}}).$$

We will denote an identity matrix by I. A subscript is sometimes added to indicate its dimension, as we did above, although this is often left out. We will see shortly that the identity matrix plays a role in matrix algebra akin to the role played by the number "1" in the real number system.

A **symmetric matrix** is a square matrix A such that $(A)_{ij} = (A)_{ji}$ for all $i, j = 1, 2, ..., n$.

Example 5.3. The matrix $\begin{bmatrix} 1 & 3 & 2 \\ 3 & 4 & 6 \\ 2 & 6 & 3 \end{bmatrix}$ is symmetric, $\begin{bmatrix} 1 & 3 & 2 \\ 7 & 4 & 6 \\ 2 & 6 & 3 \end{bmatrix}$ is not.

[1] A square zero matrix is therefore technically also a diagonal matrix.

5.1.1 Addition, Scalar Multiplication and Transpose

Addition: Matrix addition is defined as element-by-element addition, i.e.., for two matrices $A = (a_{ij})_{m\times n}$ and $B = (b_{ij})_{m\times n}$, we define

$$(A+B)_{ij} = (A)_{ij} + (B)_{ij} \text{ for all } i = 1,\dots,m;\ j = 1,\dots,n\,.$$

Matrix addition is defined only for matrices of the same dimensions.

Example 5.4. $\begin{bmatrix}1 & 4\\ 3 & 2\\ 6 & 5\end{bmatrix} + \begin{bmatrix}6 & 9\\ 1 & 2\\ 1 & 10\end{bmatrix} = \begin{bmatrix}1+6 & 4+9\\ 3+1 & 2+2\\ 6+1 & 5+10\end{bmatrix} = \begin{bmatrix}7 & 13\\ 4 & 4\\ 7 & 15\end{bmatrix}.$

It should also be obvious that

$$A + B = B + A\,,$$
$$(A+B)+C = A+(B+C)\,.$$

This means that *as far as addition is concerned*, we can manipulate matrices in the same way we manipulate ordinary numbers (as long as the matrices being added have the same dimensions).
Scalar Multiplication: For a scalar α and matrix $A = (a_{ij})_{m\times n}$, we define

$$(\alpha A)_{ij} = (A\alpha)_{ij} = \alpha(A)_{ij} \text{ for all } i = 1,\dots,m\ ;\ j = 1,\dots,n\,.$$

i.e., the product of a scalar and a matrix is defined to be the multiplication of each element of the matrix by the scalar.

Example 5.5. $b\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\\ a_{31} & a_{32}\end{bmatrix} = \begin{bmatrix}ba_{11} & ba_{12}\\ ba_{21} & ba_{22}\\ ba_{31} & ba_{32}\end{bmatrix}.$

We can use scalar multiplication to define **matrix subtraction**:

$$A - B = A + (-1)B.$$

Transpose: When we transpose a matrix, we write its rows as its columns, and its columns as its rows. That is, the transpose of an $(m\times n)$ matrix A, denoted either by A^{T} or A', is defined by

$$(A^{\mathrm{T}})_{ij} = (A)_{ji} \text{ for all } i = 1,2,...,m \text{ and } j = 1,2,...,n.$$

Example 5.6. $\begin{bmatrix}1 & 4\\ 3 & 2\\ 6 & 5\end{bmatrix}^{\mathrm{T}} = \begin{bmatrix}1 & 3 & 6\\ 4 & 2 & 5\end{bmatrix}.$

In order to use space more efficiently, we will often write a column vector $x = \begin{bmatrix}x_1\\ \vdots\\ x_m\end{bmatrix}$ as

$$x = \begin{bmatrix}x_1 & x_2 & \dots & x_m\end{bmatrix}^{\mathrm{T}} \text{ or } x^{\mathrm{T}} = \begin{bmatrix}x_1 & x_2 & \dots & x_m\end{bmatrix}\,.$$

We can use the transpose operator to define symmetric matrices: a symmetric matrix is simply a square matrix where $A^{\mathrm{T}} = A$.

5.1.2 Exercises

Exercise 5.1. What is the dimension of $A = \begin{bmatrix} 7 & 13 \\ 4 & 4 \\ 7 & 15 \end{bmatrix}$? What is $(A)_{1,2}$ and $(A)_{3,1}$?

Exercise 5.2. Suppose $A = (a_{ij})_{2\times 4}$ where $a_{ij} = i + j$. Write out the matrix in full.

Exercise 5.3. Express the following matrices in full:

(a) $(a_{ij})_{4\times 4}$ where $a_{ij} = 1$ when $i = j$, 0 otherwise.

(b) $(a_{ij})_{4\times 4}$ where $a_{ij} = 0$ if $i \neq j$ (fill the rest of the entries with "$*$").

(c) $(a_{ij})_{4\times 4}$ where $a_{ij} = 0$ if $i < j$ (fill the rest of the entries with "$*$").

(d) $(a_{ij})_{4\times 4}$ where $a_{ij} = 0$ if $i > j$ (fill the rest of the entries with "$*$").

These are all square matrices. Matrix (c) is a "lower triangular matrix" and (d) is an "upper triangular matrix" (so we have in (c) and (d) matrices that are square and triangular!). Matrix (b) is diagonal, which is both upper and lower triangular.

Exercise 5.4. What is u and v if

$$\begin{bmatrix} u+2v & 1 & 3 \\ 9 & 0 & 4 \\ 3 & 4 & 7 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 3 \\ 9 & 0 & u+v \\ 3 & 4 & 7 \end{bmatrix}?$$

Exercise 5.5. Let v_1, v_2, v_3, v_4 represent cities and suppose there are one-way flights from v_1 to v_2 and v_3, from v_2 to v_3 and v_4, and two-way flights between v_1 and v_4. Write out a matrix A such that $(A)_{ij} = 1$ if there is a flight from v_i to v_j, and zero otherwise.

Exercise 5.6. Let $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$. Is $A = B$?

Exercise 5.7. If $2A = \begin{bmatrix} 3 & 4 \\ 2 & 8 \\ 1 & 5 \end{bmatrix}$, what is A? If $B - \frac{1}{2}\begin{bmatrix} 3 & 4 \\ 1 & 8 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 4 \\ 2 & 5 \\ 3 & 1 \end{bmatrix}$, what is B?

Exercise 5.8. Which of the following matrices are symmetric?

(a) $\begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 5 & 4 & b \\ 3 & 4 & 3 & 3 \\ 5 & b & 3 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 3 & 5 \\ 2 & 5 & 4 & b \\ 3 & 4 & 3 & 3 \\ 5 & b & 3 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$

Exercise 5.9. True or False?

i. Symmetric matrices must be square.
ii. A scalar is symmetric.
iii. If A is symmetric, then αA is symmetric.
iv. The sum of symmetric matrices is symmetric.
v. All diagonal matrices are symmetric.
vi. If $(A^{\mathrm{T}})^{\mathrm{T}} = A$, then A is symmetric.

Exercise 5.10. (a) Find A and B if they simultaneously satisfy

$$2A + B = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 3 & 0 \end{bmatrix} \quad \text{and} \quad A + 2B = \begin{bmatrix} 4 & 2 & 3 \\ 5 & 1 & 1 \end{bmatrix} .$$

(b) If $A + B = C$ and $3A - 2B = 0$ simultaneously, find A and B in terms of C.

5.2 Matrix Multiplication

Let A be $m \times n$ and B be $n \times p$ — here we require the number of columns in A and the number of rows in B to be the same. Then the product AB is defined as the $m \times p$ matrix whose (i, j)th element is

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} .$$

That is, the (i, j)th element of the product AB is defined as the sum of the product of the elements of the ith row of A with the corresponding elements in the jth column of B. Put another way, the $(i, j)th$ element of the product AB is the dot or inner product of the ith row of A with the jth column of B. For example, the $(1, 1)$th element of AB is

$$(AB)_{11} = \sum_{k=1}^{n} a_{1k} b_{k1} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} + \cdots + a_{1n}b_{n1} .$$

The $(2, 3)$th element of AB is

$$(AB)_{23} = \sum_{k=1}^{n} a_{2k} b_{k3} = a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33} + \cdots + a_{2n}b_{n3} ,$$

Visually, for a product of a 3×3 matrix and a 3×2 matrix, we have

$$\begin{bmatrix} \boxed{a_{11} & a_{12} & a_{13}} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \boxed{b_{11}} & b_{12} \\ \boxed{b_{21}} & b_{22} \\ \boxed{b_{31}} & b_{32} \end{bmatrix} = \begin{bmatrix} \boxed{a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}} & \bullet \\ \bullet & \bullet \\ \bullet & \bullet \end{bmatrix}$$

$$\begin{bmatrix} \boxed{a_{11} & a_{12} & a_{13}} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} & \boxed{b_{12}} \\ b_{21} & \boxed{b_{22}} \\ b_{31} & \boxed{b_{32}} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{3} a_{1k}b_{k1} & \boxed{a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32}} \\ \bullet & \bullet \\ \bullet & \bullet \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \boxed{a_{21} & a_{22} & a_{23}} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} \boxed{b_{11}} & b_{12} \\ \boxed{b_{21}} & b_{22} \\ \boxed{b_{31}} & b_{32} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{3} a_{1k}b_{k1} & \sum_{k=1}^{3} a_{1k}b_{k2} \\ \boxed{a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31}} & \bullet \\ \bullet & \bullet \end{bmatrix}$$

and so on.

Example 5.7. Let $A = \begin{bmatrix} 2 & 8 \\ 3 & 0 \\ 5 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 4 & 7 \\ 6 & 9 \end{bmatrix}$. Then

$$AB = \begin{bmatrix} 2 & 8 \\ 3 & 0 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 4 & 7 \\ 6 & 9 \end{bmatrix} = \begin{bmatrix} 2 \cdot 4 + 8 \cdot 6 & 2 \cdot 7 + 8 \cdot 9 \\ 3 \cdot 4 + 0 \cdot 6 & 3 \cdot 7 + 0 \cdot 9 \\ 5 \cdot 4 + 1 \cdot 6 & 5 \cdot 7 + 1 \cdot 9 \end{bmatrix} = \begin{bmatrix} 56 & 86 \\ 12 & 21 \\ 26 & 44 \end{bmatrix} .$$

Example 5.8. The system of equations

$$\begin{aligned} 2x_1 - \ x_2 &= 4 \\ x_1 + 2x_2 &= 2 \end{aligned}$$

can be written in matrix form as

$$\underbrace{\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 4 \\ 2 \end{bmatrix}}_{b}, \text{ or } Ax = b\,.$$

5.2.1 Exercises

These exercises illustrate crucial aspects of matrix multiplication. You should work through the exercises before proceeding to the next section.

Exercise 5.11. Let $A = \begin{bmatrix} 2 & 8 \\ 3 & 0 \\ 5 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 0 \\ 3 & 8 \end{bmatrix}$ and $C = \begin{bmatrix} 7 & 2 \\ 6 & 3 \end{bmatrix}$.

(a) Compute the products BC, CB, and AB. (b) Can BA even be computed?

Remark: This exercise shows that for any two matrices A and B, $AB \neq BA$ in general. That is, we have to distinguish between pre-multiplication and post-multiplication. In the product AB, we say that B is pre-multiplied by A, or that A is post-multiplied by B.

Exercise 5.12. Show that $x^{\mathrm{T}}x \geq 0$ for any vector $x = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^{\mathrm{T}}$. When will $x^{\mathrm{T}}x = 0$?
Remark: For any column vector x, the product $x^{\mathrm{T}}x$ is the sum of the squares of its elements. We call it the dot or inner product of the column vector x with itself.

Exercise 5.13.

(a) Compute $\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -2 & 4 \\ 1 & -2 \end{bmatrix}$. (b) Compute $A^2 = AA$ where $A = \begin{bmatrix} 1 & b \\ -\frac{1}{b} & -1 \end{bmatrix}$, $b \neq 0$

Remark: This exercise shows that you can multiply two non-zero matrices and end up with a zero matrix. Therefore $AB = 0$ does **not** *imply $A = 0$ or $B = 0$. It is even possible for the square of a non-zero matrix to be a zero matrix. Of course, if $A = 0$ or $B = 0$, then $AB = 0$.*

As you can see, in many ways matrix multiplication does not behave like the usual multiplication of numbers. For instance, the order of multiplication matters, and $AB = 0$ does not imply $A = 0$ or $B = 0$. But in some ways matrix multiplication *does* behave like regular multiplication of numbers, as the next exercise shows.

Exercise 5.14. Prove that

(a) $(AB)C = A(BC)$ where A, B, and C are $m \times n$, $n \times p$ and $p \times q$ respectively.

(b) $A(B + C) = AB + AC$ where A is $m \times n$, and B and C are $n \times p$.

(c) $(A + B)C = (AC + BC)$ where A and B are $m \times n$ and C is $n \times p$.

Exercise 5.15. Let A be an $m \times n$ matrix, and let I_n and I_m be identity matrices of dimensions $n \times n$ and $m \times m$ respectively. Show that $I_m A = AI_n = A$.

Exercise 5.16. Show that

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = b_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{41} \end{bmatrix} + b_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \\ a_{42} \end{bmatrix} + b_3 \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \\ a_{43} \end{bmatrix}$$

i.e., Ab is a linear combination of the columns of A, with weights given in b.

Exercise 5.17. (a) Show that $(AB)^{\mathrm{T}} = B^{\mathrm{T}}A^{\mathrm{T}}$ for any $m \times n$ matrix A and any $n \times p$ matrix B. Verify this equality for the matrices

$$A = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{bmatrix}.$$

(b) Prove that $(ABC)^{\mathrm{T}} = C^{\mathrm{T}}B^{\mathrm{T}}A^{\mathrm{T}}$.

Exercise 5.18. Explain why $X^{\mathrm{T}}X$ is square and symmetric for any general $n \times k$ matrix X.

Remark: The matrix $X^{\mathrm{T}}X$ is encountered frequently in all statistical disciplines.

Exercise 5.19. The **trace** of an $n \times n$ matrix $A = (a_{ij})_{n\times n}$ is defined to be

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

That is, the trace of a square matrix is simply the sum of its diagonal elements. The trace of a scalar is the scalar itself.

(a) If A and B are square matrices of the same dimensions, show that $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$,.

(b) If A is a square matrix, show that $\mathrm{tr}(A^{\mathrm{T}}) = \mathrm{tr}(A)$.

(c) If A is $m \times n$ and B is $n \times m$, show that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

(d) If x is an $n \times 1$ column vector, show that $x^{\mathrm{T}}x = \mathrm{tr}(xx^{\mathrm{T}})$ by

i. direct multiplication,
ii. using (c) and the fact that the trace of a scalar is the scalar itself.

The trace operation is surprisingly useful in proofs and for deriving and simplifying matrix equations.

Exercise 5.20. Let i_n be an $n \times 1$ vector of ones, i.e., $i_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^{\mathrm{T}}$.

(a) Show that the formula for the sample mean of the elements of the column vector $y = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^{\mathrm{T}}$ can be written as $\overline{y} = (i_n^{\mathrm{T}} i_n)^{-1} i_n^{\mathrm{T}} y$.

(b) Show that $M_0 = I_n - i_n(i_n^{\mathrm{T}} i_n)^{-1} i_n^{\mathrm{T}}$ is symmetric, and that $M_0 M_0 = M_0$.

(c) Show that the sample variance of the data in y can be written as

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{y^{\mathrm{T}} M_0 y}{n-1}.$$

Exercise 5.21. Prove that $A(\alpha B) = (\alpha A)B = \alpha(AB)$.

5.3 Partitioned Matrices

We can partition the contents of an $m \times n$ matrix into blocks of submatrices. For instance, we can write

$$A = \begin{bmatrix} 1 & 3 & 2 & 6 \\ 2 & 8 & 2 & 1 \\ 3 & 1 & 2 & 4 \\ 4 & 2 & 1 & 3 \\ 3 & 1 & 1 & 7 \end{bmatrix} = \left[\begin{array}{cc|cc} 1 & 3 & 2 & 6 \\ 2 & 8 & 2 & 1 \\ \hline 3 & 1 & 2 & 4 \\ 4 & 2 & 1 & 3 \\ 3 & 1 & 1 & 7 \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where

$$A_{11} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, A_{21} = \begin{bmatrix} 3 \\ 4 \\ 3 \end{bmatrix}, A_{12} = \begin{bmatrix} 3 & 2 & 6 \\ 8 & 2 & 1 \end{bmatrix} \text{ and } A_{22} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 1 & 3 \\ 1 & 1 & 7 \end{bmatrix}.$$

Partitioned matrices are often called **block matrices**. Of course, there are many ways of partitioning any given matrix. The following is another partition of the matrix A:

$$A = \begin{bmatrix} 1 & 3 & 2 & 6 \\ 2 & 8 & 2 & 1 \\ 3 & 1 & 2 & 4 \\ 4 & 2 & 1 & 3 \\ 3 & 1 & 1 & 7 \end{bmatrix} = \left[\begin{array}{cc|cc} 1 & 3 & 2 & 6 \\ 2 & 8 & 2 & 1 \\ 3 & 1 & 2 & 4 \\ \hline 4 & 2 & 1 & 3 \\ 3 & 1 & 1 & 7 \end{array}\right].$$

It can be shown that addition and multiplication of partitioned matrices can be carried out as though the blocks are elements, as long as the matrices are partitioned conformably.

Addition of Partitioned Matrices. Consider two $m \times n$ matrices A and B partitioned in the following manner:

$$A = \begin{bmatrix} \underbrace{A_{11}}_{m_1 \times n_1} & \underbrace{A_{12}}_{m_1 \times n_2} \\ \underbrace{A_{21}}_{m_2 \times n_1} & \underbrace{A_{22}}_{m_2 \times n_2} \end{bmatrix} \text{ and } B = \begin{bmatrix} \underbrace{B_{11}}_{m_1 \times n_1} & \underbrace{B_{12}}_{m_1 \times n_2} \\ \underbrace{B_{21}}_{m_2 \times n_1} & \underbrace{B_{22}}_{m_2 \times n_2} \end{bmatrix}$$

where $n_1 + n_2 = n$ and $m_1 + m_2 = m$. We emphasize that A and B must be of the same size and partitioned identically. Then

$$A + B = \begin{bmatrix} \underbrace{A_{11} + B_{11}}_{m_1 \times n_1} & \underbrace{A_{12} + B_{12}}_{m_1 \times n_2} \\ \underbrace{A_{21} + B_{21}}_{m_2 \times n_1} & \underbrace{A_{22} + B_{22}}_{m_2 \times n_2} \end{bmatrix}. \tag{5.1}$$

Multiplication of Partitioned Matrices. Now consider two matrices A and B with dimensions $m \times p$ and $p \times n$ respectively, are partitioned as follows:

$$A = \begin{bmatrix} \underbrace{A_{11}}_{m_1 \times p_1} & \underbrace{A_{12}}_{m_1 \times p_2} \\ \underbrace{A_{21}}_{m_2 \times p_1} & \underbrace{A_{22}}_{m_2 \times p_2} \end{bmatrix} \text{ and } B = \begin{bmatrix} \underbrace{B_{11}}_{p_1 \times n_1} & \underbrace{B_{12}}_{p_1 \times n_2} \\ \underbrace{B_{21}}_{p_2 \times n_1} & \underbrace{B_{22}}_{p_2 \times n_2} \end{bmatrix}.$$

In particular, the partition is such that the column-wise partition of A matches the row-wise

partition of B. Then

$$AB = \begin{bmatrix} \underbrace{A_{11}}_{m_1\times p_1} & \underbrace{A_{12}}_{m_1\times p_2} \\ \underbrace{A_{21}}_{m_2\times p_1} & \underbrace{A_{22}}_{m_2\times p_2} \end{bmatrix} \begin{bmatrix} \underbrace{B_{11}}_{p_1\times n_1} & \underbrace{B_{12}}_{p_1\times n_2} \\ \underbrace{B_{21}}_{p_2\times n_1} & \underbrace{B_{22}}_{p_2\times n_2} \end{bmatrix} = \begin{bmatrix} \underbrace{A_{11}B_{11}+A_{12}B_{21}}_{m_1\times n_1} & \underbrace{A_{11}B_{12}+A_{12}B_{22}}_{m_1\times n_2} \\ \underbrace{A_{21}B_{11}+A_{22}B_{21}}_{m_2\times n_1} & \underbrace{A_{21}B_{12}+A_{22}B_{22}}_{m_2\times n_2} \end{bmatrix}. \tag{5.2}$$

Transposition of Partitioned Matrices. It is straightforward to show that

$$A = \begin{bmatrix} \underbrace{A_{11}}_{m_1\times n_1} & \underbrace{A_{12}}_{m_1\times n_2} \\ \underbrace{A_{21}}_{m_2\times n_1} & \underbrace{A_{22}}_{m_2\times n_2} \end{bmatrix} \quad \Rightarrow \quad A^{\mathrm{T}} = \begin{bmatrix} \underbrace{A_{11}^{\mathrm{T}}}_{n_1\times m_1} & \underbrace{A_{21}^{\mathrm{T}}}_{n_1\times m_2} \\ \underbrace{A_{12}^{\mathrm{T}}}_{n_2\times m_1} & \underbrace{A_{22}^{\mathrm{T}}}_{n_2\times m_2} \end{bmatrix}. \tag{5.3}$$

Remark on Matrix Multiplication: So far we have spoken of inner products of vectors, scalar multiplication (multiplication of of matrices and vectors with a scalar), and regular matrix multiplication. There are yet other kinds of matrix multiplication concepts. For instance, the **Hadamard product**, denoted $\circ$ or $\odot$, refers to element-wise multiplication, e.g.,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \odot \begin{bmatrix} 2 & 3 \\ 4 & 5 \\ 6 & 7 \end{bmatrix} = \begin{bmatrix} 1\cdot 2 & 2\cdot 3 \\ 3\cdot 4 & 4\cdot 5 \\ 5\cdot 6 & 6\cdot 7 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 12 & 20 \\ 30 & 42 \end{bmatrix}.$$

The **Kronecker product**, denoted $\otimes$, of an $m \times n$ matrix A with a $p \times q$ matrix B is the $mp \times nq$ block matrix formed by multiplying each element of A by the entire B matrix. For example

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \left[\begin{array}{cc|cc|cc} a_{11} & 0 & a_{12} & 0 & a_{13} & 0 \\ 0 & a_{11} & 0 & a_{12} & 0 & a_{13} \\ \hline a_{21} & 0 & a_{22} & 0 & a_{23} & 0 \\ 0 & a_{21} & 0 & a_{22} & 0 & a_{23} \end{array}\right].$$

5.3.1 Exercises

Exercise 5.22. Let

$$A = \left[\begin{array}{c|ccc} 1 & 3 & 2 & 6 \\ 2 & 8 & 2 & 1 \\ \hline 3 & 1 & 2 & 4 \\ 4 & 2 & 1 & 3 \\ 3 & 1 & 1 & 7 \end{array}\right] \quad \text{and} \quad B = \left[\begin{array}{c|cc} 2 & 0 & 1 \\ \hline 3 & 1 & 3 \\ 1 & 5 & 4 \\ 4 & 1 & 1 \end{array}\right].$$

Verify the partitioned matrix multiplication formulas by computing AB in the usual way, then compute AB using (5.2). Verify the transposition formula (5.3) for matrix A.

Exercise 5.23. Let A be a $m \times n$ matrix and b be a $n \times 1$ vector. We have shown earlier that Ab is a linear combination of the columns of A. In terms of partitioned matrices, we have

$$Ab = \left[\begin{array}{c|c|c|c} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array}\right] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} A_{*1} & A_{*2} & \cdots & A_{*n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = A_{*1}b_1 + A_{*2}b_2 + \cdots + A_{*n}b_n$$

Let $c = \begin{bmatrix} c_1 & c_2 & \ldots & c_m \end{bmatrix}^{\mathrm{T}}$. Show that $c^{\mathrm{T}}A$ is a linear combination of the rows of A.

Exercise 5.24. Let X be a $n \times 3$ data matrix containing n observations of three variables:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

where x_{ij} represents the ith observation of variable j. We can partition this matrix to emphasize the variables by writing X as $X = \begin{bmatrix} X_{*1} & X_{*2} & X_{*3} \end{bmatrix}$ where

$$X_{*1} = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{n1} \end{bmatrix}, \; X_{*2} = \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ \vdots \\ x_{n2} \end{bmatrix} \text{ and } \; X_{*3} = \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \\ \vdots \\ x_{n3} \end{bmatrix}.$$

Alternatively, we can partition the data matrix to emphasize the observations:

$$X = \begin{bmatrix} X_{1*} \\ X_{2*} \\ X_{3*} \\ \vdots \\ X_{n*} \end{bmatrix}$$

where $X_{i*} = \begin{bmatrix} x_{i1} & x_{i2} & x_{i3} \end{bmatrix}$ is the row vector containing the ith observations of all three variables, $i = 1, 2, ..., n$. Show that the matrix $X^{\mathrm{T}}X$ can be written as

$$X^{\mathrm{T}}X = \begin{bmatrix} X_{*1}^{\mathrm{T}}X_{*1} & X_{*1}^{\mathrm{T}}X_{*2} & X_{*1}^{\mathrm{T}}X_{*3} \\ X_{*2}^{\mathrm{T}}X_{*1} & X_{*2}^{\mathrm{T}}X_{*2} & X_{*2}^{\mathrm{T}}X_{*3} \\ X_{*3}^{\mathrm{T}}X_{*1} & X_{*3}^{\mathrm{T}}X_{*2} & X_{*3}^{\mathrm{T}}X_{*3} \end{bmatrix} = \sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i1}x_{i3} \\ \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & \sum_{i=1}^{n} x_{i2}x_{i3} \\ \sum_{i=1}^{n} x_{i1}x_{i3} & \sum_{i=1}^{n} x_{i2}x_{i3} & \sum_{i=1}^{n} x_{i3}^2 \end{bmatrix}$$

5.4 Introduction to Inverses and Determinants

5.4.1 The Inverse Matrix

The $n \times m$ matrix B is said to be a **left-inverse** of a $m \times n$ matrix A if $BA = I_n$. The $n \times m$ matrix C is a **right-inverse** of A if $AC = I_m$. If A is $n \times n$, and $BA = AC = I_n$, then it must be the case that $B = C$ since

$$BA = I_n \Rightarrow BAC = I_nC \Rightarrow BI_n = C \Rightarrow B = C.$$

In this case, we call $B = C$ the **two-sided inverse**, or simply the **inverse of A, and give it the special notation A^{-1}. That is, the inverse of a $n \times n$ matrix A, *if it exists*, is the unique matrix A^{-1} such that

$$A^{-1}A = I_n = AA^{-1}.$$

We could leave out the second equality from the definition, since as we have already shown, $A^{-1}A = I \Rightarrow AA^{-1} = I$.

Example 5.9. The inverse of the matrix

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \quad \text{is} \quad A^{-1} = -\frac{1}{2}\begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}.$$

This can be verified by direct multiplication:

$$A^{-1}A = -\frac{1}{2}\begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We do not have to show $AA^{-1} = I_2$, since it is implied. You may wish to do so nonetheless, as an exercise.

Example 5.10. Let A and B be the matrices

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -1 & 0.2 & 0.4 \\ 2 & -0.2 & -0.4 \end{bmatrix}.$$

You can easily verify (by direct multiplication) that

$$BA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{but} \quad AB = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.2 & 0.4 \\ 0 & 0.4 & 0.8 \end{bmatrix}.$$

The matrix B is a left-inverse of A. We give left-inverses the special notation A^{-1}_{left}. Likewise, right-inverses are given the special notation A^{-1}_{right}. We will say more about left- and right-inverses in a later chapter. For this chapter we will focus on (two-sided) inverses. The term "inverse" will always mean a two-sided inverse.

We emphasize that A has a (two-sided) inverse only if it is square. Furthermore, not all square matrices have an inverse. The inverse of an arbitrary 2×2 matrix $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, if it exists, is

$$A^{-1} = \frac{1}{\det(A)}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \quad \text{where } \det(A) = a_{11}a_{22} - a_{12}a_{21}. \tag{5.4}$$

You can easily verify this by direct multiplication. It is worth your while to commit formula (5.4) to memory.

The expression $\det(A)$ in (5.4) is called the **determinant** of the 2×2 matrix A. Notice that the inverse exists only if $\det(A) \neq 0$. If the inverse of A does not exist, we say that A is **singular**. If the inverse exists, we say that A is **non-singular**. An alternative notation for $\det(A)$ is $|A|$. We will use both notations in this book. In particular, we use the latter when indicating the determinant of a matrix written out in full. For instance, the determinant of the matrix $(a_{ij})_{2\times 2}$ is

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}\,.$$

Example 5.11. The inverse of the matrix $A = \begin{bmatrix} 1 & 4 \\ 5 & 6 \end{bmatrix}$ is

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} 6 & -4 \\ -5 & 1 \end{bmatrix} = -\frac{1}{14} \begin{bmatrix} 6 & -4 \\ -5 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{3}{7} & \frac{2}{7} \\ \frac{5}{14} & -\frac{1}{14} \end{bmatrix}.$$

Example 5.12. The determinant of the matrix $A = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}$ is $\det(A) = 1 \cdot 6 - 2 \cdot 3 = 0$, so A does not have an inverse.

When will $\det(A) = 0$? Examining the expression for $\det(A)$ in (5.4), we see that the determinant will be zero if one or both rows or columns are all zero, or if one row is a multiple of the other, or if one column is a multiple of the other.

The inverse of a scalar is obviously just its reciprocal. The following example shows the inverse of a particular 3×3 matrix.

Example 5.13. The inverse of $A = \begin{bmatrix} 0 & 2 & 4 \\ 3 & 1 & 2 \\ 6 & 2 & 1 \end{bmatrix}$ is $A^{-1} = \begin{bmatrix} -\frac{1}{6} & \frac{1}{3} & 0 \\ \frac{1}{2} & -\frac{4}{3} & \frac{2}{3} \\ 0 & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}$. This can be seen by direct multiplication:

$$\begin{bmatrix} -\frac{1}{6} & \frac{1}{3} & 0 \\ \frac{1}{2} & -\frac{4}{3} & \frac{2}{3} \\ 0 & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 0 & 2 & 4 \\ 3 & 1 & 2 \\ 6 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We'll omit from these notes any discussion as to how to find the inverse of a general $n \times n$ square matrix, see Tay, Preve, and Baydur (2025) for details. Nonetheless, even without seeing the formula or algorithms for computing the inverse of a matrix, we are able to prove the following general statements. Suppose the $n \times n$ matrices A and B are non-singular, i.e., their inverses exist. Then

$$\text{i. } (A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}, \quad \text{ii. } (AB)^{-1} = B^{-1}A^{-1}.$$

Proof: For i., start with $AA^{-1} = I$. Transpose both sides to get $(A^{-1})^{\mathrm{T}}A^{\mathrm{T}} = I$. Finally post-multiply both sides by $(A^{\mathrm{T}})^{-1}$ to get

$$(A^{-1})^{\mathrm{T}}A^{\mathrm{T}}(A^{\mathrm{T}})^{-1} = I(A^{\mathrm{T}})^{-1} \Rightarrow (A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}.$$

For ii., pre-multiply AB first by A^{-1} and then by B^{-1}. This gives

$$A^{-1}AB = B$$
$$B^{-1}A^{-1}AB = B^{-1}B = I.$$

This says that $B^{-1}A^{-1}$ is the inverse of AB since multiplying the two gives the identity matrix.

One implication of the first result is that the inverse of a symmetric matrix is symmetric: if A is symmetric, then $A^{\mathrm{T}} = A$, so we have

$$(A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1} = A^{-1}$$

which says that A^{-1} is symmetric. For the second result, it is important to keep in mind that this result holds only if A and B are both square. It is possible for A to be $n \times k$ and B to be $k \times n$ such that the square matrix AB is non-singular. But since A and B are not square, they do not have inverses. In that case the statement $(AB)^{-1} = B^{-1}A^{-1}$ is obviously meaningless.

5.4.2 Systems of Linear Equations

One application of matrix inverses is to find solutions to systems of linear equations. Consider a system of n equations in n unknowns x_1, x_2, ..., x_n,

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array} \tag{5.5}$$

which can be written as

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \quad \text{or} \quad Ax = b\,.$$

To be clear, we are speaking here of systems where there are as many equations as there are unknowns. If the inverse of A exists, then the system has a unique solution:

$$Ax = b \;\Rightarrow\; A^{-1}Ax = A^{-1}b \;\Rightarrow\; x = A^{-1}b\,.$$

Example 5.14. Consider the following systems of equations

$$\text{(i)}\ \begin{array}{l} 2x_1 - x_2 = 4 \\ x_1 + 2x_2 = 2 \end{array} \qquad \text{(ii)}\ \begin{array}{l} 2x_1 + x_2 = 4 \\ 6x_1 + 3x_2 = 12 \end{array} \qquad \text{(iii)}\ \begin{array}{l} 2x_1 + x_2 = 4 \\ 6x_1 + 3x_2 = 10 \end{array} \tag{5.6}$$

You can see that system (i) has a unique solution. System (ii) has infinitely many solutions (the graphs of the two equations coincide). System (iii) has no solution; the graphs of the two equations are parallel. The three systems can be written in the matrix form $Ax = b$:

$$\text{(i)}\ \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \text{(ii)}\ \begin{bmatrix} 2 & 1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 12 \end{bmatrix} \quad \text{(iii)}\ \begin{bmatrix} 2 & 1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \end{bmatrix}$$

Since

$$\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}^{-1} = \frac{1}{5} \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$$

the unique solution for system (i) is

$$x = A^{-1}b = \frac{1}{5}\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

For systems (ii) and (iii), we find that the coefficient matrix A does not have an inverse, since

$$\det\begin{bmatrix} 2 & 1 \\ 6 & 3 \end{bmatrix} = 2 \cdot 3 - 1 \cdot 6 = 0\,.$$

Notice that non-existence of the coefficient matrix inverse does not imply that there are no solutions. It could be that there are multiple solutions.

5.4.3 The Determinant and Cramer's Rule

Consider now the general 2×2 system

$$\begin{array}{rcrcl} a_{11}x_1 & + & a_{12}x_2 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & = & b_2 \end{array} \quad \text{or} \quad Ax = b \tag{5.7}$$

Solving this system gives

$$x_1 = \frac{a_{22}\,b_1 - a_{12}\,b_2}{a_{11}a_{22} - a_{12}a_{21}} \quad \text{and} \quad x_2 = \frac{a_{11}\,b_2 - a_{21}\,b_1}{a_{11}a_{22} - a_{12}a_{21}}\,.$$

Of course, this is the solution only if the (common) denominator in both expressions is not zero. The denominator is just the determinant of the matrix A. Notice also that the numerators of the solutions for x_1 and x_2 are, respectively, the determinants of the matrices

$$A_1(b) = \begin{bmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{bmatrix} \quad \text{and} \quad A_2(b) = \begin{bmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{bmatrix}\,.$$

These are just the matrix A with one column replaced by b. This is **Cramer's Rule** for systems of two equations in two unknowns: for system (5.7), the solutions are

$$x_1 = \frac{\det(A_1(b))}{\det(A)} = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} \quad \text{and} \quad x_2 = \frac{\det(A_2(b))}{\det(A)} = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}\,.$$

The idea extends to larger systems of equations with as many equations as unknowns. If you work out the solutions for the general three-equations three-unknowns system

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & b_3 \end{array}$$

you will find the solutions to be

$$x_1 = \frac{b_1a_{22}a_{33} + a_{12}a_{23}b_3 + a_{13}b_2a_{32} - a_{13}a_{22}b_3 - b_1a_{23}a_{32} - a_{12}b_2a_{33}}{a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}}$$

$$x_2 = \frac{a_{11}b_2a_{33} + b_1a_{23}a_{31} + a_{13}a_{21}b_3 - a_{13}b_2a_{31} - a_{11}a_{23}b_3 - b_1a_{21}a_{33}}{a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}}$$

$$x_3 = \frac{a_{11}a_{22}b_3 + a_{12}b_2a_{31} + b_1a_{21}a_{32} - b_1a_{22}a_{31} - a_{11}b_2a_{32} - a_{12}a_{21}b_3}{a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}}$$

You do not want to memorize this solution, at least not in this form. But notice two things: first, the denominator is the same for all three expressions. We define the expression in the denominator to be the determinant of the 3×3 coefficient matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

We must have $\det(A) \neq 0$ in order for there to be a unique solution. Second, using this definition for the determinant, the numerators in the solutions for x_1, x_2 and x_3 are, respectively, the determinants of the matrices

$$A_1(b) = \begin{bmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{bmatrix}, \; A_2(b) = \begin{bmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{bmatrix} \text{ and } A_3(b) = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{bmatrix}.$$

This gives **Cramer's Rule** for systems of three equations in three unknowns:

$$x_1 = \frac{\det(A_1(b))}{\det(A)}, \; x_2 = \frac{\det(A_2(b))}{\det(A)} \text{ and } x_3 = \frac{\det(A_3(b))}{\det(A)}.$$

The determinant for larger square matrices can be thought of in a similar way, as the (common) denominator in the solutions to the general n-equations in n-unknowns system $Ax = b$. Furthermore, the solution to such a system is

$$x_i = \frac{\det(A_i(b))}{\det(A)}, \; i = 1, 2, \ldots, n$$

where $A_i(b)$ is the determinant of the matrix A with the ith column replaced by b. See Tay, Preve, and Baydur (2025) for details on how determinants can be computed.

The following properties of determinants are useful:

i. if A has a row of zeros or a column of zeros, then $\det(A) = 0$.
ii. if a single row or column of A is multiplied by some constant α, then its determinant is multiplied by α.
iii. The determinant of a triangular matrix is the product of its diagonal elements.
iv. $\det(A^{\mathrm{T}}) = \det(A)$.
v. Every time we swap the rows of a matrix, its determinant changes sign. Same for columns.
vi. Adding a multiple of one row to another row does not change the determinant. Same for columns.
vii. If A and B are two square matrices, then $\det(AB) = \det(A)\det(B)$.

5.4.4 Exercises

Exercise 5.25. Find the inverse of the transpose of $A = \begin{bmatrix} 0 & 2 & 4 \\ 3 & 1 & 2 \\ 6 & 2 & 1 \end{bmatrix}$. (Hint: see Example 5.13.)

Exercise 5.26. Show that the inverse of a diagonal matrix $A = \text{diag}(a_{11}, \ldots, a_{nn})$ is the diagonal matrix

$$A^{-1} = \text{diag}\left(\frac{1}{a_{11}}, \ldots, \frac{1}{a_{nn}}\right).$$

Exercise 5.27. Suppose one row of a (square) matrix is a multiple of another row. Explain why this matrix has no inverse.

Exercise 5.28. Consider the following system of equations

$$\begin{array}{rcrcrcl} 4x_1 & + & & + & x_3 & = & 4 \\ 8x_1 & + & x_2 & + & -3x_3 & = & 3 \\ 12x_1 & + & x_2 & + & & = & 1 \end{array}$$

a. Express this system in the form $Ax = b$ and solve it by finding A^{-1} and then computing $A^{-1}b$.

b. Verify your solution in a. by solving the system using Cramer's Rule.

Exercise 5.29. Suppose A is an $m \times m$ matrix and b and c are $m \times 1$ vectors. Does $Ab = Ac$ imply that $b = c$? If no, give a counterexample.

5.5 Matrix Definiteness

A $n \times n$ symmetric matrix A is said to be **positive definite** if

$$x^{\mathrm{T}}Ax > 0 \text{ for all } n\text{-vectors } x \neq 0_n. \tag{5.8}$$

If the inequality in (5.8) is non-strict, then A is **positive semidefinite**. If the inequality in (5.8) is reversed, A is **negative definite**. If it is reversed and made non-strict, then A is called **negative semidefinite**. We emphasize that the conditions must hold for *all* non-zero vectors x. Expressions of the form $x^{\mathrm{T}}Ax$ where x is $n \times 1$ and A is $n \times n$ and symmetric are called quadratic forms.

Example 5.15. The matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ is positive definite since

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2(x_1^2 + x_1x_2 + x_2^2) = 2[(x_1 + 0.5x_2)^2 + 0.75x_2^2] > 0$$

as long as x_1 and x_2 are not both zero.

Example 5.16. The matrix $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ is indefinite (not definite) since

$$Q = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 4x_1x_2 + x_2^2.$$

If $x_1 = 1$ and $x_2 = 1$, then $Q > 0$. If $x_1 = 1$ and $x_2 = -1$, then $Q < 0$.

We will see later that "variance-covariance matrices" are always at least positive semidefiniteness, often positive definite (Section 5.7). The positive or negative definiteness of the "Hessian" of a multivariable function is also an indicator of whether a function is convex or concave, which in turn plays an important role in function optimization. Definiteness of matrices also play an important role in matrix factorizations, dynamic systems, and many other areas where matrix algebra is used. One method for checking the definiteness of matrices uses the determinants of certain submatrices of the matrix, called principal minors. Another uses eigenvalues. See Tay, Preve, and Baydur (2025) for details. Often we are able to surmise the definiteness of a matrix from its structure, as in Exercise 5.30.

5.5.1 Exercises

Exercise 5.30. Suppose X is $n \times k$. Explain why the matrix $X^{\mathrm{T}}X$ is positive semidefinite. Explain why it is positive definite if $Xc \neq 0$ for all k-vectors $c \neq 0_k$. (The next section explains the significance of the condition $Xc \neq 0$ for all k-vectors $c \neq 0_n$.) *Hint: Consider the expression* $c^{\mathrm{T}}X^{\mathrm{T}}Xc$.

5.6 The Rank of a Matrix

A point x in $\mathbb{R}^m$ can be thought of as a m-dimensional vector, or "m-vector". If $X = \{x_1, x_2, \dots, x_n\}$ is a set of n m-vectors, and if at least one of these vectors can be written as a linear combination of the others, i.e., if

$$x_i = c_1x_1 + \cdots + c_{i-1}x_{i-1} + c_{i+1}x_{i+1} + \cdots + c_nx_n\,,$$

then we say that the vectors are linearly dependent. Another way of saying this is that we can find $c_1, c_2, \dots, c_n$, not all equal to zero, such that

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n = 0\,.$$

If we cannot express any vector in X as a linear combination of the other vectors, then the vectors in X are linearly independent. In that case, the vectors in X will satisfy the condition

$$c_1x_1 + c_2x_2 + \cdots + c_nx_n = 0 \quad \Rightarrow \quad c_1 = c_2 = \cdots = c_n = 0\,.$$

A vector space or subspace is a set of vectors such that linear combinations of vectors in the space always result in a vector in the space. Every vector space or subspace must contain the zero vector. The set of *all* linear combinations of the vectors in X is a vector subspace of $\mathbb{R}^m$. The dimension of this subspace cannot exceed $\min\{m, n\}$. Finally, recall that two vectors are orthogonal if their inner product is zero.

Consider an $m \times n$ matrix A, where possibly $m \neq n$. We can view the columns of A as a collection of n m-vectors:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \left[\begin{array}{c|c|c|c} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array}\right].$$

Linear combinations of the column vectors of A can be written as Ax where x is some n-vector. If we consider the function

$$y = f(x) = Ax\,,\ x \in \mathbb{R}^n \tag{5.9}$$

mapping n-vectors into m-vectors, then the range of this function is the set of all linear combinations of the columns of A, spanning a vector subspace of $\mathbb{R}^m$ of dimension $r \leq \min\{m, n\}$. We call this subspace the **column space** of A and refer to r as the **column rank** of A.

Likewise, we can view the rows of A as a collection of m n-vectors, i.e.,

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \left[\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline a_{21} & a_{22} & \cdots & a_{2n} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline a_{m1} & a_{m2} & \cdots & a_{mn} \end{array}\right].$$

Linear combinations of the m row vectors can be written as $y^T A$ or $A^{\mathrm{T}} y$ where y is an m-vector. The range of the function

$$x = g(y) = A^{\mathrm{T}} y\,,\ y \in \mathbb{R}^m \tag{5.10}$$

is the column space of A^{T}, which is also the **row space** of A, since the columns of A^{T} are the rows of A. The dimension of the row space is called the **row rank** of A.

It turns out that for any matrix A, the row and column ranks of A are the same. Suppose the column rank of A is r. This means we can find r linearly independent columns in A. Gather these columns into a $m \times r$ matrix C. Since every column of A can be written as a linear combination of the r columns in C, we can write $A = CR$ where R is $r \times n$, each column containing the necessary weights to generate the corresponding columns of A as a linear combination of the vectors in C. However, the fact that $A = CR$ also means that every *row* of A is a linear combination of the rows of R, the necessary weights appearing in the corresponding rows of C. Since R has r rows, the row rank of A also cannot exceed r, i.e.,

$$\text{row rank}(A) \leq r = \text{column rank}(A)\,.$$

Applying a similar argument to A^{T} shows that the row rank of A^{T} must be less than or equal to the column rank of A^{T}. But since the rows of A^{T} are the columns of A, we have

$$\text{column rank}(A) \leq \text{row rank}(A)\,.$$

It follows that

$$\text{column rank}(A) = \text{row rank}(A). \tag{5.11}$$

We can therefore speak unambiguously of the "rank" of a matrix A, and simply write $\text{rank}(A)$, where $0 \leq \text{rank}(A) \leq \min\{m, n\}$. If $\text{rank}(A) = \min\{m, n\}$, then we say that A has **full rank**. If this coincides with the number of columns n, $r = n \leq m$, we can also say that the matrix has **full column rank**. If the rank coincides with the number of rows, $r = m \leq n$, we say that it has **full row rank**.

A square $n \times n$ matrix has an inverse if (and only if) A has full rank. The following are three further results regarding matrix rank:

i. For any matrices A and B such that AB exists, we have

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}.$$

This result holds because the columns of AB are linear combinations of the columns of A, therefore $\text{rank}(AB) \leq \text{rank}(A)$. Likewise, the rows of AB are linear combinations of the rows of B, therefore $\text{rank}(AB) \leq \text{rank}(B)$. It follows that $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$.

ii. If A is a full rank $m \times m$ matrix and B is $m \times p$ of rank r, then $\text{rank}(AB) = r$.

iii. For any matrix A, we have

$$\text{rank}(A^{\mathrm{T}}A) = \text{rank}(AA^{\mathrm{T}}) = \text{rank}(A)\,.$$

For a proof, see Tay, Preve, and Baydur (2025).

5.7 Vectors and Matrices of Random Variables

Organizing large numbers of random variables using matrix algebra provides convenient formulas for manipulating their expectations, variances and covariances, and for expressing their joint pdf.

5.7.1 Expectations and Variance-Covariance Matrices

The expectation of a vector x of m random variables $x = \begin{bmatrix} X_1 & X_2 & ... & X_m \end{bmatrix}^{\mathrm{T}}$ is defined as the vector of their expectations, i.e.,

$$E(x) = \begin{bmatrix} E(X_1) & E(X_2) & ... & E(X_m) \end{bmatrix}^{\mathrm{T}}.$$

Likewise, if X is a matrix of random variables, then

$$X = \begin{bmatrix} X_{11} & X_{12} & ... & X_{1n} \\ X_{21} & X_{22} & ... & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & ... & X_{mn} \end{bmatrix} \Leftrightarrow E(X) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & ... & E(X_{1n}) \\ E(X_{21}) & E(X_{22}) & ... & E(X_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{m1}) & E(X_{m2}) & ... & E(X_{mn}) \end{bmatrix}.$$

With these definitions, we can define the **variance-covariance matrix** of a vector x of random variables. Let

$$\tilde{x} = x - E(x) = \begin{bmatrix} X_1 - E(X_1) \\ X_2 - E(X_2) \\ \vdots \\ X_m - E(X_m) \end{bmatrix} = \begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \vdots \\ \tilde{X}_m \end{bmatrix}.$$

Then

$$E(\tilde{x}\tilde{x}^{\mathrm{T}}) = E((x - E(x))(x - E(x))^{\mathrm{T}})$$

$$= E\begin{bmatrix} \tilde{X}_1^2 & \tilde{X}_1\tilde{X}_2 & \dots & \tilde{X}_1\tilde{X}_m \\ \tilde{X}_2\tilde{X}_1 & \tilde{X}_2^2 & \dots & \tilde{X}_2\tilde{X}_m \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{X}_m\tilde{X}_1 & \tilde{X}_m\tilde{X}_2 & \dots & \tilde{X}_m\tilde{X}_m \end{bmatrix}$$

$$= \begin{bmatrix} E(\tilde{X}_1^2) & E(\tilde{X}_1\tilde{X}_2) & \dots & E(\tilde{X}_1\tilde{X}_m) \\ E(\tilde{X}_2\tilde{X}_1) & E(\tilde{X}_2^2) & \dots & E(\tilde{X}_2\tilde{X}_m) \\ \vdots & \vdots & \ddots & \vdots \\ E(\tilde{X}_m\tilde{X}_1) & E(\tilde{X}_m\tilde{X}_2) & \dots & E(\tilde{X}_m\tilde{X}_m) \end{bmatrix} \tag{5.12}$$

$$= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_m) \\ Cov(X_1, X_2) & Var(X_2) & \dots & Cov(X_2, X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_m) & Cov(X_2, X_m) & \dots & Var(X_m) \end{bmatrix}.$$

In other words, $E((x - E(x))(x - E(x))^{\mathrm{T}})$ is a symmetric matrix containing the variances of all of the variables in x, and their covariances. We denote the variance-covariance matrix of a vector of random variables x by $Var(x)$:

$$Var(x) = E((x - E(x))(x - E(x))^{\mathrm{T}}).$$

Example 5.17. Let X_1, X_2 and X_3 be random variables with

$$E(X_1) = 1, E(X_2) = 3, E(X_3) = 5,$$
$$Var(X_1) = 2,\ Var(X_2) = 3,\ Var(X_3) = 2, \text{ and}$$
$$Cov(X_1, X_2) = 1, Cov(X_1, X_3) = 0, Cov(X_2, X_3) = 2$$

and let x be the 3×1 vector $\begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix}^{\mathrm{T}}$. Then

$$E(X) = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad \text{and} \quad Var(X) = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

Recall that if X is a (univariate) random variable, then $E(aX+b) = aE(X)+b$, $Var(aX+b) = a^2\,Var(X)$, and $Var(X) = E(X^2) - E(X)^2$. The following are the matrix analogues of these results. Suppose x is an $m \times 1$ vector of random variables, $A = (a_{ij})_{km}$ is a $k \times m$ matrix of constants, and b is a $k \times 1$ vector of constants. Then

(i) $E(Ax + b) = AE(x) + b$,

(ii) $Var(Ax + b) = A\,Var(x)A^{\mathrm{T}}$,

(iii) $Var(x) = E(xx^{\mathrm{T}}) - E(x)E(x)^{\mathrm{T}}$.

To show (i), we note that the ith element of the $k \times 1$ vector $Ax + b$ is $\sum_{j=1}^{m}(a_{ij}X_j + b_i)$, and the expectation of this term is

$$E\left(\sum_{j=1}^{m}(a_{ij}X_j + b_i)\right) = \sum_{j=1}^{m} a_{ij}E(X_j) + b_i\,,$$

which in turn is the ith element of the vector $AE(x) + b$. For (ii), since $Ax + b - E(Ax + b) = A(x - E(x)) = A\tilde{x}$, we have

$$Var(Ax + b) = E((A\tilde{x})(A\tilde{x})^{\mathrm{T}}) = E(A\tilde{x}\tilde{x}^{\mathrm{T}}A^{\mathrm{T}}) = AE(\tilde{x}\tilde{x}^{\mathrm{T}})A^{\mathrm{T}} = A\ Var(x)A^{\mathrm{T}}\,.$$

You are asked to prove (iii) in Exercise 5.31.

Example 5.18. Given a vector of random variables x, the linear combination $c^{\mathrm{T}}x$ of the random variables in x has variance-covariance matrix

$$Var(c^{\mathrm{T}}x) = c^{\mathrm{T}}\,Var(x)c\,.$$

Since variances cannot be negative, we have $c^{\mathrm{T}}\,Var(x)c \geq 0$ for all c, i.e., $Var(x)$ is a positive semidefinite matrix. If there is a linear combination of the random variables in x that has zero variance, then at least one or more of the variables in x is actually a constant (a "degenerate random variable"), or at least one of the variables in x is a linear combination of the others. Otherwise we have $c^{\mathrm{T}}\,Var(x)c > 0$ for all $c \neq 0$, i.e., $Var(x)$ is positive definite.

5.7.2 The Multivariate Normal Distribution

We presented the pdf of a bivariate normal distribution in Section 3.6. We present here the pdf of a general multivariate normal distribution and some associated results. A $k \times 1$ vector of random variables x is said to have a multivariate normal distribution with mean μ and positive definite variance-covariance matrix Σ, denoted $\text{Normal}_k(\mu, \Sigma)$, if its pdf has the form

$$f(x) = (2\pi)^{-\frac{k}{2}}\det(\Sigma)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\}\,.$$

We list a few results below, omitting proofs:

(a) If Σ is diagonal, then $X_1, X_2, \ldots, X_k$ are independent random variables.

(b) If $x \sim \text{Normal}_k(\mu, \Sigma)$, then for $A_{m\times k}$ and $b_{m\times 1}$,

$$Ax + b \ \sim\ \text{Normal}_m(A\mu + b, A\Sigma A^{\mathrm{T}}).$$

(c) If we partition x as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \text{Normal}_k\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

where x_1 is $k_1 \times 1$ and x_2 is $k_2 \times 1$, with $k_1 + k_2 = k$, then the marginal distribution of x_1 is $\text{Normal}_{k_1}(\mu_1, \Sigma_{11})$, and the conditional distribution of x_2 given x_1 is

$$x_2 \mid x_1 \sim \text{Normal}_{k_2}(\mu_{2|1}, \Sigma_{22|1})$$

where $\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ and $\Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

(d) If $x \sim \text{Normal}_k(0, I)$ and A is a rank v symmetric matrix such that $AA = A$, then the scalar $x^{\mathrm{T}}Ax$ is distributed $\chi^2(v)$:

$$x^{\mathrm{T}}Ax \ \sim \ \chi^2(v)\,.$$

Matrices A such that $AA = A$ are said to be **idempotent**.

(e) If $x \sim \text{Normal}_k(\mu, \Sigma)$, then $(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \sim \chi^2(k)$.

5.7.3 Exercises

Exercise 5.31. Show that $Var(x) = E(xx^{\mathrm{T}}) - E(x)E(x)^{\mathrm{T}}$.

Exercise 5.32. Show that $E(\text{trace}(X)) = \text{trace}(E(X))$ where $X = (X_{ij})_{n\times n}$ is a matrix of random variables.

5.8 Differentiation of Matrix Forms

There are useful differentiation formulas available when the expression for the function to be differentiated has certain matrix forms. The following are a few particularly important examples.

Example 5.19. If $y = x^{\mathrm{T}}Ax$ where $A = (a_{jk})_{nn}$ is $n \times n$ and x is $n \times 1$, then

$$\nabla y = \frac{\partial y}{\partial x} = \frac{\partial}{\partial x}\,(x^{\mathrm{T}}Ax) = (A + A^{\mathrm{T}})x. \tag{5.13}$$

Proof: $y = x^{\mathrm{T}}Ax = \sum_{j=1}^n \sum_{k=1}^n a_{jk}x_jx_k$. The derivative $\partial y/\partial x$ is the $n \times 1$ vector whose *i-th* element is

$$\frac{\partial}{\partial x_i}\left(\sum_{j=1}^n\sum_{k=1}^n a_{jk}x_jx_k\right) = \underbrace{\sum_{k=1}^n a_{ik}x_k}_{\text{when } j=i} + \underbrace{\sum_{j=1}^n a_{ji}x_j}_{\text{when } k=i}\,.$$

The first sum after the equality is the product of the *ith* row of A into x. The second sum after the equality is the product of the *ith* row of A^{T} into x. In other words, $\partial y/\partial x = (A + A^{\mathrm{T}})x$.

It may be helpful to you to verify (5.13) by direct differentiation for a special case, say, where A is 2×2. You are asked to do this in an exercise. Note that if A is symmetric, then (5.13) becomes

$$\nabla y = \frac{\partial y}{\partial x} = \frac{\partial}{\partial x}x^{\mathrm{T}}Ax = (A + A^{\mathrm{T}})x = 2Ax. \tag{5.14}$$

This result is the matrix analogue of the univariate differentiation rule $f(x) = ax^2 \Rightarrow f'(x) = 2ax$.

Example 5.20. Let $y = f(x) = Ax$ where $A = (a_{ij})_{mn}$ is $m \times n$ and $x = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^{\mathrm{T}}$ is $n \times 1$. This is an example of a vector-valued function, mapping $x \in \mathbb{R}^n$ into $y \in \mathbb{R}^m$. We have

$$Df = \frac{\partial y}{\partial x^{\mathrm{T}}} = A\,. \tag{5.15}$$

This is the matrix analogue of the univariate differentiation rule $f(x) = ax \Rightarrow f'(x) = a$.

Proof: The product Ax is an $m \times 1$ vector whose *i-th* element is $\sum_{k=1}^n a_{ik}x_k$. Therefore the $(i,j)th$ element of $\partial y/\partial x^{\mathrm{T}}$ is $(\partial/\partial x_j)\sum_{k=1}^n a_{ik}x_k = a_{ij}$. This says that $\partial y/\partial x^{\mathrm{T}} = A$.

Example 5.21. The previous two examples show that if $y = x^{\mathrm{T}}Ax$ where A is an $n \times n$ symmetric matrix of constants and x is an $n \times 1$ vector of variables, then the Hessian is

$$\frac{d^2y}{dxdx^{\mathrm{T}}} = D(\nabla y) = D(2Ax) = 2A\,.$$

5.8.1 Exercises

Exercise 5.33. (a) Show that if $y = f(x) = x^{\mathrm{T}}A$ where $A = (a_{ij})_{mn}$ is a $m \times n$ matrix of constants and $x^{\mathrm{T}} = \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix}$, then

$$\partial y/\partial x = A\,.$$

(b) If c and x are $n \times 1$ vectors, show that

$$\frac{\partial}{\partial x}c^{\mathrm{T}}x = c\,.$$

Exercise 5.34. Let $A = (a_{ij})_{2,2}$ be a 2×2 matrix of constants, and x be a 2×1 vector of variables. Multiply out $x^{\mathrm{T}}Ax$ in full, and show by direct differentiation that

$$\frac{\partial y}{\partial x} = (A + A^{\mathrm{T}})x\,.$$

Chapter 6

Least Squares with Matrix Algebra

The mathematics of least squares is best expressed in matrix form. Proofs of results are much more concise and more general, and we can draw on the insights of linear algebra to understand least squares at a deeper level. Furthermore, the same mathematics applies to a vast number of advanced linear models including multiple equation models, and is useful even for non-linear ones, so it is well worth the time and effort to master the mathematics of least squares estimation expressed using matrix algebra.

6.1 The Setup

Suppose you are interested in estimating the conditional expectation $E(Y \mid X_1, ..., X_{K-1})$ which is assumed to have the form

$$E(Y \mid X_1, ..., X_{K-1}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{K-1} X_{K-1} .$$

Here $X_1, ..., X_{K-1}$ represent $K-1$ different variables, not observations of a single variable. We define the noise term ϵ as

$$\epsilon = Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - ... - \beta_{K-1} X_{K-1}$$

so we can write

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{K-1} X_{K-1} + \epsilon \,,\; E(\epsilon \mid X_1, ..., X_{K-1}) = 0 \,.$$

Suppose that you have a representative iid sample $\{Y_i, X_{i1}, X_{i2}, ..., X_{i,K-1}\}_{i=1}^n$ from this population, where X_{ik} denote the ith observation of variable X_k, then we can write

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{K-1} X_{i,K-1} + \epsilon_i \,,\; i = 1, 2, ..., n. \tag{6.1}$$

We can also write (6.1) in matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & ... & X_{1,K-1} \\ 1 & X_{21} & X_{22} & ... & X_{2,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & ... & X_{n,K-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{K-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{6.2}$$

or simply

$$y = X\beta + \varepsilon \tag{6.3}$$

where y is the $n \times 1$ vector $\begin{bmatrix} Y_1 & Y_2 & ... & Y_n \end{bmatrix}^{\mathrm{T}}$, X is the $n \times K$ matrix of regressors, β is the $K \times 1$ coefficient vector, and ε is the $n \times 1$ vector of noise terms. We assume that $n > K$.

We can partition the regressor matrix by observation:

$$X = \left[\begin{array}{ccccc} 1 & X_{11} & X_{12} & ... & X_{1,K-1} \\ \hline 1 & X_{21} & X_{22} & ... & X_{2,K-1} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline 1 & X_{n1} & X_{n2} & ... & X_{n,K-1} \end{array}\right] = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \tag{6.4}$$

and write the model as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$Y_i = X_{i*}\beta + \epsilon_i \,,\ i = 1, 2, ..., n.$$

It is sometimes helpful to partition the regressor matrix by variable:

$$X = \left[\begin{array}{c|c|c|c|c} 1 & X_{11} & X_{21} & ... & X_{1,K-1} \\ 1 & X_{12} & X_{22} & ... & X_{2,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & ... & X_{n,K-1} \end{array}\right] = \begin{bmatrix} i_n & X_{*1} & X_{*2} & ... X_{*,K-1} \end{bmatrix} \tag{6.5}$$

where i_n is the $n \times 1$ vector of ones, and X_{*k} is the vector of observations of variable X_k. Feasibility of OLS estimation will require X to have full column rank, i.e.,

$$Xc = c_0 + c_1 X_{*1} + c_2 X_{*2} + \cdots + c_{K-1} X_{*,K-1} = 0_{n\times 1} \iff c = 0_{K\times 1}$$

where $c = \begin{bmatrix} c_0 & ... & c_{K-1} \end{bmatrix}^{\mathrm{T}}$. In other words, we must assume that there is variation in each of the variables (apart from the constant vector), and that no one variable can be written as a linear combination of the other variables. The full column rank assumption implies that $X^{\mathrm{T}}X$ is non-singular (i.e., has an inverse).

Since $E(\epsilon \mid X_1, ..., X_{K-1}) = 0$ in population, and we have an iid representative sample from the population, we can assume that:

$$E(\epsilon_i \mid X_{*1}, X_{*2}, ..., X_{*,K-1}) = 0 \ \text{ for all } \ i = 1, 2, ..., n$$

and

$$Cov(\epsilon_i \epsilon_j \mid X_{*1}, X_{*2}, ..., X_{*,K-1}) = 0 \quad \text{for all} \quad i \neq j \,,\ i, j = 1, 2, ..., n.$$

We continue to assume homoskedastic errors

$$Var(\epsilon_i \mid X_{*1}, X_{*2}, ..., X_{*,K-1}) = \sigma^2 \quad \text{for all} \quad i = 1, 2, ..., n,$$

In matrix form, we can write these assumptions even more simply as

$$E(\varepsilon \mid X) = 0_{n\times 1} \ \text{ and } \ Var(\varepsilon \mid X) = E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$$

Remember that $Var(\varepsilon \mid X)$ is the conditional variance-covariance matrix of ε

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1 \mid X) & Cov(\varepsilon_1, \varepsilon_2 \mid X) & ... & Cov(\varepsilon_1, \varepsilon_n \mid X) \\ Cov(\varepsilon_2, \varepsilon_1 \mid X) & Var(\varepsilon_2 \mid X) & ... & Cov(\varepsilon_2, \varepsilon_n \mid X) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_n, \varepsilon_1 \mid X) & Cov(\varepsilon_n, \varepsilon_2 \mid X) & ... & Var(\varepsilon_n \mid X) \end{bmatrix}.$$

It is equal to $E(\varepsilon\varepsilon^{\mathrm{T}} \mid X)$ because $E(\varepsilon \mid X) = 0$. We have

$$E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2 I$$

because of our assumption of iid samples and homoskedasticity. If the errors are conditionally heteroskedastic with $Var(\epsilon_i \mid X) = \sigma_i^2$ but uncorrelated, then the conditional variance-covariance matrix of ε becomes

$$E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{bmatrix} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2).$$

If there is correlation between the errors, then the var-cov matrix of ε will no longer be diagonal.

6.2 Ordinary Least Squares

Let $\hat{\beta}$ denote some estimator for β. Then the fitted values associated with these estimators are

$$\underset{n\times 1}{\hat{y}} = \underset{n\times K}{X}\ \underset{K\times 1}{\hat{\beta}}$$

and the residuals are

$$\underset{n\times 1}{\hat{\varepsilon}} = \underset{n\times 1}{y} - \underset{n\times 1}{\hat{y}} = \underset{n\times 1}{y} - \underset{n\times K}{X}\ \underset{K\times 1}{\hat{\beta}}\ .$$

The residual sum of squares is then

$$\begin{aligned} \underset{1\times 1}{RSS} &= \underset{1\times n}{\hat{\varepsilon}^{\mathrm{T}}}\ \underset{n\times 1}{\hat{\varepsilon}} = \underset{---1\times n---}{(y - X\hat{\beta})^{\mathrm{T}}}\ \underset{--n\times 1--}{(y - X\hat{\beta})} \\ &= \underset{1\times n}{y^{\mathrm{T}}}\ \underset{n\times 1}{y} - \underset{1\times K}{\hat{\beta}^{\mathrm{T}}}\ \underset{K\times n}{X^{\mathrm{T}}}\ \underset{n\times 1}{y} - \underset{1\times n}{y^{\mathrm{T}}}\ \underset{n\times K}{X}\ \underset{K\times 1}{\hat{\beta}} + \underset{1\times K}{\hat{\beta}^{\mathrm{T}}}\ \underset{K\times n}{X^{\mathrm{T}}}\ \underset{n\times K}{X}\ \underset{K\times 1}{\hat{\beta}} \\ &= \underset{1\times 1}{y^{\mathrm{T}}y} - 2\,\underset{--1\times 1--}{\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y} + \underset{---1\times 1---}{\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\hat{\beta}} \end{aligned}$$

where we have used the fact that $\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y$ is the transpose of $y^{\mathrm{T}}X\hat{\beta}$, and the transpose of a scalar is the scalar itself. I have indicated the dimensions of the matrices in the expressions above to help you read the matrix equations. I will not do so from this point, but I strongly recommend that you continue this practice until you feel comfortable reading matrix expressions.

The OLS estimators are those that minimize RSS:

$$\hat{\beta}^{ols} = \text{argmin}_{\hat{\beta}} RSS\,.$$

The first-order conditions are

$$\left.\frac{\partial RSS}{\partial \hat{\beta}}\right|_{\hat{\beta}^{ols}} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\hat{\beta}^{ols} = 0. \tag{6.6}$$

This implies

$$X^{\mathrm{T}}X\hat{\beta}^{ols} = X^{\mathrm{T}}y$$

which, given our assumption that X is full column rank, can be solved for $\hat{\beta}^{ols}$:

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y. \tag{6.7}$$

The second partial derivatives of the RSS is positive definite:

$$\frac{\partial^2 RSS}{\partial\hat{\beta}\partial\hat{\beta}^{\mathrm{T}}} = 2X^{\mathrm{T}}X$$

since X is full column rank, which means that $Xc \neq 0$ for all $c \neq 0$, and therefore

$$c^{\mathrm{T}}X^{\mathrm{T}}Xc = (Xc)^{\mathrm{T}}Xc > 0\,.$$

This guarantees that (6.7) solves the minimization problem.

The FOC can also be written as

$$X^{\mathrm{T}}(y - X\hat{\beta}^{ols}) = X^{\mathrm{T}}\hat{\varepsilon}^{ols} = 0. \tag{6.8}$$

Partitioning X^{T} "by variable" as in (6.5), we can see that (6.8) says that OLS residuals sum to zero, and are orthogonal to each of the regressors:

$$X^{\mathrm{T}}\hat{\varepsilon}^{ols} = \begin{bmatrix} i_n^{\mathrm{T}} \\ X_{*1}^{\mathrm{T}} \\ X_{*2}^{\mathrm{T}} \\ \vdots \\ X_{*,K-1}^{\mathrm{T}} \end{bmatrix} \hat{\varepsilon}^{ols} = \begin{bmatrix} i_n^{\mathrm{T}}\hat{\varepsilon}^{ols} \\ X_{*1}^{\mathrm{T}}\hat{\varepsilon}^{ols} \\ X_{*2}^{\mathrm{T}}\hat{\varepsilon}^{ols} \\ \vdots \\ X_{*,K-1}^{\mathrm{T}}\hat{\varepsilon}^{ols} \end{bmatrix} = 0_{K\times 1}$$

In other words, we have

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^{ols} = 0 \;\text{ and }\; \sum_{i=1}^{n}X_{ik}\hat{\varepsilon}_i^{ols} = 0 \;\text{ for all }\; k = 1,\dots,K-1\,. \tag{6.9}$$

We can also view our estimators as arising from a "method of moments" perspectively. The assumption that $E(\epsilon \mid X_1,\dots,X_{K-1}) = 0$ implies $E(\epsilon) = 0$ and $E(X_k\epsilon) = 0$ for all $k = 1,\dots,K-1$. By choosing our estimators to solve (6.9), we are choosing our estimators to satisfy the sample moment conditions corresponding to these population moments.

6.3 Algebraic Properties of OLS Estimators

We list here some algebraic properties of OLS estimators. From this point onwards, we drop the 'OLS' superscript from the OLS estimators, fitted values and residuals, and write $\hat{\beta}$, $\hat{Y}$ and $\hat{\varepsilon}$ for $\hat{\beta}^{ols}$, $\hat{Y}^{ols}$ and $\hat{\varepsilon}^{ols}$. We will reinstate the superscript whenever context demands it so.

1. OLS estimators are **linear estimators**, by which we mean:

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Ay\,.$$

   This says that each OLS estimator $\hat{\beta}_k$, $k = 0, 1, ..., K-1$ can be written as

$$\hat{\beta}_k = \sum_{i=1}^{n} a_{ki}Y_i$$

   where a_{ki} , $i = 1, 2, ..., n$ are the elements of the kth row of $A = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

2. The OLS estimators can also be written as

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + \varepsilon) = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon.$$

   This form of the OLS estimator is useful for deriving its statistical properties, since it expresses $\hat{\beta}$ in terms of the actual parameter value β.

3. The OLS fitted values can be written as

$$\hat{y} = X\hat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

   The matrix $X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is sometimes called the "hat" matrix (because it puts a "hat" on y). It is also called the "projection" matrix, since it projects y onto the space spanned by the columns of X (see Tay, Preve, and Baydur (2025) Chapter 10). It is often denoted P. It has the convenient property that it is symmetric:

$$P^{\mathrm{T}} = (X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} = (X^{\mathrm{TT}}((X^{\mathrm{T}}X)^{-1})^{\mathrm{T}}X^{\mathrm{T}}) = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = P$$

   where we have used the fact that $(X^{\mathrm{T}}X)^{-1}$ is symmetric (why is $(X^{\mathrm{T}}X)^{-1}$ symmetric?).

4. The matrix P is also idempotent, meaning that $PP = P$:

$$PP = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = P.$$

   Symmetric and idempotent matrices have the convenient property that their rank is equal to their trace, which is easy to compute. Since

$$\mathrm{tr}(P) = \mathrm{tr}(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}) = \mathrm{tr}(X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}) = \mathrm{tr}(I_K) = K,$$

   the rank of P is K.

5. The OLS residuals can be written as

$$\hat{\varepsilon} = y - \hat{y} = (I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})y.$$

The matrix $I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is also symmetric and idempotent, and its trace, and therefore its rank, is $n-K$ (see exercises). It is often denoted by M, and has the property that it "eliminates X" in the sense that

$$MX = (I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})X = X - X = 0.$$

As a consequence of this, we have

$$MP = MX(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = 0.$$

Of course, you can also see this from $MP = (I-P)P = P - PP = P - P = 0$.

6. We have already noted from the FOC that the OLS residuals sum to zero, and are orthogonal to each of the regressors. Since $y = \hat{y} + \hat{\varepsilon}$, it follows that $\overline{Y} = \overline{\hat{Y}}$. Furthermore, $\hat{y}^{\mathrm{T}}\hat{\varepsilon} = 0$. That is, the fitted values and the residuals are orthogonal. We can also use the fact that $MP = PM = 0$:

$$\hat{y}^{\mathrm{T}}\hat{\varepsilon} = y^{\mathrm{T}}PMy = 0.$$

For those who are not uncomfortable thinking about n-dimensional vectors in geometric terms, this means the fitted values and residuals are at "right-angles" in n-dimensional space. The length of a vector y is $\sqrt{y^{\mathrm{T}}y}$. Using orthogonality of the fitted values and residuals, we get

$$\begin{aligned} y^{\mathrm{T}}y &= \hat{y}^{\mathrm{T}}\hat{y} + 2\hat{y}^{\mathrm{T}}\hat{\varepsilon} + \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} \\ &= \hat{y}^{\mathrm{T}}\hat{y} + \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}. \end{aligned} \tag{6.10}$$

This is just Pythagoras's Theorem (in n-dimensional space).

7. The $TSS = FSS + RSS$ equality continues to hold in the multiple regression case

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2. \tag{6.11}$$

See Exercise 6.3 and Exercise 6.4 for a matrix algebra version of this identity.

8. One useful application of the fact that M eliminates X is to derive a formula linking the residuals to the noise terms. We have

$$\hat{\varepsilon} = My = M(X\beta + \varepsilon) = M\varepsilon$$

This result, and the fact that M is symmetric and idempotent, means that the sum of squared residuals can be written as

$$\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = (M\varepsilon)^{\mathrm{T}}M\varepsilon = \varepsilon^{\mathrm{T}}M^{\mathrm{T}}M\varepsilon = \varepsilon^{\mathrm{T}}M\varepsilon.$$

This expression is also very useful for deriving properties of OLS estimators.

6.4 Statistical Properties of OLS Estimators.

6.4.1 Unbiasedness

Under our assumptions, $\hat{\beta}$ is unbiased. Using $\hat{\beta} = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon$, we have

$$E(\hat{\beta} \mid X) = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(\varepsilon \mid X) = \beta$$

which implies $E(\hat{\beta}) = \beta$. The key assumption delivering unbiasedness is, of course, $E(\varepsilon \mid X) = 0$. Note that the structure of the variance-covariance matrix of ε is irrelevant as far as unbiasedness of OLS estimators is concerned

6.4.2 Variance-Covariance Matrices

The variances and covariances of all of the OLS coefficient estimators can be obtained by computing the (conditional) variance-covariance matrix of $\hat{\beta}$:

$$\begin{aligned} Var(\hat{\beta} \mid X) &= E((\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))^{\mathrm{T}} \mid X) \\ &= E((\hat{\beta} - \beta)(\hat{\beta} - \beta)^{\mathrm{T}} \mid X) \\ &= E((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon\varepsilon^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1} \mid X) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(\varepsilon\varepsilon^{\mathrm{T}} \mid X)X(X^{\mathrm{T}}X)^{-1}. \end{aligned}$$

If we further assume homoskedastic and uncorrelated errors, then $E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I$, and $Var(\hat{\beta} \mid X)$ simplifies to

$$\begin{aligned} Var(\hat{\beta} \mid X) &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\sigma^2 I)X(X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2(X^{\mathrm{T}}X)^{-1} \end{aligned} \tag{6.12}$$

Example 6.1. In the simple linear regression $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$, $i = 1, 2, ..., n$, we have

$$X = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix} \quad \text{and} \quad X^{\mathrm{T}}X = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 \end{bmatrix}.$$

The OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found from the 2×1 vector

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\,.$$

The formulas obtained are the same as the ones previously derived. With homoskedastic and uncorrelated errors, the variances and covariances of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found from

$$Var(\hat{\beta} \mid X) = \begin{bmatrix} Var(\hat{\beta}_0 \mid X) & Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X) \\ Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X) & Var(\hat{\beta}_1 \mid X) \end{bmatrix} = \sigma^2(X^{\mathrm{T}}X)^{-1}. \tag{6.13}$$

Previously we derived the expression for the variance of $\hat{\beta}_1$ only. The expression in (6.13) contains both the variance of $\hat{\beta}_1$ and $\hat{\beta}_0$ and their covariance. You will explore these expressions in the exercises.

In order to operationalize (6.13), we have to estimate σ^2. Under homoskedasticity, an unbiased estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{n-K} = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n-K}.$$

To prove unbiasedness of this estimator, we note that

$$\begin{aligned} E(\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} \mid X) &= E(\varepsilon^{\mathrm{T}} M \varepsilon \mid X) = E(\mathrm{tr}(\varepsilon^{\mathrm{T}} M \varepsilon) \mid X) \\ &= E(\mathrm{tr}(\varepsilon\varepsilon^{\mathrm{T}} M) \mid X) = \mathrm{tr}(E(\varepsilon\varepsilon^{\mathrm{T}} M \mid X)) \\ &= \mathrm{tr}(E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) M) = \mathrm{tr}(\sigma^2 M) \\ &= \sigma^2(n-K). \end{aligned}$$

Unbiasedness of $\widehat{\sigma^2}$ follows. We therefore estimate $Var(\hat{\beta})$ using

$$\widehat{Var}(\hat{\beta} \mid X) = \widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}.$$

Example 6.2. In Example 3.2 we regressed ln *earn* on *educ* using data from `earnings2019.csv` using the `lm()` function, repeated here for reference.

```
library(tidyverse)
dat1 <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
lm(log(earn)~educ, data=dat1) %>% summary %>% coefficients
```

```
             Estimate  Std. Error  t value      Pr(>|t|)
(Intercept) 1.3199894 0.057541299 22.93986 9.401431e-111
educ        0.1279965 0.003978753 32.17000 2.442343e-206
```

For illustration, we compute the same results using the formulas derived in this chapter.

```
mlr <- function(y, X){
  n <- dim(X)[1]
  K <- dim(X)[2]
  XTXinv <- solve(t(X)%*%X)
  betahat <- XTXinv %*% t(X)%*%y
  yhat <- X %*% betahat
  ehat <- y - yhat
  sigmasqhat <- sum(ehat^2)/(dim(X)[1]-dim(X)[2])
  betahatvar <- sigmasqhat * XTXinv
  betahatse  <- sqrt(diag(betahatvar))
  betahat_t  <- betahat/betahatse
  results <- cbind(
    betahat, betahatse, betahat_t, 2*pt(-abs(betahat_t), n-K)
  )
  colnames(results) <- c("coef.", "s.e.", "t-stat", "p-value")
  model_return <- list("results"=results, "sigmasqhat"=sigmasqhat, "betahatvar"=betahatvar)
    return(model_return)
  }
y = log(dat1$earn)
X = cbind("intercept"=1, "educ"=dat1$educ)
```

```
model1 <- mlr(y,X)
model1$results
```

```
              coef.        s.e.   t-stat       p-value
intercept 1.3199894 0.057541299 22.93986 9.401431e-111
educ      0.1279965 0.003978753 32.17000 2.442343e-206
```

The code above works for larger regressions as well. Below we regress $\ln earn$ on $educ$, $tenure$, age and age^2:

```
y = log(dat1$earn)
X = cbind("intercept"=1, "educ"=dat1$educ, "tenure"=dat1$tenure, "age"=dat1$age, "agesq"=dat1$age^2)
model2 <- mlr(y,X)
model2$results
```

```
                 coef.          s.e.     t-stat       p-value
intercept -0.1408533857 1.133321e-01  -1.242838  2.139865e-01
educ       0.1258950407 3.790677e-03  33.211754 1.689882e-218
tenure     0.0154510891 1.082961e-03  14.267442  2.741774e-45
age        0.0625061595 4.776315e-03  13.086690  1.702263e-38
agesq     -0.0006692738 5.241977e-05 -12.767583  9.422808e-37
```

The estimate variance-covariance matrix is

```
model2$betahatvar
```

```
              intercept          educ        tenure           age         agesq
intercept  1.284416e-02 -1.880320e-04  1.206658e-05 -4.679158e-04  4.957007e-06
educ      -1.880320e-04  1.436923e-05 -2.315549e-08 -9.040899e-07  1.083268e-08
tenure     1.206658e-05 -2.315549e-08  1.172805e-06 -6.690121e-07  2.934586e-09
age       -4.679158e-04 -9.040899e-07 -6.690121e-07  2.281319e-05 -2.471022e-07
agesq      4.957007e-06  1.083268e-08  2.934586e-09 -2.471022e-07  2.747833e-09
```

In the general case $E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \Omega$, the formula for the variance-covriance matrix of $\hat{\beta}$ is

$$Var(\hat{\beta} \mid X) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Omega X(X^{\mathrm{T}}X)^{-1}.$$

To get any further we would have to put more structure on Ω. If we have uncorrelated but possibly heteroskedastic noise terms, then

$$\begin{aligned}
&Var(\hat{\beta} \mid X) \\
&= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)X(X^{\mathrm{T}}X)^{-1} \\
&= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \ldots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \qquad (6.14) \\
&= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \sigma_i^2 X_{i*}^{\mathrm{T}}X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}.
\end{aligned}$$

Expression (6.14) is not "operational" because we do not know the σ_i^2 nor can we estimate them — there are n of them and we only have n observations. Nonetheless it is still possible to develop an estimator for $Var(\hat{\beta})$ from (6.14). We will see how this can be done when we

discuss asymptotic properties of OLS estimators. In the meantime, we continue our discussion under the assumption of homoskedasticity. In particular, we show that if the noise terms are homoskedastic, then OLS estimators are the most precise estimators among all linear unbiased estimators, or "best linear unbiased" in the sense that

$$Var(c^{\mathrm{T}}\hat{\beta} \mid X) \leq Var(c^{\mathrm{T}}\tilde{\beta} \mid X) \tag{6.15}$$

for all $K \times 1$ vectors c, and for all unbiased estimators of the form $\tilde{\beta} = By$. Then we discuss hypothesis testing under homoskedastic errors, and then finally the asymptotic properties of OLS estimators.

6.4.3 Best Linear Unbiasedness

We elaborate on (6.15) a little bit. Recall that OLS estimators are linear estimators, meaning that it can be written in the form $\hat{\beta} = Ay$. In the case of OLS estimators, $A = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Suppose someone came up with a different linear unbiased estimator for β, call it $\tilde{\beta} = By$ where $B \neq A$. Which is better, $\hat{\beta}$ or $\tilde{\beta}$? They are both unbiased, so are equally good from that perspective. But given two unbiased estimators we should prefer the one with smaller variance. The inequality (6.15) says that each individual $\hat{\beta}_k$ will have a variance less than, or at worst equal to, the variance of $\tilde{\beta}_k$ (choose c such that $c_k = 1$, $c_j = 0$ for all $j \neq k$). It also says that any linear combinations of $\hat{\beta}$ will have variance less than, or at worst equal to, the variance of the same linear combination of $\tilde{\beta}$.

We say that under homoskedasticity, OLS estimators of the linear regression model are **best linear unbiased**. This result is known as the Gauss-Markov theorem. We also say OLS estimators under homoskedasticity are **efficient**.

Example 6.3. For a regression $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1}X_{i,K-1} + \epsilon_i$ with homoskedastic ϵ_i, the OLS prediction for Y at $(X_1, \ldots, X_{K-1}) = (X_{0,1}, \ldots, X_{0,K-1})$ is

$$\hat{Y}(X_0) = \hat{\beta}_0 + \hat{\beta}_1 X_{01} + \cdots + \hat{\beta}_{K-1}X_{0,K-1}\,.$$

This is a linear combination of OLS estimators, and by (6.15) it is the best linear unbiased prediction of Y at X_0.

Example 6.4. Recall the dataset `multireg_eg.csv` containing data on three variables X, Y and Z. The first two are continuous random variables whereas Z is discrete, taking values in $\{1, 2, 3, 4, 5\}$. See Fig. 4.3. If it helps, imagine that Y is test score for a certain course, X is study time per week a student spends on the course, and Z is the student's background for the course. To estimate the effect of study time X on test scores Y controlling for student prior preparedness for the course, you estimate the regression

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

using OLS, with interest in the parameter β_1. You find little evidence of heteroskedasticity, so you know that the OLS estimator $\hat{\beta}_1$ gives you the best linear unbiased estimator.

Suppose someone now suggests to you to estimate five simple linear regressions of Y on X, one for each value of Z, i.e., estimate the regressions

$$Y_i = \beta_{0j} + \beta_{1j} X_i + \epsilon_i \,, \quad \text{for all } \; i \; \text{ such that } \; z_i = j \,, \; j \in \{1,2,3,4,5\}$$

and then average the five estimates, i.e., let $\tilde{\beta}_1 = \frac{1}{5}\sum_{j=1}^{5} \hat{\beta}_i^j$. You are asked in an exercise to show that $\tilde{\beta}_1$ is in fact a linear unbiased estimator for β_1. However, we should still prefer the OLS estimator, since the OLS estimator is best among all linear unbiased estimators.

To prove (6.15), let $\tilde{\beta} = By$ where B comprises constants and elements of X, and such that $\tilde{\beta}$ is unbiased. Write $B = D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, so

$$\tilde{\beta} = DX\beta + D\varepsilon + \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon.$$

We have already assumed $E(\varepsilon \mid X) = 0$. To ensure unbiasedness of $\tilde{\beta}$, we have also to assume that $DX = 0$. We make this assumption. Then

$$\begin{aligned} Var(\tilde{\beta} \mid X) &= E((\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^{\mathrm{T}} \mid X) \\ &= E((D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\varepsilon\varepsilon^{\mathrm{T}}(D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \mid X) \\ &= \sigma^2((X^{\mathrm{T}}X)^{-1} + DD^{\mathrm{T}}) \\ &= Var(\hat{\beta}) + \sigma^2 DD^{\mathrm{T}} \,. \end{aligned}$$

It follows that for any $c \neq 0$,

$$\begin{aligned} Var(c^{\mathrm{T}}\tilde{\beta} \mid X) &= c^{\mathrm{T}}\, Var(\tilde{\beta} \mid X)c \\ &= c^{\mathrm{T}}\, Var(\hat{\beta})c + \sigma^2 c^{\mathrm{T}} DD^{\mathrm{T}} c \\ &= Var(c^{\mathrm{T}}\hat{\beta}) + \sigma^2 (D^{\mathrm{T}}c)^{\mathrm{T}}(D^{\mathrm{T}}c) \,. \end{aligned}$$

The second term on the right-hand side is non-negative, so result (6.15) follows.

6.4.4 Hypothesis Testing

We have already seen how to test single and joint hypotheses in Section 4.5. In this section, we develop matrix algebra formulas for the t and F tests.

If in the regression $y = X\beta + \varepsilon$, $E(\varepsilon \mid X) = 0$ and $Var(\varepsilon \mid X) = \sigma^2 I$, we add the assumption that the noise terms are normally distributed, i.e.,

$$y = X\beta + \varepsilon \,, \;\; \varepsilon \mid X \sim \text{Normal}_n(0, \sigma^2 I) \,,$$

then we have

$$\hat{\beta} \mid X \sim \text{Normal}_K(\beta, \sigma^2 (X^{\mathrm{T}}X)^{-1}) \tag{6.16}$$

since $\hat{\beta} = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon$ is a constant plus a linear combination of normally distributed noise terms. This can be used to develop t and F tests of linear hypotheses concerning elements of β. A general single linear hypothesis can be written as

$$H_0 : r^{\mathrm{T}}\beta = r_0 \quad \text{vs} \quad r^{\mathrm{T}}\beta \neq r_0.$$

Example 6.5. To test $\beta_1 + \beta_2 = 1$ in the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i,$$

set $r^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$ and $r_0 = 1$.

From (6.16), we have

$$r^{\mathrm{T}}\hat{\beta} \mid X \sim \text{Normal}(r^{\mathrm{T}}\beta, r^{\mathrm{T}}(\sigma^2(X^{\mathrm{T}}X)^{-1})r).$$

If the null hypothesis $r^{\mathrm{T}}\beta = r_0$ holds, then

$$r^{\mathrm{T}}\hat{\beta} \mid X \sim \text{Normal}(r_0, r^{\mathrm{T}}(\sigma^2(X^{\mathrm{T}}X)^{-1})r)$$

and

$$\frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}(\sigma^2(X^{\mathrm{T}}X)^{-1})r}} \sim \text{Normal}(0,1).$$

Furthermore, it can be shown that if we replace σ^2 with $\widehat{\sigma^2}$, then

$$\begin{aligned} t &= \frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}(\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1})r}} \\ &= \frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}\widehat{Var}(\hat{\beta} \mid X)r}} \sim \text{t}(n-K). \end{aligned} \tag{6.17}$$

This can be used to test the hypothesis $H_0 : r^{\mathrm{T}}\beta = r_0$ in the usual way.

To test multiple hypotheses jointly, write the hypotheses as

$$H_0 : \mathcal{R}\beta = r_0 \quad \text{vs} \quad H_A : \mathcal{R}\beta \neq r_0$$

where now $\mathcal{R}$ is a $(J \times K)$ matrix, and r_0 is a $(J \times 1)$ vector.

Example 6.6. To test the hypotheses

$$H_0 : \beta_1 + \beta_2 = 1 \text{ and } \beta_3 = 0 \quad \text{vs} \quad H_A : \beta_1 + \beta_2 \neq 1 \text{ or } \beta_3 \neq 0 \text{ (or both)},$$

in the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

set the matrices $\mathcal{R}$ and r_0 to

$$\mathcal{R} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad r_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

To test multiple hypotheses jointly, we can again compare the residual sum of squares from the restricted and unrestricted regressions, as explained in Section 4.5. In particular, it can be

shown that if the null is true, then

$$F = \frac{(\hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})/J}{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}/(n-K)} \sim \mathrm{F}(J, n-K) \tag{6.18}$$

where J is the number of restrictions being tested and $\hat{\varepsilon}$ is the vector of unrestricted OLS residuals. You would reject H_0 if the F-statistic is "improbably large", meaning that $F > \mathrm{F}_{1-\alpha}(J, n-K)$ where $\mathrm{F}_{1-\alpha}(J, n-K)$ is the $1-\alpha$ percentile of the $\mathrm{F}(J, n-K)$ distribution. Typically $\alpha = 0.01$, 0.05 or 0.1.

It turns out that in practice, one does not actually have to compute the restricted regression. It can be shown that

$$\hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0) \tag{6.19}$$

where $\hat{\beta}$ is the unrestricted OLS estimators (we will show this shortly). Furthermore, since the denominator of the F-statistic is $\widehat{\sigma^2}$, we can write the F-statistic as

$$\begin{aligned} F &= (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}(\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1})\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)/J \\ &= (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}\,\widehat{Var}(\hat{\beta} \mid X)\,\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)/J \\ &\sim \mathrm{F}(J, n-K) \end{aligned} \tag{6.20}$$

If the error terms are not normally distributed, we can use the asymptotically valid chi-sq version of the test, namely

$$JF \overset{a}{\sim} \chi^2(J)\,.$$

The following is the proof of (6.19). Let the regression model be $y = X\beta + \varepsilon$, and let $\hat{\beta}^{rls}$ be the least squares estimator for β subject to the restriction that $\mathcal{R}\beta = r$. We first show that

$$\hat{\beta}^{rls} = \hat{\beta}^{ols} + (X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}}(\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}})^{-1}(r - \mathcal{R}\hat{\beta}^{ols})$$

where $\hat{\beta}^{ols}$ is the usual unrestricted OLS estimator. The restricted RSS minimization problem is

$$\hat{\beta}^{rls} = \mathrm{argmin}_{\hat{\beta}}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) \text{ subject to } \mathcal{R}\hat{\beta} - r = 0.$$

The Lagrangian and FOC are

$$L = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + 2(r^{\mathrm{T}} - \hat{\beta}^{\mathrm{T}}\mathcal{R}^{\mathrm{T}})\lambda\,.$$

$$\begin{aligned} \left.\frac{\partial L}{\partial \hat{\beta}}\right|_{\hat{\beta}^{rls},\hat{\lambda}} &= -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\hat{\beta}^{rls} - 2\mathcal{R}^{\mathrm{T}}\hat{\lambda} = 0 \\ \left.\frac{\partial L}{\partial \hat{\lambda}}\right|_{\hat{\beta}^{rls},\hat{\lambda}} &= \quad 2(r - \mathcal{R}\hat{\beta}^{rls}) = 0 \end{aligned}$$

The second equation in the FOC merely says that the restriction must hold. The first equation in the FOC implies

$$\hat{\beta}^{rls} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y + (X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}}\hat{\lambda} = \hat{\beta}^{ols} + (X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}}\hat{\lambda}\,.$$

Multiplying throughout by $\mathcal{R}$ gives

$$\mathcal{R}\hat{\beta}^{rls} = \mathcal{R}\hat{\beta}^{ols} + \mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}}\hat{\lambda}\,.$$

If follows that

$$\begin{aligned}\hat{\lambda} &= (\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta}^{rls} - \mathcal{R}\hat{\beta}^{ols}) \\ &= (\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}})^{-1}(r - \mathcal{R}\hat{\beta}^{ols})\,,\end{aligned}$$

and therefore

$$\hat{\beta}^{rls} = \hat{\beta}^{ols} + (X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}}(\mathcal{R}(X^{\mathrm{T}}X)^{-1}\mathcal{R}^{\mathrm{T}})^{-1}(r - \mathcal{R}\hat{\beta}^{ols})\,. \tag{6.21}$$

Now let

$$\begin{aligned}\hat{\varepsilon}_{rls} &= y - X\hat{\beta}^{rls} \\ &= y - X\hat{\beta}^{ols} + X\hat{\beta}^{ols} - X\hat{\beta}^{rls} = \hat{\varepsilon}_{ols} + X(\hat{\beta}^{ols} - \hat{\beta}^{rls})\,.\end{aligned} \tag{6.22}$$

Since (unrestricted) OLS residuals are orthogonal to the regressors, we have

$$\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{rls} = \hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols} + \hat{\varepsilon}_{ols}^{\mathrm{T}}X(\hat{\beta}^{ols} - \hat{\beta}^{rls}) = \hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}.$$

Therefore

$$(\hat{\varepsilon}_{rls} - \hat{\varepsilon}_{ols})^{\mathrm{T}}(\hat{\varepsilon}_{rls} - \hat{\varepsilon}_{ols}) = \hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}. \tag{6.23}$$

Finally, use (6.21), (6.22) and (6.23) to show (6.19).

Example 6.7. Previously we estimated the equation

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon$$

by OLS using the `lm` function, and used the `linearHypothesis` function from the `car` package to carry out a t-test of the hypothesis $H_0 : \beta_2 = \beta_3$ vs $H_A : \beta_2 \neq \beta_3$, and an F-test of the joint hypotheses $H_0 : \beta_2 = \beta_3$ and $\beta_4 = 0$. The results are repeated below:

```
library(car)
dat2<-read_csv("data\\earnings2019.csv",show_col_types=FALSE) %>%
  mutate(female=1-male,
         white=if_else(race=="White", 1,0),
         black=if_else(race=="Black", 1,0),
         other=if_else(race=="Other", 1,0)) %>% select(-race)
mdl_unres <- lm(log(earn) ~ educ + black*female, data=dat2)
linearHypothesis(mdl_unres, c("black=female"))
linearHypothesis(mdl_unres, c("black=female", "black:female=0"))
```

```
Linear hypothesis test:
black - female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   4942 1603.7
2   4941 1603.5  1   0.18778 0.5786 0.4469
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   4943 1606.5
2   4941 1603.5  2    2.9507 4.5461 0.01065 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As an illustration, we compute the two tests using the formulas derived in this section

```
y = log(dat2$earn)
X = cbind("intercept"=1, "educ"=dat2$educ, "black"=dat2$black, "female"=dat2$female,
          "black.female" = dat2$black * dat2$female)
n = dim(X)[1]
K = dim(X)[2]
model1 <- mlr(y,X)
model1$results
# t-test of "black=female"
r = matrix(c(0,0,1,-1,0), ncol=1)
r0 = 0
betahat <- as.matrix(model1$results[,1])
tstat <- (t(r) %*% betahat - r0)/sqrt(t(r) %*% model1$betahatvar %*% r)
tstat_pval <- 2*pt(-abs(tstat), n-K)
cat("\n Test: b2=b3")
cat("\n tstat:", round(tstat,5), "   p-value:", round(tstat_pval,5), "\n")
# F-test of "black=female" and "black.female=0"
R <- matrix(c(0,0,1,-1,0,
              0,0,0, 0,1), nrow=2, byrow=TRUE)
J <- dim(R)[1]
r0 <- matrix(c(0,0), nrow = 2)
Fstat <- t(R%*%betahat - r0) %*% solve(R %*% model1$betahatvar %*% t(R)) %*% (R %*% betahat - r0)/J
Fstat_pval <- 1-pf(Fstat, J, n-K)
cat("\n Test: b2=b3, b4=0")
cat("\n Fstat:", round(Fstat,5), "   p-value:", round(Fstat_pval,5),"\n")
```

```
                  coef.        s.e.     t-stat       p-value
intercept     1.53685588 0.057085622  26.921943 4.368569e-149
educ          0.12785815 0.003879464  32.957687 1.656104e-215
black        -0.26004998 0.026957206  -9.646771  7.892653e-22
female       -0.28066123 0.019594037 -14.323808  1.259945e-45
black.female  0.08729863 0.035485808   2.460100  1.392382e-02

 Test: b2=b3
 tstat: 0.76067    p-value: 0.44689

 Test: b2=b3, b4=0
 Fstat: 4.54611    p-value: 0.01065
```

Notice that the `linearHypothesis()` function returns an F-statistic even when testing a single hypothesis. Of course, even though the F-test is designed with testing joint hypothesis, there is no reason why is can't be used to test a single hypothesis. It can be shown (see exercises) that in general the F-statistic for a test of a single hypothesis is the square of the corresponding t-statistic. You can verify that this is the case in the example above. The two tests will produce identical p-values.

6.5 Some Asymptotic Results

In this section, we provide rough arguments for the consistency and asymptotic normality of the OLS estimators. We also provide an asymptotically valid estimator for the variance-covariance matrix of the OLS estimator under conditional heteroskedasticity. For more complete arguments, please see any advanced econometrics texts, such as Hayashi (2000).

6.5.1 Consistency

Partitioning X by observation, as in (6.4), we can write the OLS estimator as

$$\begin{aligned}
\hat{\beta}^{ols} &= \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\epsilon \\
&= \beta + \left\{ \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \dots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \right\}^{-1} \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \dots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
&= \beta + \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \right)^{-1} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} \epsilon_i \\
&= \beta + \left(\frac{1}{n} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} \epsilon_i \right).
\end{aligned} \tag{6.24}$$

Note that

$$\frac{1}{n} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} = \begin{bmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n} X_{i1} & \dots & \frac{1}{n}\sum_{i=1}^{n} X_{i,K-1} \\ \frac{1}{n}\sum_{i=1}^{n} X_{i1} & \frac{1}{n}\sum_{i=1}^{n} X_{i1}^2 & \dots & \frac{1}{n}\sum_{i=1}^{n} X_{i1}X_{i,K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_{i,K-1} & \frac{1}{n}\sum_{i=1}^{n} X_{i1}X_{i,K-1} & \dots & \frac{1}{n}\sum_{i=1}^{n} X_{i,K-1}^2 \end{bmatrix}$$

Since we have iid samples, we should expect each of the component sample means to converge to the corresponding population moment, e.g., $\frac{1}{n}\sum_{i=1}^{n} X_{i1} \xrightarrow{p} E(X_1)$, $\frac{1}{n}\sum_{i=1}^{n} X_{i1}^2 \xrightarrow{p} E(X_1^2)$, $\frac{1}{n}\sum_{i=1}^{n} X_{i1}X_{i,K-1} \xrightarrow{p} E(X_1 X_{K-1})$ and so on. That is,

$$\frac{1}{n} \sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \xrightarrow{p} \begin{bmatrix} 1 & E(X_1) & \dots & E(X_{K-1}) \\ E(X_1) & E(X_1^2) & \dots & E(X_1 X_{K-1}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{K-1}) & E(X_1 X_{K-1}) & \dots & E(X_{K-1}^2) \end{bmatrix} = "\Sigma_{XX}"$$

which we assume is invertible. Likewise, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}\epsilon_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \\ \frac{1}{n}\sum_{i=1}^{n} X_{i1}\epsilon_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_{i,K-1}\epsilon_i \end{bmatrix} \xrightarrow{p} \begin{bmatrix} E(\epsilon) \\ E(X_1\epsilon) \\ \vdots \\ E(X_{K-1}\epsilon) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Therefore

$$\hat{\beta}^{ols} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}\epsilon_i\right) \quad \xrightarrow{p} \beta + \Sigma_{XX}^{-1}0_K = \beta\,.$$

The key assumptions for consistency are the assumptions $E(\epsilon) = 0$, $E(X_1\epsilon) = 0$, ..., $E(X_{K-1}\epsilon) = 0$, i.e., that the noise term ϵ has zero mean and is uncorrelated with each of the regressors in population.

This is basically the general version of the simple linear regression argument that

$$\hat{\beta}_1 = \beta_1 + \frac{\text{sample cov}(X_i, \epsilon_i)}{\text{sample var}(X_i)} \xrightarrow{p} \beta_1$$

if $Cov(X_i, \epsilon_i) = 0$ and $Var(X_i)$ is finite, and if their sample counterparts converge in probability to them.

6.5.2 Asymptotic Normality

Since $\hat{\beta}_1$ is consistent, its distribution is essentially degenerate in the limit. To talk about limiting distributions, we have to scale $\hat{\beta}_1$. Rearranging 6.24 we have

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}\epsilon_i\right),$$

If in addition to our previous assumptions we add the assumption that

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 X_{i*}^{\mathrm{T}}X_{i*} \xrightarrow{p} S \;\text{ finite and non-singular}$$

then a CLT allows us to claim asymptotic normality. We have:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}\epsilon_i \xrightarrow{d} \text{Normal}_K(0, S),$$

therefore

$$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}}_{\xrightarrow{p}\,\Sigma_{XX}^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}\epsilon_i\right)}_{\xrightarrow{d}\,\text{Normal}_K(0,S)} \to_d \text{Normal}_K(0, \Sigma_{XX}^{-1}S\Sigma_{XX}^{-1})$$

That is, $\hat{\beta}$ is consistent, with asymptotic variance $Avar(\hat{\beta}) = \Sigma_{XX}^{-1}S\Sigma_{XX}^{-1}$.

6.5.3 Heteroskedasticity-Robust Standard Errors

This result justifies the approximation

$$Var(\hat{\beta}) \approx (1/n)\Sigma_{XX}^{-1} S \Sigma_{XX}^{-1}.$$

An obvious estimator for Σ_{XX} is

$$\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} = \frac{1}{n} X^{\mathrm{T}} X\,.$$

Some additional assumptions (see Hayashi (2000), for example) guarantee

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*} \;\to_p\; S \tag{6.25}$$

where $\hat{\epsilon}_i$ are the OLS residuals. This allows us to consistently estimate the asymptotic variance of $\hat{\beta}$ by

$$\widehat{Avar}(\hat{\beta}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1},$$

and justifies the use of the following as an estimator for the variance of $\hat{\beta}$:

$$\begin{aligned}
\widehat{Var}_{HC0}(\hat{\beta}) &= \frac{1}{n}\widehat{Avar}(\hat{\beta}) \\
&= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \\
&= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}.
\end{aligned} \tag{6.26}$$

Note that we have so far assume iid samples but we have not assumed homoskedasticity. In other words, 6.26 is valid under heteroskedasticity. The variance estimator in (6.26) is called a "Heteroskedasticity-Consistent Variance Estimator". There are several versions. The version presented in (6.26) is often referred to as "HC0", which is why we label it as such. Other versions will be discussed in the exercises. Because of its form, (6.26) is often called a "sandwich" estimator.

If there is conditional heteroskedasticity in the noise terms, the usual OLS variance estimator $\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}$ is not appropriate since $\widehat{\sigma^2}((1/n)X^{\mathrm{T}}X)^{-1}$ is not a consistent estimator for the asymptotic variance. On the other hand, the variance estimator (6.26) remains consistent even if in fact the noise terms are conditionally homoskedastic. In this sense it is safer to use (6.26) for estimating the estimator variance if there is any possibility of conditional heteroskedasticity.

We have already noted that OLS estimators are efficient if there is conditional homoskedasticity. If there is conditional heteroskedasticity, then OLS estimators are no longer efficient. (The formula (6.26) allows us to estimate the estimator *variance* consistently, but doesn't do anything about the efficiency of $\hat{\beta}$ itself.) We have already addressed how we might try to get efficient estimators in the previous chapter.

Example 6.8. We adapt our earlier `mlr` function to give HC0 Heteroskedasticity Robust standard errors and apply it to the regression

$$\ln earn = \beta_0 + \beta_1 educ + \beta_2 black + \beta_3 female + \beta_4 black.female + \epsilon\,.$$

```
mlr_hc0 <- function(y, X){
  n <- dim(X)[1]
  K <- dim(X)[2]
  XTXinv <- solve(t(X)%*%X)
  betahat <- XTXinv %*% t(X)%*%y
  yhat <- X %*% betahat
  ehat <- y - yhat
  sigmasqhat <- sum(ehat^2)/(dim(X)[1]-dim(X)[2])
  hatS = 0
  for (i in 1:n){
    xi = as.matrix(X[i,])
    hatS = hatS + ehat[i]^2*xi%*%t(xi)
  }
  betahatvar_hc0 <- XTXinv %*% hatS %*% XTXinv ## XTXinv computed earlier
  betahatse_hc0  <- sqrt(diag(betahatvar_hc0))
  betahat_t_hc0  <- betahat/betahatse_hc0
  results <- cbind(
    betahat, betahatse_hc0, betahat_t_hc0, 2*pnorm(-abs(betahat_t_hc0), 0, 1)
  )
  colnames(results) <- c("coef.", "s.e.(hc0)", "t-stat", "p-value")
  model_return <- list("results"=results, "betahatvar"=betahatvar_hc0)
    return(model_return)
  }
y = log(dat1$earn)
X = cbind("intercept"=1, "educ"=dat2$educ, "black"=dat2$black, "female"=dat2$female,
          "black.female"=dat2$black * dat2$female)
model2_hc0 <- mlr_hc0(y,X)
cat("Estimation results:\n")
model2_hc0$results
cat("\n variance-covariance matrix (hc0) of betahat:\n")
model2_hc0$betahatvar %>% round(6)
```

```
Estimation results:
                   coef.   s.e.(hc0)     t-stat       p-value
intercept     1.53685588 0.056516372  27.193109 7.835611e-163
educ          0.12785815 0.003927411  32.555323 1.760392e-232
black        -0.26004998 0.026926439  -9.657793  4.555715e-22
female       -0.28066123 0.020169608 -13.915056  5.131978e-44
black.female  0.08729863 0.034640247   2.520150  1.173047e-02

 variance-covariance matrix (hc0) of betahat:
             intercept      educ     black    female black.female
intercept     0.003194 -0.000214 -0.000290 -0.000070     0.000188
educ         -0.000214  0.000015  0.000005 -0.000011     0.000003
black        -0.000290  0.000005  0.000725  0.000224    -0.000724
female       -0.000070 -0.000011  0.000224  0.000407    -0.000400
black.female  0.000188  0.000003 -0.000724 -0.000400     0.001200
```

The function `hccm()` from the `car` package also calculates this, as does the `vcovHC()` function from the `sandwich` package.

```
model3 <- lm(log(earn) ~ educ + black*female, data=dat2)
betahatvar_HC0_car = hccm(model3,type="hc0")
round(betahatvar_HC0_car,6)
```

```
             (Intercept)      educ     black    female black:female
(Intercept)     0.003194 -0.000214 -0.000290 -0.000070     0.000188
educ           -0.000214  0.000015  0.000005 -0.000011     0.000003
black          -0.000290  0.000005  0.000725  0.000224    -0.000724
female         -0.000070 -0.000011  0.000224  0.000407    -0.000400
black:female    0.000188  0.000003 -0.000724 -0.000400     0.001200
```

```
betahatvar_HC0_sando = sandwich::vcovHC(model3,type="HC0")
round(betahatvar_HC0_sando,6)
```

```
             (Intercept)      educ     black    female black:female
(Intercept)     0.003194 -0.000214 -0.000290 -0.000070     0.000188
educ           -0.000214  0.000015  0.000005 -0.000011     0.000003
black          -0.000290  0.000005  0.000725  0.000224    -0.000724
female         -0.000070 -0.000011  0.000224  0.000407    -0.000400
black:female    0.000188  0.000003 -0.000724 -0.000400     0.001200
```

We can use the heteroskedasticity consistent variance estimator to construct heteroskedasticity robust t and F statistics, by replacing $\widehat{Var}(\hat{\beta} \mid X)$ in (6.17) and (6.20) with the heteroskedasticity robust variance estimator in (6.26). For the `linearHypothesis` function from `car`, you can use the robust F-test as shown below, where we use the chi-square version of the test (since the robust standard errors are only valid asymptotically):

```
linearHypothesis(model3, c("black=female", "black:female=0"),
                 vcov=betahatvar_HC0_sando,
                 test="Chisq")
```

```
Linear hypothesis test:
black - female = 0
black:female = 0

Model 1: restricted model
Model 2: log(earn) ~ educ + black * female

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1   4943
2   4941  2 9.6196    0.00815 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.6 Exercises

Exercise 6.1. Let $y = X\beta + \varepsilon\,,\ E(\varepsilon \mid X) = 0\,,\ E(\varepsilon\varepsilon^{\mathrm{T}}) = \sigma^2 I$ represent the simple linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i\,,\ i = 1, 2, ..., n$$

(a) Use (6.7) to find expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ and show that they are the same as those obtained in (3.21).

(b) Use (6.13) to find expressions for $Var(\hat{\beta}_0 \mid X)$ and $Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X)$ in the simple linear regression. What is the sign of $Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X)$?

Remark: For intuition regarding the sign of $Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X)$, consider the fact that estimated regression lines always pass through the point $(\overline{X}, \overline{Y})$.

Exercise 6.2. Let X be a $n \times K$ matrix with full column rank. Show that

$$M = I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$

is symmetric and idempotent, with rank $n - K$.

Exercise 6.3. Let $y = X\beta + \varepsilon$ represent the regression $Y_i = \beta_0 + \epsilon_i$, $i = 1, 2, ..., n$.

(a) Show that the residuals from this regression can be written as

$$\hat{\varepsilon} = M_0 y$$

where $M_0 = I - i_n(i_n^{\mathrm{T}} i_n)^{-1} i_n^{\mathrm{T}}$ and i_n is the $n \times 1$ vector of ones.

(b) Show by direct computation that $M_0 y$ is the vector of deviations from means, i.e.,

$$M_0 y = (I - i_n(i_n^{\mathrm{T}} i_n)^{-1} i_n^{\mathrm{T}})y = \begin{bmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{bmatrix}.$$

(c) Show that $y^{\mathrm{T}} M_0 y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$

Exercise 6.4. Show that for the general linear regression model (with intercept) that

$$y^{\mathrm{T}} M_0 y = \hat{y}^{\mathrm{T}} M_0 \hat{y} + \hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon}.$$

This is the $TSS = FSS + RSS$ equality that forms the basis of the R^2 goodness-of-fit measure.

Exercise 6.5. Consider the regression $e^i = X\beta + \varepsilon$, where e^i is the $n \times 1$ vector comprising all zeros except for a '1' in the ith position. Let the matrix X be $n \times K$ with full column rank.

(a) Show that the fitted values $\widehat{e^i}$ has the expression

$$\widehat{e^i} = X(X^{\mathrm{T}}X)^{-1} X_{i*}^{\mathrm{T}}$$

where X_{i*} is the ith row of the X matrix.

(b) Define the 'leverage' of observation i to be

$$h_i = X_{i*}(X^{\mathrm{T}}X)^{-1}X_{i*}^{\mathrm{T}}.$$

Show that $0 \leq h_i \leq 1$. *Hint: Use part (a) and the "Pythagoras's Theorem" result in (6.10).*

(c) Explain why $\sum_{i=1}^{n} h_i$ is the trace of the matrix $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Show that $\sum_{i=1}^{n} h_i = K$. (In other words, the "average value" of h_i is K/n.)
Remark: It can be shown that

$$\hat{\beta} - \hat{\beta}_{-i} = \left(\frac{1}{1-h_i}\right)(X^{\mathrm{T}}X)^{-1}X_{i*}^{\mathrm{T}}\,\epsilon_i$$

where $\hat{\beta}_{-i}$ is the OLS estimator obtained when observation i is left out. An observation with h_i close to 1 therefore has very high leverage, or is "influential".

Exercise 6.6. Consider the linear regression $y = X\beta + \varepsilon$ where $E(\varepsilon \mid X) = 0$ and $E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I$. The fact that $\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$ is biased implies that each individual $\hat{\epsilon}_i^2$ must in general be a biased estimator for σ^2. Show that

$$E(\hat{\epsilon}_i^2) = (1-h_i)\sigma^2.$$

Hint: Use $\hat{\epsilon}_i^2 = (e^i)^{\mathrm{T}}\hat{\varepsilon}\hat{\varepsilon}^{\mathrm{T}}e^i$ where e^i is as defined in the previous question, and $\hat{\varepsilon} = M\varepsilon$ where $M = I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

Exercise 6.7. The "HC0" version of the heteroskedasticity-consistent standard errors given in (6.26) is sometimes criticized for not taking into consideration the fact that K degrees of freedom are used in the computation of $\hat{\epsilon}_i$. Another version proposes to take this into account by estimating S with

$$\hat{S}_1 = \frac{1}{n-K}\sum_{i=1}^{n}\hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}}X_{i*}$$

which also consistently estimates S.

(a) Show that using $\hat{S}_1$ instead of $\hat{S}$ in (6.25) results in the Heteroskedasticity-Consistent variance estimator

$$\widehat{Var}_{HC1}(\hat{\beta}) = \left(\sum_{i=1}^{n}X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}\left(\frac{n}{n-K}\sum_{i=1}^{n}\hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}}X_{i*}\right)\left(\sum_{i=1}^{n}X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}. \tag{6.27}$$

Amend the code in (Example 6.8) to use the HC1 version of the variance estimator, and verify your results using the `vcovHC()` function.

(b) Another version, based on the result in (Exercise 6.6), is

$$\widehat{Var}_{HC2}(\hat{\beta}) = \left(\sum_{i=1}^{n}X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\hat{\epsilon}_i^2}{1-h_i}X_{i*}^{\mathrm{T}}X_{i*}\right)\left(\sum_{i=1}^{n}X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}. \tag{6.28}$$

Amend the code in (Example 6.8) to use the HC2 version of the variance estimator, and verify your results using the `vcovHC()` function.

(c) The result in (Exercise 6.6), of course, assumes conditional homoskedasticity. Yet another version is

$$\widehat{Var}_{HC3}(\hat{\beta}) = \left(\sum_{i=1}^{n}X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\hat{\epsilon}_i^2}{(1-h_i)^2}X_{i*}^{\mathrm{T}}X_{i*}\right)\left(\sum_{i=1}^{n}X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}. \tag{6.29}$$

This version puts more weight on observations that are more influential. Amend the code in Example 6.8 to use the HC3 version of the variance estimator, and verify your results using the `vcovHC()` function.

Exercise 6.8. (a) Show that the F-statistic for testing a single restriction, i.e., when $J = 1$, is equal to the square of the t-statistic for testing the same hypothesis. (*Hint: Compare (6.17) and (6.20) after setting $R = r^{\mathrm{T}}$ in the F-statistic.*)

(b) Suppose you have the regressions

$$\text{(A)} \qquad Y_i = \beta_0 + \beta_1 X_{i1} + \ldots \beta_{K-1} X_{i,K-1} + \epsilon_i \,,$$

$$\text{(B)} \qquad Y_i = \beta_0 + \beta_1 X_{i1} + \ldots \beta_{K-1} X_{i,K-1} + \beta_K X_{iK} + \epsilon_i \,.$$

Two possible ways of deciding whether or not to include variable X_{iK} in the regression is to do a t-test (or an F-test) of the hypothesis $\beta_K = 0$. Another way is to see whether or not the Adjusted-R^2 in (B) is greater than the Adjusted-R^2 in (A). Show that the latter method is equivalent to including $X_{K,i}$ if the absolute value of the t-statistic for $\hat{\beta}_K$ in (B) is greater than 1. *Hint: use the version of the F-statistic in (6.18).*

Exercise 6.9. Suppose you estimated the regression equation $y = X\beta + \varepsilon$, $E(\varepsilon \mid X) = 0$, $E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I$ using the iid sample $\{Y_i, X_{i*}\}_{i=1}^n$. Let

$$\widehat{Y}(X_{0*}) = X_{0*}\hat{\beta}$$

be the prediction of Y for an independent draw from the population with X characteristics equal to X_{0*}. The prediction error is

$$\hat{e}(X_{0*}) = Y(X_{0*}) - \hat{Y}(X_{0*}) = X_{0*}\beta + \epsilon_0 - X_{0*}\hat{\beta} = X_{0*}(\beta - \hat{\beta}) + \epsilon_0 \,.$$

(a) Derive an expression for the prediction error variance.

(b) Specialize your answer in (a) to the simple linear regression $Y = \beta_0 + \beta_1 X_1 + \epsilon$, predicting Y at $X_1 = X_{01}$. Show that the prediction mean squared error is

$$\hat{e}(X_{01}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_{01} - \overline{X}_1)^2}{\sum_{i=1}^n (X_{01} - \overline{X}_1)^2} \right).$$

Explain why this is also the prediction error variance.

Chapter 7

Topics in OLS Estimation of the Linear Regression Model

This chapter serves two purposes: One, it is a summary of the theory we have covered to this point. Second, we discuss can some of the things that can go wrong in least squares estimation of the linear regression model. As to what to do when such-and-such problem arises, this will only be partly answered in this chapter. Many of the issues discussed here require more advanced theory presented in the next few chapter.

The R code in this chapter uses the following package (and one call to the `sandwich` package).

```
library(tidyverse)   # For data handling and visualization
```

7.1 Recap

The linear regression model estimates population conditional expectation which is assumed to be linear-in-parameters:

$$E(Y \mid X_1, \dots, X_{K-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1}.$$

The regressors $X_1, \dots, X_{K-1}$ may be distinct variables, but some may be transformations of or interactions between other regressors, e.g., we can have

$$E(\ln earn \mid age, educ) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 male + \beta_5 male \cdot educ.$$

We can write the regression model as

$$Y = E(Y \mid X_1, \dots, X_{K-1}) + \epsilon = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \epsilon\,.$$

As long as the conditional expectation is correctly specified, we have

$$E(\epsilon \mid X_1, \dots, X_{K-1}) = 0\,.$$

Follow on implications are that $E(\epsilon) = 0$, and $E(\epsilon X_k) = 0$ for $k = 0, \dots, K-1$. Another way of describing the latter set of conditions is that the noise variable ϵ is uncorrelated with each of the regressors $X_1, \dots, X_{K-1}$.

The classical linear regression model adds the assumption $Var(\epsilon \mid X_1, \dots, X_{K-1}) = \sigma^2$. Since ϵ has zero conditional mean, this "homoskedasticity" assumption can be (and often is) written as $E(\epsilon^2 \mid X_1, \dots, X_{K-1}) = \sigma^2$.

If you have a representative iid sample $\{Y_i, X_{i1}, \dots, X_{i,K-1}\}_{i=1}^n$ from the population (we assume cross-sectional samples for the moment), we can write

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i\,, i = 1, \dots, n\,,$$

where

$$E(\epsilon_i \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = 0$$
$$E(\epsilon_i^2 \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = \sigma^2$$
$$E(\epsilon_i \epsilon_j \mid X_{11}, \dots, X_{n1}; \dots; X_{1,K-1}, \dots, X_{n,K-1}) = 0$$

for all $i, j = 1, \dots, n$, $i \neq j$. All this can be written in matrix form as

$$y = X\beta + \varepsilon\,,\ E(\varepsilon \mid X) = 0\,,\ E(\varepsilon\varepsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n\,.$$

The OLS estimator for β is

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\,.$$

This requires $X^{\mathrm{T}}X$ to be invertible, which will be the case if the columns of X are linearly independent, meaning that

$$Xc = 0_n \iff c = 0_K\,.$$

If there is a $c \neq 0_K$ such that $Xc = 0_n$, than you have a "perfect collinearity" situation, and OLS is infeasible. In practice, your OLS estimates can be unreliable or infeasible if the columns of X are "nearly perfectly collinear", a situation we call "multicollinearity". OLS might be infeasible because the computations (particularly the inverting of $X^{\mathrm{T}}X$) become numerically unstable. Even where OLS is feasible, you will generally get large standard errors where there is strong multicollinearity.

Having obtained OLS estimates, the OLS fitted values can be calculated as $\hat{y}_{ols} = X\hat{\beta}^{ols}$ and the OLS residuals can be calculated as $\hat{\varepsilon}_{ols} = y - \hat{y}_{ols}$.

If $E(\varepsilon \mid X) = 0$, the OLS estimator is unbiased. If the homoskedasticity assumption holds, then $Var(\hat{\beta}_{ols} \mid X) = \sigma^2(X^{\mathrm{T}}X)^{-1}$. The noise variance σ^2 can be estimated using

$$\widehat{\sigma^2} = \frac{\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}}{n-K}\,.$$

An estimator of the variance-covariance matrix of $\hat{\beta}^{ols}$ is then

$$\widehat{Var}(\hat{\beta}^{ols} \mid X) = \widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}\,.$$

If errors are homoskedastic, than the OLS estimators of the linear regression model turns out to be "Best Linear Unbiased". That is, given any other linear unbiased estimator $\tilde{\beta}$ of the linear regression model, we have

$$Var(c^{\mathrm{T}}\hat{\beta}^{ols} \mid X) \leq Var(c^{\mathrm{T}}\tilde{\beta} \mid X)$$

for any $K \times 1$ vector $c \neq 0_K$. For goodness-of-fit, the R^2 statistic

$$R^2 = 1 - \frac{\hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols}}{y^{\mathrm{T}}M_0 y} = 1 - \frac{RSS}{TSS}$$

measures the proportion of variation in Y_i is accounted for by the explanatory variables. It has the feature that $0 \leq R^2 \leq 1$, as long as an intercept is included and estimation was done by *ordinary* least squares. The R^2 never decreases (and usually increases) as more regressors are

included in the regression. To counter this, the adjusted-R^2

$$\text{adj.-}R^2 = 1 - \frac{RSS/n - K}{TSS/n - 1}\,.$$

is sometimes reported.

For testing the single hypothesis $H_0 : r^{\mathrm{T}}\beta = r_0 \quad \text{vs} \quad r^{\mathrm{T}}\beta \neq r_0$, where r is $K \times 1$ and r_0 is a scalar, we can use the t-test

$$t = \frac{r^{\mathrm{T}}\hat{\beta} - r_0}{\sqrt{r^{\mathrm{T}}\widehat{Var}(\hat{\beta}^{ols} \mid X)r}}\,.$$

For multiple hypotheses $H_0 : \mathcal{R}\beta = r_0 \quad \text{vs} \quad H_A : \mathcal{R}\beta \neq r_0$, where $\mathcal{R}$ is a $(J \times K)$ matrix and r_0 is a $(J \times 1)$ vector, you can use the F-test

$$\begin{aligned} F &= \frac{(\hat{\varepsilon}_{rls}^{\mathrm{T}}\hat{\varepsilon}_{rls} - \hat{\varepsilon}_{ols}^{\mathrm{T}}\hat{\varepsilon}_{ols})/J}{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}/(n-K)} \\ &= \frac{(R^2_{ur} - R^2_r)/J}{(1 - R^2_{ur})/(n-K)} \\ &= (\mathcal{R}\hat{\beta} - r_0)^{\mathrm{T}}(\mathcal{R}\,\widehat{Var}(\hat{\beta}^{ols} \mid X)\,\mathcal{R}^{\mathrm{T}})^{-1}(\mathcal{R}\hat{\beta} - r_0)/J \end{aligned}$$

where $\hat{\varepsilon}_{rls}$ are the restricted least squares estimator of β, R^2_r is the R^2 of this restricted regression, and R^2_{ur} is the R^2 of the unrestricted regression. If the noise terms ϵ_i are normally distributed, then

$$t \sim \mathrm{t}(n-K) \quad \text{and} \quad F \sim \mathrm{F}(J, n-K)\,.$$

Otherwise, we can usually rely on large-sample approximate tests

$$t \overset{a}{\sim} \text{Normal}(0,1) \quad \text{and} \quad JF \overset{a}{\sim} \chi^2(J)\,.$$

Over the next few sections, we mention some of the things that can go wrong in a regression analysis, some of which we have already mentioned previously. Where feasible, we will say something about how we might detect that there is a problem in the first place, and how to fix it. In other cases (particularly the later examples in this chapter), the "fixes" will be discussed in later chapters. We begin with two fairly innocuous assumptions, normality of the noise term and homoskedasticity.

7.2 Normality of Noise Term

Non-normality of errors does not affect unbiasedness or consistency of the OLS estimator. Most of the theory of OLS stands without this assumption. Its primary function is to provide the finite sample distribution for the t- and F-statistics. One way to test for normality is to use the fact that a normal random variate has skewness coefficient is zero (because it is symmetric), and its kurtosis coefficient is three: if $X \sim \text{Normal}(\mu, \sigma^2)$, then

$$S = E((X-\mu)^3)/\sigma^3 = 0 \quad \text{and} \quad Kur = E((X-\mu)^4)/\sigma^4 = 3\,.$$

The kurtosis coefficient, being the expectation of a fourth moment, emphasizes larger deviations from mean over small deviations from mean (deviations from mean less than one become very small when raised to the fourth power). A kurtosis coefficient greater than 3 suggests higher probability of large deviations from mean, relative to a comparable normally distributed random variable. The skewness and kurtosis coefficients can be estimated using

$$\widehat{S} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{3/2}} \quad \text{and} \quad \widehat{Kur} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{2}}.$$

The Jarque-Bera statistic applies this idea to regression residuals, using the statistic

$$JB = \frac{n-K}{6}\left(\widehat{S}^2 + \frac{1}{4}(\widehat{Kur} - 3)^2\right)$$

which is approximately $\chi^2(2)$ in large samples under the null. Some implementations ignore the degree-of-freedom correction and use n in the numerator instead of $n-K$.

We test for normality of the residuals in the regression

$$\ln earn_i = \beta_0 + \beta_1 \ln wexp_i + \beta_2 \ln tenure_i + \epsilon_i$$

```
Skew <- function(x){
  return(mean((x-mean(x))^3)/(mean((x-mean(x))^2)^(3/2)))
}
Kurt <- function(x){
  return(mean((x-mean(x))^4)/(mean((x-mean(x))^2)^2))
}
JB <- function(mdl){
  # requires lm object, returns JB Stat, p-val, Skewness and Kurtosis Coef.
  N <- nobs(mdl)
  K <- summary(mdl)$df[1]
  ehat <- residuals(mdl)
  JBSkew <- Skew(ehat)
  JBKurt <- Kurt(ehat)
  JBstat <- ((N-K)/6*(JBSkew^2 + (1/4)*(JBKurt-3)^2))
  JBpval <- 1-pchisq(JBstat,2)
  return(list("JBstat"=JBstat, "JBpval"=JBpval, "Skewness"=JBSkew, "Kurtosis"=JBKurt))
}

df_earn <-read_csv("data\\earnings2019.csv",show_col_types=FALSE)
mdl <- lm(log(earn)~log(wexp)+log(tenure), data=df_earn)
JBtest <- JB(mdl)
fmt <- function(x){format(round(x,4),nsmall=4)}
cat("JB:", fmt(JBtest$JBstat), "  p-val:", fmt(JBtest$JBpval),
    "  Skewness:", fmt(JBtest$Skewness), "  Kurtosis:", fmt(JBtest$Kurtosis),"\n")
```

```
JB: 338.6182   p-val: 0.0000   Skewness: 0.0816   Kurtosis: 4.2718
```

The null of normality is rejected. The distribution is quite symmetric, but there is 'excess kurtosis' (kurtosis in excess of 3). A histogram of the OLS residuals is shown below.

```
hist(residuals(mdl), 20)
```

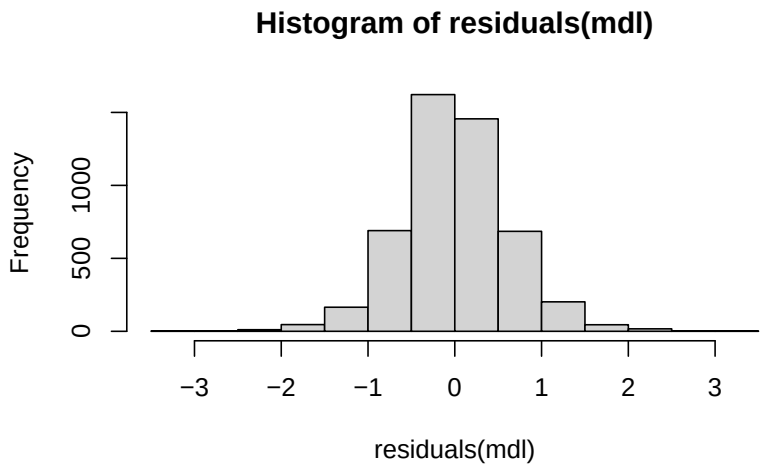


Figure 7.1: Residual histogram.

The problem with highly skewed and very fat-tailed noise distributions is that the CLT may converge slowly, so that the asymptotic tests may be poor approximations unless sample size is sufficiently large. Here we do not have a skewness problem, but the estimated kurtosis is above four, which is probably large enough that we would want a reasonably large sample size. We have almost 5000 observations which should be more than enough. Sometimes the Y variable is transformed to make the noise term "more normal". In fact, in our example our log transformation of *earn* contributes substantially to the normality of the noise term.

7.3 Heteroskedasticity

We know that heteroskedasticity of the noise terms does not affect the unbiasedness or consistency of OLS estimators of the linear regression model. As explained in previous chapters, the only concern regarding heteroskedasticity when using OLS is to calculate the standard errors properly. In particular, we ought to use

$$\widehat{Var}_{HC0}(\hat{\beta}^{ols}) = \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}$$

or one of its variants, instead of the usual $\widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}$. The benefit of using the heteroskedasticity-robust variance-covariance matrix is that it remains consistent for $Var(\hat{\beta}^{ols})$ even under homoskedasticity, so one can argue that standard errors reported by programs ought to be obtained from it by default.[1] The heteroskedasticity-robust version can also be used in the expressions for the t and F statistics, replacing $\widehat{Var}(\hat{\beta}^{ols})$, to get heteroskedasticity-robust t and F tests.

We know, however, that OLS estimators are efficient when the noise terms are homoskedastic, and may not be so otherwise. The question, then, is whether or not we can do better than OLS

[1] Unfortunately, default standard errors in virtually all econometric software libraries are based on the homoskedastic version of the variance-covariance matrix of $\widehat{\beta}^{ols}$.

in terms of getting more precise estimators when the noise terms are heteroskedastic.

We begin with a simple illustrative example where we can directly show the inefficiency of OLS under heteroskedasticity.

Example 7.1. Suppose

$$Y_i = \beta_1 X_{i1} + \epsilon_i \,,\; i = 1, 2, ..., n \,. \tag{7.1}$$

We assume all of the usual OLS assumptions continue to hold, except that

$$E(\epsilon_i^2 \mid X_{11}, \ldots, X_{n1}) = \sigma^2 X_{i1}^2 \quad \text{for all} \;\; i = 1, \ldots, n \,.$$

The OLS estimator for β_1 in this example is

$$\hat{\beta}_1^{ols} = \frac{\sum_{i=1}^N X_{i1} Y_i}{\sum_{i=1}^N X_{i1}^2} \tag{7.2}$$

which is unbiased for β_1 (see exercises). Under our assumptions, the variance of the OLS estimator is

$$\begin{aligned} Var(\hat{\beta}_1^{ols} \mid X_{11}, \ldots, X_{n1}) &= \frac{\sum_{i=1}^N X_{i1}^2 \, Var(\epsilon_i \mid X_{11}, \ldots, X_{n1})}{\left(\sum_{i=1}^N X_{i1}^2\right)^2} \\ &= \frac{\sum_{i=1}^N X_{i1}^2 (\sigma^2 X_{i1}^2)}{\left(\sum_{i=1}^N X_{i1}^2\right)^2} \\ &= \frac{\sigma^2 \sum_{i=1}^N X_{i1}^4}{\left(\sum_{i=1}^N X_{i1}^2\right)^2} \,. \end{aligned} \tag{7.3}$$

We will show that this estimator is inefficient, by presenting a more efficient linear unbiased estimator. First, weight each observation by $1/X_{i1}$ and run the regression

$$\frac{Y_i}{X_{i1}} = \beta_1 + \frac{\epsilon_i}{X_{i1}} = \beta_1 + \epsilon_i^* \,. \tag{7.4}$$

That is, simply regress Y_i/X_{i1} on a constant. The modified noise terms in this regression will continue to have zero conditional expectation

$$E(\epsilon_i / X_{i1} \mid X_{i1}, \ldots, X_{n1}) = (1/X_{i1}) E(\epsilon_i \mid X_{i1}, \ldots, X_{n1}) = 0$$

and remain uncorrelated (exercise). Furthermore, its conditional variance is now constant:

$$Var(\epsilon_i / X_{i1} \mid X_{i1}, \ldots, X_{n1}) = (1/X_{i1}^2) \, Var(\epsilon_i \mid X_{i1}, \ldots, X_{n1}) = \sigma^2 \,.$$

OLS estimation applied to this modified regression model (7.4) gives the estimator

$$\hat{\beta}_1^{wls} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{X_{i1}} \,. \tag{7.5}$$

The 'wls' superscript stands for "weighted least squares". It is easy to show that (7.5) is unbiased (exercise). Since (7.5) can be written as

$$\widehat{\beta}_1^{wls} = \sum_{i=1}^{n} \left(\frac{1}{nX_{i1}} \right) Y_i \,,$$

it is also a linear estimator. In fact, it is BLU since the regression satisfies all the necessary conditions for OLS to be BLU. The fact that $\widehat{\beta}_1^{wls}$ is BLU suggests that $\widehat{\beta}_1^{ols}$ is not. In our simple example, it is straightforward to directly demonstrate that

$$Var(\widehat{\beta}_1^{wls} \mid X_{11}, \dots, X_{n1}) \leq Var(\widehat{\beta}_1^{ols} \mid X_{11}, \dots, X_{n1}) \,.$$

Since $\widehat{\beta}_1^{wls}$ is a sample mean of n observations of a random variable with variance σ^2, its variance is

$$Var(\widehat{\beta}_1^{wls} \mid X_{11}, \dots, X_{n1}) = \frac{\sigma^2}{n} \,. \tag{7.6}$$

It can be shown (see exercises) that

$$\frac{\sum_{i=1}^{n} X_{i1}^4}{\left(\sum_{i=1}^{n} X_{i1}^2\right)^2} \geq \frac{1}{n} \,, \tag{7.7}$$

therefore

$$Var(\widehat{\beta}_1^{wls} \mid X_{11}, \dots, X_{n1}) = \frac{\sigma^2}{n} \leq \frac{\sigma^2 \sum_{i=1}^{N} X_{i1}^4}{\left(\sum_{i=1}^{N} X_{i1}^2\right)^2} = Var(\widehat{\beta}_1^{ols} \mid X_{11}, \dots, X_{n1}) \,.$$

The reason OLS estimators are inefficient when there is conditional heteroskedasticity is that the observations with large noise variances are less informative about the population regression line than the ones with smaller noise variances, but OLS makes no use of this fact. Information ignored leads to inefficiency. The weighted least squares approach, on the other hand, uses this information directly by assigning less weight to noisier observations, and more weight to observations whose noise terms have lower variance.

7.3.1 Weighted Least Squares

Suppose our linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i \,, \; i = 1, \dots, n,$$

with $E(\epsilon_i \mid X) = 0 \,, \; i = 1, \dots, n$, $E(\epsilon_i \epsilon_j \mid X) = 0$, $i, j = 1, \dots, n$ and $i \neq j$, and

$$E(\epsilon_i^2 \mid X) = \sigma_i^2 = \sigma^2 \eta_i \,, \; i = 1, \dots, n$$

where η_i is a **completely known** function of the regressors. In other words, in $\sigma^2 \eta_i$, the only unknown parameter is σ^2. For instance,

$$\sigma_i^2 = \sigma^2 |X_i| \quad \text{or} \quad \sigma_i^2 = \sigma^2 X_i^2 \quad \text{or} \; \sigma_i^2 = \sigma^2 \exp(X_i) \,.$$

Put yet another way, we assume the form of conditional heteroskedasticity is completely known up to a constant factor.

The idea of weighted least squares is to weight each observation so that the weighted noise terms are no longer heteroskedastic. That is, we modify the regression equation to

$$\frac{Y_i}{\sqrt{\eta_i}} = \beta_0 \frac{1}{\sqrt{\eta_i}} + \beta_1 \frac{X_{i1}}{\sqrt{\eta_i}} + \cdots + \beta_{K-1} \frac{X_{i,K-1}}{\sqrt{\eta_i}} + \frac{\epsilon_i}{\sqrt{\eta_i}}$$

which we can write as

$$Y_i^* = \beta_0 X_{0,i}^* + \beta_1 X_{i1}^* + \cdots + \beta_{K-1} X_{i,K-1}^* + \epsilon_i^* \,. \tag{7.8}$$

Since η_i is fixed conditional on the regressors, we have

$$E(\epsilon_i^* \mid X) = E\left(\left(\frac{\epsilon_i}{\sqrt{\eta_i}} \right) \,\middle|\, X \right) = \frac{1}{\sqrt{\eta_i}} E(\epsilon_i \mid X) = 0 \,;$$

$$E({\epsilon_i^*}^2 \mid X) = E\left(\left(\frac{\epsilon_i}{\sqrt{\eta_i}} \right)^2 \,\middle|\, X \right) = \frac{1}{\eta_i} E(\epsilon_i^2 \mid X) = \frac{1}{\eta_i} \sigma^2 \eta_i = \sigma^2 \,;$$

$$E(\epsilon_i^* \epsilon_j^* \mid X) = E\left(\left(\frac{\epsilon_i \epsilon_j}{\sqrt{\eta_i}\sqrt{\eta_j}} \right) \,\middle|\, X \right) = \frac{1}{\sqrt{\eta_i}\sqrt{\eta_j}} E(\epsilon_i \epsilon_j \mid X) = 0 \,.$$

OLS estimation of (7.8) will produce BLU estimators of the coefficients since all the necessary conditions for OLS to be BLU are met.

Applying OLS on the transformed regression equation (7.8) is equivalent to choosing $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_{K-1}$ to minimize

$$\begin{aligned} &\sum_{i=1}^{n} (Y_i^* - \hat{\beta}_0 X_{i0}^* - \hat{\beta}_1 X_{i1}^* - \cdots - \hat{\beta}_{K-1} X_{i,K-1}^*)^2 \\ =&\sum_{i=1}^{n} (Y_i/\sqrt{\eta_i} - \hat{\beta}_0 (1/\sqrt{\eta_i}) - \hat{\beta}_1 X_{i1}/\sqrt{\eta_i} - \cdots - \hat{\beta}_{K-1} X_{i,K-1}/\sqrt{\eta_i})^2 \\ =&\sum_{i=1}^{n} \frac{1}{\eta_i} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_{K-1} X_{i,K-1})^2 \,. \end{aligned} \tag{7.9}$$

That is, we are minimizing a sum of *weighted* squared residuals $\sum_{i=1}^{n} \omega_i \hat{\epsilon}_i^2$ where the weights are

$$\omega_i = 1/\eta_i \,.$$

and where the heteroskedasticity is of the form $\sigma_i^2 = \sigma^2 \eta_i$. You can also think of this as the reason for the name "Weighted Least Squares".

After obtaining $\hat{\beta}_0^{wls}, \ldots, \hat{\beta}_{K-1}^{wls}$, you should report your results as

$$\hat{Y} = \hat{\beta}_0^{wls} + \hat{\beta}_1^{wls} X_1 + \ldots \hat{\beta}_{K-1}^{wls} X_{K-1} \,,$$

with standard errors, of course. The WLS fitted values and residuals are computed as

$$\hat{Y}_i^{wls} = \hat{\beta}_0^{wls} + \hat{\beta}_1^{wls} X_{i1} + \cdots + \hat{\beta}_{K-1}^{wls} X_{i,K-1}$$

$$\hat{\epsilon}_{i,wls} = Y_i - \hat{Y}_i^{wls} = Y_i - \hat{\beta}_0^{wls} - \hat{\beta}_1^{wls} X_{i1} - \cdots - \hat{\beta}_{K-1}^{wls} X_{i,K-1}.$$

For assessing goodness-of-fit, we should use the WLS residuals to construct the R^2:

$$R^2_{wls} = 1 - \frac{\sum_{i=1}^N \hat{\epsilon}^2_{i,wls}}{\sum_{i=1}^N (Y_i - \overline{Y})^2}. \tag{7.10}$$

The R^2_{wls} will generally be less than the R^2 from OLS estimation (why?) and may even be negative.

Example 7.2. We estimate a simple linear regression on the values y and x (with intercept) in the data set *heterosk.csv*. Fig. 7.2 displays a plot of y on x that shows heteroskedasticity, with the conditional variance is increasing with x.

```
df_het <- read_csv("data\\heterosk.csv",col_types = c("n","n","n"))
ggplot(data=df_het) + geom_point(aes(x=x,y=y), size=1) + theme_classic()
```

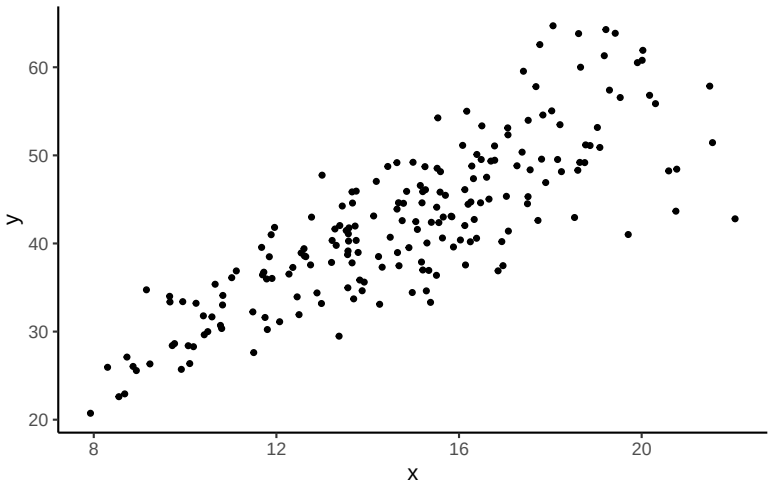


Figure 7.2: A data set with heteroskedasticity.

OLS estimation of this regression gives the following output.

```
ols <- lm(y~x, data=df_het)
sum_ols <- summary(ols)
coef(sum_ols)
cat("R-squared: ",sum_ols$r.squared,"\n")
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 6.285792  1.7817576  3.52786 5.207066e-04
x           2.430744  0.1177712 20.63955 3.034896e-51
R-squared:  0.6826877
```

Because there is obvious heteroskedasticity in this example, we should not trust the standard errors, t-statistics and p-values presented above. If we wish to stick with OLS, then we have to calculate the heteroskedasticity-robust standard error, which we do below using the `vcovHC()`

function from the `sandwich` package. The heteroskedasticity-robust standard errors are the square root of the diagonal elements of the matrix `rbst_V` in the example below.

```
rbst_V <- sandwich::vcovHC(ols, type="HC0")
rbst_se <- sqrt(diag(rbst_V))
rbst_output <- coef(sum_ols)
colnames(rbst_output) <- c("Estimate", "rbst-se", "rbst-t", "p-val")
rbst_output[,'rbst-se'] <- rbst_se
rbst_output[,'rbst-t'] <- rbst_output[,'Estimate']/rbst_se
rbst_output[,'p-val'] <- 2*(1-pt(abs(rbst_output[,'rbst-t']),sum_ols$df[2]))
round(rbst_output,6)
```

```
            Estimate  rbst-se    rbst-t    p-val
(Intercept) 6.285792 1.863926  3.372339 0.000896
x           2.430744 0.136242 17.841435 0.000000
```

Now we assume that $Var(\epsilon_i \mid X) = \sigma^2 X_i^2$, which seems reasonable given the scatterplot in Fig. 7.2. We run WLS using the `weights` option in `lm()` function. Note that the option `weights` refer to weights on the squared residuals, as in the ω_i in (7.9), i.e., it is $1/\eta_i$ if $\sigma_i^2 = \sigma^2 \eta_i$.

```
df_het$wt <- 1/df_het$x^2
wls2 <- lm(y~x,data=df_het, weights=wt)
sum_wls2 <- summary(wls2)
coef(sum_wls2)
cat("R-squared: ", sum_wls2$r.squared,"\n")
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  4.82571  1.4065451  3.430896 7.321779e-04
x            2.53166  0.1023702 24.730430 1.839271e-62
R-squared:  0.755433
```

The standard errors are lower than in the OLS regression, which is not unexpected. Note that the R^2 reported above is not the R^2_{wls} that was recommended in (7.10). We calculate R^2_{wls} below

```
ehat <- df_het$y - coef(wls2)[1] - coef(wls2)[2]*df_het$x
rss <- sum(ehat^2)
tss <- sum((df_het$y - mean(df_het$y))^2)
R2 <- 1 - rss/tss
cat("R-square: ", R2,"\n")
```

```
R-square:  0.6814966
```

We see that R^2_{wls} is lower than in the OLS regression, as expected, but only slightly so. The R^2 provided by `lm()` when using `weights` is a "weighted R-squared", obtained in the following way: let $\overline{Y}_{wls}$ be the WLS estimator (with the same weights as previously) of β_0 from the regression $Y_i = \beta_0 + \epsilon_i$. In other words, $\overline{Y}_{wls}$ is the weighted mean of $\{Y_i\}_{i=1}^N$. Then

$$\text{weighted-}R^2 = \frac{\sum_{i=1}^N w_i(\hat{Y}_i^{wls} - \overline{Y}_{wls})^2}{\sum_{i=1}^N w_i(Y_i^{wls} - \overline{Y}_{wls})^2}.$$

In other words, it is the weighted fitted sum of squares divided by the weighted total sum of squares, centered on the weighted mean of $\{Y_i\}_{i=1}^N$. We replicate the `lm()` weighted R-squared below:

```
wls0 <- lm(y~1, data=df_het, weights=wt)  # Regression on intercept only
tss_wtd <- sum(df_het$wt * (df_het$y - coef(wls0))^2)
fss_wtd <- sum(df_het$wt*(wls2$fitted.values - coef(wls0))^2)
WeightedR2 <- fss_wtd/tss_wtd
WeightedR2
```

```
[1] 0.755433
```

One difficulty with WLS is that we generally do not know the form of the heteroskedasticity. In the simple linear regression case, should η_i be $|X_{i1}|$ or X_{i1}^2 or some other function of the regressors? The problem of specifying the form of the heteroskedasticity becomes more challenging in the multivariate regression case

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_{K-1,i} X_{K-1,i} + \epsilon_i \,, i = 1, \ldots, n \,.$$

Any heteroskedasticity is likely to depend on more than one regressor, to different degrees, so there will be parameters to estimate. E.g., we might have something like

$$\eta_i = \exp(\alpha_1 X_{i1} + \cdots + \alpha_{K-1} X_{i,K-1})$$

(the exponentiation is to ensure the variance is positive). Then to implement WLS, we first have to estimate the parameters in the variance equation. One way to do this is to first estimate the main equation using OLS, and obtain the OLS residuals. Then regress the log of the squared residuals on a constant and the regressors to estimate σ^2 and the α parameters, and then finally compute the "fitted variances" $\widehat{\sigma_i^2}$ and weight the squared residuals by $1/\widehat{\sigma_i^2}$.

7.3.2 Testing for Heteroskedasticity

It may be of interest to test whether heteroskedasticity is an issue in the first place. The following are some possible tests. All involve first estimating the main regression by OLS and obtaining the OLS residuals $\hat{\epsilon}_{i,ols}$. First, run the regression

$$\hat{\epsilon}_{i,ols}^2 = \alpha_0 + \alpha_1 X_{i1} + \ldots \alpha_{K-1} X_{i,K-1} + u_i$$

and test $H_0 : \alpha_1 = \cdots = \alpha_{K-1} = 0$ using an F test. An alternative is to use an "LM" test after running the regression above: under the null hypothesis, we have

$$N R_\epsilon^2 \overset{a}{\sim} \chi^2(K).$$

To allow for possible non-linear forms we can include powers of regressors and interaction terms between them in the variance regression:

$$\hat{\epsilon}_{i,ols}^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_{K-1} X_{i,K-1} + \delta_1 X_{i1}^2 + \cdots + \delta_{K-1} X_{i,K-1}^2 + \gamma_{12} X_{i1} X_{i2} + \cdots + u_i$$

then testing if all of the coefficients (not including the intercept) are zero. Obviously you lose degrees of freedom quickly as the number of regressors grow. One way around this problem is to run the regression

$$\hat{\epsilon}_{i,ols}^2 = \alpha_0 + \alpha_1 \hat{Y}_{i,ols} + \alpha_2 \hat{Y}_{i,ols}^2 + u_i$$

where $\hat{Y}_{i,ols}$ refers to the OLS fitted values from the main equation. The hypothesis that there is no heteroskedasticity is $H_0 : \alpha_1 = \alpha_2 = 0$ using an F test or an LM test.

The first approach is often referred to as Breusch-Pagan tests for heteroskedasticity. The approach using OLS fitted values is called the White test for heteroskedasticity.

Example 7.3. We apply the Breusch-Pagan test for heteroskedasticity to the regression

$$\ln earn = \beta_0 + \beta_1 \ln wexp_i + \beta_2 \ln tenure_i + \epsilon_i.$$

```
mdl <- lm(log(earn)~log(wexp)+log(tenure), data=df_earn)  #--Main Equation
cat("Main Regression\n")                                  #--Main Regr Output Title
round(summary(mdl)$coefficients,4)                        #--Main Regr Output
```

```
Main Regression
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   2.9167     0.0229 127.3595        0
log(wexp)    -0.0411     0.0095  -4.3294        0
log(tenure)   0.1746     0.0091  19.1215        0
```

```
df_earn$ehat <- residuals(mdl)                                #--Get OLS Residuals
heteq <- lm((ehat^2)~log(wexp)+log(tenure), data=df_earn)  #--BP-Test Regression
cat("Heteroskedasticity Test Regression\n")                   #--Test Regr Output Title
round(summary(heteq)$coefficients,4)                          #--Test Regr Output
```

```
Heteroskedasticity Test Regression
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4668     0.0261 17.9070   0.0000
log(wexp)    -0.0224     0.0108 -2.0728   0.0382
log(tenure)  -0.0167     0.0104 -1.6028   0.1090
```

```
## BP, F-version
f_het <- summary(heteq)$fstatistic  #--Retrieve F-Stat (stat, df1, df2)
cat("BP-F Stat: ", f_het[1], "    p-val: ", 1-pf(f_het[1], f_het[2], f_het[3]), "\n")
```

```
BP-F Stat:  4.291278     p-val:  0.01373845
```

```
## BP, LM-version
lm_het <- nobs(heteq)*summary(heteq)$r.squared        # Calc. LM Stat.
lm_pval <- 1 - pchisq(lm_het, 2)  # 2 restrictions
cat("BP-LM Stat: ", lm_het, "     p-val: ", lm_pval, "\n", sep="")
```

```
BP-LM Stat: 8.57288     p-val: 0.0137538
```

There is some evidence of heteroskedasticity, and the individual t-tests suggest that the noise variance decreases with work experience.

7.4 Misspecification of Conditional Expectation

If you specify that $E(Y \mid X_1) = \beta_0 + \beta_1 X_1$ but the conditional expectation does not have this form, then parameters become hard to interpret, if not meaningless, and predictions from the model will usually be biased. Fortunately, as we have seen, the multiple linear regression model provides one with substantial flexibility in specify the functional form, by transforming variables (including transforming the dependent variable), adding quadratic terms, using piecewise linear or piecewise quadratic forms, adding interaction terms, and many other possibilities. Of course, it may well be that the conditional expectation cannot be written in linear-in-parameters form,

but often the multiple linear regression framework provides sufficient flexibility to at least get what appears to be an appropriate specification.

As a rough check for functional form adequacy, one can always plot residuals against individual regressors. If the functional specification is adequate, the residuals should look like random noise. For instance, if $E(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$ but you specify $E(Y \mid X) = \beta_0 + \beta_1 X_1$, then the residuals will display a quadratic pattern with respect to X_1. Two more formal checks for functional form adequacy include the RESET test and tests of non-nested alternatives.

7.4.1 RESET test for functional form misspecification

Given a regression specification

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \epsilon_i \,,$$

the Regression Equation Specification Error Test (or "RESET Test") checks if adding powers ($X_{1,i}^2$, $X_{2,i}^2$,...) and interaction terms ($X_{1,i}X_{2,i}$, $X_{1,i}X_{3,i}$, etc.) of the regressors can significantly improve the fit. Of course, this can lead to severe loss of degrees-of-freedom, and may cause multicollinearity issues. The trick employed by the RESET test is to add the squares, cubes, and possibly higher powers of the original OLS fitted values $\hat{Y}_{i,ols}$ into the regression specification and tests if these additions have significant explanatory power. For instance, the squared fitted value is a linear combination of the squares of all the regressions and the interaction terms between them, so the hope is that if the squares of some of the regressors are important, this will get reflected as a significant coefficient on $\widehat{Y}_{i,ols}^2$ when it is added to the test equation. The test equation is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{K-1} X_{i,K-1} + \alpha_2 \hat{Y}_{i,ols}^2 + \cdots + \alpha_p \hat{Y}_{i,ols}^p + \epsilon_i \,, \tag{7.11}$$

and the hypothesis of adequacy of the functional form specification is

$$H_0 : \alpha_2 = \cdots = \alpha_p = 0.$$

The test equation cannot include $\hat{Y}_{i,ols}$ (see exercises), and often only the second or second and third powers are included. An F-test (or t-test, if only the second power is included) can be used to test the hypothesis. We illustrate the RESET test in the next example. Note that the RESET test is interpreted as a test of adequacy of the *functional form specification* of the original regression, and not as saying anything about the whether or not certain variables should or should not be included.

Example 7.4. We apply the RESET test to the regression

$$\ln earn_i = \beta_0 + \beta_1 wexp_i + \beta_2 tenure_i + \epsilon_i.$$

using data in `earnings2019.csv`.

```
mdl_base <- lm(log(earn)~wexp+tenure, data=df_earn)
df_earn$yhat <- fitted(mdl_base)
mdl_test <- lm(log(earn)~wexp+tenure+I(yhat^2), data=df_earn)
cat("Base Regression:\n")
```

```
round(summary(mdl_base)$coefficients, 4)
cat("\nTest Regression:\n")
round(summary(mdl_test)$coefficients, 4)
```

```
Base Regression:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   3.0249     0.0156 194.1339        0
wexp         -0.0049     0.0011  -4.3724        0
tenure        0.0187     0.0011  17.1674        0

Test Regression:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3285     2.3951  7.2350        0
wexp         -0.0531     0.0082 -6.5134        0
tenure        0.2092     0.0319  6.5528        0
I(yhat^2)    -1.5680     0.2625 -5.9722        0
```

The hypothesis of adequacy of the functional form in the base regression is rejected.

7.4.2 Testing Nonnested Alternatives

Regression specifications such as

$$\text{[A]} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$$\text{and} \quad \text{[B]} \quad Y_i = \beta_0 + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2} + \epsilon_i$$

are "non-nested alternatives", i.e., one is not a special case of the other. One way of testing which specification fits better is to construct a "super-model" that includes both [A] and [B] as restricted cases, i.e.,

$$\text{[A]} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 \ln X_{i1} + \beta_4 \ln X_{i2} + \epsilon_i$$

and to test for coefficient significance. This approach is often plagued by multicollinearity problems. An alternative is to fit both models separately, collect their fitted values, and include each fitted value series as a regressor in the other specification, i.e., regress

$$\text{[A']} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \delta_1 \hat{Y}_i^B + \epsilon_i$$

$$\text{and} \quad \text{[B']} \quad Y_i = \beta_0 + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2i} + \delta_2 \hat{Y}_i^A + \epsilon_i$$

and test (separately) if the coefficients on the fitted values are statistically significant. The idea is to see if each specification has anything to add to the other. If $\delta_1 = 0$ is rejected and $\delta_2 = 0$ is not, then this suggests that [B] is a better specification (the result does not suggest [B] is the *best* specification, just better than [A].) Likewise, [A] is preferred to [B] if $\delta_2 = 0$ is rejected and $\delta_1 = 0$ is not. It may be that both are rejected, which suggests that neither specification is adequate. If neither are rejected, then it appears that there is little in the data to distinguish between the two specifications. Note that the dependent variable in both alternatives must be the same.

We compare the specifications

$$\begin{aligned} &[\text{A}] \quad \ln earn_i = \beta_0 + \beta_1 wexp_i + \beta_2 tenure_i + \epsilon_i \\ \text{and} \quad &[\text{B}] \quad \ln earn_i = \beta_0 + \beta_1 \ln wexp_i + \beta_2 \ln tenure_i + \epsilon_i \,. \end{aligned}$$

```
mdlA <- lm(log(earn)~wexp+tenure, data=df_earn)
mdlB <- lm(log(earn)~log(wexp)+log(tenure), data=df_earn)
df_earn$yhatA <- fitted(mdlA)
df_earn$yhatB <- fitted(mdlB)
cat("Model A plus yhatB:\n")
mdlAplusB <- lm(log(earn)~wexp+tenure+yhatB, data=df_earn)
round(summary(mdlAplusB)$coefficients,4)
cat("\nModel B plus yhatA:\n")
mdlBplusA <- lm(log(earn)~log(wexp)+log(tenure)+yhatA, data=df_earn)
round(summary(mdlBplusA)$coefficients,4)
```

```
Model A plus yhatB:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2825     0.3331  0.8482   0.3963
wexp         -0.0012     0.0012 -0.9975   0.3186
tenure        0.0022     0.0023  0.9532   0.3406
yhatB         0.9075     0.1101  8.2430   0.0000

Model B plus yhatA:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5557     0.3606  7.0876   0.0000
log(wexp)    -0.0373     0.0103 -3.6330   0.0003
log(tenure)   0.1577     0.0192  8.2204   0.0000
yhatA         0.1219     0.1215  1.0033   0.3158
```

It appears that the specification [B] is preferred over specification [A].

7.5 Omitted Variables

We have already discussed the problem of omitted variables at length. Basically the problem is that you estimate

$$E(Y \mid X_1, \ldots, X_{K-1}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} \tag{7.12}$$

when you really want to be estimating

$$\begin{aligned} &E(Y \mid X_1, \ldots, X_{K-1}, X_K, \ldots, X_M) \\ &\quad = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \beta_K X_K + \cdots + \beta_{K+M} X_{K+M} \,. \end{aligned} \tag{7.13}$$

Maybe there are factors $X_1, \ldots, X_M$ that directly affect Y. You are interested in the direct effect of X_1 on Y, and you think to control for the factors X_2, ..., X_{K-1}, but not the others. If there are correlations between the omitted variables $X_K, \ldots, X_M$ and the included variables $X_1, \ldots, X_{K-1}$, then the coefficients on $X_1, \ldots, X_{K-1}$ in (7.12) will generally not be the same as the coefficients on $X_1, \ldots, X_{K-1}$ in (7.13), including the parameter β_1 of particular interest. If none of the variables in $X_K, \ldots, X_M$ are correlated with any of the variables in $X_1, \ldots, X_{K-1}$, then

the coefficients on $X_1, \ldots, X_{K-1}$ in (7.12) will be the same as the coefficients on $X_1, \ldots, X_{K-1}$ in (7.13), but even then, including the omitted variables will usually help reduce standard errors.

The solution to the omitted variable problem is, of course, to include the missing variables, but this assumes that the missing variables are available. In applications, often we want to include variables in a regression that are simply not available, or perhaps of low quality. As an example, in the returns to schooling equation

$$\ln earn = \beta_0 + \beta_1 educ + (\text{controls}) + \epsilon$$

Controls may include race and sex variables, tenure, age, etc. One control one would like to put in would be some measure of innate ability, but this is a notoriously difficult variable to estimate, an proxies such as IQ or grades may not even be available for the individuals in your sample.

There are a few solutions available (depending on the specifics of the problem, structure of the data, etc.). One solution is to use "instrumental variables" and "instrumental variable estimation". We explore this in the next chapter.

7.6 Sampling issues

So far we have assumed an iid random sample from the population. The problems we have discussed arise from misspecification issues (e.g., omitted variables, incorrect function form for the conditional expectation). Now we consider two examples where you may have the correct specification, but *sampling issues* lead to biased estimates.

7.6.1 Truncated Sampling

Consider the simple linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

and suppose β_1 is positive so you have a positively sloped PRF. Suppose you have a "truncated sample" where you cannot observe any observation where $Y_i > c$. This means that the only observations with larger values of X_i that are included in your sample will be the ones with lower or negative values of ϵ_i, since a large X_{i1} together with large positive ϵ_i makes $Y_i > c$ more likely. This implies a negative correlation between X and ϵ, and invalidates the assumption $E(\epsilon_i \mid X_{11}, \ldots, X_{n1}) = 0$. The following is an empirical illustration where the PRF has a positive slope, and observations with $Y_i > 1500$ are unavailable. The plot in Fig. 7.3 shows the full (black circles) and truncated (red x's) samples. The estimated OLS sample regression line for the full data set (black) and the truncated data set (red) are shown, illustrating the downward bias in $\hat{\beta}_1$.

```
set.seed(13)
X <- rnorm(100, mean=50, sd=20)
Y <- 1220 + 4*X + rnorm(100, mean=0, sd=50)
df_notrunc <- data.frame(X, Y)
df_trunc <- filter(df_notrunc, Y<=1500)
ggplot() +
  geom_point(data=df_notrunc,aes(x=X,y=Y), pch=1, size=2) +
```

```
  geom_smooth(data=df_notrunc,aes(x=X,y=Y), method="lm", se=FALSE, col="black") +
  geom_point(data=df_trunc, aes(x=X,y=Y), pch=4, size=2,col='red') +
  geom_smooth(data=df_trunc,aes(x=X,y=Y), method="lm", se=FALSE, col="red") +
  theme_minimal()
```

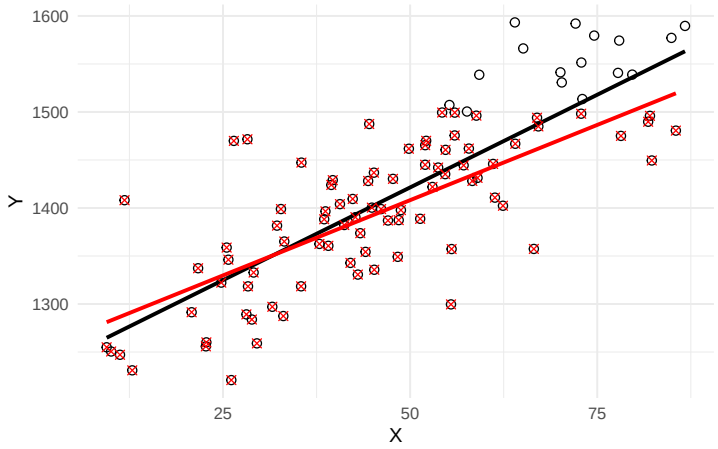


Figure 7.3: A truncated data set.

A related problem with similar outcome is **censored sampling**. Imagine, for instance, that the observations with Y above 1500 are not missing, but the actual value of Y is replaced with $Y = 1500$. The solutions to this and the truncated sampling problem requires us to go beyond the least squares/linear regression framework, and is covered in the chapter on limited dependent variables.

7.6.2 Measurement Error

Another kind of sampling issue is measurement error. Suppose $Y = \beta_0 + \beta_1 X + \epsilon$ describes the relationship between Y and X, but X is only observed with error, i.e., you observe $X^* = X + u$. Assume that the measurement error u is independent of X. Then

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon \\ &= \beta_0 + \beta_1 (X^* - u) + \epsilon \\ &= \beta_0 + \beta_1 X^* + (\epsilon - \beta_1 u) \\ &= \beta_0 + \beta_1 X^* + v \end{aligned}$$

where $v = \epsilon - \beta_1 u$. You proceed with what appears to be the only feasible option to you, which is to run the regression

$$Y = \beta_0 + \beta_1 X^* + v\,,$$

but since u is correlated with X^*, the assumption $E[v|X^*] = 0$ does not hold. In the simulated example below, we have a positively sloped PRF, shown in red, and measurement error in the regressor. Since β_1 is positive, X^* and v are negatively correlated, meaning that the error term v will tend to be positive for smaller $X*$ and negative for larger $X*$. This tendency is visible in Fig. 7.4 below. The red circles are the sample you would have observed with no measurement error. The black circles are the same sample points, but with measurement error in the X_i's.

```
set.seed(13)
X <- rnorm(100, mean=50, sd=20)
Y <- 1220 + 4*X + rnorm(100, mean=0, sd=10)
Xstar <- X + rnorm(100, mean=0, sd=10)
df_measerr <- data.frame(X, Y, Xstar)
ggplot() +
  geom_point(data=df_measerr,aes(x=Xstar,y=Y), pch=1, size=2) +
  geom_point(data=df_measerr,aes(x=X,y=Y), pch=1, size=2, col="red") +
  geom_abline(intercept=1220, slope=4, col='red') +
  geom_smooth(data=df_measerr, aes(x=Xstar, y=Y), method="lm", col='black',
              se=FALSE, linewidth=0.8) + xlab("X, Xstar") +
  theme_minimal()
```

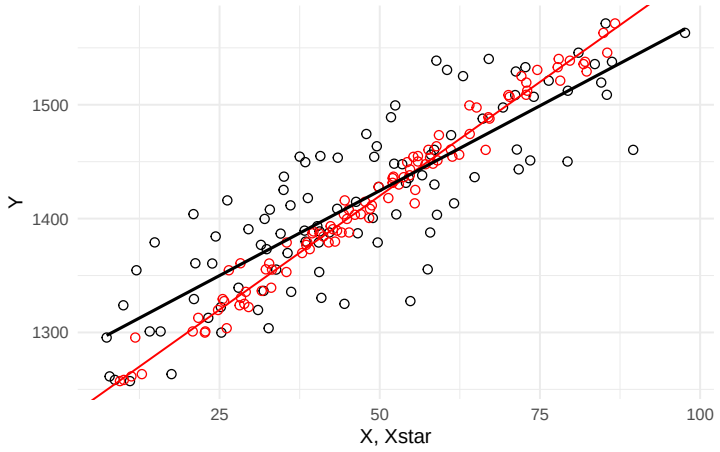


Figure 7.4: A data set with measurement error in the regressor.

The instrumental variable approach covered in the chapter on that topic may be able to handle this problem.

7.7 Simultaneity Bias

Suppose now that we actually don't want to estimate a conditional expectation, but we want instead to estimate a "structural equation". Consider a demand-and-supply example. Suppose the market for a good is governed by the following demand and supply equations

$$\begin{aligned} Q_t^d &= \delta_0 + \delta_1 P_t + \epsilon_t^d && \text{(Demand Eq } \delta_1 < 0) \\ Q_t^s &= \alpha_0 + \alpha_1 P_t + \epsilon_t^s && \text{(Supply Eq } \alpha_1 > 0) \\ Q_t^s &= Q_t^d && \text{(Market Clearing)} \end{aligned}$$

where Q and P represent log quantities and log prices respectively, so δ_1 and α_1 represent price elasticities of demand and supply respectively. Suppose the demand shock ϵ_t^d and supply shock ϵ_y^s are iid noise terms with zero means and variances σ_d^2 and σ_s^2 respectively, and are mutually uncorrelated. Market clearing means that observed quantity and prices occur at the intersection of the demand and supply equations, i.e., observed prices are such that

$$\delta_0 + \delta_1 P_t + \epsilon_t^d = \alpha_0 + \alpha_1 P_t + \epsilon_t^s$$

which we can solve to get

$$P_t = \frac{\alpha_0 - \delta_0}{\delta_1 - \alpha_1} + \frac{\epsilon_t^s - \epsilon_t^d}{\delta_1 - \alpha_1}. \tag{7.14}$$

Substituting this expression for prices into either the demand or supply equation gives

$$Q_t = \left(\delta_0 + \delta_1 \frac{\alpha_0 - \delta_0}{\delta_1 - \alpha_1}\right) + \frac{\delta_1 \epsilon_t^s - \alpha_1 \epsilon_t^d}{\delta_1 - \alpha_1}. \tag{7.15}$$

Equations 7.14 and 7.15 imply

$$Var(P_t) = \frac{\sigma_s^2 + \sigma_d^2}{(\delta_1 - \alpha_1)^2} \quad \text{and} \quad Cov(P_t, Q_t) = \frac{\delta_1 \sigma_s^2 + \alpha_1 \sigma_d^2}{(\delta_1 - \alpha_1)^2}.$$

This means that in a regression of $Q_t = \beta_0 + \beta_1 P_t + \epsilon_t$, we will get

$$\hat{\beta}_1 \xrightarrow{p} \frac{Cov(Q_t, P_t)}{Var(P_t)} = \frac{\delta_1 \sigma_s^2 + \alpha_1 \sigma_d^2}{\sigma_s^2 + \sigma_d^2} \tag{7.16}$$

which is neither the price elasticity of demand nor the price elasticity of supply, but a linear combination of the two.

Fig. 7.5 illustrates this issue. We simulate the demand and supply example with a series of random demand and supply shocks. Each demand shock shifts the demand curve and each supply shock shifts the supply curve resulting in a new intersection point. The set of equilibrium prices and quantity does not reflect either the demand or the supply function.

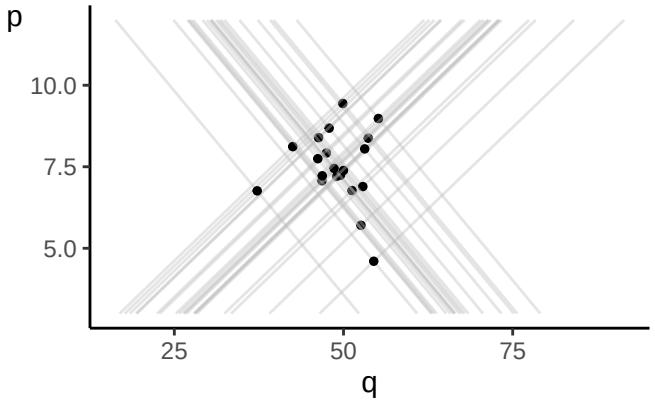


Figure 7.5: Simultaneity Bias.

The problem here is that prices and quantities are simultaneously determined by the intersection of the demand and supply functions; both prices and quantities are "endogenous" variables. The consequence of this is that regardless of whether you view the regression of Q_t on P_t as estimating the demand or supply equation, the noise term in the regression will be correlated with P_t. A supply shock shifts the supply function and changes both Q_t and P_t. Likewise, a demand shock shifts the demand function and again changes both Q_t and P_t. The use of the term "endogeneity" comes from applications like these, but it is now used for all situations where the noise term is correlated with one or more of the regressors.

7.8 Exercises

Exercise 7.1. Show that the OLS estimator (7.2) for β_1 in Example 7.1 is unbiased. Show also that the modified noise terms are uncorrelated, i.e.,

$$E(\epsilon_i^* \epsilon_j^* \mid X) = 0$$

for all $i \neq j$; $i, j = 1, 2, ..., N$.

Exercise 7.2. In the notes we claimed that

$$\frac{\sum_{i=1}^N X_i^4}{\left(\sum_{i=1}^N X_i^2\right)^2} \geq \frac{1}{N}.$$

Prove this by showing that for any set of values $\{z_i\}_{i=1}^N$, we have

$$N \sum_{i=1}^N z_i^2 - \left(\sum_{i=1}^N z_i\right)^2 \geq 0.$$

(*Hint: start with the fact that* $\sum_{i=1}^N (z_i - \bar{z})^2 \geq 0$.) Then substitute X_i^2 for z_i. When will equality hold?

Exercise 7.3.

a. Calculate the heteroskedasticity-robust standard errors for the regression in Example 7.3. Compare the robust standard errors with the OLS standard errors.

b. Estimate (using `lm()`) the main equation in Example 7.3 using WLS, assuming

$$Var(\epsilon_i \mid \{wexp\}_{i=1}^n, \{tenure\}_{i=1}^n) = \sigma^2 wexp_i\,.$$

Compare the WLS estimation results with the OLS estimation results (with heteroskedasticity robust standard errors).

Exercise 7.4.

a. Why is it that we cannot include $\hat{Y}_{i,ols}$ in the RESET test equation?

b. Add the third power of the OLS fitted value in the RESET test equation in Example 7.4. What happens to the statistical significance of the original regressors? Can you explain the likely cause? *Hint: what is the correlation between* `yhat^2` *and* `yhat^3`*?*

Exercise 7.5. We have shown that measurement error in the regressor leads to inconsistent estimators. Does measurement error in the regressand Y also result in inconsistent estimators?

Chapter 8

Instrumental Variables and GMM

We present the concept of instrumental variables, and an estimation method called Generalized Method of Moments (GMM). This extended framework enables consistent estimation of economic relationships in situations where there are endogeneity problems, i.e., where (for whatever reason) there are correlations between the noise term and one or more regressors. However, it requires the availability of good instrumental variables, which may not be so easy to find.

The R code in this chapter uses the packages

```
library(tidyverse);
library(lmtest); library(sandwich); library(car)
library(ivreg); library(gmm)
```

8.1 Using Instruments

Suppose you wish to estimate the simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon\,. \tag{8.1}$$

You are interested in estimating the effect that a change in X will have on Y, but you strongly suspect that ϵ is correlated with X. This may be because of omitted factors (which for some reason you are unable to include, or because of measurement error, or perhaps there is simultaneity bias. This correlation between

Suppose, however, that there exists another variable Z such that $Cov(X, Z) \neq 0$ and $Cov(Z, \epsilon) = 0$. Such a variable is called an instrumental variable. It turns out that, if at least one such variable exists, we can use it to help us estimate β_1, though there are trade-offs involved. We demonstrate this from two perspectives.

8.1.1 A Method of Moments Perspective

Suppose that in population we have

$$Y = \beta_0 + \beta_1 X + \epsilon\,. \tag{8.2}$$

where ϵ is zero mean but ϵ is correlated with X, but there there is another variable Z that is (i) correlated with X but (ii) uncorrelated with ϵ. We'll make the slightly stronger assumption that $E(\epsilon \mid Z) = 0$. This gives us the following "Population Moment Conditions":

$$\begin{aligned} E(\epsilon) &= E(Y - \beta_0 - \beta_1 X) = 0 \\ E(\epsilon Z) &= E((Y - \beta_0 - \beta_1 X)Z) = 0 \end{aligned} \tag{8.3}$$

Suppose that we have a iid sample $\{Y_i, X_i, Z_i\}_{i=1}^n$ from the population. The method of moments approach is to choose as our estimators of β_0 and β_1 those values $\widehat{\beta}_0^{mm}$ and $\widehat{\beta}_1^{mm}$ such that the

sample moments corresponding to (8.3), that is,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm}X_i) = 0 \qquad [A] \\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm}X_i)Z_i = 0 \qquad [B] \end{aligned} \tag{8.4}$$

Solving gives the method of moments (MM) estimator. From [A] we get

$$\overline{Y} = \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm}\overline{X} \Rightarrow \hat{\beta}_0^{mm} = \overline{Y} - \hat{\beta}_1^{mm}\overline{X}\,.$$

Substituting into [B] gives

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - (\overline{Y} - \hat{\beta}_1^{mm}\overline{X}) - \hat{\beta}_1^{mm}X_i)Z_i = 0 \\ \sum_{i=1}^{n}((Y_i - \overline{Y}) - \hat{\beta}_1^{mm}(X_i - \overline{X}))Z_i = 0 \\ \sum_{i=1}^{n}(Y_i - \overline{Y})Z_i - \hat{\beta}_1^{mm}\sum_{i=1}^{n}(X_i - \overline{X})Z_i = 0\,. \end{aligned}$$

Solving for $\hat{\beta}_0^{mm}$ and $\hat{\beta}_1^{mm}$ gives

$$\begin{aligned} \hat{\beta}_0^{mm} &= \overline{Y} - \hat{\beta}_1^{mm}\overline{X} \\ \hat{\beta}_1^{mm} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})Z_i}{\sum_{i=1}^{n}(X_i - \overline{X})Z_i} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})} \end{aligned} \tag{8.5}$$

These estimators are consistent. Focusing on $\hat{\beta}_1^{mm}$, we have

$$\begin{aligned} \hat{\beta}_1^{mm} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})Z_i}{\sum_{i=1}^{n}(X_i - \overline{X})Z_i} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})Y_i}{\sum_{i=1}^{n}(Z_i - \overline{Z})X_i} \\ &= \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(\beta_0 + \beta_1 X_i + \epsilon_i)}{\sum_{i=1}^{n}(Z_i - \overline{Z})X_i} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})\epsilon_i}{\sum_{i=1}^{n}(Z_i - \overline{Z})X_i} \xrightarrow{p} \beta_1 + \frac{Cov(Z,\epsilon)}{Cov(Z,X)} = \beta_1 \end{aligned}$$

since $Cov(Z,\epsilon) = 0$ and (just as importantly) $Cov(Z,X) \neq 0$

Alternatively, you can show that

$$E(Y - \beta_0 - \beta_1 X) = 0 \text{ and } E((Y - \beta_0 - \beta_1 X)Z) = 0$$

implies $\beta_1 = Cov(Z,Y)/Cov(Z,X)$ (see exercises). Then

$$\hat{\beta}_1^{mm} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum_{i=1}^{n}(X_i - \overline{X})(Z_i - \overline{Z})} \xrightarrow{p} \frac{Cov(Y,Z)}{Cov(X,Z)} = \beta_1$$

8.1.2 A Two-Stage Least Squares Approach

Another way of using the instrumental variable is via "two-stage least squares" or 2SLS. We know that given X_i and Z_i, we can decompose X_i into two perfectly uncorrelated parts by regressing X_i on Z_i by OLS, giving

$$X_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i + r_{i,x|z} = \hat{X}_i + r_{i,x|z} \,. \tag{8.6}$$

Since $\hat{X}_i$ is simply a linear function of Z_i, and Z_i is uncorrelated with ϵ_i, so $\hat{X}_i$ is uncorrelated with ϵ_i. In other words, all of the movements in X_i that are correlated with ϵ_i are "concentrated" into $r_{i,x|z}$. Alternatively, you can think of $\hat{X}_i$ as that part of X_i that remains after filtering out the movements in X_i that are correlated with ϵ_i.

The idea is to use only the movements in X_i that are uncorrelated with ϵ_i when determining the effect of X_i on Y_i. In other words, we regress Y_i on $\hat{X}_i$ instead of X_i. This gives

$$\hat{\beta}_1^{2sls} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(\hat{X}_i - \overline{\hat{X}})}{\sum_{i=1}^n (\hat{X}_i - \overline{\hat{X}})^2} \,. \tag{8.7}$$

This turns out to be equivalent to the MM estimator that we derived earlier. Since $\hat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i$, we have $\overline{\hat{X}} = \hat{\delta}_0 + \hat{\delta}_1 \overline{Z}$ and

$$\hat{X}_i - \overline{\hat{X}} = \hat{\delta}_1 (Z_i - \overline{Z})$$

and

$$\sum_{i=1}^n (\hat{X}_i - \overline{\hat{X}})^2 = \hat{\delta}_1^2 \sum_{i=1} (Z_i - \overline{Z})^2 \,.$$

Substituting into (8.7) gives

$$\hat{\beta}_1^{2sls} = \frac{\hat{\delta}_1 \sum_{i=1}^T (Z_i - \overline{Z})(Y_i - \overline{Y})}{\hat{\delta}_1^2 \sum_{i=1}^T (\hat{X}_t - \overline{\hat{X}})^2} = \frac{\sum_{i=1}^T (Z_i - \overline{Z})(Y_i - \overline{Y})}{\hat{\delta}_1 \sum_{i=1}^T (\hat{X}_t - \overline{\hat{X}})^2} \,. \tag{8.8}$$

Since $\hat{\delta}_1$ is the OLS estimator for the coefficient on Z_i in a regression of X_i and Z_i, we have

$$\hat{\delta}_1 = \frac{\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X})}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \,.$$

Substituting (8.8) gives

$$\hat{\beta}_1^{2sls} = \frac{\sum_{i=1}^T (Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X})} \,. \tag{8.9}$$

which is the same as the MM estimator $\hat{\beta}_1^{mm}$ derived earlier. For the moment, we shall use $\hat{\beta}_1^{iv}$ to refer to both. In this simple one endogenous regressor one instrument case, both MM and 2SLS approaches give the same result. As we will see, this will not necessarily be the case in more elaborate situations.

Example 8.1. In Section 7.7 we gave a demand and supply example to demonstrate simultaneity bias. A quick recap: suppose

$$\begin{aligned} Q^d &= \delta_0 + \delta_1 P + \epsilon^d && \text{(Demand Eq } \delta_1 < 0) \\ Q^s &= \alpha_0 + \alpha_1 P + \epsilon^s && \text{(Supply Eq } \alpha_1 > 0) \\ Q^s &= Q^d && \text{(Market Clearing)} \end{aligned}$$

and you are interested in estimating the demand equation. You observed equilibrium quantities and prices, but these occur at the intersection of demand and supply, as shown in Fig. 7.5, reproduced here as Fig. 8.1, and do not reflect the demand equation (nor the supply equation). We showed a regression of Q on P produces an estimate that is some average of α_1 and δ_1.

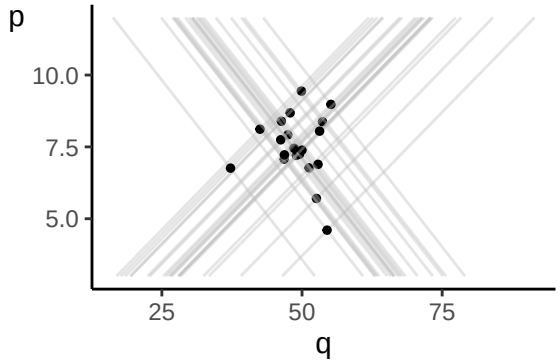


Figure 8.1: Simultaneity bias.

The issue here is that both prices and quantity are simultaneously determined, so both are endogenous. Price, in particular, is not exogenous: all variation in the data come from both demand and supply shocks. Both demand and supply shocks shift demand/supply functions. Shifts in either demand or supply function changes both quantity and price. The regressor (price) is therefore correlated with the regression noise term.

Now **suppose** there is some observable variable r_t that shifts the supply function but not the demand function. That is, the market is represented by

$$\begin{aligned} Q^d &= \delta_0 + \delta_1 P + \epsilon^d && \text{(Demand Eq } \delta_1 < 0) \\ Q^s &= \alpha_0 + \alpha_1 P + \alpha_2 r + \epsilon^s && \text{(Supply Eq } \alpha_1 > 0) \\ Q^s &= Q^d && \text{(Market Clearing)} \end{aligned}$$

where $\alpha_2 \neq 0$ and r are uncorrelated with the demand shocks. Imagine further that we can "shut down" the demand and supply shocks, allowing only r to change. Then only the supply curve changes (as r changes) and intersections of demand and supply maps out the demand function.

This illustrates the point that variation in r helps to "identify" the demand function. The problem is that in reality we *cannot* "shut down" the demand shocks, so how do we isolate variation in P due to r only? The two stage least squares perspective show that we can do so by regressing P_i on r_i, and then regress Q_i on *fitted* P_i, i.e., $\hat{P}_i$, instead of P_i, since $\hat{P}_i$ contains variation in Y_i due to r_i only. We say that we use r_i as "instrument" to identify demand function.

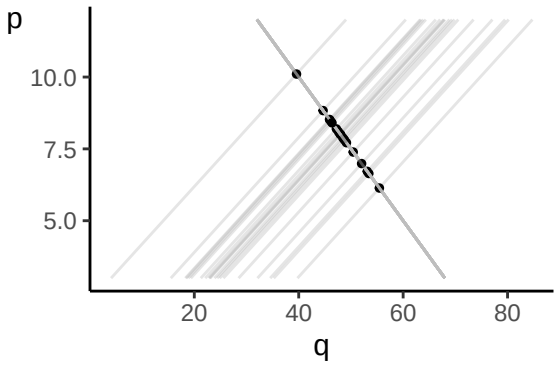


Figure 8.2: Variation in r identifies the demand function.

There are a few important points to note. First, $iv/2sls/mm$ estimators are in general *biased* estimators:

$$\hat{\beta}_1^{iv} = \beta_1 + \frac{\sum_{i=1}^n (Z_i - \overline{Z})\epsilon_i}{\sum_{i=1}^n (Z_i - \overline{Z})X_i}$$

$$E(\hat{\beta}_1^{iv} \mid X_1, \dots, X_n, Z_1, \dots, Z_n) = \beta_1 + \frac{\sum_{i=1}^n (Z_i - \overline{Z})E(\epsilon_i \mid X_1, \dots, X_n, Z_1, \dots, Z_n)}{\sum_{i=1}^n (Z_i - \overline{Z})X_i}$$

but since ϵ is correlated with X, $E(\epsilon_i \mid X_1, \dots, X_n, Z_1, \dots, Z_n)$ will be some function of X_i, $i = 1, \dots, n$, and cannot come out of the summation. Despite the bias, the $mm/tsls/iv$ estimator is consistent. The bias disappears, and the estimator converges in probability to the desired parameter value.

Second, there is a trade-off in obtaining the consistency, which is that in general our estimators will be less precise (have larger standard errors). Suppose $Var(\epsilon \mid X, Z) = \sigma^2$, then $Var(\epsilon_i \mid X_1, \dots, X_n, Z_1, \dots, Z_n) = \sigma^2$ and we have

$$\begin{aligned}
Var(\hat{\beta}_1^{iv} \mid \dots) &= \frac{\sum_{i=1}^n (Z_i - \overline{Z})^2 Var(\epsilon_i \mid \dots)}{(\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X}))^2} \\
&= \frac{\sigma^2 \sum_{i=1}^n (Z_i - \overline{Z})^2}{(\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X}))^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2 \left(\frac{(\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X}))^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2 \sum_{i=1}^n (X_i - \overline{X})^2} \right)} \\
&= \frac{\sigma^2}{R_{X|Z}^2 \sum_{i=1}^n (X_i - \overline{X})^2}
\end{aligned}$$

where $R_{X|Z}^2$ is the R^2 from the regression of X_i on Z_i. Since $0 \leq R_{X|Z}^2 < 1$, $Var(\hat{\beta}_1^{iv} \mid \dots)$ is larger than the variance of the OLS estimator. The two are the same if $R_{X|Z} = 1$, but in this case, X_i and Z_i are perfectly correlated, and Z_i must also be endogenous, so not a valid instrument. (If X is correlated with ϵ, and Z is not correlated with ϵ, then Z cannot be perfectly correlated with X.)

The problem with IVs is that often, the correlation between X and Z is too low. If $R^2_{X|Z}$ is near zero, then the variance of the IV estimator will be very large, to the point that the IV estimator may be effectively useless. Furthermore, although technically consistent, in finite samples the estimator may behave quite poorly. Recall the consistency proof:

$$\hat{\beta}_1^{mm} = \beta_1 + \frac{\sum_{i=1}^n (Z_i - \overline{Z})\epsilon_i}{\sum_{i=1}^n (Z_i - \overline{Z})X_i} \xrightarrow{p} \beta_1 + \frac{Cov(Z,\epsilon)}{Cov(Z,X)} = \beta_1$$

In finite samples, $Cov(Z,\epsilon)$ is unlikely to be exactly zero. If $Cov(Z,X)$ is close to zero in population, the sample covariance between Z_i and X_i is likely to be small, resulting in a large finite sample bias. We call an instrumental variable that is poorly correlated with the endogenous regressor a "weak instrument".

To estimate σ^2, use:

$$\widehat{\sigma^2_{iv}} = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}^2_{i,iv}$$

where $\hat{\epsilon}_{i,iv} = Y_i - \hat{\beta}_0^{iv} + \hat{\beta}_1^{iv} X_i$. Note that $\hat{\epsilon}_{i,iv}$ uses X_i rather that $\widehat{X}_i$. Furthermore, it is conventional to divide by $n-2$ even though IV/MM/2SLS only has large sample justifications. For R^2, use:

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}^2_{i,iv}}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

which will be less than the OLS R^2.

Example 8.2. We illustrate the theory with a simulated version of the demand and supply example. Suppose

$$\begin{aligned} Q_i^d &= 8 - 2P_i + \epsilon_i^d && \text{(Demand Eq } \delta_1 < 0) \\ Q_i^s &= -2 + 2P_i + 0.8r_i + \epsilon_i^s && \text{(Supply Eq } \alpha_1 > 0) \\ Q_i^s &= Q_i^d && \text{(Market Clearing)} \end{aligned}$$

where $r_i \sim N(3,4)$ is observed, but the $e_i^d \sim N(0,1)$ and $e_i^s \sim N(0,1)$ are not. The code below simulates and plots 100 observations of the observed equilibrium prices and quantities.

```
set.seed(13)
a0 <- 8; a1 <- -2; b0 <- -2; b1 <- 2; b2 = 0.8
n <- 100; p <- seq(0.5, 4, 0.1)
pstar <- qstar <- r <- rep(0, n)
p1 <- ggplot()
for (i in 1:n){
  r[i]  <- rnorm(1,3,2)
  es <- rnorm(1,0,1); ed <- rnorm(1,0,1)
  plotdat <- tibble(
              p = p,
              qd = a0 + a1*p + ed,
              qs = b0 + b1*p + b2*r[i] + es)
  pstar[i] <- (b0 - a0)/(a1-b1) + b2*r[i]/(a1-b1) + (es-ed)/(a1-b1)
  qstar[i] <- a0 + a1*pstar[i] + ed
  p1 <- p1 + geom_line(data=plotdat, aes(x=p, y=qs), color='gray', alpha=0.3) +
```

```
    geom_line(data=plotdat, aes(x=p, y=qd), color='gray', alpha=0.3)
}
dat <- tibble(p=pstar, q=qstar)
p1 <- p1 + geom_point(data=dat, aes(x=pstar, y=qstar), size = 1, color="blue") +
  ylab("q") + theme_bw()
p1
```

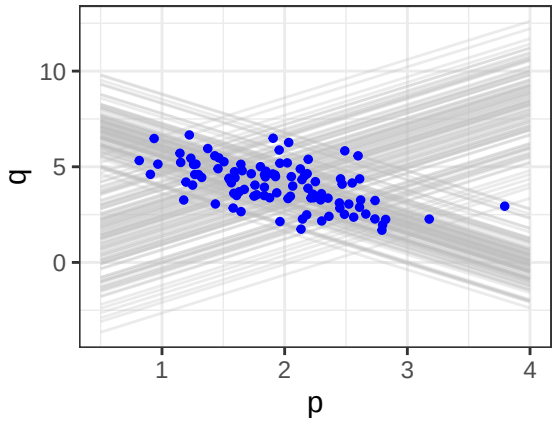


Figure 8.3: Simulated demand and supply example.

The OLS Estimates

```
dat <- tibble(p=pstar, q=qstar, r=r)
ddss_ols <- lm(q~p, data=dat)
summary(ddss_ols)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  6.341252  0.3660495 17.323483 1.366065e-31
p           -1.175239  0.1816905 -6.468357 3.906490e-09
```

The IV Estimates (using `ivreg` package, default standard errors) are

```
ddss_iv <- ivreg(q ~ p | r, data=dat)
ddss_coef <- summary(ddss_iv)$coef
attr(ddss_coef, "df") <- NULL
attr(ddss_coef, "nobs") <- NULL
ddss_coef
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  8.492688  0.5937838 14.302662 9.954936e-26
p           -2.283065  0.3000241 -7.609603 1.710061e-11
```

The IV Estimates (using `ivreg` package, heteroskedasticity-robust standard errors) are

```
coeftest(ddss_iv, vcov=vcovHC(ddss_iv, type="HC0"))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  8.49269    0.57835 14.6845 < 2.2e-16 ***
p           -2.28306    0.28389 -8.0421 2.063e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, the IV estimate is much closer to true value of the demand elasticity parameter than OLS estimates. Default and heteroskedastic estimates for IV estimates are similar (not surprising, since data is homoskedastic), and the standard error of the IV estimator is larger than standard error of the OLS estimator.

Fig. 8.4 contains plot of OLS (purple) and IV (red) estimated regression lines.

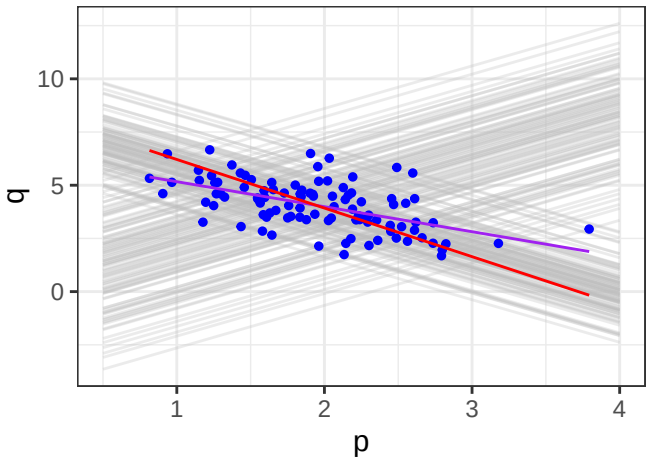


Figure 8.4: Simulated demand and supply example, IV vs OLS estimates

As an illustration, we manually compute below the IV estimate, with non-heteroskedasticity-robust and heteroskedasticity-robust standard errors:

```
y <- as.matrix(dat$q)                                  #
X <- as.matrix(cbind(rep(1, length(dat$p)), dat$p))  # Assemble data
Z <- as.matrix(cbind(rep(1, length(dat$r)), dat$r))  #
b_iv <- solve(t(Z) %*% X) %*% (t(Z) %*% y)    # calculate IV estimate
# var-cov assuming homoskedasticity
y_iv_hat <- X %*% b_iv
e_iv <- y - y_iv_hat
s2 <- sum(e_iv^2) / (n-2)
var_b <- s2 * solve(t(Z) %*% X) %*% (t(Z) %*% Z) %*% solve(t(X) %*% Z)
# Robust variance-covariance
S <- matrix(c(0,0,0,0), nrow=2)
for (i in 1:n){
  zi <- matrix(Z[i,], nrow=2)
  S <- S + e_iv[i]^2 * zi %*% t(zi)
}
var_b_rb <- solve(t(Z) %*% X) %*% S %*% solve(t(X) %*% Z)

iv_coefs <- cbind(b_iv, sqrt(diag(var_b)), sqrt(diag(var_b_rb)))
colnames(iv_coefs) <- c("Estimate", "default s.e.", "robust s.e.")
rownames(iv_coefs) <- c("(Intercept)", "X")
iv_coefs
```

```
             Estimate default s.e. robust s.e.
(Intercept)  8.492688    0.5937838   0.5783452
X           -2.283065    0.3000241   0.2838908
```

We have successfully replicated estimates from `ivreg` package.

8.2 Using Matrix Algebra

As an intermediate step towards more elaborate scenarios, which will require matrix algebra, we express the simple case from the previous sections (one endogenous regressor, one instrument, no exogenous regressors) using matrix algebra. Suppose the regression equation of interest is:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

where X_1 is endogenous (correlated with the noise term). Suppose Z_1 is correlated with X_1 and uncorrelated with ϵ. Note that the variables previously labelled as X and Z are now labelled as X_1 and Z_1. The population moment conditions are:

$$E(\epsilon) = E(Y - \beta_0 - \beta_1 X_1) = 0 \quad \text{and} \quad E(\epsilon Z_1) = E((Y - \beta_0 - \beta_1 X_1)Z_1) = 0\,.$$

The sample moment conditions are

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_{i1}) = 0$$
$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_{i1})Z_{i1} = 0$$

In matrix algebra, we can write this as

$$Z^{\mathrm{T}}(y - X\hat{\beta}^{mm}) = Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta}^{mm} = 0$$

where

$$y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix},\ \varepsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},\ Z = \begin{bmatrix} 1 & Z_{11} \\ \vdots & \vdots \\ 1 & Z_{n1} \end{bmatrix}$$

A few remarks: first, note that $Z^{\mathrm{T}}X$ is 2×2. Second, Z_{i1} correlated with X_{i1} means that $Z^{\mathrm{T}}X$ is invertible. Solving the moment conditions then gives

$$\hat{\beta}^{mm} = (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y \tag{8.10}$$

Now consider the two-stage least squares approach. In step 1, regress X on Z. As X is $n \times 2$, this means regressing each column of X on Z:

$$i_n = Zb_0 + u_{*0} \ \text{ and } \ X_{*1} = Zb_1 + u_{*1}$$

We can put into this one single matrix:

$$\begin{bmatrix} i_n & X_{*1} \end{bmatrix} = Z \begin{bmatrix} b_0 & b_1 \end{bmatrix} + \begin{bmatrix} u_{*0} & u_{*1} \end{bmatrix}$$

or $X = ZB + U$. We have $\hat{B} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X$. The fitted value from this step is

$$\hat{X} = Z\hat{B} = Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X\,.$$

In step 2, regress y on $\hat{X}$, which gives

$$\begin{aligned}\hat{\beta}^{2sls} &= (\hat{X}^{\mathrm{T}}\hat{X})^{-1}\hat{X}^{\mathrm{T}}y \\ &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y \\ &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y \\ &= (Z^{\mathrm{T}}X)^{-1}(Z^{\mathrm{T}}Z)(X^{\mathrm{T}}Z)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y \\ &= (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y\,.\end{aligned} \tag{8.11}$$

As expected, we get the same formula, which we will refer to as the IV estimator

$$\beta^{iv} = (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y\,. \tag{8.12}$$

Showing consistency is straightforward:

$$\begin{aligned}\hat{\beta}^{iv} &= (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y = (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}(X\beta+\epsilon) \\ &= \beta + (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}\epsilon \\ &= \beta + (\tfrac{1}{n}Z^{\mathrm{T}}X)^{-1}(\tfrac{1}{n}Z^{\mathrm{T}}\epsilon) \xrightarrow{p} \beta\end{aligned}$$

since $\frac{1}{n}Z^{\mathrm{T}}\epsilon \xrightarrow{p} 0_{2\times 1}$ and we assume $\frac{1}{n}Z^{\mathrm{T}}X$ converges to a non-singular matrix. We omit the arguments for asymptotic normality.

For asymptotically valid variance-covariance matrices, we have for homoskedastic errors:

$$\widehat{Var}(\hat{\beta}_{iv}) = \widehat{\sigma^2}(Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}Z(X^{\mathrm{T}}Z)^{-1} \tag{8.13}$$

where $\widehat{\sigma^2}$ is as previously defined. The heteroskedasticity-robust version is:

$$\widehat{Var}_{HC0}(\hat{\beta}_{iv}) = (Z^{\mathrm{T}}X)^{-1}\left(\sum_{i=1}^{n}\hat{\epsilon}_{i,iv}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)(X^{\mathrm{T}}Z)^{-1} \tag{8.14}$$

where Z_{i*} are the i-rows of Z

8.3 Generalization

Suppose now that in your regression you have multiple regressors, some of which are correlated with the error term (i.e., some are endogenous) and some are not correlated with the error term. That is, you have both endogenous and exogenous regressors. We allow for the possibility that there are more instrument variables than there are endogenous variables. To start, we assume 1 exogenous regressor, 1 endogenous regressor, and 2 valid instruments for endogenous regressor:

$$Y = \beta_0 + \beta_1 X_1^k + \beta_2 X_2^g + \epsilon$$

where X_1^k is exogenous (not correlated with the noise term), X_2^g is endogenous (correlated with the noise term), and there exists Z_2 and Z_3 both correlated with X_2^g and uncorrelated with ϵ.

We must have at least as many instruments as we do endogenous variables. When we have more instruments than endogenous variables, we say we have "overidentification". If we have the same number of instruments as we do endogenous variables, we say we have "just-identification".

8.3.1 Method of Moments Approach:

In our specific example (1 exog, 1 endog, 2 instruments), the population satisfies $E(\epsilon) = 0$, $Cov(X_1^k, \epsilon) = 0$, $Cov(Z_2, \epsilon) = 0$, $Cov(Z_3, \epsilon) = 0$. That is,

$$E(\epsilon) = E(Y - \beta_0 - \beta_1 X_1^k - \beta_2 X_2^g) = 0$$
$$E(\epsilon X_1^k) = E((Y - \beta_0 - \beta_1 X_1^k - \beta_2 X_2^g) X_1^k) = 0$$
$$E(\epsilon Z_2) = E((Y - \beta_0 - \beta_1 X_1^k - \beta_2 X_2^g) Z_2) = 0$$
$$E(\epsilon Z_3) = E((Y - \beta_0 - \beta_1 X_1^k - \beta_2 X_2^g) Z_3) = 0$$

Suppose we have a representative iid sample $\{X_{i1}^k, X_{i2}^g, Z_{i2}, Z_{i3}, Y_i\}_{1=i}^n$ from the population. The sample moments corresponding to the population moments are

$$\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^{mm} = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_{i1}^k - \hat{\beta}_2^{mm} X_{i2}^g) = 0$$
$$\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^{mm} X_{i1}^k = \frac{1}{n}\sum_{i=1}^n ((Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_{i1}^k - \hat{\beta}_2^{mm} X_{i2}^g) X_{i1}^k) = 0$$
$$\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^{mm} Z_{i2} = \frac{1}{n}\sum_{i=1}^n ((Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_{i1}^k - \hat{\beta}_2^{mm} X_{i2}^g) Z_{i2}) = 0$$
$$\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^{mm} Z_{i3} = \frac{1}{n}\sum_{i=1}^n ((Y_i - \hat{\beta}_0^{mm} - \hat{\beta}_1^{mm} X_{i1}^k - \hat{\beta}_2^{mm} X_{i2}^g) Z_{i3}) = 0$$

We want to "solve" these sample moments for $\hat{\beta}_0^{mm}$, $\hat{\beta}_1^{mm}$, $\hat{\beta}_2^{mm}$, but we have 4 equations in 3 unknowns, so the equations cannot be solved exactly. Instead, we choose $\hat{\beta}_0^{mm}$, $\hat{\beta}_1^{mm}$, $\hat{\beta}_2^{mm}$ to make "LHS as close to zero" as possible. In particular, choose $\hat{\beta}_0^{mm}$, $\hat{\beta}_1^{mm}$, $\hat{\beta}_2^{mm}$ to minimize the sum of squared moments:

$$\left(\sum_{i=1}^n \hat{\epsilon}_i^{mm}\right)^2 + \left(\sum_{i=1}^n \hat{\epsilon}_i^{mm} X_{i1}^k\right)^2 + \left(\sum_{i=1}^n \hat{\epsilon}_i^{mm} Z_{i2}\right)^2 + \left(\sum_{i=1}^n \hat{\epsilon}_i^{mm} Z_{i3}\right)^2 .$$

In matrix algebra terms, if we let

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & X_{11}^k & X_{12}^g \\ 1 & X_{21}^k & X_{22}^g \\ \vdots & \vdots & \vdots \\ 1 & X_{n1}^k & X_{n2}^g \end{bmatrix}, \; \hat{\beta}^{mm} = \begin{bmatrix} \hat{\beta}_0^{mm} \\ \hat{\beta}_1^{mm} \\ \hat{\beta}_2^{mm} \end{bmatrix}, \; Z = \begin{bmatrix} 1 & X_{11}^k & Z_{12} & Z_{13} \\ 1 & X_{21}^k & Z_{22} & Z_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1}^k & Z_{n2} & Z_{n3} \end{bmatrix}$$

then sample moments (dropping the $1/n$) are:

$$\underbrace{Z^{\mathrm{T}}}_{4\times n} \underbrace{(y - X\hat{\beta}^{mm})}_{n\times 1} = \underbrace{Z^{\mathrm{T}}}_{4\times n} \underbrace{y}_{n\times 1} - \underbrace{Z^{\mathrm{T}}}_{4\times n} \underbrace{X}_{n\times 3} \underbrace{\hat{\beta}^{mm}}_{3\times 1} = \underbrace{0}_{4\times 1} .$$

We choose $\hat{\beta}_0^{mm}$, $\hat{\beta}_1^{mm}$, $\hat{\beta}_2^{mm}$ to minimize the "sum of squared moments"

$$\begin{aligned} &\underbrace{(Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta})^{\mathrm{T}}}_{1\times 4}\underbrace{(Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta})}_{4\times 1} \\ =\ & y^{\mathrm{T}}ZZ^{\mathrm{T}}y - 2\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}ZZ^{\mathrm{T}}y + \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}ZZ^{\mathrm{T}}X\hat{\beta} \end{aligned} \tag{8.15}$$

Minimizing this gives

$$\hat{\beta}^{mm} = (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}y \tag{8.16}$$

which requires that the 3×3 matrix $X^{\mathrm{T}}ZZ^{\mathrm{T}}X$ be invertible.

To show consistency of the MM estimator:

$$\begin{aligned} \hat{\beta}^{mm} &= (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}y \\ &= (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}(X\beta + \epsilon) \\ &= (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}X\beta + (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}\epsilon \\ &= \beta + ((\tfrac{1}{n}X^{\mathrm{T}}Z)((\tfrac{1}{n}Z^{\mathrm{T}}X))^{-1}((\tfrac{1}{n}X^{\mathrm{T}}Z)(\tfrac{1}{n}Z^{\mathrm{T}}\epsilon) \quad \xrightarrow{p} \beta \end{aligned}$$

which requires that $\frac{1}{n}Z^{\mathrm{T}}\epsilon \xrightarrow{p} 0_{4\times 1}$ and $\frac{1}{n}Z^{\mathrm{T}}X \xrightarrow{p} \Sigma_{ZX}$ full column rank. The variance-covariance matrix under homoskedasticity is:

$$\widehat{Var}(\hat{\beta}^{mm}) = \sigma^2(X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)Z^{\mathrm{T}}X(X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}$$

Estimate σ^2 with $\widetilde{\sigma^2} = \dfrac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_{i,iv}^2$. The heteroskedasticity-robust version is

$$\widehat{Var}(\hat{\beta}^{mm}) = (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z\left(\sum_{i=1}^{n}\hat{\epsilon}_{i,iv}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)Z^{\mathrm{T}}X(X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}\,.$$

What happens if we apply these formulas to the **just-identified** case, where the number of endogenous is equal to the number of instruments? For instance, suppose in our example that we only have Z_2 to instrument for X_2^g. Then $Z^{\mathrm{T}}X$ is square 3×3, and we get

$$\begin{aligned} \hat{\beta}_{mm} &= (X^{\mathrm{T}}ZZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}y = (Z^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}Z)^{-1}X^{\mathrm{T}}ZZ^{\mathrm{T}}y \\ &= (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y\,. \end{aligned}$$

The corresponding variance-covariance matrices, under homoskedasticity, reduces to:

$$\widehat{Var}(\hat{\beta}_{mm}) = \widehat{\sigma^2}(Z^{\mathrm{T}}X)^{-1}(Z^{\mathrm{T}}Z)(X^{\mathrm{T}}Z)^{-1}\,.$$

The heteroskedasticity-robust version reduces to

$$\widehat{Var}(\hat{\beta}_{mm}) = (Z^{\mathrm{T}}X)^{-1}\left(\sum_{i=1}^{n}\hat{\epsilon}_{i,iv}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)(X^{\mathrm{T}}Z)^{-1}$$

The MM estimator and its var-cov. matrices all collapse to the IV case derived in Section 8.2.

8.3.2 Two-Stage Least Squares Approach

We now consider the 2SLS approach for this example. In step 1, regress X on Z. The coefficient estimates are $\hat{B} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X$ (this is a 4×3 matrix). The fitted values are $\hat{X} = Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X$.

The matrix of fitted values is $n \times 3$ in size. The first column is a vector of 1's, the second column is X^k_{i1}, and the third column is $\hat{X}^g_{i2}$ obtained from a regression of X^g_{i2} on intercept, X^k_{i1}, Z_{i2} and Z_{i3}. In step 2, we regress y on $\hat{X}$, which gives

$$\begin{aligned}\hat{\beta}^{2sls} &= (\hat{X}^{\mathrm{T}}\hat{X})^{-1}\hat{X}^{\mathrm{T}}y \\ &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y \\ &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y\,.\end{aligned} \tag{8.17}$$

This requires that the 4×3 matrix $Z^{\mathrm{T}}X$ has full column rank and that Z has full column rank. Notice that the 2SLS approach gives us a different formula from what we obtained in the MM approach.

The proof of consistency is as follows:

$$\begin{aligned}\hat{\beta}^{2sls} &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y \\ &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}(X\beta+\epsilon) \\ &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X\beta \\ &\qquad + (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}\epsilon \\ &= \beta + (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}\epsilon \\ &= \beta + (\tfrac{1}{n}X^{\mathrm{T}}Z(\tfrac{1}{n}Z^{\mathrm{T}}Z)^{-1}\tfrac{1}{n}Z^{\mathrm{T}}X)^{-1}\tfrac{1}{n}X^{\mathrm{T}}Z(\tfrac{1}{n}Z^{\mathrm{T}}Z)^{-1}\tfrac{1}{n}Z^{\mathrm{T}}\epsilon \xrightarrow{p} \beta\end{aligned}$$

The variance-covariance matrices are (details of proofs omitted), under homoskedasticity:

$$\widehat{Var}(\hat{\beta}_{2sls}) = \widehat{\sigma^2}(X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1} \tag{8.18}$$

The heteroskedasticity-robust is

$$\begin{aligned}&\widehat{Var}(\hat{\beta}_{2sls}) = \\ &(X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}\left[\sum_{i=1}^{n}\hat{\epsilon}^2_{i,iv}Z^{\mathrm{T}}_{i*}Z_{i*}\right](Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X(X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}\end{aligned} \tag{8.19}$$

In the **just-identified** case, where the number of endogenous variables is equal to the number of instruments, $Z^{\mathrm{T}}X$ is square, the 2SLS estimator reduces to:

$$\begin{aligned}\hat{\beta}_{2sls} &= (X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y \\ &= (Z^{\mathrm{T}}X)^{-1}(Z^{\mathrm{T}}Z)(X^{\mathrm{T}}Z)^{-1}X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y = (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y\end{aligned}$$

which is the same as IV estimator from Section 8.2. The variance formulas also converge to IV

variance formulas.

Example 8.3. We'll use data from `earnings2019.csv` to estimate the equation

$$\ln earn = \beta_0 + \beta_1 age + \beta_2 \ln tenure + \beta_3 educ + \epsilon$$

The concern is that a measure of *ability* is unavailable, and omitted, resulting in endogeneity of the *educ* variable. We assume *age* and $\ln tenure$ exogenous. We assume that *feduc* and *meduc* are valid instruments. Ignoring potential endogeneity, the OLS estimates are

```
dat <- read_csv("data\\earnings2019.csv", show_col_types=FALSE) %>%
  mutate(ln_earn = log(earn), ln_tenure=log(tenure), const = 1)

cat("Assuming homoskedasticity (default)\n")
mdl2_ols <- lm(ln_earn ~ age + ln_tenure + educ, data=dat) # Estimate OLS
summary(mdl2_ols)$coefficients %>% round(4) # Print default coefficients and standard errors

cat("\nUsing heteroskedasticity-robust standard errors")
coeftest(mdl2_ols, vcov=vcovHC, type="HC") %>% round(4)# Robust standard errors
```

```
Assuming homoskedasticity (default)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9699     0.0628 15.4349    0e+00
age           0.0025     0.0008  3.3341    9e-04
ln_tenure     0.1477     0.0090 16.4018    0e+00
educ          0.1268     0.0038 33.1159    0e+00

Using heteroskedasticity-robust standard errors
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9699     0.0655 14.8101   <2e-16 ***
age           0.0025     0.0008  3.1315   0.0017 **
ln_tenure     0.1477     0.0092 16.0494   <2e-16 ***
educ          0.1268     0.0039 32.1078   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the formulas derived in this section, the MM estimates are as follows:

```
## Assemble data for MM / 2SLS
y <- dat %>% select(c(ln_earn)) %>% as.matrix()
X <- dat %>% select(c(const, age, ln_tenure, educ)) %>% as.matrix()
Z <- dat %>% select(c(const, age, ln_tenure, feduc, meduc)) %>% as.matrix()
n <- length(y)
Zcol <- dim(Z)[2]
ZTX <- t(Z) %*% X ; XTZ <- t(X) %*% Z ; ZTZ <- t(Z) %*% Z ; ZTy <- t(Z) %*% y

#--MM--
beta_MM <- solve(XTZ %*% ZTX) %*% XTZ %*% ZTy
ehat_IV <- y - X %*% beta_MM
s2hat <- sum(ehat_IV^2)/n
eZZ <- matrix(0, nrow=Zcol, ncol=Zcol)
for (i in 1:n){
```

```
  eZZ <- eZZ + ehat_IV[i]^2 * t(Z[i,,drop=F]) %*% Z[i,,drop=F]
}
vbeta_MM <-  s2hat * solve(XTZ%*%ZTX) %*% XTZ %*% ZTZ %*% ZTX %*% solve(XTZ%*%ZTX)
vbeta_MM_rob <- solve(XTZ%*%ZTX) %*% XTZ %*% eZZ %*% ZTX %*% solve(XTZ%*%ZTX)

MM_results <- cbind(estimates = beta_MM,
                    s.e. = sqrt(diag(vbeta_MM)),
                    s.e.robust = sqrt(diag(vbeta_MM_rob)))
MM_results %>% round(4)
```

```
          ln_earn   s.e. s.e.robust
const     -1.7976 0.5844     0.5911
age        0.0096 0.0026     0.0027
ln_tenure  0.1386 0.0108     0.0111
educ       0.2991 0.0336     0.0341
```

Using the formulas derived in this section, the 2SLS estimator is

```
#--2SLS--
beta_TSLS <- solve(XTZ %*% solve(ZTZ) %*% ZTX) %*% XTZ %*% solve(ZTZ) %*% ZTy
ehat_TSLS <- y - X %*% beta_TSLS
s2hat_TSLS <- sum(ehat_TSLS^2)/n
eZZ_TSLS <- matrix(0, nrow=Zcol, ncol=Zcol)
for (i in 1:n){
  eZZ_TSLS <- eZZ_TSLS + ehat_TSLS[i]^2 * t(Z[i,,drop=F]) %*% Z[i,,drop=F]
}
vbeta_TSLS <-  s2hat_TSLS * solve(XTZ %*% solve(ZTZ) %*% ZTX)
vbeta_TSLS_rob <- solve(XTZ %*% solve(ZTZ) %*% ZTX) %*% XTZ %*% solve(ZTZ) %*%
  eZZ_TSLS %*% solve(ZTZ) %*% ZTX %*% solve(XTZ %*% solve(ZTZ) %*% ZTX)

TSLS_results <- cbind(estimates = beta_TSLS,
                    s.e. = sqrt(diag(vbeta_TSLS)),
                    s.e.robust = sqrt(diag(vbeta_TSLS_rob)))
TSLS_results %>% round(4)
```

```
          ln_earn   s.e. s.e.robust
const     -0.3843 0.1915     0.2088
age        0.0032 0.0008     0.0009
ln_tenure  0.1399 0.0096     0.0098
educ       0.2205 0.0131     0.0143
```

We can get the 2SLS estimates from the `ivreg` package:

```
mdl2_iv <- ivreg(ln_earn ~ age + ln_tenure + educ |
                   age + ln_tenure + feduc + meduc, data=dat)
mdl2_iv_coef <- summary(mdl2_iv)$coef
attr(mdl2_iv_coef,"df")<-NULL
attr(mdl2_iv_coef,"nobs")<-NULL
mdl2_iv_coef %>% round(4)
coeftest(mdl2_iv,vcov=vcovHC(mdl2_iv,type="HC0"))  %>% round(4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3843     0.1916 -2.0058   0.0449
age           0.0032     0.0008  3.9523   0.0001
ln_tenure     0.1399     0.0096 14.5964   0.0000
educ          0.2205     0.0131 16.8675   0.0000
```

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3843     0.2088 -1.8406   0.0657 .
age           0.0032     0.0009  3.7345   0.0002 ***
ln_tenure     0.1399     0.0098 14.2368   <2e-16 ***
educ          0.2205     0.0143 15.4326   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

which validates our results computed directly from the formulas.

All formulas and results continue to apply to general case with K exogenous regressors, G endogenous regressors and M instruments, $M \geq G$

$$Y = \beta_0 + \beta_1 X_1^k + \cdots + \beta_K X_K^k + \beta_{K+1} X_{K+1}^g + \cdots + \beta_{K+G} X_{K+G}^g + \epsilon$$

with instrumental variables $Z_1, \ldots Z_K, Z_{K+1}, \ldots Z_{K+M}$, where $Z_1 = X_1^k, \ldots, Z_K = X_K^k$, satisfying $cov(Z_j, \epsilon) = 0$ for all $j = 1, \ldots, K+M$. The only change is that Z is now $n \times (K+M+1)$ and X is $n \times (K+G+1)$.

8.4 (Optimal) Generalized Method of Moments

We now consider the **generalized method of moments** approach, which encompasses both the MM and 2SLS estimators. The GMM estimator minimizes *weighted* sum of squared moments, i.e.,

$$\hat{\beta}_W^{gmm} = \text{argmin}_{\hat{\beta}} \underbrace{(Z^{\mathrm{T}} y - Z^{\mathrm{T}} X\hat{\beta})^{\mathrm{T}} W (Z^{\mathrm{T}} y - Z^{\mathrm{T}} X\hat{\beta})}_{\text{"}J(W)\text{"}} \tag{8.20}$$

where W is some symmetric positive-definite weight matrix (may change with n and may be data dependent). We will assume this matrix is known for the moment. The matrix X is the $n \times K+G+1$ matrix of regressors (exogenous and endogenous) and Z is the $n \times K+M+1$ matrix of exogenous variables (exogenous regressors and instruments).

Minimizing $J(W)$ gives

$$\hat{\beta}_W^{gmm} = (X^{\mathrm{T}} Z W Z^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Z W Z^{\mathrm{T}} y \qquad \text{(Exercise!)} \tag{8.21}$$

Obviously MM is GMM with $W = I_n$ and 2SLS is GMM with $W = (Z^{\mathrm{T}} Z)^{-1}$. The proof of consistency is straightforward:

$$\begin{aligned}
\hat{\beta}_W^{gmm} &= (X^{\mathrm{T}} Z W Z^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Z W Z^{\mathrm{T}} y \\
&= \beta + (X^{\mathrm{T}} Z W Z^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Z W Z^{\mathrm{T}} \epsilon \\
&= \beta + [(\tfrac{1}{n} X^{\mathrm{T}} Z) W (\tfrac{1}{n} Z^{\mathrm{T}} X)]^{-1} (\tfrac{1}{n} X^{\mathrm{T}} Z) W (\tfrac{1}{n} Z^{\mathrm{T}} \epsilon) \quad \xrightarrow{p} \quad \beta
\end{aligned}$$

The variance-covariance matrix under homoskedasticity is:

$$\widehat{Var}(\hat{\beta}_W^{gmm}) = \widehat{\sigma^2}(X^{\mathrm{T}}ZWZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZW(Z^{\mathrm{T}}Z)^{-1}WZ^{\mathrm{T}}X(X^{\mathrm{T}}ZWZ^{\mathrm{T}}X)^{-1} \tag{8.22}$$

whereas the heteroskedasticity-robust version is

$$\begin{aligned}&\widehat{Var}(\hat{\beta}_W^{gmm})\\ &= (X^{\mathrm{T}}ZWZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZW\left(\sum_{i=1}^{n}\hat{\epsilon}_{i,gmm}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)WZ^{\mathrm{T}}X(X^{\mathrm{T}}ZWZ^{\mathrm{T}}X)^{-1}\end{aligned} \tag{8.23}$$

The question is what the weight matrix should be? It turns out (proof omitted) that an optimal choice of weights is

$$W^* = \left(\sum_{i=1}^{n}\hat{\epsilon}_{i,gmm}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)^{-1}$$

This is usually implemented with a two-step approach: First, compute $\hat{\beta}_W^{gmm}$ for some (non-optimal) weighting matrix W. The common choice is to use $W = (Z^{\mathrm{T}}Z)^{-1}$, which gives the (inefficient but consistent) 2SLS estimator $\hat{\beta}^{2sls}$, calculate $\hat{\epsilon}_{i,2sls}$. Then calculate $W^* = \left(\sum_{i=1}^{n}\hat{\epsilon}_{i,2sls}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)^{-1}$. Finally, calculate the optimal GMM estimator as

$$\hat{\beta}^{gmm} = (X^{\mathrm{T}}ZW^*Z^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZW^*Z^{\mathrm{T}}y. \tag{8.24}$$

The variance-covariance matrix under homoskedasticity is

$$\widehat{Var}(\hat{\beta}^{gmm}) = \widehat{\sigma^2}\left(X^{\mathrm{T}}Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X\right)^{-1} \tag{8.25}$$

where $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_{i,gmm}^2$. The heteroskedasticity-robust version is

$$\widehat{Var}(\hat{\beta}^{gmm}) = \left(X^{\mathrm{T}}Z\left(\sum_{i=1}^{N}\hat{\epsilon}_{i,gmm}^2 Z_{i*}^{\mathrm{T}}Z_{i*}\right)^{-1}Z^{\mathrm{T}}X\right)^{-1} \tag{8.26}$$

The form of the variance of the optimal GMM estimator *under homoskedasticity* is the same as that of 2SLS, so 2SLS is as good as optimal GMM under homoskedasticity (2SLS and two-step implementation of optimal GMM are not numerically identical – they give different estimates – but both are asymptotically efficient).

Example 8.4. We calculate GMM estimates and standard errors for the `earnings2019` example. We will need the following items, all of which were calculated in an earlier example: 2SLS residuals as `ehat_2SLS`, $\sum_{i=1}^{n}\hat{\epsilon}_{i,2sls}^2 Z_{i*}^{\mathrm{T}}Z_{i*}$ as `eZZ_TSLS`, and $X^{\mathrm{T}}Z$, $Z^{\mathrm{T}}X$, $Z^{\mathrm{T}}Z$ and $Z^{\mathrm{T}}y$ as `XTZ`, `ZTX`, `ZTZ` and `ZTy`.

```
W <- solve(eZZ_TSLS)
beta_GMM <- solve(XTZ %*% W %*% ZTX) %*% XTZ %*% W %*% ZTy
ehat_GMM <- y - X %*% beta_GMM
s2_GMM <- mean(ehat_GMM^2)
```

```
vbeta_GMM <- s2_GMM * solve(XTZ %*% solve(ZTZ) %*% ZTX)
eZZ_GMM <- matrix(0, nrow=Zcol, ncol=Zcol)
for (i in 1:n){
  eZZ_GMM <- eZZ_GMM + ehat_GMM[i]^2 * t(Z[i,,drop=F]) %*% Z[i,,drop=F]
}
vbeta_GMM_rob <- solve(XTZ %*% solve(eZZ_GMM) %*% ZTX)
# Compile results
GMM_results <- cbind(estimates = as.vector(beta_GMM),
                     s.e. = sqrt(diag(vbeta_GMM)),
                     s.e.robust = sqrt(diag(vbeta_GMM_rob)))
# Present results
GMM_results %>% round(4)
```

```
           estimates   s.e. s.e.robust
const        -0.3690 0.1913     0.2086
age           0.0032 0.0008     0.0009
ln_tenure     0.1404 0.0096     0.0098
educ          0.2194 0.0130     0.0143
```

The entries in the "estimates" column give the optimal GMM estimates. The standard errors given in the column "s.e." is appropriate for opt. GMM under the assumption of homoskedasticity. Under homoskedasticity, you can either use 2SLS or Optimal GMM. They are numerically different, but both are consistent and asymptotically efficient. Under heteroskedasticity, optimal GMM is asymptotically more efficient, but you should use the standard errors under the "s.e.robust" column. In this example, there is not much difference between the 2SLS and GMM standard errors.

In the code below, we use the GMM package to obtain optimal GMM with heteroskedasticity-robust standard errors.

```
GMM_results_pkg <- gmm(
  ln_earn ~ age + ln_tenure + educ, ~ age + ln_tenure + feduc + meduc,
  data = dat, wmatrix = "optimal", vcov = "MDS", type = "twoStep")
summary(GMM_results_pkg)$coef[,1:2] %>% round(4)
```

```
             Estimate Std. Error
(Intercept)  -0.3690     0.2086
age           0.0032     0.0009
ln_tenure     0.1404     0.0098
educ          0.2194     0.0143
```

8.5 GMM Inference

8.5.1 Testing Linear Restrictions

We can do the usual t and F tests after GMM estimation. If your regression has $K-1$ regressors plus an intercept, then the "Wald" statistic for jointly testing J number of linear hypotheses, $H_0 : \mathcal{R}\beta = r_0$, where $\mathcal{R}$ is $J \times K$ and r_0 is $K \times 1$, is

$$W = (R\hat{\beta}^{gmm} - r)^{\mathrm{T}}(R\,\widehat{Var}(\hat{\beta}^{gmm})R^{\mathrm{T}})^{-1}(R\hat{\beta}^{gmm} - r) \overset{a}{\sim} \chi^2_{(J)}$$

This is the usual asymptotic version of F test, using the GMM estimators and variance-covariance matrix in place of the OLS ones. Continuing with our example

$$\ln earn = \beta_0 + \beta_1 age + \beta_2 \ln tenure + \beta_3 educ + \epsilon\,,$$

we test $H_0 : \beta_1 = 0$ and $\beta_2 = \beta_3$ or $(\beta_2 - \beta_3 = 0)$

```
R = matrix(c(0,1,0,0,0,0,1,-1), nrow=2, byrow=TRUE)
r = matrix(c(0,0), ncol=1)
b = beta_GMM
V = vbeta_GMM_rob
F_stat = t(R %*% b-r) %*% solve(R %*% V %*% t(R)) %*% (R %*% b-r)
cat("F:",F_stat,", p-value:", 1-pchisq(F_stat,nrow(R)))
```

```
F: 23.78748 , p-value: 6.833054e-06
```

8.5.2 Testing for Weak Instruments

Weak instruments (those poorly correlated with the endogenous regressors) will result in estimators with poor finite sample properties (high variance, possibly large finite sample biases). To check for weak instruments, run the "first stage regression" (as though doing 2SLS manually), i.e.,

- Regress each endogenous regressor on all exogenous regressors and instruments
- Test for significance of the instruments in the first stage regressions
- F-statistics should be large (on the order of 20 or so)

The "First Stage Regression" in our example is

$$educ_i = \delta_0 + \delta_1 age_i + \delta_2 \ln tenure_i + \delta_3 feduc_i + \delta_4 meduc_i$$

and the hypothesis of invalid instrument is $H_0 : \delta_3 = \delta_4 = 0$. We test this in the code below:

```
mdl_firststage <- lm(educ ~ age+ln_tenure+feduc+meduc, data=dat)
linearHypothesis(mdl_firststage, c('feduc=0','meduc=0'),
                 vcov=vcovHC(mdl_firststage,type="HC1"))
```

```
Linear hypothesis test:
feduc = 0
meduc = 0

Model 1: restricted model
Model 2: educ ~ age + ln_tenure + feduc + meduc

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   4943
2   4941  2 252.69 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that $feduc_i$ and $meduc_i$ are not weak instruments.

8.5.3 Tests of Overidentifying Restrictions

Recall that the

- GMM objective function is: $J(W) = (Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta})^{\mathrm{T}}W(Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta})$
- General GMM estimator is: $\hat{\beta}_W^{gmm} = (X^{\mathrm{T}}ZWZ^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}ZWZ^{\mathrm{T}}y$

If $Z^{\mathrm{T}}X$ is square (the just-identified case) and invertible, then the GMM estimator reduces to $\hat{\beta}_W^{gmm} = (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y$. This, of course, is just the MM/2SLS/IV estimator. The objective function becomes:

$$J(W) = (Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta}^{gmm})^{\mathrm{T}}W(Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta}^{gmm}) = 0$$

since

$$Z^{\mathrm{T}}y - Z^{\mathrm{T}}X\hat{\beta}^{gmm} = Z^{\mathrm{T}}y - Z^{\mathrm{T}}X(Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y = 0\,.$$

In the over-identified case, on the other hand, we will have $J(W) > 0$ in general. However, if moment conditions **do** hold, then sample moment conditions should hold approximately, and $J(W)$ will still be close to zero. It can be shown, in that case, that

$$J \overset{a}{\sim} \chi^2(M - G)$$

$M - G$ is the number of "overidentifying restrictions" (number of excess instruments). This is the "test of overidentified restrictions" or J-test. A significant J-stat indicates that one or more of the moment conditions do not hold, perhaps one (or more) of the presumed exogenous regressors is actually endogenous, or perhaps one of the instruments is not exogenous, or some combination of these situations.

In our example we have one extra moment restriction so we can carry out the J-test:

```
J <- t(ZTy - ZTX %*% beta_GMM) %*% solve(eZZ_GMM) %*% (ZTy - ZTX %*% beta_GMM)
Jpval <- 1-pchisq(J,ncol(Z)-ncol(X))
cat("J-stat:", J, "   p-value:", Jpval)
```

```
J-stat: 8.168    p-value: 0.004263589
```

The J-statistic indicates some misspecification

8.6 Testing Endogeneity

If we have valid instruments, we can test if one or more (or all) of the endogenous regressors can be treated as exogenous. In the regression $Y = X\beta + \epsilon$ suppose

$$X = \begin{bmatrix} 1_n & X_{*1}^k & \dots & X_{*K}^k & X_{*,K+1}^g & \dots & X_{*,K+G}^g \end{bmatrix}$$
$$Z = \begin{bmatrix} 1_n & X_{*1}^k & \dots & X_{*K}^k & Z_{*,K+1} & \dots & Z_{*,K+M} \end{bmatrix}$$

The population moment conditions are $E(Z^{\mathrm{T}}\epsilon) = 0$. If X_{K+1}^g is in fact not endogenous, we can add it to the vector Z, i.e.,

$$\tilde{Z} = \begin{bmatrix} 1_n & X_{*1}^k & \dots & X_{*K}^k & X_{*,K+1}^g & Z_{*,K+1} & \dots & Z_{*,K+M} \end{bmatrix}$$

and the moment condition $E(\tilde{Z}^{\mathrm{T}}\epsilon) = 0$ will still hold. The idea of the test then is

- Estimate the regression equation using instrument set Z, get J_Z
- Estimate the regression equation using instrument set $\tilde{Z}$, get $J_{\tilde{Z}}$

If in fact X^g_{K+1} is exogenous, then both J-statistics should be close in value (with $J_{\tilde{Z}}$ larger than J_Z since more moment conditions are involved when using $\tilde{Z}$). If X^g_{K+1} is in fact not exogenous, then there should be significant difference between J_Z and $J_{\tilde{Z}}$.

Under the null that $X^g_{K+1,i}$ is exogenous, the "difference-in-J" statistic is

$$C = J_{\tilde{Z}} - J_Z \overset{a}{\sim} \chi^2(Q)$$

where Q is the nunmber of endogenous variables being tested for exogeneity (here $Q = 1$).

We continue with our example, and test if *educ* can be treated as exogenous. One complication is that to ensure $C > 0$, the weight matrix used in computing $J(Z)$ has to be the appropriate sub-matrix of the weight matrix used in computing $J(\tilde{Z})$. This is implemented in the R code below.

```
# C-Statistic, checking if "educ" is endogenous

#-- GMM when "educ" is exogenous
#-- set up matrices
Zr <- dat %>% select(c(const, age, ln_tenure, educ, feduc, meduc)) %>% as.matrix()
Zrcol <- dim(Zr)[2]
ZrTX <- t(Zr) %*% X ; XTZr <- t(X) %*% Zr ; ZrTZr <- t(Zr) %*% Zr ; ZrTy <- t(Zr) %*% y

#-- Get the necessary weight matrices
beta_TSLS_a <- solve(XTZr %*% solve(ZrTZr) %*% ZrTX) %*% XTZr %*% solve(ZrTZr) %*% ZrTy
ehat_TSLS_a <- y - X %*% beta_TSLS_a
eZZ_TSLS_a <- matrix(0, nrow=Zrcol, ncol=Zrcol)
for(i in 1:n){
  eZZ_TSLS_a <- eZZ_TSLS_a + ehat_TSLS_a[i]^2 * t(Zr[i,,drop=F]) %*% Zr[i,,drop=F]
}
W_a <- solve(eZZ_TSLS_a)
W_b <- W_a[-4,-4]  # fourth row and column associated with educ

#--GMM with educ exog
beta_GMM_a <- solve(XTZr %*% W_a %*% ZrTX) %*% XTZr %*% W_a %*% ZrTy
ehat_GMM_a <- y - X %*% beta_GMM_a
J_a <- t(t(Zr) %*% ehat_GMM_a) %*% W_a %*% t(Zr) %*% ehat_GMM_a

#--GMM with educ endo
beta_GMM_b <- solve(XTZ %*% W_b %*% ZTX) %*% XTZ %*% W_b %*% ZTy
ehat_GMM_b <- y - X %*% beta_GMM_b
J_b <- t(t(Z) %*% ehat_GMM_b) %*% W_b %*% t(Z) %*% ehat_GMM_b

#--Calculate C stat
C_stat <- J_a - J_b
cat("C:",C_stat,", p-value:", 1-pchisq(C_stat,ncol(Zr)-ncol(Z)))
```

```
C: 46.38576 , p-value: 9.711898e-12
```

We soundly reject the null that *educ* is exogenous.

8.7 Exercises

Exercise 8.1. Suppose that the variable Z is a valid instrument for X in the regression $Y = \beta_0 + \beta_1 X + \epsilon$. Show that

$$E(Y - \beta_0 - \beta_1 X) = 0 \text{ and } E((Y - \beta_0 - \beta_1 X)Z) = 0$$

implies $\beta_1 = Cov(Z, Y)/Cov(Z, X)$.

Exercise 8.2. Show that

$$\hat{\beta}_{mm} = (X^{\mathrm{T}} Z Z^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Z Z^{\mathrm{T}} .$$

by minimizing the sum of squared moments in (8.15)

(You can skip the second order conditions.)

Exercise 8.3. Show that solving the minimization problem

$$\hat{\beta}_W^{gmm} = \text{argmin}_{\hat{\beta}} \underbrace{(Z^{\mathrm{T}} y - Z^{\mathrm{T}} X\hat{\beta})^{\mathrm{T}} W (Z^{\mathrm{T}} y - Z^{\mathrm{T}} X\hat{\beta})}_{\text{"}J(W)\text{"}}$$

produces

$$\hat{\beta}_W^{gmm} = (X^{\mathrm{T}} Z W Z^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Z W Z^{\mathrm{T}} y .$$

(You can skip the second-order conditions).

Exercise 8.4. Show for the regression with G endogenous regressors and M instruments, that if $M = G$, then $\hat{\beta}_W^{gmm}$, $\hat{\beta}^{mm}$ and $\hat{\beta}^{2sls}$ given in equations (8.21), (8.16) and (8.17) respectively all reduce to the IV estimator formula (8.12).

Chapter 9
Introduction to Time Series

To come.

References

Auguie, Baptiste. 2015. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. http://CRAN.R-project.org/package=gridExtra.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Hayashi, Fumio. 2000. *Econometrics*. Princeton University Press.

Karline, Soetaert. 2015. *plot3D*. https://cran.r-project.org/web/packages/plot3D/index.html.

Meschiari, Stefano. 2023. *Latex2exp: Use LaTeX Expressions in Plots*.

Pedersen, Thomas Lin. 2023. *Patchwork: The Composer of Plots*.

Tay, Anthony, Daniel Preve, and Ismail Baydur. 2025. *Mathematics and Programming for the Quantitative Economist*. World Scientific.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*.

Zeileis, Achim, Susanne Köll, and Nathaniel Graham. 2020. "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R." *J. Stat. Softw.* 95 (1). https://doi.org/10.18637/jss.v095.i01.