Social-economic analysis with `Python`

Thomas de Graaff

2025-10-07

Table of contents

Preface

What

This syllabus provides the course material for the first three weeks of the course **Sociaaleconomische Analyse** (in Dutch). With this course we would like to bridge the gap between, on the one hand, applied statistics and (micro-economic) modeling and on the other hand putting all this in practice when performing empirical research. As such this course can as well be seen as preparation for the bachelor thesis. But above all, the course aims to provide students with some tools that we see as very useful for research; not only in the socio-economic sciences but outside them as well.

As we only have a limited amount of time available for this course, the amount of topics we can deal with is by its nature restricted. We decided to focus in the first three week on the basics of applied econometrics and as such this first part builds upon the foundations of the statistics course in the first period of the second year. But now we challenge the student to build more elaborate statistical models where specific attention is given to *presentation* and *interpretation* of the results. The last three parts of the course move on to economic welfare from a behavioral perspective. Not only is the concept of economic welfare central to almost all economic theories, the behavioral perspective allows for more *reflection* on the neo-classical assumptions typically made in introductionary economic courses.

Why

Although there are many and very good introductory textbooks on economic models and applied econometrics, the combination of the two is seldom seen. Apart from that there are two reasons why we wanted to write our own material. First, usually less time is spent on why certain, and at first sight very restrictive, *assumptions* are made. We want to bridge that gap and provide the student with more intuition on where models, evidence, and finally perhaps the “truth” (if there is such a thing) comes from. Second, how to *present* statistical evidence and the *interpretation* of that evidence is very important but usually not given much attention.

For Whom

This syllabus assumes that the reader has a basic working knowledge of statistics, data science and some calculus (typically those method courses Earth, Economics and Sustainability students enjoy in their first year and in period I of the second year). The syllabus can however be read as stand-alone, although that requires some more attentive reading and practising. Where we think it is necessary we provide (references to) background material. For the course **Sociaaleconomische analyse** both types of syllabi should be read **in total** and it might be wise to read the relevant material *before* the respective lecture. The big advantage of course is that lectures and reading material now really go one to one.

More specifically the courses *Wiskunde voor AED*, *Inleiding Economie*, *Modeling Future Earth* and *Statistiek en Data-analyse*.

1 Introduction

In this first introductory chapter, I will lay out the relation between theory development and theory testing as they are the cornerstones of scientific progress. After all, a theory or idea can only be scientific if the theory can be tested and, if need be, refuted. If the theory cannot be tested then it is not science. I will also explain the basic workflow of scientific research and the tools needed with specific emphasis on research in the social sciences. This chapter ends with a reading guide where we discuss each chapter in this syllabus and the relations between the chapters.

1.1 Theory, Models and Hypotheses

In 2021, Guido Imbens, Joshua Angrist and David Card received the Nobel price for economics (officially *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel*). The field they work in is applied econometrics with specific focus on finding **causal** relations. That means that with data they want to test whether phenomenon X has an effect on phenomenon Y. More, in detail, with a causal effect we mean that when we change X, there will be an effect on Y, *ceteris paribus*, and **not** necessarily the other way. So, when we change Y, X will not necessarily change.

Some students find the term *applied econometrics* daunting, but is no more or less than the application of statistic techniques in the socio-economic field.

And identifying causal effects is what most applied econometric work nowadays is focused on. And we will focus on that as well in Chapter 2, Chapter 3 and Chapter 4. But how do you know what to test, or, in other words, where do phenomena X and Y come from? Those phenomena and possible relations originate from scientific theories as you will have in all disciplines. And those theories are typically casts in models—usually in a very abstract manner. Models come in the form of computer simulations (such as with agent-based modeling), real physical models (as with displays), but often models are formulated in mathematical notation with the aim of being as precise, lucid and clear as possible. But note that these models are not necessarily theory. Theory is the underlying set of relations and assumptions that can say something about the specific structure of models. But very often one theory can lead to **multiple** models, each perhaps highlighting different aspects of the underlying theory. An example of such a theory is the Law of Diminishing Marginal Utility: each additional unit of the same good is appreciated less by consumers. This theory can be expressed in many mathematical ways but the underlying concept as displayed in Figure 1.1 always remains the same. This type of function, increasing but slower and slower, belongs to the family of *concave* functions. A function with exponential growth (such as e^{gt} belongs to the family of *convex* functions). But how such a function should exactly be defined is not a-priori clear.

So how to relate this with each other in scientific research? Well, when doing research you are interested in something that is not yet known (the research gap). Your aim is to (partly) fill this research gap by answering a research question. To answer this research question you need theory (a theoretical framework); what do you need to assume, what are the most important (moderator) variables, how do they relate with each other, and so on and so forth. From this theory you construct a model. Not necessarily a mathematical one. For example, you can also make a model in a Geographical Information System environment where you visualize layers of information that you think are most relevant based upon theory (in this case often previous scientific literature). Or you make a simulation model examining risks of flooding by rivers. The final step is the stage where your model should provide you with some answers. Sometimes they are concerned with optimality (what is the best location of a new road in a GIS environment), prediction (where are river dikes most vulnerable), or with establishing a (causal) relation. And it is the latter that this course deals with. How can we know that there is a relation between phenomenon X and phenomenon Y and how do we know whether that relation is causal?

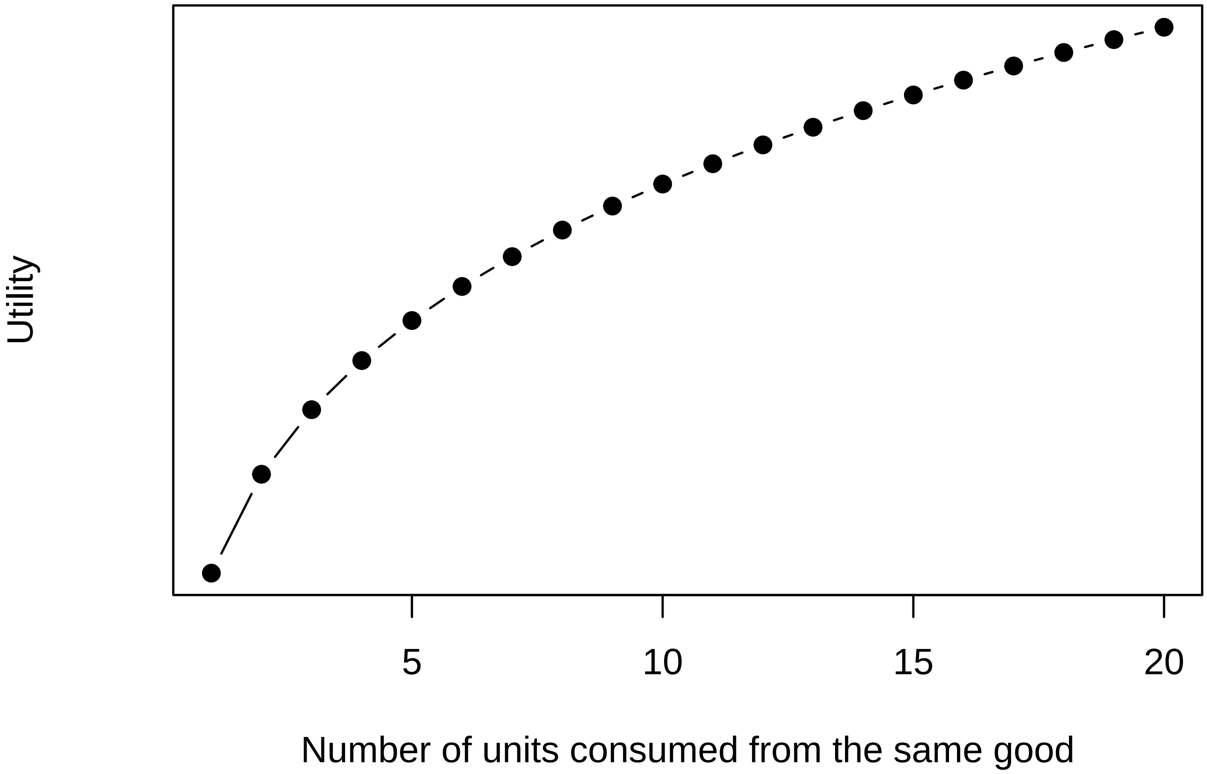


Figure 1.1: Law of Diminishing Marginal Utility

For that we use applied econometrics (which is a form of applied statistics but then in the social-economic sciences domain—the exact difference will be discussed in Chapter 2). And to establish a, hopefully causal, relation, we test our models with empirical data. Be aware, though, that the applied econometrics materials we teach in this course (and in all introductory courses of applied statistics and econometrics all over the world—the "101" courses) is based on so-called frequentist statistics (To freshen up you knowledge about the basics of statistics you might want to read Appendix A). The exact definition is not important for now but know that it is intrinsically related with hypothesis testing.

And hypothesis testing is most often associated with that scientific philosopher—and perhaps the only one you know—Karl Popper (as displayed in Figure 1.2).

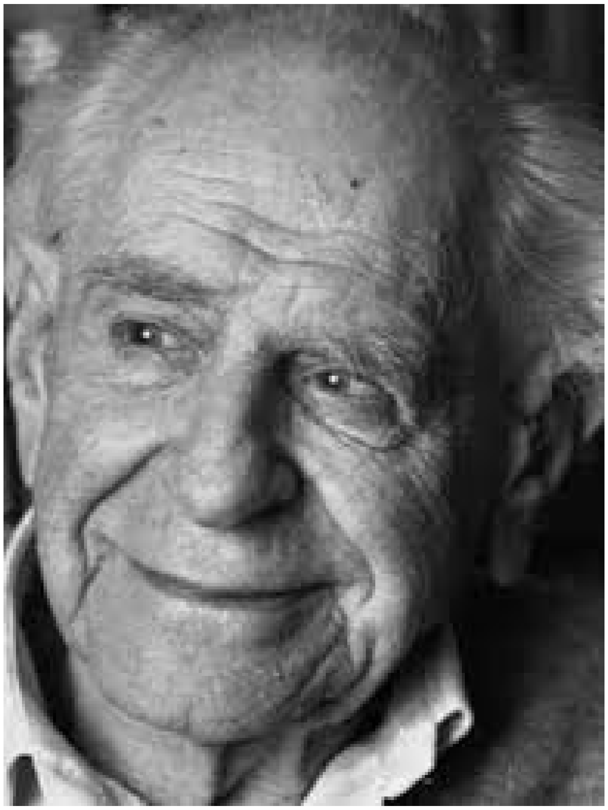


Figure 1.2: Karl Popper

Popper was a so-called empiricist and claimed that theories in the empirical sciences (that includes most of the social sciences) can never be proven, only rejected. That is why you can reject a null-hypothesis (H_0), but **never** accept the alternative hypothesis (H_a). And this is *highly* related with the theoretical framework behind frequentist statistics. Loosely speaking, in frequentist statistics you construct a world where H_0 is true and you try to reject that world with data (we will come back to this in Chapter 2)—but that world does not say anything about the validity of the alternative hypothesis. Now, Popper never claimed that rejecting one null-hypothesis will reject a whole theory. For that you need a larger body of evidence, including results from all sorts of studies—not only statistical ones, leading to a *general* consensus amongst the *whole* scientific community (Popper 2005).

In truth, although the scientific approach of working with null-hypotheses is a very valuable one (and remarkably practical), it does not always lead to definitive answers. That is because contradicting models can lead to similar null-hypotheses. Moreover, often the connection between research question and null-hypothesis is not a direct one. Consider that you want to know the effect of X on Y, so your research question is: "What is the effect of X on Y?". But, in a frequentist world you only reject hypotheses—thus leading to results in the line of: "The effect of X on Y is *not* ...". This makes the evidence for your research question at least circumstantial and in the best case indirect. The bottom-line here is that one needs to be careful in drawing conclusions based on null-hypotheses (and in a broader sense based on models in general). Scientific research typically advances very slowly—but hopefully in a robust and parsimonious way!

1.2 Doing Research (in the Social Sciences)

It is remarkable that, although in principle students are (should be) prepared for scientific research, they receive little guidance in *how* they should do scientific research. What are the tips & tricks of the trade and what—and, more importantly why—should you use with respect to specific (types of) applications and what is the relation between them. In our view most of this should belong in your first course upon entering the university (with the appropriate course title “Research Methods 101”). And some of it you indeed have learned in your first year, but in our experience students still lack “operational” knowledge. Therefore, we discuss below the four elements we think are among the most important—at least for this course. There are others, but for now this will do.

1.2.1 Work tidy

Our first and most important tip is to work tidy. Try to make your work look **good**. And with work we mean everything you submit (such as tutorials, papers, examinations, and theses). And that is because lecturers are just like people and often think from primary instincts with their reptilian brain: if it doesn’t look good, not much time is spent on arguing and thinking as well! Moreover, when your work is difficult to read, lecturers get annoyed. Making your work look good and in the same time more lucid and transparent also serves a higher purpose as it is then easier to detect mistakes. Namely, everyone makes mistakes. The important thing is to detect them early, learn from them and remedy them. This advances science in general and is a very important feature of the scientific process. Chapter 4 will spent additional time on working tidy and making it looking good.

1.2.2 Know where your stuff is

A second very straightforward tip is to be organised and to know where your stuff is. Often, students come to us for help with all their files piled on a stack on their desktop and facing difficulties finding where their work is. It is always advisable to use a folder structure and have one folder for one project (or for one course). And to the use sub-folders for data, text, code, pdf’s and so forth. A second tip for organisation is to think about versioning. As the well-known Figure Figure 1.3 shows the number of versions of one file very quickly can get out of hand. Think at least about a consistent naming structure (perhaps with the date involved such as `paper_20221215.doc`).

1.2.3 Make notes

One skill that in our opinion is given not too much attention is making *useful* notes. It has been proven that writing things down is beneficial; not only for remembering but also for understanding. And that seems to be best just by using a pen as this slows writing down and you have to think about what to write down. Underlining or marking is useful less beneficial than writing accompanying notes. But when should you write notes? Well, when attending lectures of course but also when reading. To leverage your notes as much as possible it is important that you have a system where you can *retrieve* your notes and compare them with other notes. The latter is the hardest part, but is in the long run the most rewarding as new connections are created between lectures, courses, books, and years. You do not need any fancy tools for this (there is literally a ton of applications to be found on internet), Microsoft's Onenote or Evernote are more than good enough. Where the workflow typically is to first use pen and paper to *capture* notes and thereafter rewrite and organise your notes in a notes system.

See for a theoretical underpinning of this statement Ahrens (2022).

1.2.4 Use a reference manager!

Perhaps the tool that has the quickest pay-off is a reference manager. For those of you who are not using one yet: **do** it. Why? Because you never have to think about your reference list again. All reference managers come with plugins for Word or other text-editors (or type-setters such as LaTeX that enable you to *automatically* generate reference list based upon in-text citations which the reference manager can also provide. You only need less than an hour to set it up, but you very quickly become more efficient (and thus *save time* in future work). There are many reference managers out there, but we advise Zotero as it is open source. There is both a cloud and desktop version and it comes with a handy tutorial. It also provides a plugin for

Note that references that do not exists or in-text citations not listed later in a reference list is considered a form of **fraud**. You always want (or need) a clear correspondence from your in-text citations to your reference list and vice versa.

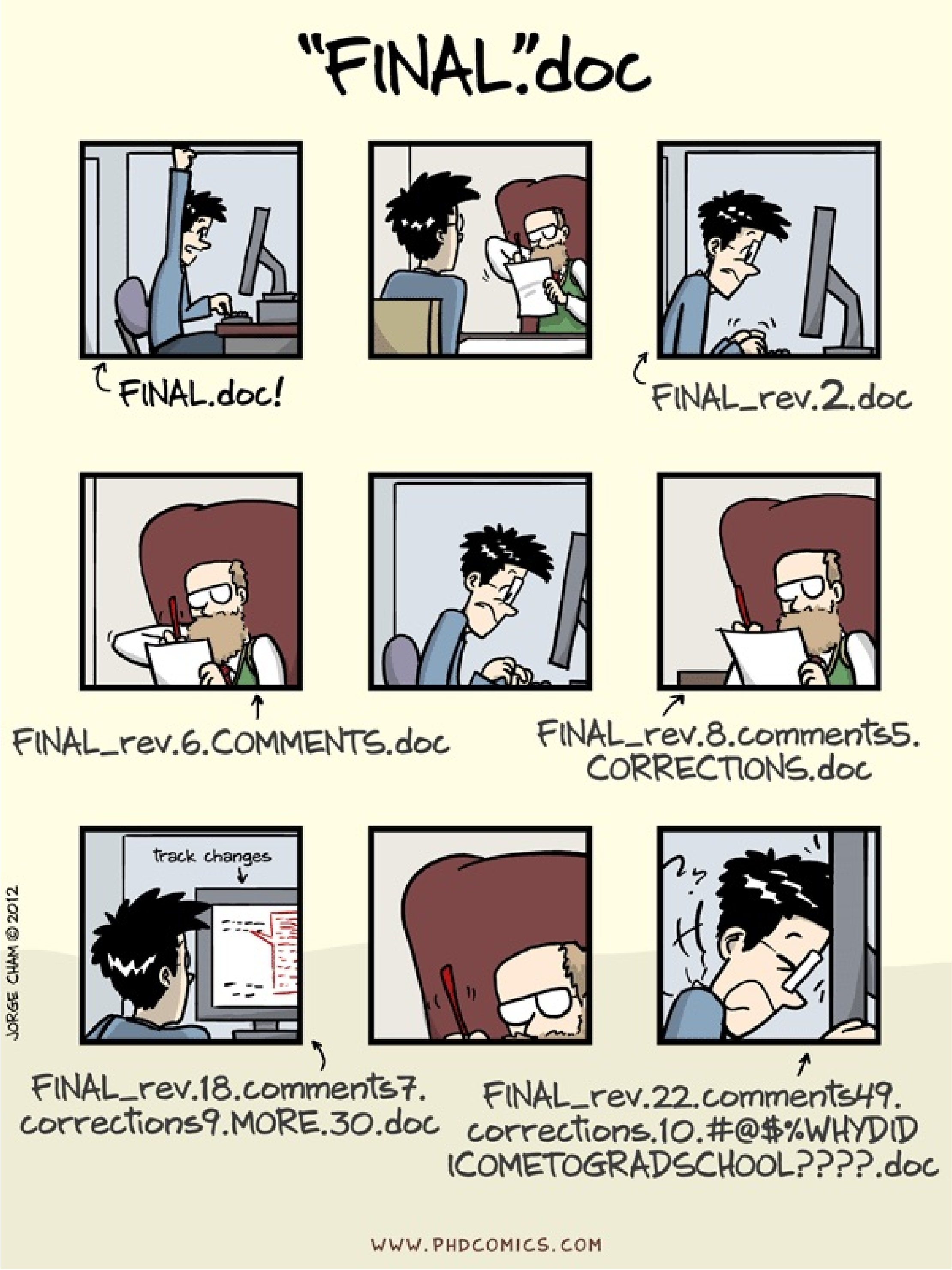


Figure 1.3: Version confusion

your browser to automatically import the bibliographic details of the paper you are reading at the moment.

1.3 Statistical software `Python`

As quantitative research becomes more and more important in the social sciences you need software to **manage** your data and provides statistical and applied econometric **analyses**. We will use `Python` in this course. However, note that `Python` has a steeper learning curve than e.g., `STATA`, and does not work immediately out-of-the-box. But the user base is large and that is important, because for each problem there is much material to be found on internet (including videos). In this syllabus we will give specific attention to `Python`-scripts, including why you code it like such and what the intuition is behind that.

1.3.1 Where to get `Python`

First, you need to install `Python` itself. We advise to use the Anaconda platform. You can do this by downloading this from Anaconda. Now choose your appropriate operating system, choose the `base` system, download `Anaconda` and install it. That's it!

Note that this setup is similar to "Statistiek en Data Analyse" as to be as consistent as possible.

In the `Anaconda` environment we will mainly use `Jupyter notebooks`. To view, edit and run `Python` code. Note, although `Jupyter notebooks` are wonderful for educational purposes, that sometimes you want to run `Python` within your own editor. From `Anaconda you` can download for instance the editor `VS code`, which is a very good and all purpose editor.

To be quite honest most editors nowadays are good. Best advise is to choose one and stick to it as you get used to it and gain in efficiency.

1.3.2 Why use `Python` and not any other application?

Ask any data scientist at the moment for the software tools most used and they will most likely answer `R` or `Python`. Of course, that should not be a valid answer (many people use `Word` as well and nobody would argue that `Word` is brilliantly programmed or designed), but it indicates the popularity (and the community) that uses `Python`.

Where 15 years ago most social scientists still used `SPSS` (and the economists `Stata`), that has now changed completely (well, the economists still use `Stata`, but the rest of the world moved on). And for good reasons, namely:

1. It is open source and thus free;
2. `Python` is flexible and thus multi-purpose;
3. there is now a **very** large userbase; everything you can dream of (that is, in the context of data science/management, maps and manipulating text), somebody else most likely already programmed;
4. it generates beautiful pictures, diagram, maps, and histograms (even 3D pie diagrams for the masochists amongst us);
5. relative to `Stata` or `Excel` it is fast, which is great for larger (spatial) databases.

Hello Large Language Models!

In general, you can use `Python` for statistical analysis, simulation analysis, data management, visual display of data, creating documents (and presentations), and even GIS applications. In that respect it is far more flexible than `Stata`. Last but perhaps not least, `Python` is more and more used outside academia as well. Twitter, Facebook, Booking and Google use `Python`.

Social science students might find working with `Python` initially strange, cumbersome or even frustrating. All the lovely drop-menus that are still provided by `Excel` and `Stata` have disappeared, and the whole thing is completely script-driven. In fact, `Python` is a full-blown program language. I am aware that this needs some adaptation. However, hopefully, learning `Python` in combination with `Jupyter`-notebooks will pay-off; if not in becoming more efficient and reproducable, then at least in the fact that you start to understand a *different* way of doing things and that the office suite (Word, Excel and Powerpoint) is not the only option out there.

1.4 Reading Guide

This course will not concern itself with theory as such, but more with how to test that theory (the *applied* in applied econometrics). Chapter 2 introduces the basic concepts of applied econometrics in the form of univariate regression. Chapter 3 extends this framework to a multivariate regression setting, but in the same deals as well with the translation of theoretical (socio-economic) models to empirical models that are testable. Chapter 4 discusses how to *specify* your model—which variables should you include and which variables not—and how to present your findings to a wider audience (that includes assessors). The final chapter summarizes and provides a general discussion.

2 Regression Analysis in the Social Sciences

```
# Python packages we need for this chapter
import numpy as np
import pandas as pd
# for making nice plots in combination with matplotlib
import seaborn as sns
from scipy import stats
# for regression related stuff
import statsmodels.api as sm
# the next package allows us to formula's for ols
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
```

2.1 Introduction

Why should we have something as applied econometrics in the social sciences? That is because we have theories and those theories contain variables such as in the direct utility function of model (2.1):

$$U(x_1, x2) = x_1^{\alpha} \cdot x_2^{\beta} \tag{2.1}$$

Here, the quantities of the goods x_1 and x_2 are considered to be **known**—also often referred to as **data**. In theoretical work they are fictional or sometimes simulated. The parameters α and β are **not known** and we often want to **estimate** them. If we know what the parameters α and β should be, then we can calculate what utility a certain amount of consumption of both goods gives and then, for example, assess which combination of goods gives maximum utility.

This course is about using data to quantify (socio-economic) parameters. Moreover, we focus on measuring **causal** effects, instead of mere correlations. Note, that in an ideal world, we would like conducting experiments as to measure a causal relation of a phenomenon X on Y. However, we almost always only have observational data on, for example, demand and prices. Therefore, this course and syllabus deals with (i) difficulties arising from using observational data to estimate these causal effects and (ii) rewriting models as (2.1) such that we can actually *use* data to tease out 'reasonable' values for—in this case—α and β.

Moreover, very often experiments are not feasible or highly unethical.

This chapter is organised as follows. The next section addresses the problem of finding a *relation* between some X and some Y. Here, we follow an example from the well-known textbook of Stock, Watson, et al. (2003) where we look at the relation between school class size and school class performance. At the same time, we also introduce some `Python` commands. To do so, this section deals as well with the statistical framework that is needed for applied econometrics. Note that we assume that the reader already had a course in introductionary statistics and that we provide only the basic concepts most important for this course in Appendix A.

Such as Statistiek en Data-analyse

2.2 So, what is the problem?

As explained in the introduction above, applied econometrics aims to give the policy maker well-informed, and evidence based, values for variables she needs. She needs these variables basically for two separate things:

1. **Causal inference**: The policy maker wants to assess the effect of a change in one variable (typically called X) on another variables (often called Y).
2. **Prediction**: If you know what variable X is, what should Y then be?

Nowadays, most applied econometric techniques are concerned with causal inference, not so much with prediction. Even more, the techniques often applied are beneficial for correct causal inference, but might harm prediction. However, see that without correct causal inference (so knowing the **true** causal mechanism) prediction is always cumbersome. That is why current methods such as machine learning methods first focus on finding the correct causal mechanism (even without sometimes specifying what they may be) and then optimize prediction.

Thus, finding (causal) mechanism helps the scientist or policy maker in assessing the outcomes of a particular (policy) intervention. In the economic realm one could think about trying to assess the following quantities:

- To what extent do people eat less meat if we increase the prices with 1% (using a meat-tax)?
- If we increase Dutch dikes with one meter, how much less flood risk will there be?
- How much do classes perform better is we reduce class-size with one student?

2.2.1 A first encounter with `Python`

For this section, we will focus on the last question. And this is an important question for policy as teachers are costly, but parents value school performance very highly. To start answering this question we *use* data. Data can be in many formats, but usually it is in a text file, often with the extension `.csv`. The good thing about this format is that it can always be read with every computer and every operating system.

`.csv` stands for comma separated value and is just a text file with commas denoting the columns.

–>

Now suppose you have the subdirectory `data` in your course folder and in that data directory you have a file called `CASchools.csv`. Now, this dataset describes 420 school districts in California and, amongst other things, their average performance (measured by a test score) and their financial constraints (measured by the amount of students per teacher).

Often it is as well advisable to make a distinction between *derived* and *raw* data. Raw data is original data and derived data is data you have transformed or worked on. In principle, you do not want to change the original data!

This dataset can be downloaded from Canvas as well, with already the data transformed.

To import the data in Python you make use of the read_csv() function from pandas, as follows:

```
# read the data
# Note that the pathname is a string and should be in parentheses
data = pd.read_csv('/Users/tomba/projects/syllabus_python/data/CASchools.csv')
```

Note that we can comment on our code by the use of the #-sign. Everything after the #-sign will not be used for computation. And I advise you to regularly comment on your code as that makes it easier for somebody and future you to understand the code. Moreover, the = symbol indicated that the output of the command on the right hand side is poured into a new object (somethings we conveniently in this case call data) on the left hand side of this symbol. This object is now a so-called dataframe which you can open use further.

When working with Jupyter notebooks, you do not have to specify the path as long as the data is in the same directory as the notebook.

Now, we would like to know how the data set looks like, for example what kind of variables it contains. We do this by invoking the command head which give the first six lines of the data set:

```
data.head()
```

```
   Unnamed: 0  district  ...        read        math
0           1     75119  ...  691.599976  690.000000
1           2     61499  ...  660.500000  661.900024
2           3     61549  ...  636.299988  650.900024
3           4     61457  ...  651.900024  643.500000
4           5     61523  ...  641.799988  639.900024

[5 rows x 15 columns]
```

This provides information about the variables and the names and types of variables. In this case variables are either a float (a real number) or a string (text). Note as well, that this kind of output is cumbersome and ugly and not fit directly for reporting. Later, we will try to make this look better in an automatic way.

The pandas function .descibe() gives descriptive statistics.

```
data.describe()
```

```
       Unnamed: 0      district  ...        read        math
count  420.000000    420.000000  ...  420.000000  420.000000
mean   210.500000  67472.809524  ...  654.970477  653.342619
std    121.387808   3466.994655  ...   20.107980   18.754202
min      1.000000  61382.000000  ...  604.500000  605.400024
25%    105.750000  64307.750000  ...  640.400024  639.375015
50%    210.500000  67760.500000  ...  655.750000  652.449982
75%    315.250000  70419.000000  ...  668.725006  665.849991
max    420.000000  75440.000000  ...  704.000000  709.500000

[8 rows x 12 columns]
```

Suppose that in this case we are only interested in the variable `math` (average scores for mathematics by district). Then we invoke this by:

```
data[['math']].describe()
```

```
             math
count  420.000000
mean   653.342619
std     18.754202
min    605.400024
25%    639.375015
50%    652.449982
75%    665.849991
max    709.500000
```

Note that we invoke the function `describe` on one variable `math` of the dataframe `data`.

Now we see descriptive statistics for the variable `math`, containing the mean, the standard deviation, the minimum and maximum, and the first, second and third quartile of this variable. For this exercise and also for the remainder of this syllabus we are actually interested in two variables that are not in the data set, but that we have to compute ourselves; namely, the average test scores of reading and mathematics and the size of the classes (or the student-teacher ratio). We do so as follows:

```
# compute new variable STR and append it to data dataframe
data['STR'] = data['students'] / data['teachers']

# compute new variable TestScore and append it to data dataframe
data['testscr'] = (data['read'] + data['math']) / 2
```

Here, the `=` symbol again means you are creating something new on the left hand side of the symbol (in this case a variable in the dataframe `data`, but in general you always make a new *object*).

`python` is a so-called object-oriented language and (almost) everything is an object including dataframes and variables.

For a first insight in the relation between class size and class performance we might want to draw a so-called scatter plot. These types of plots relate the values of two variables in a two-dimensional way by giving the values as coordinates. The following syntax will do so.

Here we use the `seaborn` library as it is easier to draw scatterplots and trendlines. Note that the first input of the function specifies the data, the second what should be on the x-axis and the third what should be on the y-axis.

```
sns.scatterplot(data = data, x = "STR" , y = "testscr")
```

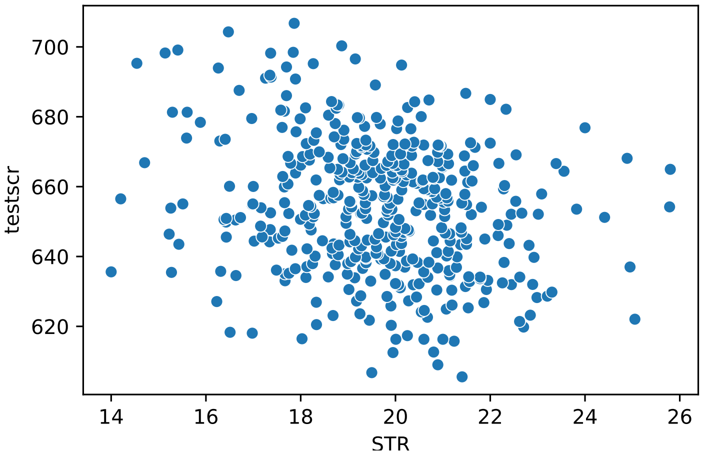


Figure 2.1: The relation between average number of pupils in a class room and average test scores in a scatter-plot

This “cloud” of dots does not yield a clearly visible relation between class performance and class size. However, this can be deceptive. Often it is difficult to discern clear relations from raw data only. Therefore we need to resort to numerical evidence.

2.2.2 Numerical evidence

To assess whether there is a relation between class performance and class size as displayed in Figure 2.1 we need numerical or statistical evidence. Before we start to engage in regression analysis, we first perform a rather simple analysis, but the underlying mechanisms are identical to that of regression analysis. We first create two groups: namely, districts with “small” (number of students per teacher is below 20 or STR < 20) and “large” (number of students per teacher is equal or above 20 or STR ≥ 20) class sizes. Then we can adopt three relatively straightforward strategies here:

1. Estimation

   - Here, we compare the average test scores in districts with low student-teacher ratios to those with high student-teacher ratios. So, we basically try to assess whether **average** behaviour is different.

2. Hypothesis testing

   - Now, we aim to **test** the “null” hypothesis that the mean test scores in the two types of districts are the same, against the “alternative” hypothesis that they differ.

3. Confidence intervals

   - This strategy estimates an interval for the **difference** in the mean test scores, so small versus large student-teacher ratio districts.

In `Python` we can make a start with this data analysis by executing the following two commands:

```
data['large'] = data['STR'] >= 20
data.groupby('large')['testscr'].agg(['mean', 'std', 'count'])
```

```
             mean        std  count
large
False  657.246233  19.385417    239
True   650.076798  17.853782    181
```

The first command *generates* a new variable called `large` and denotes an indicator being 0 (or False) if STR < 20 and 1 (or True) if STR ≥ 20. The second command summarizes the testscore variable again, but now only gives the mean, standard deviation and the number of observations and does this by each value of the new variable `large` (the `groupby` function). Of course, this output is rather ugly and it is better to make a nice table such as Table 2.1).

Table 2.1: Descriptive statistics of small and large classes

Class size	Average score	Standard deviation	Observations
Small	657.2	19.4	239
Large	650.1	17.9	181

Now, for all three strategies (estimation, testing, confidence intervals) we want to know something about the difference—usually denoted as Δ. Or, specifically:

1. For estimation: determine the Δ or the difference between group means
2. For hypothesis testing: can we *reject* the null-hypothesis that the difference is zero, or$\Delta = 0$
3. For confidence intervals: can we construct a confidence interval for Δ

2.2.2.1 Estimation

In his case the concept of estimation (that is to determine the difference between the two groups' average scores) is rather straightforward as we need to calculate the *difference* between the mean test scores within each group, or:

$$\begin{aligned}\bar{Y}_{small} - \bar{Y}_{large} &= \frac{1}{n_{small}} \sum_{i=1}^{n_{small}} Y_i - \frac{1}{n_{large}} \sum_{i=1}^{n_{large}} Y_i \\ &= 657.2 - 650.1 \\ &= 7.2 \end{aligned} \tag{2.2}$$

This basically means subtracting the average scores of Table 2.1 (later we see how to do this automatically in R). Now, the difference—Δ—equals 7.2. We then have to ask ourselves whether this is a large difference in a real-world sense. Note that, from Figure 2.1, test scores seem to range from 600 to 800 and do not really have a direct meaning for us. A useful trick then is to look at the standard deviation (note that if things are normally distributed, 95% of all probability mass is within the range mean plus or minus two times the standard deviation). In this case, the difference is about 1/3 of the standard deviation. A different way of looking at this is looking at the percentiles of test scores. In python we can do with the percentile function from numpy:

```python
percentiles = [10, 25, 50, 75, 90]
np.percentile(data['testscr'], percentiles)
```

```
array([630.39502258, 640.05000305, 654.44999695, 666.66249084,
       678.8599884 ])
```

where the function percentile asks for a tabulation of certain statistics and gives the x-th percentile of that statistic. Now note that between the 50th and 75th percentile there is only 12 points. So given this information, the difference of 7.4 is rather sizable. But whether this difference is big enough to be important for school reform discussions, for parents, or for a school committee is a question we cannot answer with this analysis.

2.2.2.2 Hypothesis testing

An alternative is to test the null-hypothesis that the difference $\Delta = 0$. For that we need a so-called difference-in-means test and compute the corresponding t-statistic,[1]

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)} \tag{2.3}$$

where $SE(\bar{Y}_s - \bar{Y}_l)$ is the *standard error* of $(\bar{Y}_s - \bar{Y}_l)$, the subscripts s and l refer to "small" and "large" STR districts, and $s_s^2 = \frac{1}{n_{small}} \sum_{i=1}^{n_s} (Y_i - \bar{Y}_s)^2$

We can compute this difference-of-means t-statistic by filling this in with the numbers of Table 2.1:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.2 - 650.1}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.2}{1.83} = 3.92 \tag{2.4}$$

But then what? Well, recall that we **reject** a null-hypothesis when the critical value is below a certain threshold (usually 5%). In this case that is equivalent with stating that $|t| > 1.96$. So, we reject (at the 5% significance level) the null hypothesis that the two means are the same. We will come back to this procedure in Section 2.3.2.2.

2.2.2.3 Confidence interval

Finally, we can construct a 95% confidence interval for the difference between the means, which is:

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l) = 7.2 \pm 1.96 \times 1.83 = (3.6, 10.8) \tag{2.5}$$

So, what does this mean again. Well, two things. First, the 95% confidence interval for Δ doesn't include 0 and, second, the hypothesis that $\Delta = 0$ is rejected at the 5% level. We will come back to confidence intervals as well, but for now a confidence interval can be seen as an interval of numbers that will not be rejected as null-hypothesis.

[1] This is something that is extensively dealt with in the *statistiek en data analyse* course.

2.2.3 Always be smart (and a bit lazy)

So, why give this rather simple procedure so much attention. That is because all “classical” statistics are centered around these three elements and statistical computer output will always give, at least, these three. And they are as well very much related with each other. Once you know two of them, you know the third one as well.

Now, although the procedure is rather straightforward, it is also a bit cumbersome and prone to errors. Therefore, it is much easier to let `python` do it:

```python
small_classes = data[data['large'] == 0]['testscr']
large_classes = data[data['large'] == 1]['testscr']

t_stat, p_val = stats.ttest_ind(small_classes, large_classes, equal_var=False)  # Welch's t-tes

# note that in code below :.3f states 3 digits
print(f"t-statistic: {t_stat:.3f}, P-value: {p_val:.3f}")
```

```
t-statistic: 3.927, P-value: 0.000
```

So, in this case we want to assess the difference in test score by groups (being small and large classes). Note, however that the difference in means, *t*-statistic and thus the confidence intervals are similar as in, e.g., equation 2.5.

Now try to find out for yourself that this output gives you the estimation of Δ of equation 2.2, the *t*-value and corresponding test outcome of equation 2.3 where the corresponding confidence intervals of equation 2.5 can be readily calculated.

2.3 Univariate regression

The three strategies we adopted in Section 2.2.2 for assessing the difference between groups directly translate to the case of regression analysis. Here we also look at estimation, hypothesis testing and confidence intervals. But before that we first look at the origin of the name regression in Section 2.3.1.

2.3.1 Genesis: *regression towards the mean*

The name regression seems to have a negative connotation, as progress is in general seen as good and regression as bad. And actually this is true as the name regression was deliberately given as to describe a negative process: in full *regression towards the mean*. The concept of regression was actually coined by Sir Francis Galton together with other statistical terms, such as correlation and deviation (Senn 2011). Galton was a notorious statistician who measured everything and else, including the length of french bread and the size of human skulls.

in 1886, Galton started to research the height of adult children with the height of their parents (Galton 1886). The original data can be seen in the scatterplot in Figure 2.2. What Galton expected was that the relation between the height of children and that of their parents was a one-to-one relation. On average children should receive the same height of their parents. So, in fact he expected a 45° line—the red line; a line with slope equal to 1.

However, he found consistently the blue line, a line with positive slope but lower than 1 (the blue line in Figure 2.3). That entails that, *on average* tall parents get tall children but not as tall as themselves. Of course, this goes as well the other way. Short parents get short children but not as short as themselves.

Galton coined this process *regression towards the mean*.[2] In the end we would all converge towards the mean and all look the same. For the Victorian Sir Frances Galton and his contemporaries in an age where social and income classes were highly separated this was truly a horror. Especially, because his cousin was Charles Darwin who actually claimed that species *diverged*. Of course, in Galton reasoning there is a mistake as this only models genetic influence and not *accidental* differences not influenced by genetics. Note as well that this analysis says something about the average, but not about individual differences.

This regression towards the mean is now seen as a very important characteristic of regression models, and you can easily be fooled by it. It is now stated as:

[2] Modern statisticians actually see this as a form of shrinkage.

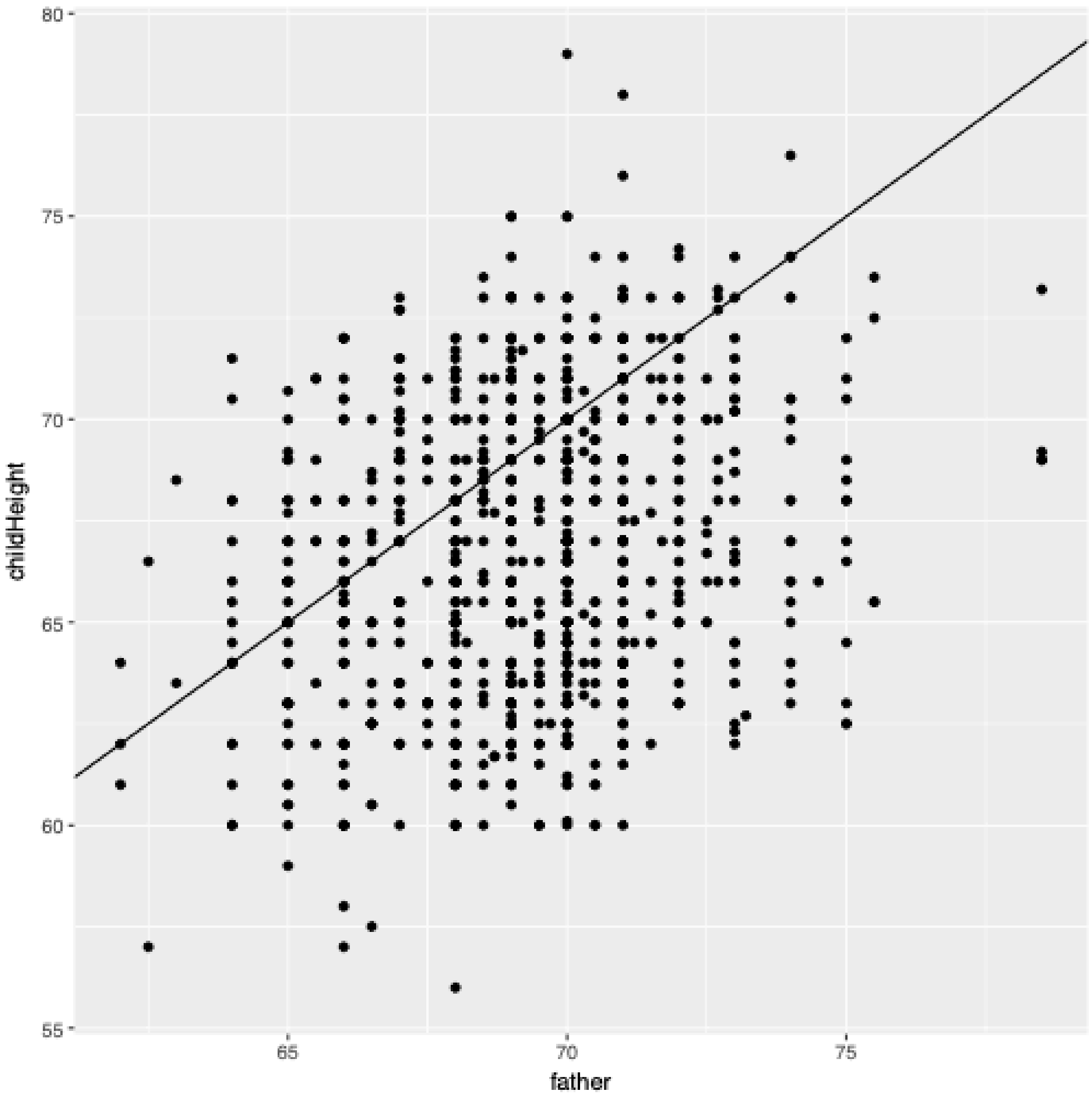


Figure 2.2: Relation between the heigh fathers and the height of children

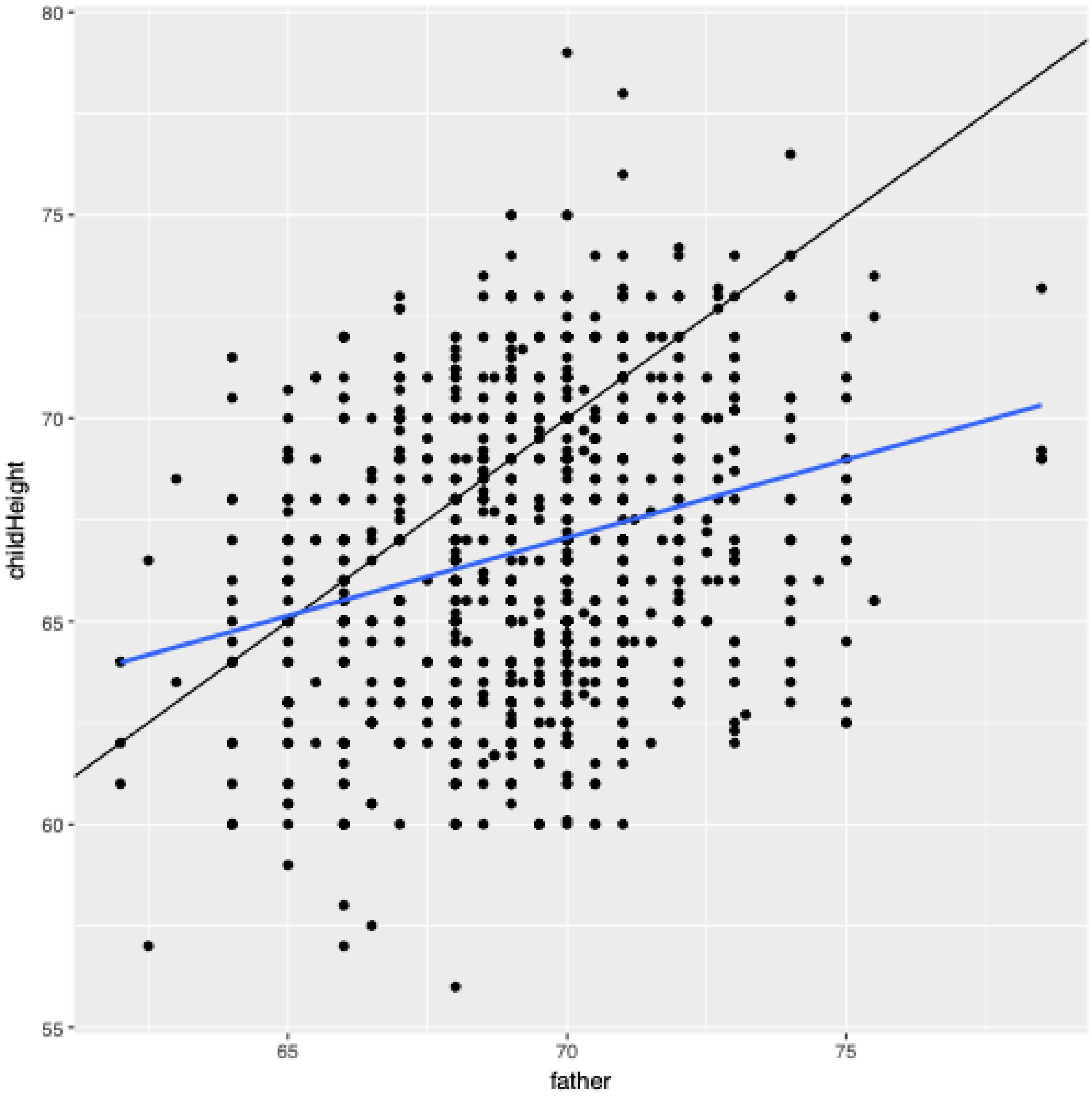


Figure 2.3: Relation heigh fathers and height children

> ... a concept that refers to the fact that if one sample of a random variable is extreme, the next sampling of the same random variable is likely to be closer to its mean

For instance, suppose that everything went really well for a course and you got a 9 for an examination. That does not mean that the next time you will do equally well (you will still do well, but not that well). Or, your favorite football club does extremely well in a particular year (Leicester City FC comes to mind who became premier league champion in 2016). That does not mean that the next year it will do equally well, and so forth and so on.

2.3.2 Regression with one regressor

So, linear regression allows us to *estimate*, and make *inferences* about, *population* slope coefficients. Inference means drawing conclusions and population refers to the fact that we do not want to say something about our sample, but instead about the whole population. Ultimately our aim is to estimate the **causal** effect on Y of a unit change in X—but for now, just think of the problem of fitting a straight line to data on two variables, Y and X.

Similar to Subsection Section 2.2.2) we have three strategies to make inferences:

- We estimation the relation:
  - This now boils down to the question how we should draw a line through the data to estimate the (population) slope using Ordinary Least Squares (OLS—a specific and most common type of regression analysis)
  - And then we have to assess the advantages and disadvantages of OLS
- We could refer to hypothesis testing:
  - Very often this comes down to testing where the slope is zero. Namely, if the slope is zero, then the data does not show a relation between Y and X.

- Using confidence intervals:
  - This is related to constructing a confidence interval for the slope

Before we look into this we first need some clarification on notation. As mentioned above, we would like to know the population regression line:

$$testscr = \beta_0 + \beta_1 STR, \tag{2.6}$$

where

$$\begin{aligned} \beta_1 &= \text{slope of population regression line} \\ &= \frac{\Delta Testscore}{\Delta STR} \\ &= \text{change in test score for a } \textbf{unit} \text{ change in STR} \end{aligned} \tag{2.7}$$

Note the definition here of β_1. It gives the **marginal effect** of a change in STR on $testscr$. So the interpretation of the parameter β_1 is very straightforward. However, we do not know the population value of β_1 and we therefore have to estimate it using data.

In general, the population linear regression *model* is different as we add element u_1.

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i \ldots n \tag{2.8}$$

Now, X denotes the independent variable or regressor, Y the dependent variable, β_0 the intercept, β_1 the slope, and u_i the regression error. The regression error consists of omitted factors, or possibly measurement error in the measurement of Y. In general, these omitted factors are other factors that influence Y, other than the variable X.

2.3.2.1 Estimating with OLS

To estimate the population linear regression model we apply the ordinary least squares estimator. Again, as Figure 2.4 shows as well, a linear regression line is a straight line through points in a scatterplot. Actually, we want to draw that line such that

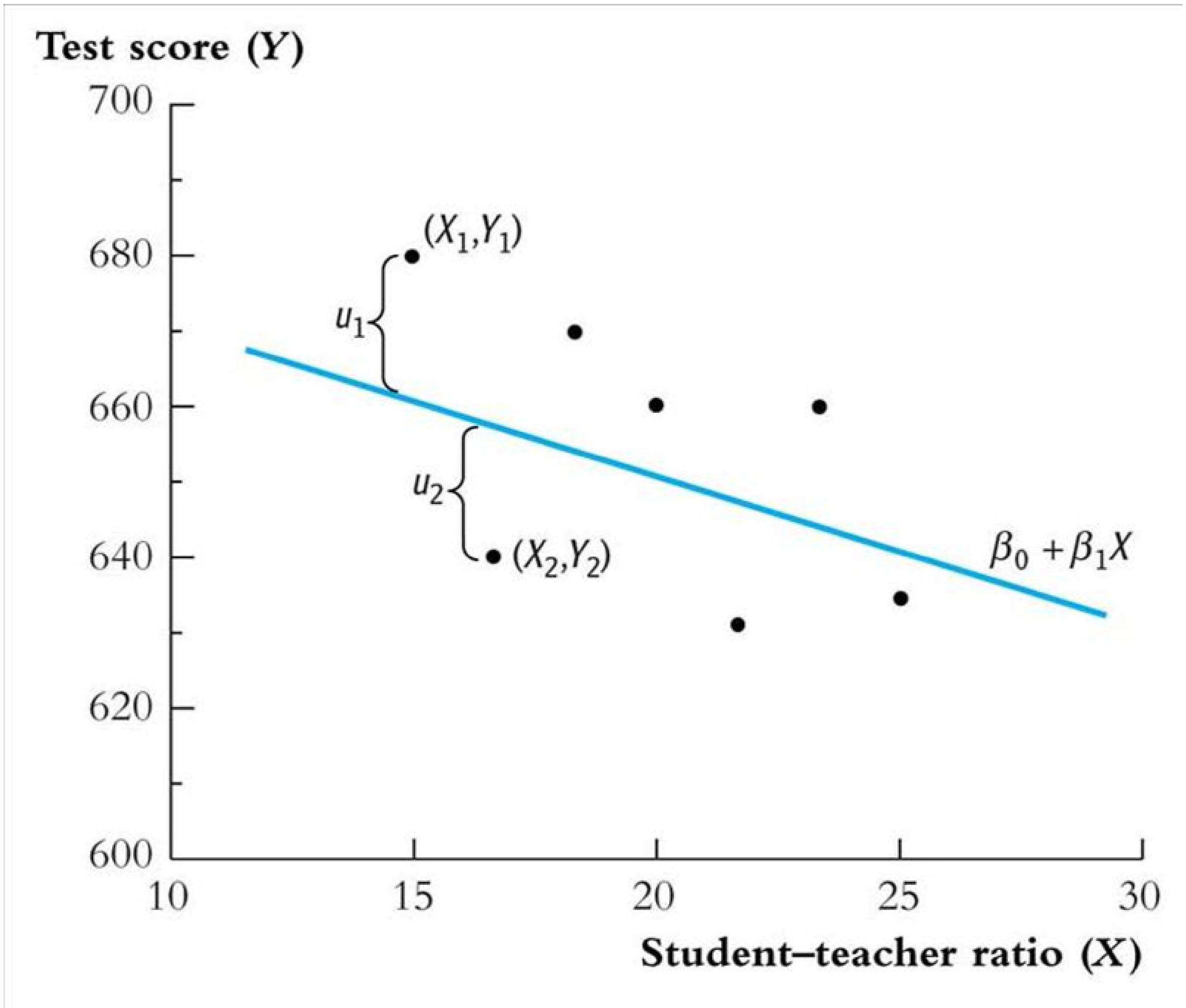


Figure 2.4: Drawing a straight line through data in a scatterplot

the distance of all points to that line is minimized. See that in Figure 2.4 the distances between the points and the line are given by the u_i's, the regression errors. So, if we somehow can minimize all u_i's we are fine. But those distances could be both positive and negative and they might cancel each other out. Therefore, we first square the regression errors and then minimize (hence the name: ordinary least *squares*). Also, see from Eq. 2.8 that:

$$u_i = Y_i - (\beta_0 + \beta_1 X_i) \longleftrightarrow (u_i)^2 = \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2 \quad (2.9)$$

But how can we estimate β_0 and β_1 from data? For that we will focus on the least squares (ordinary least squares or OLS) estimator of the unknown parameters β_0 and β_1, which solves,

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 X_i)\right]^2 \quad (2.10)$$

In fact, the OLS estimators of the slope β_1 and the intercept β_0 are:[3]

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \quad &(2.11) \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \quad &(2.12) \end{aligned}$$

Although you do **not** need to learn these formula's by heart some insightful comments can be retrieved from them. First, if a parameter is estimated then it gets a hat symbol on its top. Second, the optimal $\hat{\beta}_1$ is equal to $\frac{s_{XY}}{s_X^2}$ and this is the sampling covariance between X and Y divided by the sampling variance of X. This is not a correlation as the units still depend on X and Y and therefore the slope can be larger than 1 or smaller than -1, but it does say something about the relation between X and Y. Third, the constant is governed by the estimated parameter $\hat{\beta}_0$.

[3] This result is given but is not all too difficult to prove. However, usually you do need these types of equations in your work.

From here we can predict the values $\widehat{Y}_i$ and residuals $\hat{u}_i$ as they are:

$$\begin{aligned} \widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i, \qquad i = 1, \dots, n \\ \hat{u}_i &= Y_i - \widehat{Y}_i, \qquad i = 1, \dots, n \end{aligned} \tag{2.13}$$

When we apply this to our data cloud in Figure 2.1 then we get the following optimal population regression line:

```
sns.regplot(data = data, x = "STR" , y = "testscr", ci = None)
plt.text(x=20, y=700, s="$\widehat{TestScore} = 698.9 - 2.28 \\times STR$", fontsize=11, color
plt.xlabel("Student-teacher ratio")
plt.ylabel("Testscore")
plt.show()
```

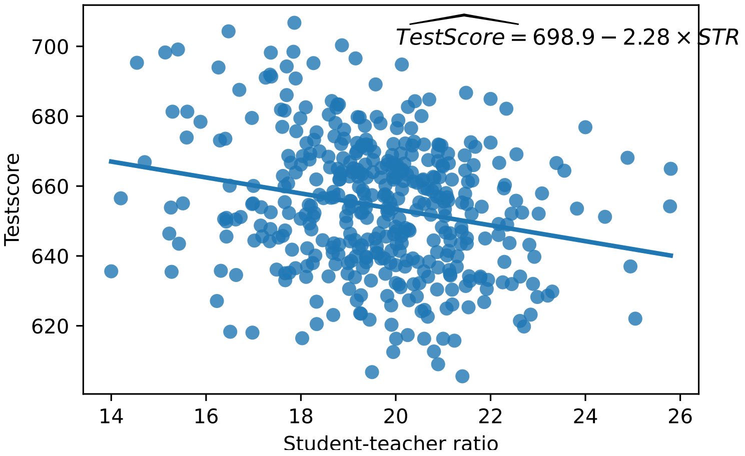


Figure 2.5: Scatterplot and estimated regression line

where the estimated slope equals $\widehat{\beta}_1 = -2.28$, the estimated intercept equals $\widehat{\beta}_0 = 698.9$ and the total population regression line can be written as: $\widehat{TestScore} = 698.9 - 2.28 \times STR$. So, how to interpret the estimated slope and intercept now? First, the slope entails that districts with one more student per teacher on average have test scores that are 2.28 points lower (that is, $\frac{\Delta TestScore}{\Delta STR} = -2.28$). Secondly, the intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. Now, this does not make any sense—it actually extrapolates the line outside the range of the data. In this case we can say that the intercept is not economically meaningful.

Now, how to fill in predictions? One of the districts in the data set is Antelope (CA) for which $STR = 19.33$ and $TestScore = 657.8$ Then the predicted value for the testscore is $\widehat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$ and the resulting residual is $\widehat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

In **python** both the constant and the slope can be easily retrieved by:

```
model_1 = smf.ols('testscr ~ STR', data=data).fit()

# Display the summary
print(model_1.summary())
```

Note as well how the formula is specified. First, the dependent variable, then a ~ sign and then the independent variable. Finally, we have to specify which dataframe to use; namely, **python** can have several dataframes in its memory.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     22.58
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           2.78e-06
Time:                        17:30:28   Log-Likelihood:                -1822.2
No. Observations:                 420   AIC:                             3648.
Df Residuals:                     418   BIC:                             3657.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    698.9329      9.467     73.825      0.000     680.323     717.543
STR           -2.2798      0.480     -4.751      0.000      -3.223      -1.337
==============================================================================
Omnibus:                        5.390   Durbin-Watson:                   0.129
Prob(Omnibus):                  0.068   Jarque-Bera (JB):                3.589
Skew:                          -0.012   Prob(JB):                        0.166
Kurtosis:                       2.548   Cond. No.                         207.
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

We will discuss some of the rest of this output later.

2.3.2.2 Hypothesis testing

We can assess the importance of the line as well with hypothesis testing. Again, recall that in in applied econometrics we will only **reject** the **null**-hypothesis, we do not accept an hypothesis based upon one statistical test only. So, we aim to test $H_0 : E(Y) = \mu_{Y,0}$ vs. $H_1 : E(Y) \neq \mu_{Y,0}$, where $\mu_{Y,0}$ is some pre-specified quantity that we are interested in. Typically $\mu_{Y,0} = 0$ as this denotes no relation, but sometimes you could be interested in, e.g., whether $\mu_{Y,0} = 1$ when testing elasticities. Or you could be interested in other quantities.

Testing statistical hypotheses is often very confusing, because of two things. First, you actually test whether the data you have corresponds with the null-hypothesis. Or, in other words:

> What is the probability that your data (D) might be right *given* the null-hypothesis (H_0): $\Pr(D|H_0)$

And that is a strange concept. You first imagine a world H_0 with the data that it *should* provide and then test that imaginary world.

Secondly, there is the notation that often works confusing. First, we have the p-value which equals the probability of drawing a statistic (e.g., $\bar{Y}$) *at least as adverse* to the null (that is: your imaginary world) as the value actually computed with your data, **assuming** again that the null-hypothesis is true—again, your imaginary world. Secondly, there is the significance level of a test which is a *pre-specified* probability of incorrectly rejecting the null, when the null is actually true.

Now, suppose that you want to calculate the p-value based on an estimated coefficient $\hat{\beta}_1$, then you construct the following test:

$$p\text{-value} = \Pr_{H_0}[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}| \tag{2.14}$$

where $\hat{\beta}_1^{act}$ is the value of $\hat{\beta}_1$ actually observed, and $\beta_{1,0}$ is the value of β_1 under the null-hypothesis (e.g., $\beta_{1,0} = 0$). Now, this is confusing, but in words it states that if you belief the null-hypothesis, what is then the *probability* that the estimated value is $\hat{\beta}_1^{act}$ or even more adverse to the value of the null hypothesis (in other words even more extreme values).

To test the null hypothesis H_0 we follow three steps. First, we need to compute the **standard error** of $\hat{\beta}_1$, which is an estimator of $\sigma_{\hat{\beta}_1}$. Using an ordinary least squares estimator, standard errors for coefficients are given by:

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 u_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}}, \tag{2.15}$$

which is a rather daunting expression.

Second, we need to compute the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sigma_{\hat{\beta}_1}} \tag{2.16}$$

Finally, we need to calculate the p-value. To do so, we need to know the sampling distribution of $\hat{\beta}_1$, which we know is complicated if n is small, but typically you have enough observations to invoke the *Central Limit Theorem.* So, if n is large, you can use the normal approximation (CLT) as follows

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}\left[\left|\hat{\beta}_1 - \beta_{1,0}\right| > \left|\hat{\beta}_1^{act} - \beta_{1,0}\right|\right] \\ &= \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{\sigma_{\hat{\beta}_1}}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{\sigma_{\hat{\beta}_1}}\right|\right] \\ &= \Pr_{H_0}[|t| > |t^{act}|] \\ &\simeq \text{probability under left + right } N(0,1) \text{ tails} \end{aligned} \tag{2.17}$$

where $\sigma_{\hat{\beta}_1}$ again equals the standard error of $\hat{\beta}_1$,

So, if you know $\hat{\beta}_1$ and $\sigma_{\hat{\beta}_1}$ you can calculate this. However, computers are much faster, in doing to. For example, suppose we want to test whether $\beta_{1,0} = 0$ using the regression output displayed above which gives $\hat{\beta}_1 = -2.28$ and $\sigma_{\hat{\beta}_1} = 0.52$. That is step 1. Note that `R` already calculated the standard error of Eq. 2.15. For the next step we need to compute the t-statistic, which is:

$$t^{act} = \frac{2.28 - \beta_{1,0}}{0.52} = \frac{2.28 - 0}{0.52} = -4.39. \qquad (2.18)$$

then the p-value can be seen from Figure 2.6:

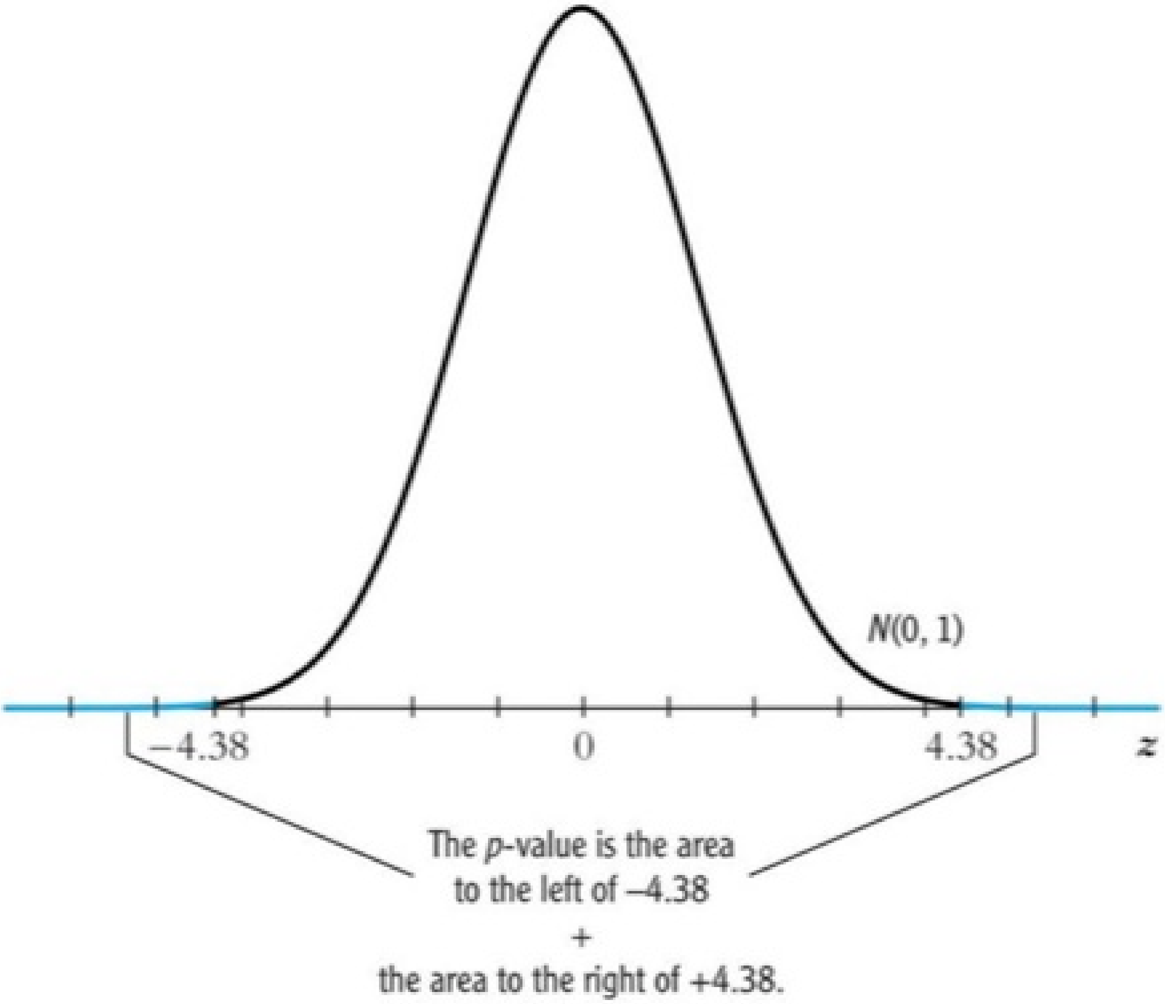


Figure 2.6: Calculating the p-value of a two-sided test when $t^{act} = -4.38$

That is, for large n (and typically we have that), the p-value is the probability that a $N(0,1)$ random variable falls outside $|\hat{\beta}_1^{act} - \beta_{1,0})/\sigma_{\hat{\beta}_1}| = |t|$. That is the blue areas under the normal distribution and they entail a probability *mass*. Now, if both surfaces on the sides are not larger than 2.5%, then we can reject the null-hypothesis against a 5% significance level. Now, the computer output above gives a p-value of 0.000, which is a bit strange. The p-value is actually not zero, but a very small number and definitely smaller than 0.05, so we can *reject* the null-hypothesis being $\beta_{1,0} = 0$ at a 5% significance level (and at a 1% and 0.1% significance level as well). Now, if the t-statistic is exactly 1.96 in absolute value, then the p-value is 0.05. So, to repeat the steps, but now using computer output for testing the hypothesis that $\beta_1 = \beta_{1,0}$

1. Get the standard error from computer output
2. Compute the t^{act}-statistics as in Eq. 2.16
3. Get the corresponding p-value. Or, reject the null-hypothesis at the 5% significance level if $|t^{act}| > 1.96$.

Now there is a link between the p-value and the significance level. The significance level is pre-specified. For example, if the pre-specified significance level is 5%, then you reject the null hypothesis if $|t| \geq 1.96$ or equivalently, you reject if $p \leq 0.05$. The p-value is sometimes called the marginal significance level. Often, it is better to communicate the p-value than simply whether a test rejects or not—the p-value contains more information than the "yes/no" statement about whether the test rejects.

But recently there has been some debate about using p-values (Amrhein, Greenland, and McShane 2019). Why should you use a 5% significance level, what is so special about that number? Is it not better just to report coefficients and standard errors? Figure 2.7 shows a figure from the journal Nature and how scientists across all fields nowadays see p-values and statistical significance. This is not to say that statistical testing does not matter, but more the reporting of that statistical testing. First of all, p-values in themselves do not contain that much information. In the regression output of above the reported p-values are being equal to 0.000 which is not informative. Secondly, the cut-off point of 5% is a bit harsh and could lead to publications being published only with p-values just below 0.05, leading to what is called publication bias.

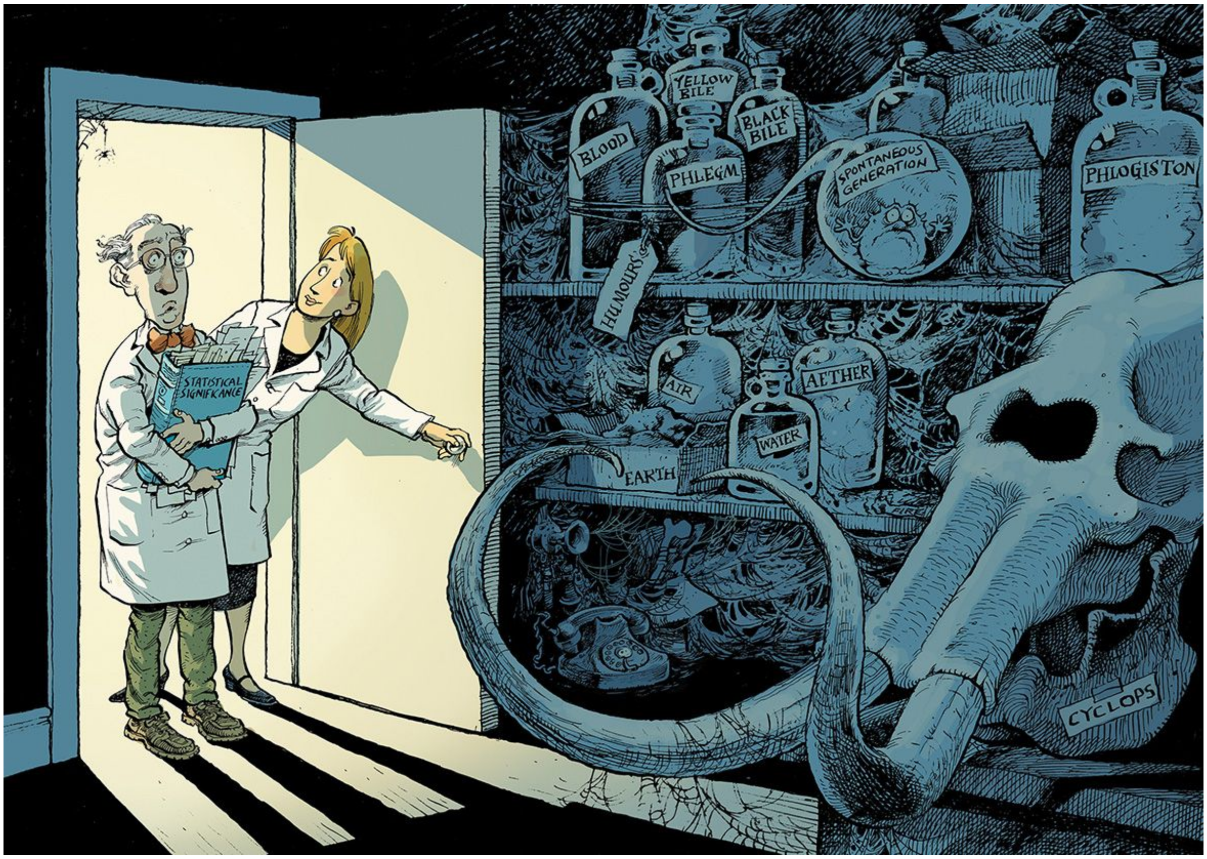


Figure 2.7: Critical review on the (mis)use of statistical significance

2.3.2.3 Confidence intervals

The exact definition of confidence intervals is a bit tricky. Namely, a 95% confidence interval for $\widehat{\beta}_1$ is an interval that contains the true value of β_1 in 95% of repeated samples. That means that a confidence interval does not give a probability (even though we would like to interpret it that way). But you can state that every value within a confidence interval would not be rejected as null-hypothesis, while every value outside the confidence interval would be rejected. Now, if we know both $\hat{\beta}_1$ and $\sigma_{\hat{\beta}_1}$ (again using computer output), then a 95% confidence interval can be very easily constructed. For our regression of output of above this entails

$$\hat{\beta}_1 \pm 1.96 \times \sigma_{\hat{\beta}_1} = -2.28 \pm 1.96 \times 0.52 = [-3.30, -1.26]. \tag{2.19}$$

So, every value between -3.30 and -1.26 will **not** be rejected as null-hypothesis, while every value outside that interval will be rejected. Note that confidence intervals are again automatically given by computer output. If one would like a confidence interval against another critical level, say against a 99% critical level, one can use the `conf_int()` function and specify as:

```
# Fit the model
model_1 = smf.ols('testscr ~ STR', data=data).fit()

# Get 99% confidence intervals
conf_int_99 = model_1.conf_int(alpha=0.01)  # 1 - 0.99 = 0.01

print(conf_int_99)
```

```
                    0           1
Intercept  674.434471  723.431428
STR         -3.521425   -1.038191
```

2.3.2.4 Regression with a dummy

Sometimes a regressor is binary, meaning an indicator or a dichotomous (0/1) variable. Let's go back again to Section 2.2.2, where we created such a binary variable with small and large class sizes ($X = 1$ if class size is small, $X = 0$ if not). Other possible examples are gender ($X = 1$ if female, $X = 0$ if male) or being treated or not ($X = 1$ if treated, $X = 0$ if not). We refer to these types of variables as being **dummy** variables—and they are very often used in the social sciences.

Now, suppose we have a population regression model that looks like:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2.20}$$

Where Y denotes, e.g., test scores, and where X, e.g., denotes a dummy variable for a large class (so, if $STR \geq 20$ then $X_i = 1$; otherwise $X_i = 0$), so there is then only two possibilities:

1. For small class size there should hold that $X_i = 0$ yielding that $Y_i = \beta_0 + u_i$. Namely $\beta_1 \times X_i = \beta_1 \times 0 = 0$. That means automatically that the expectation of Y_i is the constant, being β_0. Another way or writing is that the expectation of model 2.20 *conditional* on the fact that $X_i = 0$ is $\mathbb{E}(Y_i \mid X_i = 0) = \beta_0$.
2. For large classes there should hold that $X_i = 1$ yielding that $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$. This means that the expectation of model 2.20 *conditional* on the fact that $X_i = 1$ is $\mathbb{E}(Y_i \mid X_i = 1) = \beta_0 + \beta_1$

So a regression with a dummy as independent variable gives two different *constants*, for each group (small/large classes) one. You can interpret this as a level-effect (only the level changes, not the slope as there is none here). The interpretation of β_1 is in this case rather special and can be denoted as:

$$\beta_1 = \mathbb{E}(Y_i \mid X_i = 1) - \mathbb{E}(Y_i \mid X_i = 0), \tag{2.21}$$

Which is just the population difference in group means.

If we go back to our example with $X_i = 1$ if $STR \geq 20$ and 0 otherwise then we get the following regression output

```
model_2 = smf.ols('testscr ~ large', data=data).fit()
print(model_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.035
Model:                            OLS   Adj. R-squared:                  0.032
Method:                 Least Squares   F-statistic:                     15.07
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           0.000120
Time:                        17:30:28   Log-Likelihood:                -1825.9
No. Observations:                 420   AIC:                             3656.
```

```
Df Residuals:                     418   BIC:                             3664.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       657.2462      1.212    542.162      0.000     654.863     659.629
large[T.True]    -7.1694      1.847     -3.882      0.000     -10.799      -3.540
==============================================================================
Omnibus:                        3.120   Durbin-Watson:                   0.107
Prob(Omnibus):                  0.210   Jarque-Bera (JB):                2.483
Skew:                           0.052   Prob(JB):                        0.289
Kurtosis:                       2.638   Cond. No.                         2.49
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Now, note that this is the same output ($\Delta = -7.2$, $\sigma_\Delta = 1.85$ and t-statistic is -3.88) as when we did the difference in means test in Section 2.2.3 (using the `R` way). To conclude, this is just another way (and much easier) to do a difference-in-means analysis. And this directly carries over for the situation when we have additional regressors.

They are not exactly the same though and that is that `R` for the t-test in Section 2.2.3 corrects for the degrees of freedom, or the number of variables it uses. In large sample they coincide though.

2.4 Least squares assumptions for causal inference

As stated at the start of Section 2.2 applied econometrics focuses on finding a **causal** effect. But how do you do know that the $\hat{\beta}_1$ you estimate using the population regression model of Eq. 2.22 is indeed a causal effect. In other words, if you change X_i with one unit, will Y_i then change with β_1 **in reality**?

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1 \ldots n \tag{2.22}$$

Fortunately, there is a small set of assumptions that indeed lead to such a causal interpretation. The so-called three least squares assumptions, being:

1. The conditional distribution of u given X has mean zero, that is, $E(u \mid X = x) = 0$.
    - We also refer to this assumption as the **conditional mean independence** assumption
    - This assumption implies that $\hat{\beta}_1$ is truly *unbiased*
2. $(X_i, Y_i), i = 1 \dots n$ are i.i.d.
    - This is true if X, Y are collected by simple random sampling
    - This delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$—again with a relatively large number (say $n > 50$) the sampling distribution can very well be approximated by a normal distribution
3. Large outliers in X and/or Y are rare.
    - Outliers can result in meaningless values of $\hat{\beta}_1$

We will first discuss these three least squares assumptions and then give some other assumptions (but not directly necessary for the identification of causal effects) as well as you frequently encounter them

2.4.1 Least squares assumption 1: conditional mean independence

This first assumption states that $E(u \mid X = x) = 0$ and is conceptually the most difficult one to grasp. Loosely speaking, it states that the regression error u is *not* related with the independent variable X. They are independent of other. Another way of looking at this is displayed in Figure 2.8. Here, whatever the value of student-teacher ratio is, the expectation of the outcome variable (test scores) is always centered around the population regression line. So, on average, you always predict correctly according to *your* model, for each value of X.

So, when is this assumption violated? For example, consider again the population regression model: $TestScore_i = \beta_0 + \beta_1 STR_i + u_i$, where u_i denotes other factors. Now, these other factors can be everything and else. And you should ask yourself

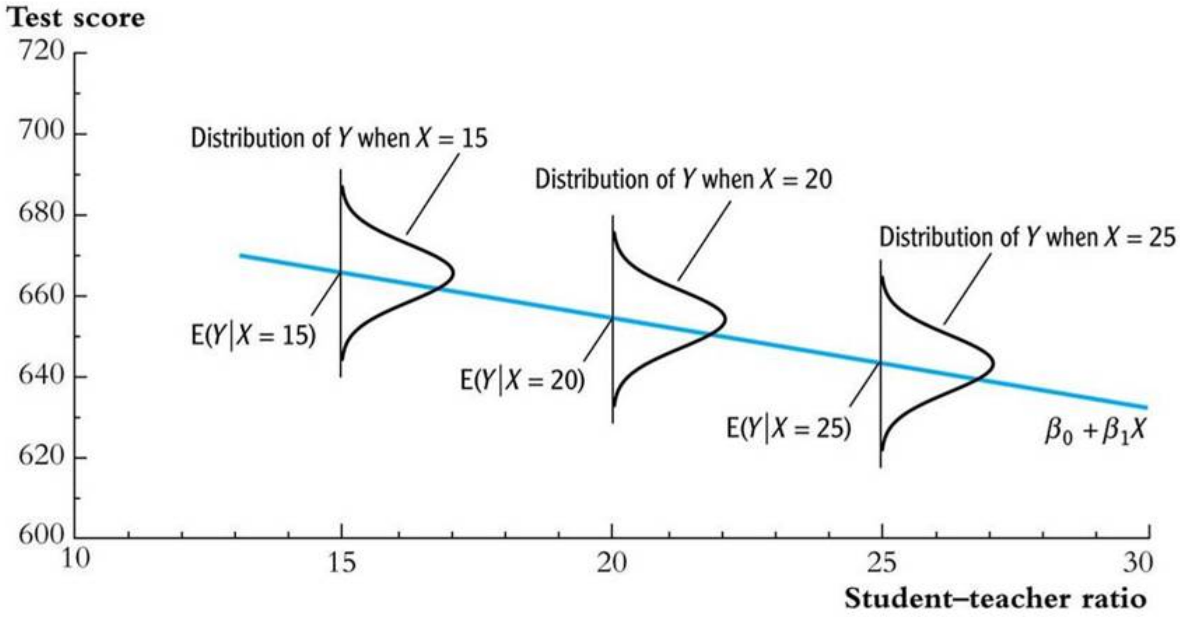


Figure 2.8: Condition mean independence assumption

whether it is plausible that $E(u|X = x) = 0$ for **all** these other factors?

This assumption lies as well at the heart of experimental settings. Namely, consider a theoretical ideal randomized controlled experiment, where:

1. X is *randomly* assigned to people (students randomly assigned to different size classes or patients randomly assigned to medical treatments).
2. Because X is assigned randomly, all other individual characteristics—the things that make up u—are *independently* distributed of X by definition.
3. Thus it automatically follows that: $E(u \mid X = x) = 0$

Now, both in actual experiments, or with **observational** data, we will need to think hard about whether $E(u|X = x) = 0$ holds. Chapter 3 and Chapter 4 provide various examples where this assumption is violated. However, if this assumption is violated it means that you have a **biased** inference, which boils down to the fact that your estimated $\widehat{\beta}_1$ is not the one that you want and that correct inference based upon this estimate cannot be done.

2.4.2 Least squares assumption 2: independenty and identically distributed

The second least squares assumptions deals with actual sampling of your data, both the dependent (Y) and independent (X) variables. That entails that $(X_i, Y_i), i = 1 \dots n$ should be *i.i.d.*. This assumptions arises automatically if the entity (individual, district) is sampled by simple *random sampling*. There are quite some possible violations to this assumption. For example, you sample via your friends on social media (snowballing), or observations are not independent but are correlated, which arises very frequently in the context of temporal correlation or spatial correlation.

The consequence of violating the *i.i.d.* assumption is less severe then violating the conditional mean independence assumption. It leads to wrong *standard errors*, not to biased estimations.

2.4.3 Least squares assumption 3: Large outliers are rare

The third and final least square assumption for causal inference is that large outliers are rare. Large outliers are not well defined and depend on the size of both Y and X, but in general it can be seen as an *extreme* value of X or Y. The problem is that such a large outlier can strongly *influence* the results and in general it can be stated that OLS can be rather sensitive to an outlier. Consider the two population regression lines in Figure 2.9. The flat one (with $\hat{\beta}_1 = 0$) does not take the isolated observation in the upper right corner into account. The one with the positive slope does. Now, clearly the one isolated observation in the upper right corner matters to a large extent and is an important driver for the results of the ordinary least squares estimator.

However, this does not automatically entail that the isolated observation should be deleted. What it does entail is that one should go back to her data and investigate whether the outlier

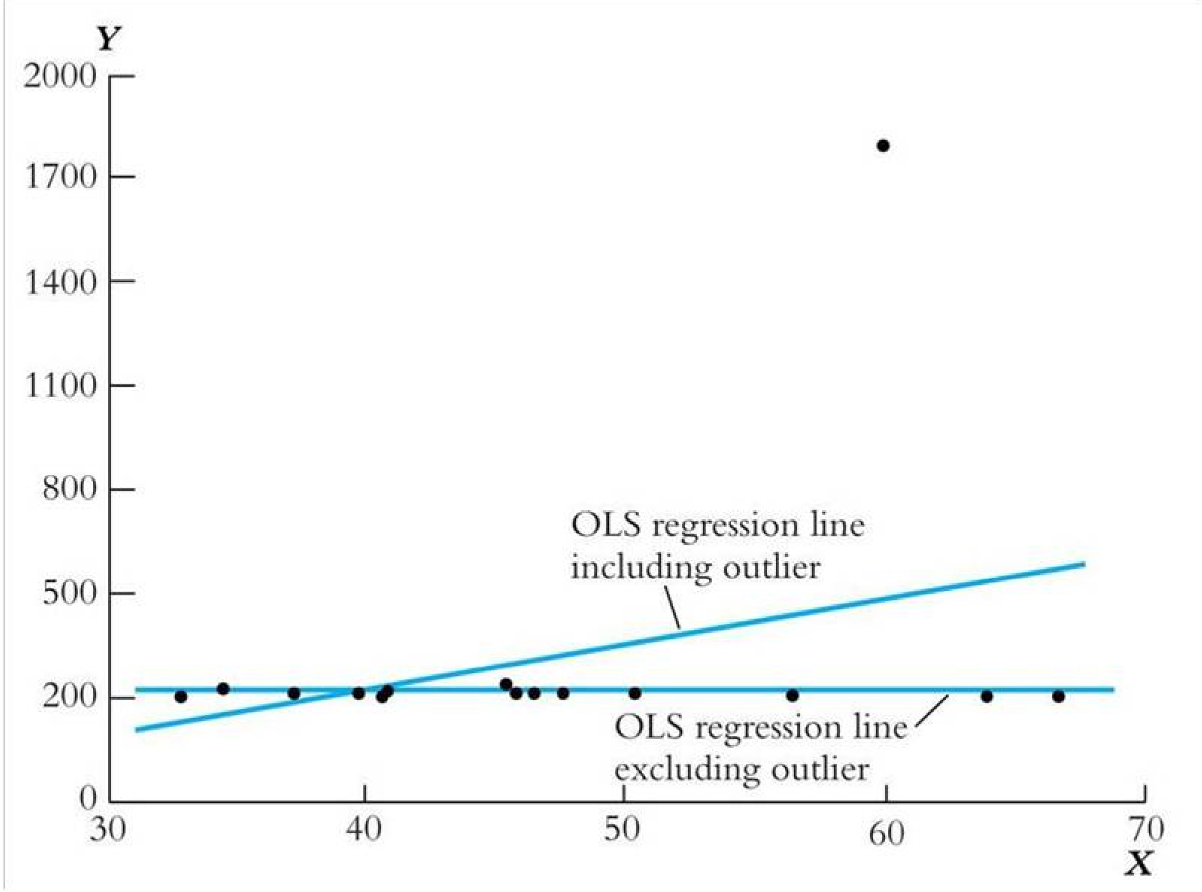


Figure 2.9: Effect of outliers on OLS estimations

could be a mistake—perhaps a typo made when preparing the data or something that went amiss when converting the data from an `Excel` format to a `Python` format.

2.5 Other least squares assumptions

Oftentimes, two other least squares assumptions are frequently encountered. However, keep in mind that you do *not* need them for causal inference. They are the assumptions of homoskedasticity and normality.

2.5.1 Homoskedasticity

Homoskedasticity is concerned with the standard errors. Its definition is if $var(u \mid X = x)$ is constant—that is, if the variance of the conditional distribution of u given X does not depend on X—then u is said to be homoskedastic. Otherwise, u is heteroskedastic.

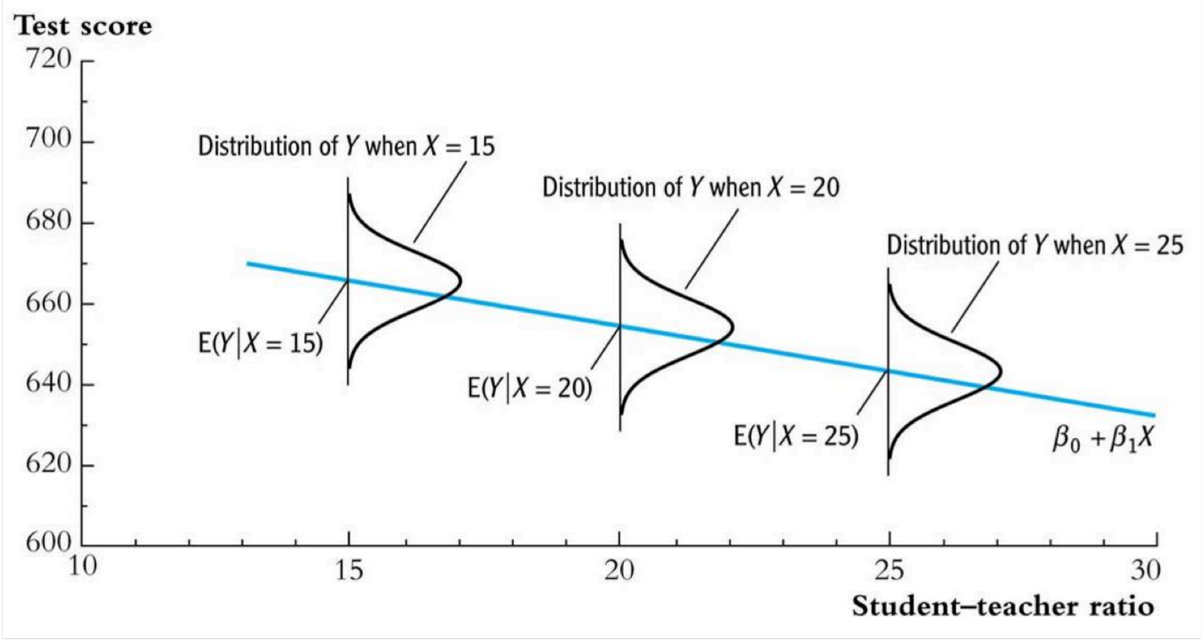


Figure 2.10: Homoskedastic standard errors

Consider Figure 2.10. Clearly the variance *around* the population regression line is everywhere the same, regardless the value of student-teacher ratio (X). Recall, that $E(u \mid X = x) = 0$ so u satisfies Least Squares Assumption 1. Now, in addition we also assume that the variance of u does not depend on x. This is the case of homoskedasticity

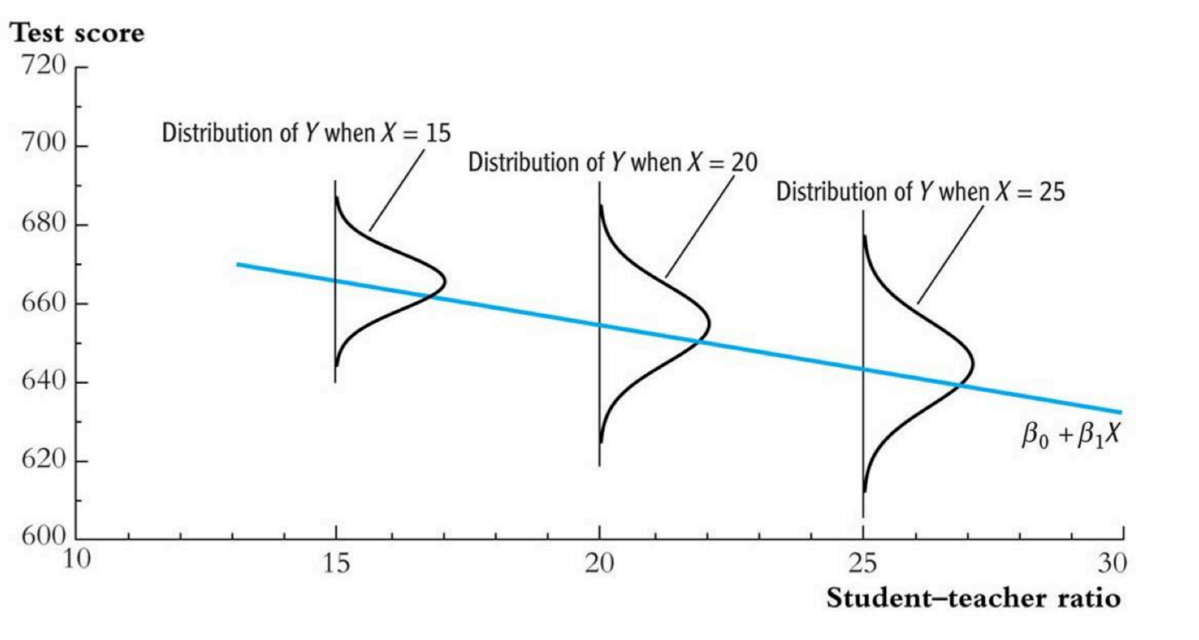


Figure 2.11: Heteroskedastic standard errors

Now consider Figure 2.11. Now clearly the variance around the population regression line increases in size of student-teacher ratio (X). So, $E(u \mid X = x) = 0$ is still satisfied, but the variance of u does now depend on x. u is now said to be heteroskedastic.

Very often data in the social sciences are heteroskedastic. For example, wages are usually heteroskedastic in the amount of education consumed. Figure 2.12 shows the relation between years of education and wages, and the larger the years of education the larger hourly earnings (wages) are—as you would assume it should be. But the variance also increases in years of education. That is because you can easily predict wages when workers enjoyed very few years of schooling—usually those are

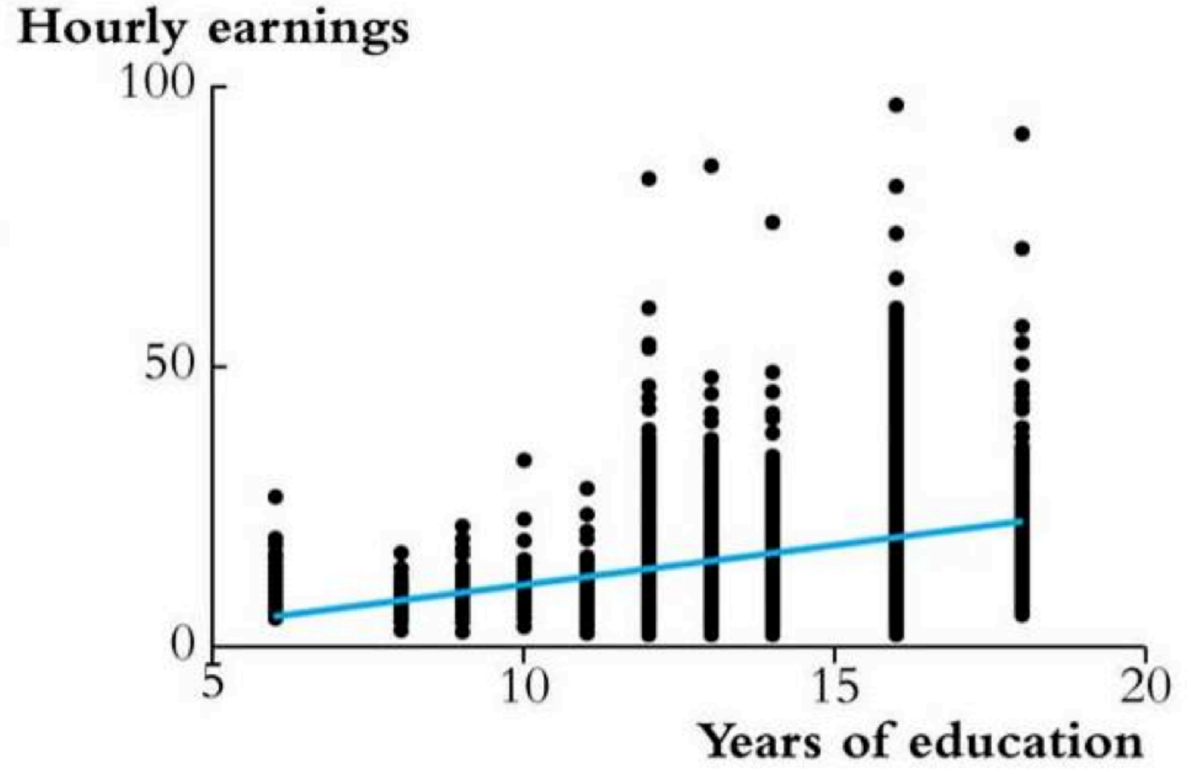


Figure 2.12: Wages versus years of education

just above minimum wages—but the spread becomes much wider when years of schooling go up.

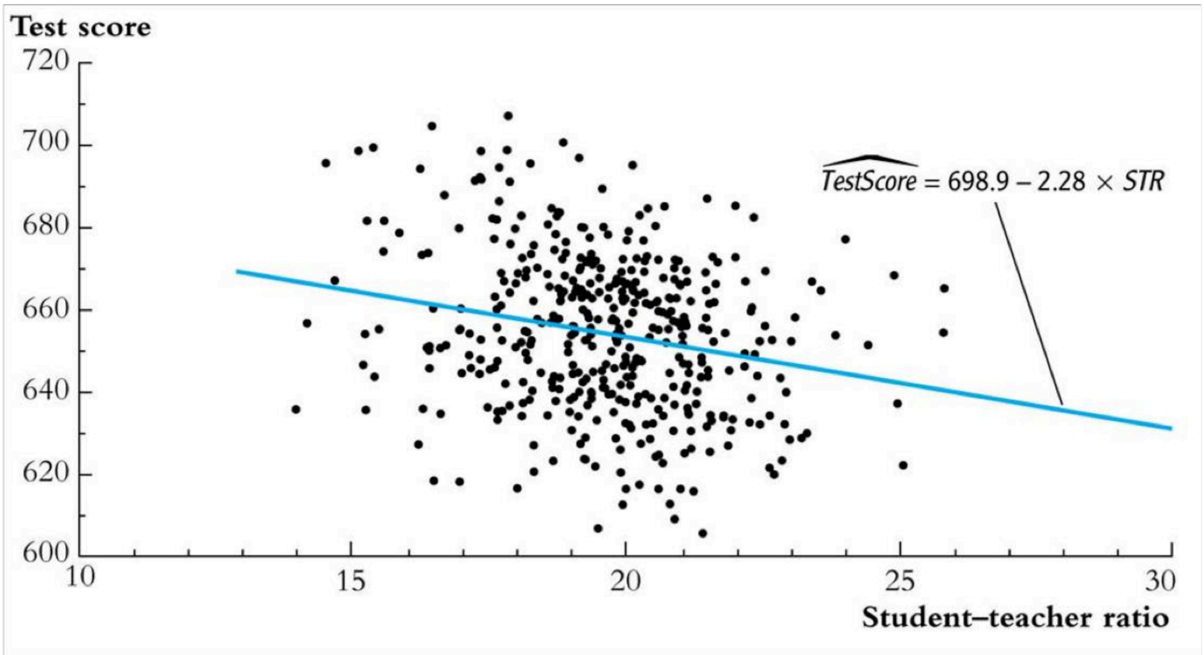


Figure 2.13: Heteroskedasticity in Californation schools?

Is this now the case for our Californian school dataset. If we look again at the scatterplot between test scores and student-teacher ratios in Figure 2.13, then that is very difficult to see. But then again, does it matter whether you face heteroskedasticity or homoskedasticity.?

Note that so far we have (without saying so) assumed that u might be heteroskedastic Recall again the three least squares assumptions:

1. $E(u \mid X = x) = 0$
2. $(X_i, Y_i), i = 1, \ldots, n$, are i.i.d.
3. Large outliers are rare

They do not say anything about homo- or heteroskedasticity and because we have not explicitly assumed homoskedastic errors, we have implicitly allowed for heteroskedasticity.

But what if the errors are in fact homoskedastic? Then in fact you can prove that OLS has the lowest variance among estimators that are *linear* in Y. The formula for the variance of $\hat{\beta}_1$ and the OLS standard error simplifies: If $var(u_i \mid X_i = x) = \sigma_u^2$, then

$$var(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_X^2} \tag{2.23}$$

which is much simpler than Eq. 2.15. Again note that $var(\hat{\beta}_1)$ is inversely proportional to $var(X)$: more spread in X means more information about $\hat{\beta}_1$—we discussed this earlier but it is clearer from this formula.

But what does this mean for estimation. Note that `Python` does not automatically apply Eq. 2.15 for its standard errors, but uses the simpler version Eq. 2.23 instead. But if we invoke the `cov_type = "HC0` command, `python` computes heteroskedasticity-robust standard errors. So if you do not, `python` computes homoskedasticity-only standard errors. So:

```
model_2 = smf.ols('testscr ~ large', data=data).fit(cov_type = "HC0")
print(model_2.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.035
Model:                            OLS   Adj. R-squared:                  0.032
Method:                 Least Squares   F-statistic:                     15.50
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           9.69e-05
Time:                        17:30:28   Log-Likelihood:                -1825.9
No. Observations:                 420   AIC:                             3656.
Df Residuals:                     418   BIC:                             3664.
Df Model:                           1                                         
Covariance Type:                  HC0                                         
=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       657.2462      1.251    525.246      0.000     654.794     659.699
large[T.True]    -7.1694      1.821     -3.936      0.000     -10.739      -3.600
==============================================================================
Omnibus:                        3.120   Durbin-Watson:                   0.107
```

```
Prob(Omnibus):                  0.210   Jarque-Bera (JB):                2.483
Skew:                           0.052   Prob(JB):                        0.289
Kurtosis:                       2.638   Cond. No.                         2.49
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

where the standard error of `large` is now 1.821.

The bottom line is that the errors are either homoskedastic or heteroskedastic and if you use heteroskedastic-robust standard errors, you are fine. Namely:

1. If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity).
2. The two formulas coincide (when n is large) in the special case of homoskedasticity.
3. So, you should **always** use heteroskedasticity-robust standard errors.

2.5.2 Normal distributed regression term

Finally, in many introductionary statistic courses, normal distributed error terms are assumed, which facilitates testing with small samples. So, u should be distributed $N(0, \sigma^2)$. If you have a reasonable amount of observations ($n > 50$), you do not need this assumption, and especially not for causal inference.

2.6 Measures of fit

A natural question that might arise is how well the population regression line fits or explains the data. For ordinary least squares estimators often two regression statistics are given that provide complementary measures of the quality of fit:

1. The regression R^2: This measures the fraction of the variance of Y that is explained by X; it is unitless and ranges between zero (no fit) and one (perfect fit). This one is almost always reported.
2. The standard error of the regression SER: This measures the magnitude of a typical regression residual in the units of Y.

2.6.1 The regression R^2

The regression R^2 is the fraction of the sample variance of Y_i "explained" by the regression. To see this, first note that $Y_i = \widehat{Y}_i + \widehat{u}_i$ or the observation is equal to the OLS prediction plus the predicted residual. In this notation, the R^2 is the ratio between the sample variance of $\widehat{Y}$ and the sample variance of Y. Here we make use of the following equity: Total sum of squares = explained "SS" + Residual "SS"—or, $TSS = ESS + RSS$—, where we can now define R^2 as:

$$\mathrm{R}^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}. \tag{2.24}$$

Now if R$^2 = 0$ then that means $ESS = 0$ and if R$^2 = 1$ then that means $ESS = TSS$. So, by definition yields $0 \leq \mathrm{R}^2 \leq 1$. There is one additional remark to make and that is that for an univariate regression model (so with one single X on the right side), R^2 equals the square of the correlation coefficient between X and Y.

2.6.2 The Standard Error of the Regression

The standard error of the regression is defined as:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2} \tag{2.25}$$

In comparison with the R^2, the SER is measured in the units of u, which are actually the units of Y. It measures the average "size" of the OLS residual (so the average 'mistake' made by the OLS regression line in absolute terms).

However, more often the *root mean squared error* ($RMSE$) is used, which is very closely related to the SER:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}, \tag{2.26}$$

where $RMSE$ only differs from the SER in the *degrees of freedom.*

If we again look at our regression output:

```
model_1 = smf.ols('testscr ~ STR', data=data).fit(cov_type = "HC0")
print(model_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     19.35
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           1.38e-05
Time:                        17:30:28   Log-Likelihood:                -1822.2
No. Observations:                 420   AIC:                             3648.
Df Residuals:                     418   BIC:                             3657.
Df Model:                           1
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    698.9329     10.340     67.597      0.000     678.668     719.198
STR           -2.2798      0.518     -4.399      0.000      -3.296      -1.264
==============================================================================
Omnibus:                        5.390   Durbin-Watson:                   0.129
Prob(Omnibus):                  0.068   Jarque-Bera (JB):                3.589
Skew:                          -0.012   Prob(JB):                        0.166
Kurtosis:                       2.548   Cond. No.                         207.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

then we see that the $R^2 = .05$. So, only 5% of all variation in test scores is explained. This of course makes sense as potential many important variables are not included in the model. However, this does not automatically mean that the impact is biased and especially that the STR is unimportant in a policy sense. Again, we focus on causal inference, not on making a good model for prediction.

```
np.sqrt(model_1.scale)
```

```
np.float64(18.580966694033535)
```

The $SER = 18.6$ (which is unfortunately not directly provided by `python` but is actually the square root of the `scale` attribute of the regression model object) indicates that the average error made is 18.6 test score units, which can be seen as sizable. In Chapter 3 we include other, and important, variables and what we then of course will see is that the R^2 increases and the SER decreases.

2.7 Conclusion

Regression analysis in the most common form of statistical analysis over the sciences. Very often is it used to model associations. However, in applied econometrics regression analysis is applied to find causal causal relations. This can be done on the basis of three assumptions, of which the first—the conditional mean assumption—is the most crucial. In fact, if this assumption holds then the data mimics the data that come out of an experiment. This assumption can unfortunately not be proven. We therefore have to think very hard whether this assumption holds. The next Chapter deals with a possible violation of this assumption, how to solve for it and at the same time we answer the question why we want multiple variables in our specification.

3 Modeling in the Social Sciences

```
# Python packages we need for this chapter
import numpy as np
import pandas as pd
import seaborn as sns
from scipy import stats
import statsmodels.api as sm
# the next package allows us to formula's for ols
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
```

In Chapter 2 we discussed the origins of, working of, and assumptions behind univariate regression. That is, a regression model with only one independent variable X on the right hand side.[1] However, and especially in the social sciences, you almost always see regressions with many independent variables. Depending on the field, these variables can be called control variables, confounding factors, mediator or moderator variables.[2] But why are these variables included? Is it only to improve model performance or are there other reasons? Section 3.1 deals with this question whereafter Section 3.2 shows how you can include additional variables in a *multivariate regression model* and especially how you should interpret them. Section 3.3 extends the multivariate regression model and shows how you can actually use this model to estimate a broad range of linear and non-linear economic models. Section 3.4 discusses the use of multiple dummy variables (see again Section 2.3.2.4) in a

[1] With right hand side we mean on the right side of the equal sign =. It is often abbreviated with RHS.

[2] There is something more to the use of these names of variables and their exact working but we leave that for another course.

way that economists refer to as *fixed effects*. The last section concludes and provides a further discussion of the benefits and limitations of multivariate regression models.

3.1 Why more independent variables?

So, why do we include more variables? One possible answer is because it makes a better predictive model. That is, a model that is able to explain the variation in the dependent variable Y better.[3] So, the R^2 increases. But, as argued in Chapter 2 we are not so much interested in prediction, but more in establishing a **causal** relation between X and Y. So, if you change X (and only X) does Y change and then with how much?

Although economists often claim that they are the only (social-)science that focuses on causality and provides a statistical framework for that, there are other approaches to causality as well. One that is often used in other sciences is the approach of the mathematican Judea Pearl (Pearl 2009). This approach focuses on the use of Directed Acyclical Graphs (DAGs), which is a graphical visualisation of causality chains (or, what impacts what). We borrow this approach for the most simple setting as explained in Figure 3.1. Here, we go back to our Californian school district dataset again, where we still are interested in the effect of class size on school performance. So, we suppose that there is an effect from student teacher ratio on test scores as displayed with an directed arrow in Figure 3.1. We also know that the R^2 of that regression model was rather low (5%), so by default there must be other but yet unknown factors, let us name them for now U (often as well referred to as unobservables), that influence test scores as well (so a directed arrow going from U to test scores).

Now we are fine with this is as long as U does **not impact** the student teacher ratio. Then, there is still an isolated effect of student teacher ratio on class size and that is exactly what we

[3] This is not entirely true. Increasing the R^2 explains **in-sample** variation better, not necessarily **out-of-sample**. The latter is really what matters for prediction and this is the focus of many machine learning techniques. Note that this argument is directly related with the regression towards the mean argument made in Section 2.3.1.

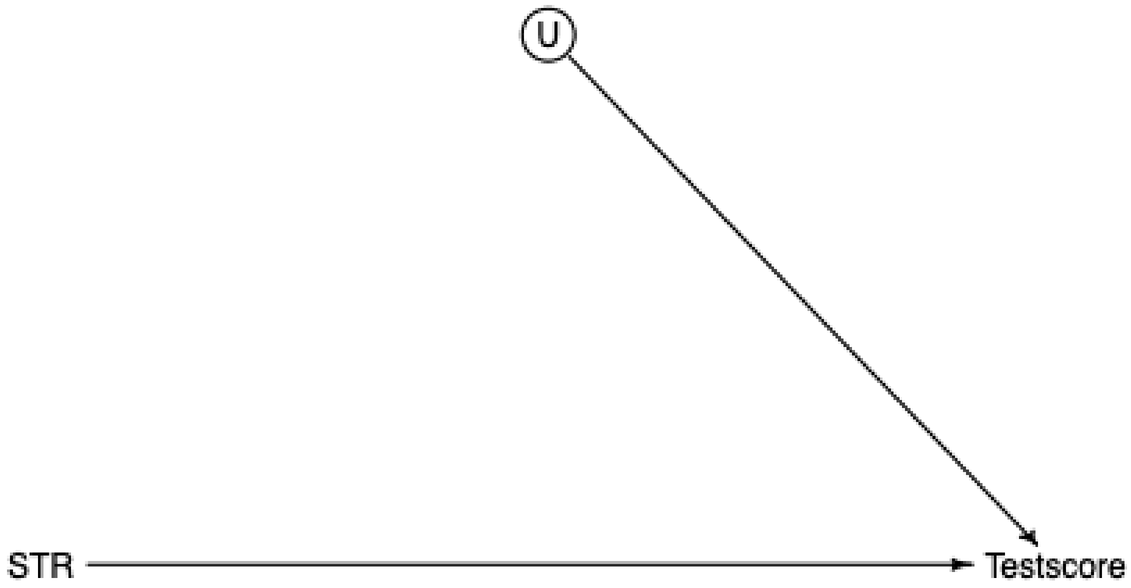


Figure 3.1: Unrelated omitted variables

want to measure. However, if there is a directed arrow going from U into STR as depicted by Figure 3.2, then the effect of student teacher ratio is not isolated anymore. Essentially, the effect of student teacher ratio on class size is composed out of two parts:

1) The **causal** effect on student teacher ratio on class size captured by the chain STR $\longrightarrow$ testscore. The one we are after.
2) The impact of the unknown variables on test scores. As we have not modeled them in our regression model, the effect is captured by the chain $U \longrightarrow$ STR $\longrightarrow$ testscore

Economists refer to this phenomenon as **omitted variable bias**, whilst in the statistical world, this is as often called confounding variables or the **confounding fork** (McElreath 2020) and it, unfortunately, occurs very often.

So, when **U** is a *common* cause for both student teacher ratio and test scores there is omitted variable bias. If we go back to our population regression model as follows:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{3.1}$$

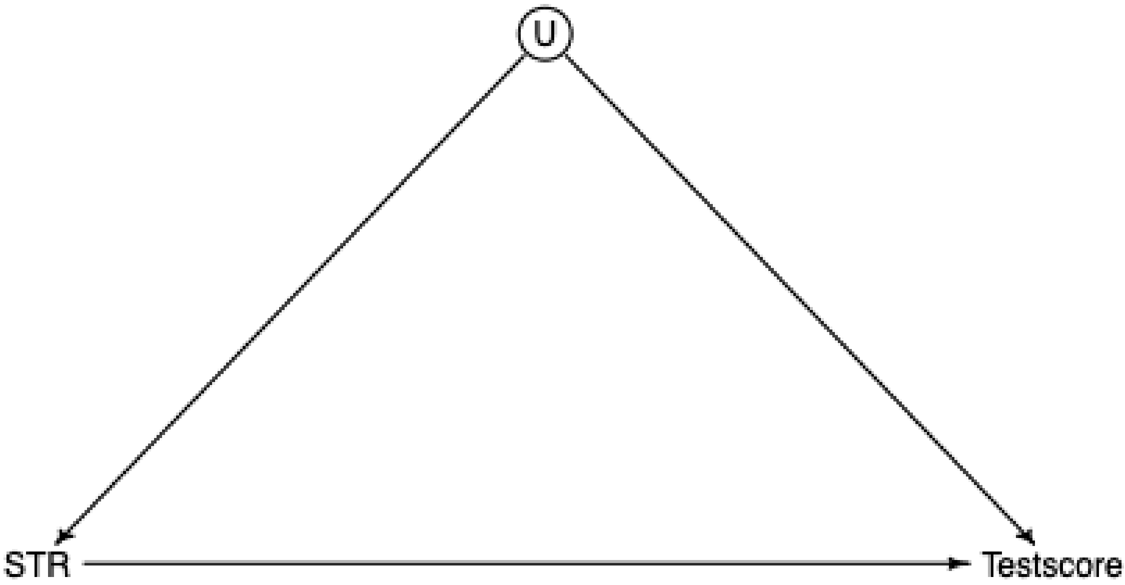


Figure 3.2: Related omitted variables

then we know that the error u arises because of factors that influence Y but are not included in the regression function; so, there are *always* omitted variables. But they do not always lead to bias. For omitted variable bias to occur, the omitted factor, let's call it Z[4], must be:

1. A **determinant** of Y (i.e. Z is part of u)
2. A **determinant** of the regressor X (*at least*, there should hold that $corr(Z, X) \neq 0$)[5]

Thus, both conditions must hold for the omission of Z to result in omitted variable bias.

Now, in our Californian district school dataset we have many more variables. One of them is a variable that measures the

[4] Z can be both known or unknown, so that is why we change from U to Z

[5] In econometric textbooks, as, e.g, in Stock, Watson, et al. (2003), this condition is weakened to only being correlation (Z and X are correlated). However, if the directed arrow goes from STR into U in Figure 3.2 then that would lead to something else than omitted variables, namely to a difference between a direct (STR $\longrightarrow$ testscore) and an indirect effect (STR $\longrightarrow U \longrightarrow$ testscore).

english language ability (whether the student has English as a second language). Note that in California there are many migrants, especially from Latin-America. Now, you can readily argue that not having English as first language plausibly affects standardized test scores: so, Z is a **determinant** of Y. Moreover, immigrant communities tend to be less affluent and thus have smaller school budgets—and, therefore, higher STR: Z is most likely as well a **determinant** of X.

So, most likely, our original estimation from Chapter 2, $\hat{\beta}_1$, is biased (so, it is not the true causal effect). But can we say something about the direction of that bias? Yes, but the argument tends to become very quickly rather complex. In this case, note that districts with more migrant communities tend to have (*i*) higher class sizes and (*ii*) lower test scores. So, to the original estimation they add a *negative* effect. Thus, following this reasoning, the "true" effect must be less negative. Now, especially with negative signs this reasoning becomes rather complex, so if common sense fails you, then there is the following formula:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\sigma_u}{\sigma_X}\rho_{Xu}, \tag{3.2}$$

where you should focus on the sign of the correlation between X and the regression residual u (all standard errors, σ, are always positive by default). Now, the first least squares assumption states that $\rho_{Xu} = 0$—no correlation between the regressor and the regression residual. But now there is correlation because of omitted variable bias. And because there is a negative relation between immigrants communities and school performance, ρ_{Xu} should be negative. Furthermore, because the original estimation from Chapter 2 was already negative to begin with the "true" β_1 should be less negative. In conclusion, districts with more English learning students (*i*) do worse on standardized tests and (*ii*) have bigger classes (smaller budgets), so ignoring the English learning factor results in overstating the class size effect (in an absolute sense).

You might wonder whether this is actually going on in the Californian district school data. To see this, Figure 3.3 offers a cross tabulation of test scores by class size and percentage English learners.

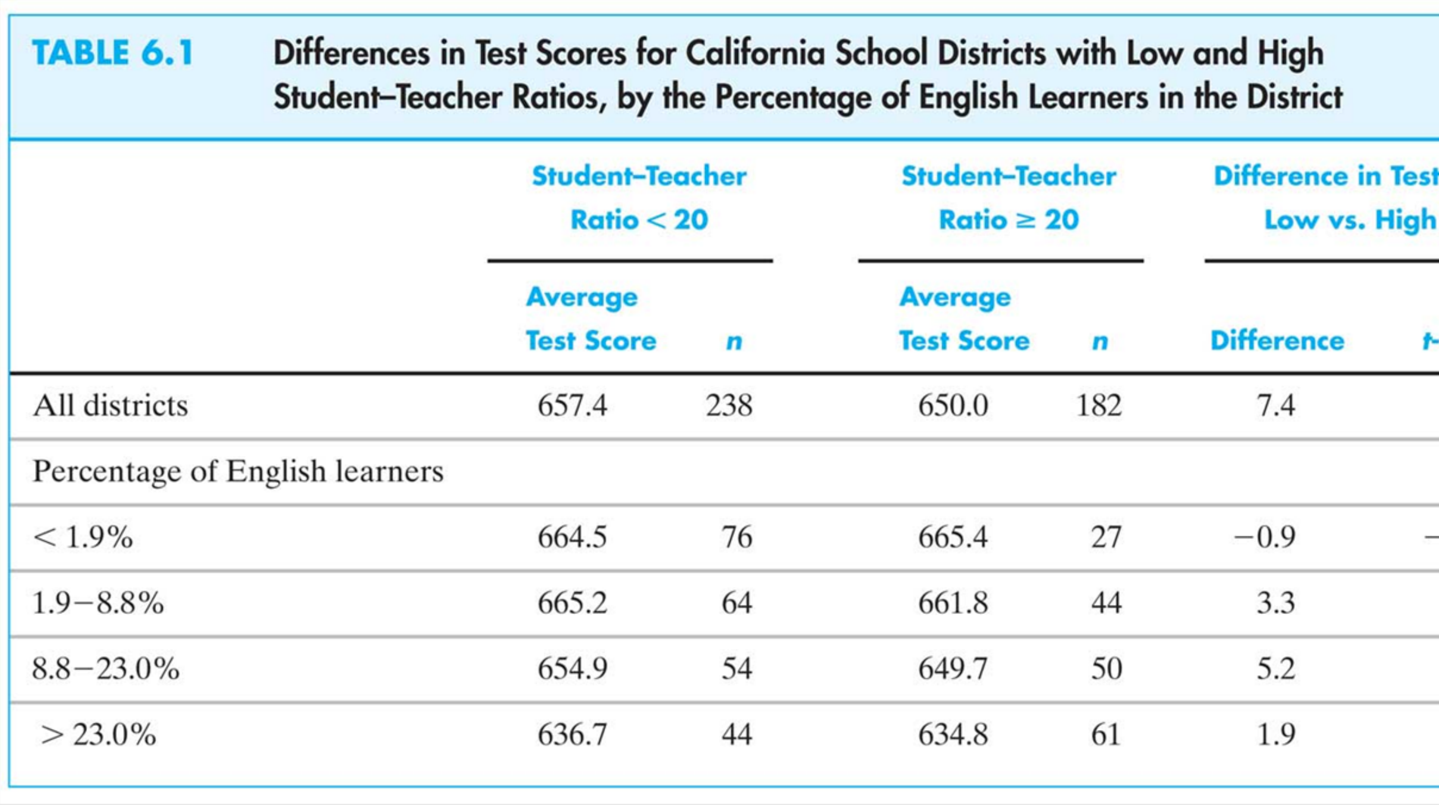


TABLE 6.1 **Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District**

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Low vs. High	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-
All districts	657.4	238	650.0	182	7.4	
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−
1.9−8.8%	665.2	64	661.8	44	3.3	
8.8−23.0%	654.9	54	649.7	50	5.2	
> 23.0%	636.7	44	634.8	61	1.9	

Figure 3.3: Cross tabulation of test scores by class size and percentage English learners

Now, the table depicted in Figure 3.3 is complex in its various dimensions. We have our two categories of class size (small and large), together with the difference in test scores, but we now stratify this by four categories of percentage English learners—that is, the percentage of pupils for whom English is *not* the native language. There are several important observations to make here:

1) districts with *fewer* English Learners (so less migrants) have on average *higher* test scores (what we assumed above);
2) districts with *fewer* English Learners (so less migrants) have *smaller* classes (what we assumed above);
3) the effect of class size with comparable percentages English learners is still (mostly negative), but not as much as we compare for all districts together (the *Difference*-column). This confirms our reasoning that our original estimate was too negative.

No, as already mentioned above, omitted variable bias occurs very often. So, how to correct for this such that the bias disappaers. In general, there are three strategies:

1. we can run a randomized controlled experiment in which treatment (STR) is randomly assigned: then percentage English learners ($PctEL$) is still a determinant of test scores, but by construction $PctEL$ should be uncorrelated with STR. Unfortunately, is it very difficult to randomize class size in reality and often this strategy is just not attainable as being too costly or unethical (this argument accounts by the way for all sciences, not only the social sciences);
2. we can adopt the cross tabulation approach of above, with finer gradations of STR and $PctEL$. Then by construction, within each group all classes have the same $PctEL$ so we control for $PctEL$. A disadvantages is that one needs many observations, especially when one wants to stratify upon other variables as well;
3. finally, and perhaps the easiest approach, we can use a population regression model in which the omitted variable ($PctEL$) is no longer omitted. We just include $PctEL$ as an additional regressor in a multiple regression model.

This is what the next section deals with. Obviously, a disadvantage of this approach is that you need observations for the omitted variable (but that also accounts for method 2).

3.2 Multivariate regression analysis

So, if we have information about an important omitted variable, as in the case of the size of migrant communities in the example above, then we can use that information in a multivariate population regression model. In the case of two regressors, that would look like:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, \dots, n \tag{3.3}$$

where:

- Y is the dependent variable
- X_1, X_2 are the two independent variables (regressors)
- (Y_i, X_{1i}, X_{2i}) denote the i$^{\text{th}}$ observation on Y, X_1, and X_2.
- β_0 is the unknown population intercept
- β_1 is the effect on Y of a change in X_1, **holding** X_2 constant
- β_2 is the effect on Y of a change in X_2, **holding** X_1 constant
- u_i is the the regression error (omitted factors)

Now, the only element that changes is the interpretation of a parameter, say β_1. In this case, it can still be seen as a 'slope' parameter, although now in 3-dimensional space, but it also states specifically that the other parameter(s) should be **held constant**. This does facilitate the interpretation of β_1. For example, consider changing X_1 by ΔX_1 while holding X_2 constant. That means that the population regression line before the change looks like:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{3.4}$$

whilst the population regression line, after the change, looks like:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 \tag{3.5}$$

And if we take the difference, then the interpretation of β_1 boils down again to the marginal effect:$\Delta Y = \beta_1 \Delta X_1$. Or, $\beta_1 = \frac{\Delta Y}{\Delta X_1}$ when holding X_2 constant and, likewise, $\beta_2 = \frac{\Delta Y}{\Delta X_2}$ when holding X_1 constant. β_0 is now the predicted value of Y when $X_1 = X_2 = 0$

If we do this for the the Californian school district data, then the original population regression line was estimated as:

$$\widehat{TestScore} = 698.9 - 2.28 STR \tag{3.6}$$

But if we now include include percent English Learners in the district ($PctEL$) to the model then the population regression 'line' becomes:

$$\widehat{TestScore} = 686.0 - 1.10 STR - 0.65 PctEL \tag{3.7}$$

Clearly, the effect of student teacher ratio becomes smaller (that is, less negative). That indicates that the original regression suffers from omitted variable bias. And this is what should happen as reasoned above. The `R` syntax for a multivariate regression model is now rather straightforward. You basically add another to the regression equation, as below:

```
model_3 = smf.ols('testscr ~ STR + english', data = data).fit(cov_type = "HC0")
print(model_3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.426
Model:                            OLS   Adj. R-squared:                  0.424
Method:                 Least Squares   F-statistic:                     225.4
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           4.27e-67
Time:                        17:22:12   Log-Likelihood:                -1716.6
No. Observations:                 420   AIC:                             3439.
Df Residuals:                     417   BIC:                             3451.
Df Model:                           2
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
Intercept    686.0322      8.697     78.882      0.000     668.986     703.078
STR           -1.1013      0.431     -2.553      0.011      -1.947      -0.256
english       -0.6498      0.031    -21.014      0.000      -0.710      -0.589
==============================================================================
Omnibus:                        0.631   Durbin-Watson:                   0.686
Prob(Omnibus):                  0.729   Jarque-Bera (JB):                0.550
Skew:                           0.088   Prob(JB):                        0.760
Kurtosis:                       3.024   Cond. No.                         301.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

```
print("The SER of model 3 is: {}".format(np.sqrt(model_3.scale)))
```

```
The SER of model 3 is: 14.464483125843136
```

Obviously, the effect of student teacher ration reduces with 50%! The interpretation of the rest of the statistical output, such as measures of fit and test statistics, follows in the subsections below.

3.2.1 Measures of fit for multiple regression

In multivariate regression models, there are four commonly used measures of fit, three of them we have seen before.

1. The standard error of regression or the SER denotes the standard deviation of $\widehat{u}_i$ and includes a degrees of freedom correction (degrees of freedom in this case denotes how many variables your have used and typically is denoted with k. The SER is defined as:

$$SER = s_{\widehat{u}} = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i^2}, \tag{3.8}$$

   where k is the number of variables (including the constant) use in the regression model. Note that in the univariate regression model $k = 2$—the slope coefficient and the constant.

2. The root mean square error (RMSE) which denotes as well the standard deviation of $\hat{u}_i$ but now without degrees of freedom. We have seen this before in Eq. 2.26 and does not change.
3. The R^2 which measures the fraction of variance of Y explained by the independent variables. Again, we have seen this one before
4. The adjusted "adjusted R^2" (or $\bar{\mathrm{R}}^2$) which is equal to the R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty. It can be formulated as:

$$\bar{\mathrm{R}}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}. \tag{3.9}$$

   Note that using this formulation, in a multivariate setting, it always should hold that $\bar{\mathrm{R}}^2 < \mathrm{R}^2$. But why do we care so much for the amount of variables that we use (denoted with k). That is because with each additional variable the R^2 always increases. And it is essential to notice that when $k = n$, the $\mathrm{R}^2 = 1$, so there is no variation left anymore. But that feels like cheating. You just have a parameter for each observation that you have, but such a model must be meaningless. Therefore, you always want to correct for the number of variables that you use.

In our Californian school district example that would amount to the following two outcomes. First for the univariate model:

$$\begin{aligned} TestScore &= 698.9 - 2.28STR && (3.10)\\ & R^2 = .05, SER = 18.6 && (3.11)\end{aligned}$$

And then for the multivariate model.

$$\begin{aligned} TestScore &= 686.0 - 1.10STR - 0.65PctEL && (3.12)\\ & R^2 = .426, \bar{R}^2 = 0.424, SER = 14.5 && (3.13)\end{aligned}$$

Note that all measures of fit increase. The $\bar{\mathrm{R}}^2$ now indicates that 42% of all variation in test scores are explained. That is a *huge* improvement compared to the 5% explanatory power of the univariate case. That indicates that the $PctEL$ strongly correlates with testscores. But again, we are not so much interested in prediction, but want to find the causal impact of class size instead. Another thing to notice here is that the R^2 and the $\bar{\mathrm{R}}^2$ are very close. That is because the number of variables is much smaller than the number of observations $k \ll n$, so that the impact of k is not very big.

3.2.2 The least squares assumptions for multivariate regression

Thus, it is easy to add other variables, so that the multivariate regression model now looks like:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i, i = 1, \dots, n \quad (3.14)$$

Suppose we are interested in β_1. How do we then know whether our estimation $\hat{\beta}_1$ is unbiased? For that we again resort to our least squares assumption, some of them will change a bit and we have to add a fourth one:

1. The first least squares assumptions changes slightly. Now, we state that the conditional distribution of u given all X_i's has mean zero, that is, $E(u|X_1 = x_1, \dots, X_k = x_k) = 0$. So, β_1 is biased even another variable X_k is correlated with u. So, only of the variables X_i has to be correlated with u and then all parameters are to a certain extent biased.
2. The second least squares assumption is more or less as before but now in a multivariate fashion, so the whole set of $(X_{1i}, \dots, X_{ki}, Y_i)$, with $i = 1, \dots, n$, should be independent and identical distributed ($i.i.d$).
3. The third least squares assumptions states again that large outliers are rare for all variables included, so for all $X_1, \dots, X_k$, and Y.
4. The fourth assumption is new and states that there is no perfect multicollinearity. We discuss this further below.

3.2.2.1 Multicollinearity

Multicollinearity comes in two flavours; perfect and imperfect. The former functions as a multivariate least squares assumptions whilst the latter oftentimes gives the largest problems. We start the discussion with perfect multicollinearity and then continue with the case of imperfect multicollinearity.

3.2.2.1.1 Perfect multicollinearity

The official definition of perfect multicollinearity is that there is a **perfect linear combination** amongst your variables. That means that there is not one optimal solution, but instead many (actually, infinitely many) more. Let us illustrate this by the following example. Suppose you include STR twice in your regression. Now, `python` produces then the following output:

```
data['STR2'] = data['STR']
model_4 = smf.ols('testscr ~ STR + STR2 + english', data = data).fit(cov_type = "HC0")
print(model_4.summary())
```

```
/Users/tomba/.virtualenvs/r-reticulate/lib/python3.13/site-packages/statsmodels/base/model.py:
  warnings.warn('covariance of constraints does not have full '
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.426
Model:                            OLS   Adj. R-squared:                  0.424
Method:                 Least Squares   F-statistic:                     225.4
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           4.27e-67
Time:                        17:22:12   Log-Likelihood:                -1716.6
No. Observations:                 420   AIC:                             3439.
Df Residuals:                     417   BIC:                             3451.
Df Model:                           2
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    686.0322      8.697     78.882      0.000     668.986     703.078
STR           -0.5506      0.216     -2.553      0.011      -0.973      -0.128
STR2          -0.5506      0.216     -2.553      0.011      -0.973      -0.128
english       -0.6498      0.031    -21.014      0.000      -0.710      -0.589
==============================================================================
Omnibus:                        0.631   Durbin-Watson:                   0.686
Prob(Omnibus):                  0.729   Jarque-Bera (JB):                0.550
```

```
Skew:                           0.088   Prob(JB):                       0.760
Kurtosis:                       3.024   Cond. No.                    6.44e+16
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The smallest eigenvalue is 1.15e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

See that `python` halves the impact of the STR variable. So both STR and $STR2$ have a similar impact half the impact of the original STR variable. Note as well that suddenly Python starts to produce warning (above the output table and below the output table in note [2]) But why is that? See that the impact of twice this variable should be equivalent to:

$$\beta_1 STR = w_1\beta_1 STR + w_2\beta_1 STR = (w_1 + w_2)\beta_1 STR, \quad (3.15)$$

where w_1 and w_2 are weights chosen such that they satisfy the condition that $w_1 + w_2 = 1$. But there is an infinite number of combinations that satisfy this condition! So, there is not one optimal solution and one of these variables should be dropped.

The violation of no perfect multicollearity often occurs when using dummies (see again Section 2.3.2.4). Suppose that we regress $TestScore$ on a constant, D, and B, where:$D_i = 1$ if $STR \leq 20$, $= 0$ otherwise ; $B_i = 1$ if $STR > 20$, $= 0$ otherwise. This example is slightly more complex as there is no perfect correlation between B and D. However, the model contains as well a constant and that create a perfect linear combination, namely $B_i + D_i = 1$ and that is the definition of a constant $(\beta_1 \times 1)$, so there is perfect multicollinearity in the model.

A different way of seeing this is to consider the following regression model and note that by definition $D_i = 1 - B_i$:

$$\begin{aligned} Testscr_i &= \beta_0 + \beta_1 D_i + \beta_2 B_i + u_i \\ &= \beta_0 + \beta_1 D_i + \beta_2(1 - D_i) + u_i \\ &= (\beta_0 + \beta_2) + (\beta_1 - \beta_2)D_i + u_i. \qquad (3.16) \end{aligned}$$

Suppose that the true constant equals 680 and the slope parameter equals 7. Then it is not difficult to see that there is an **infinite** amount of combinations possible of values for β_0, β_1 and β_2 that leads to these numbers.

Now, this example is a special case of the so-called dummy variable trap. Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive—that is, there are multiple categories and every observation falls in one and only one category (e.g., infant, child, teenager, adult). If you include all these dummy variables and a constant, you will have perfect multicollinearity—the dummy variable trap.

There are possible solutions to the dummy variable trap:

1. Omit one of the groups (e.g., the infants), or
2. Omit the intercept

In most cases you omit one of the groups (typically the one with the lowest value). This give the constant then the interpretation of the average value of that left-out category, where the dummy variables are then the relative differences to that left-out category.

Now, perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data. And, usually this is not a problem, because if you have perfect multicollinearity, your statistical software will let you know—either by crashing or giving an error message or by "dropping" one of the variables arbitrarily and very often the solution to perfect multicollinearity is to modify your list of regressors such that you no longer have perfect multicollinearity.

3.2.2.1.2 Imperfect multicollinearity

Imperfect and perfect multicollinearity are quite different despite the similarity of the names. Imperfect multicollinearity, namely, occurs when two or more regressors are very highly correlated. And if two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line—they are collinear—but unless the correlation is exactly ± 1, that collinearity is imperfect. What this implies is that one or more of the regression coefficients will be imprecisely estimated. Why is that? That is because of the definition of the coefficient in a multivariate regression model. Namely, the coefficient on X_1 is the effect of X_1 **holding** X_2 **constant**, but if X_1 and X_2 are highly correlated, then there is very little variation in X_1 once X_2 is held constant. That means that the data are pretty much uninformative about what happens when X_1 changes but X_2 doesn't, so the variance of the OLS estimator of the coefficient on X_1 will be large. And this results in large standard errors for one or more of the OLS coefficients. But often this is very hard to detect. Are standard errors high because of imperfect multicollinearity, because the number of observations is very low, or because there is large variation in the data? The answer to this unfortunately boils down to reasoning, but before you start estimating your statistical models it always good to look at scatterplots and correlations between variables.

But what is a high correlation? With a reasonable amount of observations all correlations below 0.9 can be considered fine. In practice, only correlations between variables higher than say 0.95 start to impose problems.

3.2.3 Testing with multivariate regression models

3.2.3.1 Hypothesis tests and confidence intervals for a single coefficient in multiple regression

Recall from Section 2.3.2.2 that for hypothesis testing in a classical statistical framework we make use of the fact that $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{var(\hat{\beta}_1)}}$ is approximately distributed as $N(0,1)$ according to the Central Limit theorem. Thus hypotheses on β_1 can be tested using the usual t-statistic, and confidence intervals are constructed as $\{\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)\}$. And this finding carries over to the multivariate setting where for $\beta_2, \dots, \beta_k$ we make use of the same framework. One thing to keep in mind is that $\hat{\beta}_1$ and $\hat{\beta}_2$ are generally not independently distributed—so neither are their t-statistics (more on this later).

Now, if we return to our Californian school district data set then we find that for the univariate case holds:

$$TestScore = \underbrace{698.9}_{10.4} - \underbrace{2.28}_{0.52} STR, \tag{3.17}$$

And the population regression "line" for the multivariate case is estimated as:

$$TestScore = \underbrace{686.0}_{8.7} - \underbrace{1.10}_{0.43} STR - \underbrace{0.650}_{0.031} PctEL \tag{3.18}$$

Remember, the coefficient on STR in Eq. 3.18 is the effect on $TestScores$ of a unit change in STR, holding constant the percentage of English Learners in the district. The corresponding 95% confidence interval for coefficient on STR in (2) is $\{-1.10 \pm 1.96 \times 0.43\} = (-1.95, -0.26)$. And the t-statistic testing $\beta_{STR} = 0$ is $t = -1.10/0.43 = -2.54$, so we reject the null-hypothesis at the 5% significance level. More evidence for the strength of the $PctEL$ variable can be seen from the fact that, under the null-hypothesis of $\beta_2 = 0$, the following must hold: t-statistic $= \frac{\hat{\beta_1}}{\sigma_{\hat{\beta_1}}} = \frac{0.65}{0.03} = 21.7$, which is a very high number for a t-statistic.

3.2.3.2 Tests of joint hypotheses

So, testing of single coefficients is just as before. Now in the Californian school district dataset there is as well a variable called $Expn$ denoting the expenditures per pupil. Consider the following population regression model:

$$TestScore_i = \beta 0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i \tag{3.19}$$

The null hypothesis that "school resources don't matter" and the alternative that they do, corresponds to:

- $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ vs
- $H_1 :$ either $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both

This is a joint hypothesis specifying a value for two or more coefficients. That is, it imposes a restriction on two or more coefficients. In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$. A "common sense" idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value. But this "one at a time" test isn't valid: the resulting test rejects too often under the null hypothesis (more than 5%)! That is because the t-statistics themselves are often not independent. Instead, we need a F-statistic, which tests all parts of a joint hypothesis at once. Unfortunately, these types of formulas can become quickly rather complex. Consider the F-test for the special case of the joint hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ in a regression with two regressors:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1 t_2}^2}\right) \tag{3.20}$$

where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between t_1 and t_2. Reject when F is large (typically to be determined from large statistical tables). The F-statistic is large when t_1 and/or t_2 is large and the F-statistic corrects (in just the right way) for the correlation between t_1 and t_2. The formula for more than two β's is nasty unless you use matrix algebra. There is a nice large-sample ($n > 50$) approximate distribution, which is the tail probability of the χ_q^2/q distribution beyond the F-statistic actually computed.

Now, `Python` does this in a much easier way by invoking the `f_test` method which is in the `statsmodels.api` package. So, for example, we want to test the joint hypothesis that the population coefficients on *str* and expenditures per pupil (*expenditure*) are both zero, against the alternative that at least one of the population coefficients is nonzero.

```python
model_5 = smf.ols('testscr ~ STR + expenditure + english', data = data).fit(cov_type = "HC0")
print(model_5.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.437
Model:                            OLS   Adj. R-squared:                  0.433
Method:                 Least Squares   F-statistic:                     148.6
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           1.87e-65
Time:                        17:22:12   Log-Likelihood:                -1712.8
```

```
No. Observations:                 420   AIC:                             3434.
Df Residuals:                     416   BIC:                             3450.
Df Model:                           3
Covariance Type:                  HC0
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept     649.5779     15.385     42.223      0.000     619.425     679.731
STR            -0.2864      0.480     -0.597      0.551      -1.227       0.654
expenditure     0.0039      0.002      2.459      0.014       0.001       0.007
english        -0.6560      0.032    -20.739      0.000      -0.718      -0.594
==============================================================================
Omnibus:                        0.046   Durbin-Watson:                   0.742
Prob(Omnibus):                  0.977   Jarque-Bera (JB):                0.070
Skew:                          -0.025   Prob(JB):                        0.966
Kurtosis:                       2.962   Cond. No.                     1.16e+05
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 1.16e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
hypotheses = '(STR = 0), (expenditure = 0)'
f_test = model_5.f_test(hypotheses)
print(f_test)
```

```
<F test: F=5.485972090418466, p=0.004449900096176l865, df_denom=416, df_num=2>
```

The output shows an F-statistic with $q = 2$ restrictions (degrees of freedom or Df) with outcome 5.49. Do not directly interpret this number, but know that Prob $> F = 0.0044$ gives the probability that under the null-hypothesis this outcome is produced. So the joint null-hypothesis that both types of expenditures are zero (at the same time), can be rejected at a 5% (and a 1%) significance level. Other types of joint tests can easily be constructed as well. For example, when you want to know whether both coefficient add up to 1, then you would state `hypotheses = '(STR + expenditure = 1)'`. The final point to make is the F-test in the regression output itself. Here, that is for example `F = 148.6`. This is a joint test that all variables, except the constant, have no impact. So, $\beta_i = 0$ for all i at the **same time**. It not often that you come across a general regression F-test that does not reject the null-hypothesis. It namely implies that your independent variables do not contain any information about the dependent variable.

And with the F-test, we now have discussed the most important regression outcome components displayed by `Python`. Most of this information you do not need for your report but we will come back later to this.

3.3 Non-linear specifications

The model we are using is coined the *linear* regression model, and, indeed, one of the underlying assumptions is that the relations between the independent and dependent variables are linear. Consider the relation again between test scores and class sizes in the Californian school district data. Using the following code (note we now combine a scatter plot with a population regression line):

```
sns.regplot(data = data, x = "STR" , y = "testscr", ci = None)
plt.text(x=20, y=700, s="$\widehat{TestScore} = 698.9 - 2.28 \\times STR$", fontsize=11, color=
plt.xlabel("Student-teacher ratio")
plt.ylabel("Testscore")
plt.show()
```

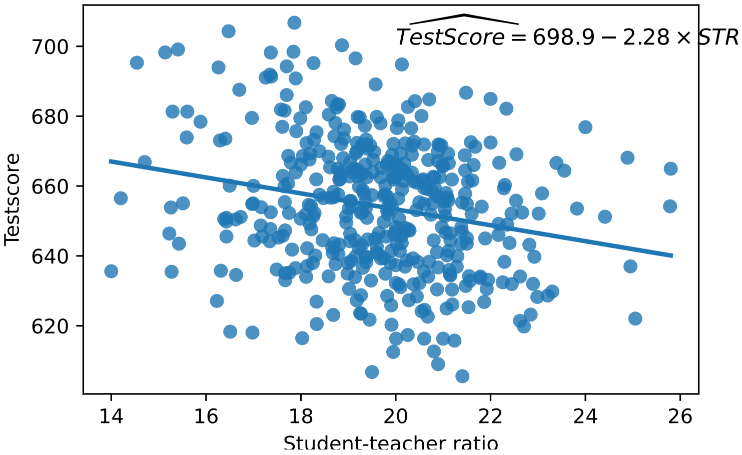


Figure 3.4: Scatterplot and estimated regression line

Indeed, there might be evidence that the relation depicted in Figure 3.4—if anything—is linear. But, clearly that is not the case for the relation between test scores and average district income. Namely, the syntax below:

```
sns.regplot(data = data, x = "income" , y = "testscr", ci = None)
plt.xlabel("Average district income")
plt.ylabel("Testscore")
plt.show()
```

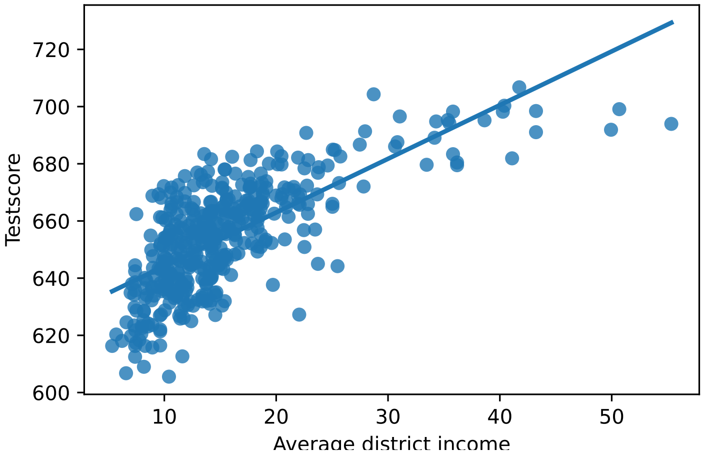


Figure 3.5: A possible non-linear relation between test scores and average district income

Figure 3.5 shows a non-linear relation, where the effect of income tapers off (note the similarity with Figure 1.1)—or, there is a marginal decreasing effect of average district income with respect to average school test scores. Thus, in affluent neighborhood test scores are higher, but increasingly less so. Of course, you can still try to estimate this with a linear population regression line as in Figure 3.5, but this introduces a well a **bias**. The estimate does not capture that what you want. Namely, it now holds that $E(u \mid X = x) \neq 0$, because for small X, say $X < 10$, the residuals are negative, for medium sized Xs most residuals are positive and for large $X > 40$ all residuals are negative again. So, there is a clear relation between X and u and they fail to be independent. This particular form of bias is coined **specification bias**. There is another issue here and that is that the effect on Y of a change in X depends on the value of X—that is, the *marginal* effect of X is not constant. Again, remember in a regression context the marginal effect is denoted by the **slope** of the population regression line.

To remedy possible specification bias, we will use nonlinear regression population regression **functions** of X, or we estimate a regression function that is nonlinear in X. Here, it is important to see that we do so by *transforming* X, so the population regression 'line'. The estimator still remains a linear regression model.

We will analyse below two complementary and often adopted approaches:

1. Using **polynomials** to transform X. That means that the effect is approximated by a quadratic, cubic, or higher-degree polynomial. This approach as well governs to an extent so-called interaction effects which is a special case, where we multiply two different variables.
2. Using **logarithmic** transformations of X, where Y and/or X is transformed by taking its logarithm. Here, the main focus is on the interpretation of the $\widehat{\beta}$s, as they change from a unit increase interpretation to a percentages interpretation which often can be found useful.

3.3.1 Polynomials

Our first approach to non-linear specification is applying polynomials of the variables that we suspect has a non-linear impact. If that is the independent variable X, the we can construct the following *linear regression* model by using polynomials:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_i^2 + ... + \beta_r X_i^r + u_i \tag{3.21}$$

Note again that this is just the linear regression model—except that the regressors are powers of X! So, in effect we transform the data—actually create new variables X^r—, but the specification in parameters remains linear. Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS. However, the coefficients are now a bit more difficult to interpret. Consider the example of above about the relation between test scores average district income, where $Income_i$ is defined as the average district income in the i^{th} district (thousands of dollars per capita). For a quadratic specification, we specify the linear regression model as below:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i \tag{3.22}$$

For a cubic specification the linear regression model becomes:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i \tag{3.23}$$

First, we focus on the estimation of the quadratic function. In Python this would look like:

```python
model_5 = smf.ols('testscr ~ income + I(income**2)', data = data).fit(cov_type = "HC0")
print(model_5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.556
Model:                            OLS   Adj. R-squared:                  0.554
Method:                 Least Squares   F-statistic:                     431.6
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          2.69e-102
Time:                        17:22:12   Log-Likelihood:                -1662.7
No. Observations:                 420   AIC:                             3331.
Df Residuals:                     417   BIC:                             3344.
Df Model:                           2
Covariance Type:                  HC0
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        607.3017      2.891    210.039      0.000     601.635     612.969
income             3.8510      0.267     14.416      0.000       3.327       4.375
I(income ** 2)    -0.0423      0.005     -8.882      0.000      -0.052      -0.033
==============================================================================
Omnibus:                        0.556   Durbin-Watson:                   0.951
Prob(Omnibus):                  0.757   Jarque-Bera (JB):                0.378
Skew:                          -0.048   Prob(JB):                        0.828
Kurtosis:                       3.111   Cond. No.                     2.23e+03
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 2.23e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Note that `I()` function here, which means that within a regression formula you can do a transformation of a variable.

Now, it is straightforward to test the null-hypothesis of linearity against the alternative that the regression function is a quadratic. Namely, we only have to consider the t-statistic of the quadratic term. And that is larger than 1.96, so against a 5% significance level we reject the null-hypothesis of linearity.

Plotting of polynomial population regression lines is rather straightforward with `seaborn`. Here, you want to combine a polynomial with a linear dimension. One way of doing this is as follows, using the `order` command:

```
sns.regplot(data = data, x = "income" , y = "testscr", ci = None, color = 'steelblue', label =
sns.regplot(data = data, x = "income" , y = "testscr", scatter = False, ci = None, color = 'ste
sns.regplot(data = data, x = "income" , y = "testscr", scatter = False, order = 2 , ci = None,
plt.xlabel("Average district income")
plt.ylabel("Testscore")
plt.legend(title= "", loc= "lower right")
plt.show()
```

An order of 2 gives a quadratic fit, an order of 3 a cubic, and so on. Note as well that I call the regplot now three times but that is to control the legend entries (labels starting with an underscore _ do not create a legend entry).

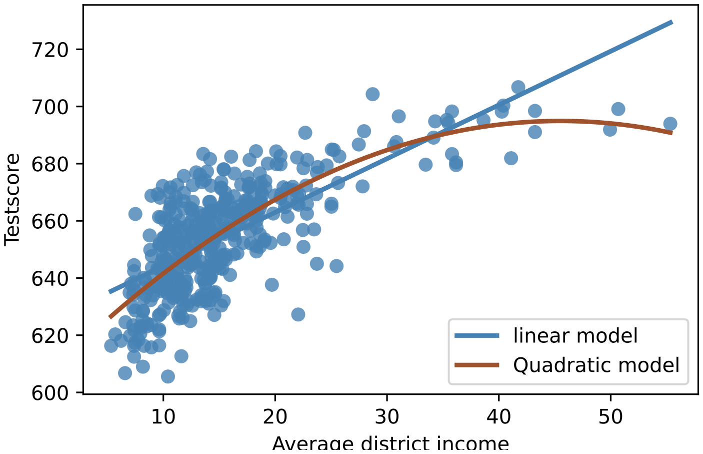


Figure 3.6: Two fitted regression lines for the relation between test scores and average district income

But what is now the marginal effect of average district income. That, now, depends on itself. Namely, $\frac{\partial \text{testscore}}{\partial \text{income}} = \beta_1 + \beta_2 \text{income}$.

Another way of seeing this is to compute the effects for different values of X

$$\widehat{TestScore}_i = 607.3 + 3.85 Income_i - 0.0423 (Income_i)^2 \quad (3.24)$$

The predicted change in test scores for a change in income from \$5,000 per capita to \$6,000 per capita then amounts to:

$$\begin{aligned} \Delta \widehat{TestScore} &= 607.3 + 3.85 \times 6 - 0.0423 \times 6^2 \\ &\quad -(607.3 + 3.85 \times 5 - 0.0423 \times 5^2) \\ &= 3.4 \end{aligned} \quad (3.25)$$

And if we calculate the predicted effects for different values of X, then we get the following table:

Table 3.1: Effect of X

Change in Income (1000 dollar per capita)	$\Delta \widehat{TestScore}$
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	0.0

Thus, the effect of a change in income is greater at low than high income levels (perhaps, a declining marginal benefit of an increase in school budgets?). But, be careful here! What is the effect of a change from 65 to 66? That is negative and already Figure 3.6 shows that a quadratic specification starts to decline after the value of about 50; and perhaps that is not the behavior that you want. So, with polynomials it is essential not to extrapolate outside the range of the data (and still interpret the outcome).

The estimation of a cubic specification is straightforward:

```
model_6 = smf.ols('testscr ~ income + I(income**2) + I(income**3)', data = data).fit(cov_type
print(model_6.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.558
Model:                            OLS   Adj. R-squared:                  0.555
Method:                 Least Squares   F-statistic:                     272.8
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           7.54e-98
Time:                        17:22:12   Log-Likelihood:                -1661.6
No. Observations:                 420   AIC:                             3331.
Df Residuals:                     416   BIC:                             3347.
Df Model:                           3
Covariance Type:                  HC0
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        600.0790      5.078    118.179      0.000     590.127     610.031
income             5.0187      0.704      7.129      0.000       3.639       6.398
I(income ** 2)    -0.0958      0.029     -3.325      0.001      -0.152      -0.039
I(income ** 3)     0.0007      0.000      1.985      0.047     8.5e-06       0.001
==============================================================================
Omnibus:                        0.567   Durbin-Watson:                   0.981
Prob(Omnibus):                  0.753   Jarque-Bera (JB):                0.385
Skew:                          -0.047   Prob(JB):                        0.825
Kurtosis:                       3.115   Cond. No.                     1.65e+05
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 1.65e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Where if we now want to test the null-hypothesis of linearity, then we have to invoke an F-test. Namely, the alternative hypothesis is that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3, so:

- H_0: Coefficients on $income^2$ **and** $income^3 = 0$
- H_1: at least one of these coefficients is nonzero.

And the outcome below shows that the null-hypothesis that the population regression is linear is rejected at the 5% (and 1%) significance level against the alternative that it is a polynomial of degree up to 3.

```
hypotheses = 'I(income ** 2) , I(income ** 3)'
f_test = model_6.f_test(hypotheses)
print(f_test)
```

```
<F test: F=38.05310787625862, p=6.655364476481898e-16, df_denom=416, df_num=2>
```

3.3.2 Interaction variables

Using interaction variables is a special case of polynomial effects. Namely, instead of multiply a variable with itself $X \times X = X^2$, you now multiple a variable with another variable. And you want to do this to take into account interactions between independent variables. Assume, for example, that a class size reduction is more effective in some circumstances than in others (which is quite conceivable). Perhaps smaller classes help more if there are many English learners (i.e., large migrant communities), who need more individual attention. That is, $\frac{\partial TestScore}{\partial STR}$ might depend on $PctEL$. More generally, this subsection looks into the fact that the marginal effect of $\frac{\partial Y}{\partial X_1}$ might depend on some other variable X_2.

3.3.2.1 Interactions between two binary variables

First, we look into the simplest (and perhaps most insightful) case of two binary (dummy) variables. Consider therefore the following linear regression model:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i, \qquad (3.26)$$

where both D_{1i} and D_{2i} are now considered to be binary. Now, of course, β_1 is the effect of changing $D_1 = 0$ to $D_1 = 1$. So, in this specification, this effect doesn't depend on the value of D_2. To allow the effect of changing D_1 to depend on D_2, we have to include the interaction term $D_{1i} \times D_{2i}$ as a regressor:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i \qquad (3.27)$$

To interpret now the coefficient β_1 we compare the two cases; for $D_1 = 0$ and for to $D_1 = 1$:

$$E(Y_i|D_{1i} = 0, D_{2i} = d_2) = \beta_0 + \beta_2 d_2 \tag{3.28}$$

$$E(Y_i|D_{1i} = 1, D_{2i} = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \tag{3.29}$$

If we now subtract them from each other:

$$E(Y_i|D_{1i} = 1, D_{2i} = d_2) - E(Y_i|D_{1i} = 0, D_{2i} = d_2) = \beta_1 + \beta_3 d_2 \tag{3.30}$$

then we have the marginal effect of D_1 which now depends on d_2. The interpretation of β_3 boils down to being incremental to the effect of D_1, when $D_2 = 1$

Let us go back to our Californian school district example with the following variables to be used: test scores, student teacher ratio, and English learners. Let:

$$HiSTR = 1 \text{ if } STR \geq 20 \text{ and } HiEL = 1 \text{ if } PctEL \geq 10 \tag{3.31}$$

$$HiSTR = 0 \text{ if } STR < 20 \text{ and } HiEL = 0 \text{ if } PctEL < 10 \tag{3.32}$$

$$\tag{3.33}$$

And if we have the estimation results we get the following outcome.

$$\widehat{TestScore} = 664.1 - 18.2 HiEL - 1.9 HiSTR - 3.5 (HiSTR \times HiEL) \tag{3.34}$$

So, how to interpret the various parameters? Perhaps the simple way is to construct the following two-by-two table:

Table 3.2: Interpretation of interaction effects with dummies

	$HiEL = 0$	$HiEL = 1$
$HiSTR = 0$	664.1	$664.1 - 18.2 = 645.9$
$HiSTR = 1$	$664.1 - 1.9 = 662.2$	$664.1 - 1.9 - 18.2 - 3.5 = 640.5$

Now, Table 3.2 specifies for each combination (and there are exactly four of them) of $HiSTR$ and $HiEL$ the average expected test score outcome. Clearly, there are different 'marginal' effects of $HiSTR$. Namely, the effect of $HiSTR$ when $HiEL = 0$ is -1.9, whilst the effect of $HiSTR$ when $HiEL = 1$ is

$-1.9 - 3.5 = -5.4$. This points out that a class size reduction is estimated to have a bigger effect when the percent of English learners is large. However, when you estimate this in R then you see that this interaction is not statistically significant, because the t-statistic equals $3.5/3.1 = 1.1$

3.3.2.2 Interactions between continuous and binary variables

The second case we consider is between a continuous and a binary variable. First assume the following regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + +u_i, \tag{3.35}$$

where D_i is a binary variable and X is a continuous variable. As specified above, the effect on Y of X (holding D constant) $= \beta_1$, which does not depend on D. To allow the effect of X to depend on D, we can include the interaction term $D_i \times X_i$ as a regressor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (D_i \times X_i) + u_i \tag{3.36}$$

What this binary-continuous interaction does is essential create two different population regression lines. Namely, for observations with $D_i = 0$ (the $D = 0$ group or the $D = 0$ regression line) there is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{3.37}$$

Whilst for observations with $D_i = 1$ (the $D = 1$ group or the $D = 1$ regression line) the regression line comes down to:

$$\begin{aligned} Y_i &= \beta_0 + \beta_2 + \beta_1 X_i + \beta_3 X_i + u_i & (3.38)\\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + u_i & (3.39) \end{aligned}$$

And these two population regression lines might both differ in the level (the constant) and in the slope of the line. So, there are three possibilities as depicted in Figure 3.7:

In the first panel (a), $\beta_3 = 0$, so there is only a level effect. In the second panel (b), both β_2 and β_3 are not 0, so there is both a level and a slope effect. The last panel indicates that

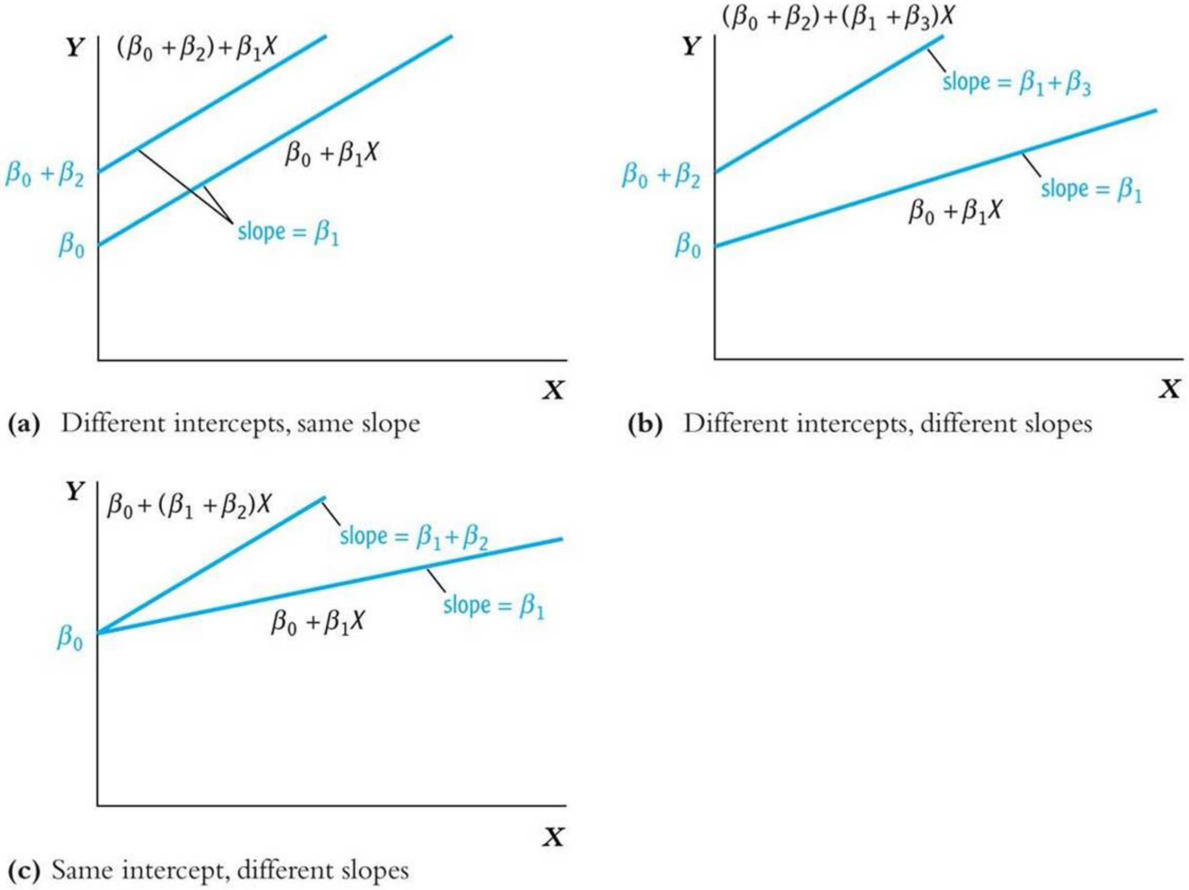


Figure 3.7: Three possible binary-continuous interaction outcomes

$\beta_2 = 0$, meaning that there is only a slope effect. But how to interpreting the coefficients now? Therefore, we take the marginal effect of

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) \tag{3.40}$$

which yields:

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 D \tag{3.41}$$

Thus, the effect of X depends on D and β_3 is the increment to the effect of X, when $D = 1$ (a slope effect).

To see this in our Californian school district example we now use the variables test scores, student teacher ratio and the as previously defined dummy variable $HiEL$ as:

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6HiEL - 1.28(STR \times HiEL) \tag{3.42}$$

Now when $HiEL = 0$ the population regression line amounts to:

$$\widehat{TestScore} = 682.2 - 0.97STR \tag{3.43}$$

And when $HiEL = 1$ the population regression line is:

$$\begin{aligned} \widehat{TestScore} &= 682.2 - 0.97STR + 5.6 - 1.28STR & (3.44) \\ &= 687.8 - 2.25STR & (3.45) \end{aligned}$$

Thus we have two regression lines: one for each $HiSTR$ group. And the conclusion is that a class size reduction is estimated to have a larger effect when the percent of English learners (migrant communities) is large.

Hypothesis testing is as before. To test whether the two regression lines have the same slope, the null-hypothesis boils down to the coefficient of $STR \times HiEL$ being zero: the t-statistic of this one become $-1.28/0.97 = -1.32$ and thus we do not reject this test. To test whether the two regression lines have the same intercept, the null-hypothesis becomes the coefficient of $HiEL$ being zero, yielding: $t = -5.6/19.5 = 0.29$, so we do not reject that null-hypothesis either. Interestingly, the null-hypothesis that the two regression lines are the same—population coefficient on $HiEL = 0$ and population coefficient on yields $STR \times HiEL = 0$: $F = 89.94(p - value < .001)$. So, we reject the joint hypothesis but neither individual hypothesis.

Finally, the question may arise how to draw such lines as in Figure 3.7. For this the following code is very useful:

```python
data['hiel'] = data['english'] >= 10
model_8 =  smf.ols('testscr ~ STR + hiel + STR : hiel', data = data).fit(cov_type = "HC0")
model_8.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.310
Model:                            OLS   Adj. R-squared:                  0.305
Method:                 Least Squares   F-statistic:                     64.28
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           3.61e-34
Time:                        17:22:13   Log-Likelihood:                -1755.3
No. Observations:                 420   AIC:                             3519.
Df Residuals:                     416   BIC:                             3535.
Df Model:                           3
Covariance Type:                  HC0
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept          682.2458     11.811     57.763      0.000     659.096     705.395
hiel[T.True]         5.6391     19.421      0.290      0.772     -32.426      43.704
STR                 -0.9685      0.586     -1.652      0.099      -2.118       0.181
STR:hiel[T.True]    -1.2766      0.962     -1.327      0.185      -3.163       0.609
==============================================================================
Omnibus:                        4.924   Durbin-Watson:                   0.585
Prob(Omnibus):                  0.085   Jarque-Bera (JB):                5.019
Skew:                           0.252   Prob(JB):                       0.0813
Kurtosis:                       2.821   Cond. No.                         527.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
"""
```

So, first, we generate a new dummy variable `hiel` as discussed above. Then we regress testscores on class size, the new `hiel` dummy variable and the interaction with the `:` operator. Finally, we can draw the figure as follows:

```
# we are going to add line with the number on the X-axis as X variable
X = np.linspace(14, 26, num=20)
# lmplot can use hue to determine different points
# also name plot because otherwise you get two plots (?)
g = sns.lmplot(x="STR", y="testscr", hue="hiel", data=data, fit_reg=False, legend = False)
plt.ylabel("Testscore")
plt.xlabel("Student teacher ratio")
plt.plot(X, 682.2 - 0.969 * X, "blue")
plt.plot(X, (682.2 + 5.639) + (-0.969 - 1.277) * X, "orange");
plt.legend(title= "Large amount of english learners", labels = ["no", "yes"])
```

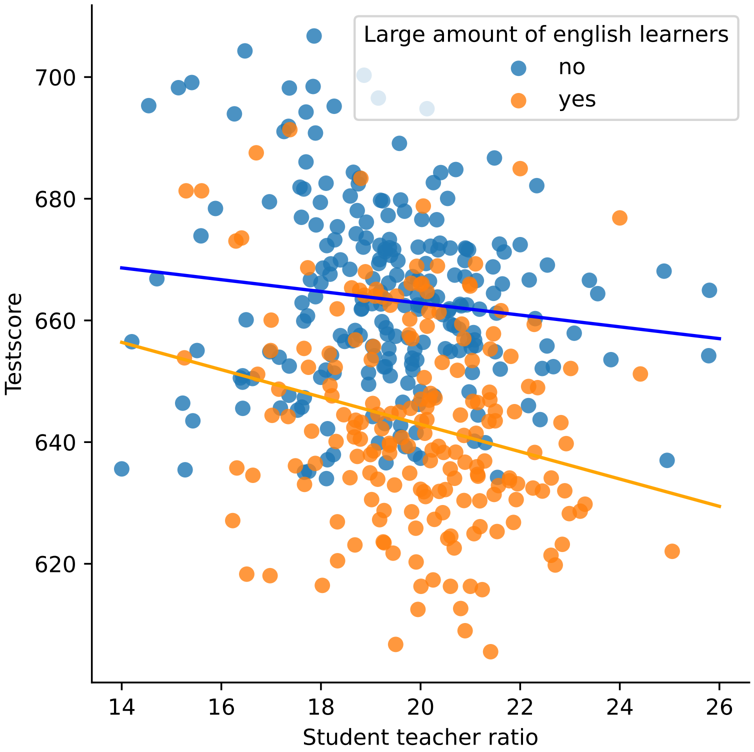


Figure 3.8: Predicted population regression lines of districts with large and small percentage english learners

Clearly, Figure 3.8 shows that districts with more English learners (containing larger migrant communities) have lower test scores overall. Above that, class size seems to have a larger negative effect on districts with more English learners as the slope is more negative.

3.3.2.3 Interactions between two continuous variables

The last case are interactions between two continuous variables and that is always a difficult case of interpret. Starting again with the model:

$$Y_i = \beta_0 + \beta 1 X_{1i} + \beta_2 X_{2i} + u_i, \tag{3.46}$$

where both X_1, X_2 are continuous and as specified, the effect of X_1 doesn't depend on X_2 and the effect of X_2 doesn't depend on X_1. Now, to allow the effect of X_1 to depend on X_2, we include the interaction term $X_{1i} \times X_{2i}$ as a regressor. Where, to interpret the coefficients, we take the first derivative of X_1 in:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \tag{3.47}$$

which yields:

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2 \tag{3.48}$$

where β_3 should be interpreted as the increment to the effect of X_1 from a unit change in X_2.

3.3.3 Logarithmic transformations

To incorporate non-linear effects, very often logarithmic transformations are used of Y and/or X, where we use $\ln(X)$, being the natural logarithm of X. One feature of logarithmic transformations is that they permit modeling relations in percentage terms (like elasticities), rather than linearly. That is because:

$$\ln(x + \Delta x) - \ln(x) = \ln(1 + \frac{\Delta x}{x}) \cong \frac{\Delta x}{x} \tag{3.49}$$

Note that this is an approximation, but from calculus we know that $\frac{d \ln(x)}{dx} = \frac{1}{x}$). And the above approximation works quite well for small numbers. For example, numerically: $\ln(1.01) = .00995 \cong .01$ and $\ln(1.10) = .0953 \cong .10$, where the latter is still rather close. Now remember the following rules for natural logarithms:

1. $\ln(a \times b) = \ln(a) + \ln(b)$
2. $\ln(\frac{a}{b}) = \ln(a) - \ln(b)$
3. $\ln(a^\alpha) = \alpha \ln(a)$
4. $\ln(e^X) = X$.

When you encounter a nonlinear model a strategy that often works is log-linearization. This works as follows for, e.g., the following Cobb-Douglas specification:

$$Y = AK^\alpha L^{1-\alpha} \rightarrow \ln(Y) = \ln(A) + \alpha \ln(K) + (1-\alpha)\ln(L). \tag{3.50}$$

Thus, you take the natural logarithm on both sides. There are three different cases of logarithmic regression models as specified in Table 3.3.

Table 3.3: Three cases of logarithmic specifications

Case	Population regression model
linear-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
log-linear	$\ln(Y_i) = \beta_0 + \beta_1 (X_i) + u_i$
log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

Though statistical testing remains the same, the interpretation of the slope coefficient differs in each case. To derive the interpretation we want to find the marginal effect of X using the first derivative. Let's do that for the three cases above.

3.3.3.1 Linear-log population regression model

The linear-log population regression model is specified as:

$$Y = \beta_0 + \beta_1 \ln(X) \tag{3.51}$$

Now take the first derivative of Y to X:

$$\frac{\partial Y}{\partial X} = \frac{\beta_1}{X} \tag{3.52}$$

so

$$\beta_1 = \frac{\partial Y}{\partial X / X} \tag{3.53}$$

In this case that means that β_1 should be interpreted as the absolute change of Y when X changes with $\beta_1/100$ percent. To illustrate this, consider the case where we take the natural logarithm of district income, so we define the new regressor as, $\ln(Income)$.

The model is now linear in $\ln(Income)$, so the linear-log model can be estimated by OLS, which yields

$$\widehat{TestScore} = 557.8 + 36.42 \times \ln(Income_i) \tag{3.54}$$

so an 1% increase in $Income$ is associated with an increase in test scores of 0.36 points on the test. And again, standard errors, confidence intervals, R^2—all the usual tools of regression apply here. But the difficulty in plottin the new regression line remains. Consider the following `Python` syntax, where we first have to define the new regressor by invoking the `generate` command.

```
data['lnincome'] = np.log(data['income'])
model_9 = smf.ols('testscr ~ lnincome', data = data).fit(cov_type = "HC0")
model_9.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.563
Model:                            OLS   Adj. R-squared:                  0.561
Method:                 Least Squares   F-statistic:                     682.9
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           6.18e-90
Time:                        17:22:13   Log-Likelihood:                -1659.7
No. Observations:                 420   AIC:                             3323.
Df Residuals:                     418   BIC:                             3331.
Df Model:                           1
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
Intercept     557.8323      3.831    145.618      0.000     550.324     565.340
lnincome       36.4197      1.394     26.133      0.000      33.688      39.151
==============================================================================
Omnibus:                        0.548   Durbin-Watson:                   0.991
Prob(Omnibus):                  0.760   Jarque-Bera (JB):                0.388
Skew:                          -0.059   Prob(JB):                        0.824
Kurtosis:                       3.091   Cond. No.                         20.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
"""
```

```
sns.regplot(data = data, x = "income" , y = "testscr", color = 'steelblue', fit_reg = False)
plt.xlabel("Average district income")
plt.ylabel("Testscore")
X = np.linspace(5, 55, num=20)
plt.plot(X, 557.8323 + 36.4197 * np.log(X), linewidth=2 )
plt.show()
```

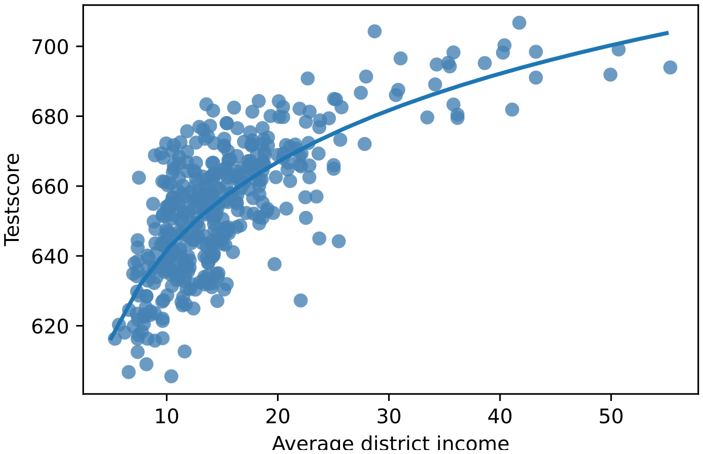


Figure 3.9: A non-linear relation with the log of average district income

When you compare Figure 3.9 with Figure 3.6 then you notice that in the case of logarithm the population remains increasing (but less and less steep). This can be considered as an advantage when you want to estimate decreasing (or increasing) returns.

3.3.3.2 Log-linear population regression model

The second case we consider is the log-linear population regression model, as specified by:

$$\ln(Y) = \beta_0 + \beta_1 X \tag{3.55}$$

To find the interpretation of β_1, we again take the first derivative $\frac{\partial Y}{\partial X}$, but first transform the model like this:

$$Y = \exp(\beta_0 + \beta_1 X) \tag{3.56}$$

then take the first derivative:

$$\frac{\partial Y}{\partial X} = \beta_1 \exp(\beta_0 + \beta_1 X) = \beta_1 Y \tag{3.57}$$

and collect terms

$$\beta_1 = \frac{\partial Y / Y}{\partial X} \tag{3.58}$$

The interpretation of β_1 now is that one unit change in X causes a β_1 percentage in Y

3.3.3.3 Log-log population regression model

Finally, we have our third case, being the log-log population regression model as specified by:

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) \tag{3.59}$$

To find the interpretation of β_1, we again take the first derivative $\frac{\partial Y}{\partial X}$, but first transform the model like this:

$$Y = \exp(\beta_0 + \beta_1 \ln(X)) \tag{3.60}$$

So

$$\frac{\partial Y}{\partial X} = \beta_1 / X \exp(\beta_0 + \beta_1 \ln(X)) = \beta_1 Y / X \tag{3.61}$$

and after collecting terms we end up with an **elasticity**:

$$\beta_1 = \frac{\partial Y / Y}{\partial X / X} \tag{3.62}$$

As an example consider the case when we want to regress ln(test scores) on ln(income). To do so, we first define a new dependent variable, ln(TestScore), and a new regressor, ln(Income) The model is now a linear regression of ln(TestScore) against ln(Income), which can be estimated by OLS as follows

$$\ln(\widehat{TestScore}) = 6.336 + 0.0554 \times \ln(Income_i), \tag{3.63}$$

where the interpretation is that an 1% increase in $Income$ is associated with an increase of .0554% in $TestScore$ ($Income$ goes up by 1%, $TestScore$ goes up by 0.06%).

Suppose that we now want to plot both the log-linear and the log-log specification, then we can use the following syntax:

```
data['lntestscr'] = np.log(data['testscr'])
model_10 = smf.ols('lntestscr ~ income', data = data).fit(cov_type = "HC0")
# notes the tables method to shrink the outcomes a bit
model_10.summary().tables[1]
```

```
<class 'statsmodels.iolib.table.SimpleTable'>
```

```
model_11 = smf.ols('lntestscr ~ lnincome', data = data).fit(cov_type = "HC0")
model_11.summary().tables[1]
```

```
<class 'statsmodels.iolib.table.SimpleTable'>
```

```
g = sns.lmplot(x="income", y="lntestscr", data=data, fit_reg=False, legend = False)
plt.xlabel("Average district income")
plt.ylabel("Testscore (in logs)")
X = np.linspace(5, 55, num=20)
plt.plot(X, 6.3363 + 0.0554 * np.log(X), linewidth=2, label = "log-log specification" )
plt.plot(X, 6.4394 + 0.0028 * X, linewidth=2, label = "log-linear specification" )
plt.legend(title= "", loc= "lower right")
```

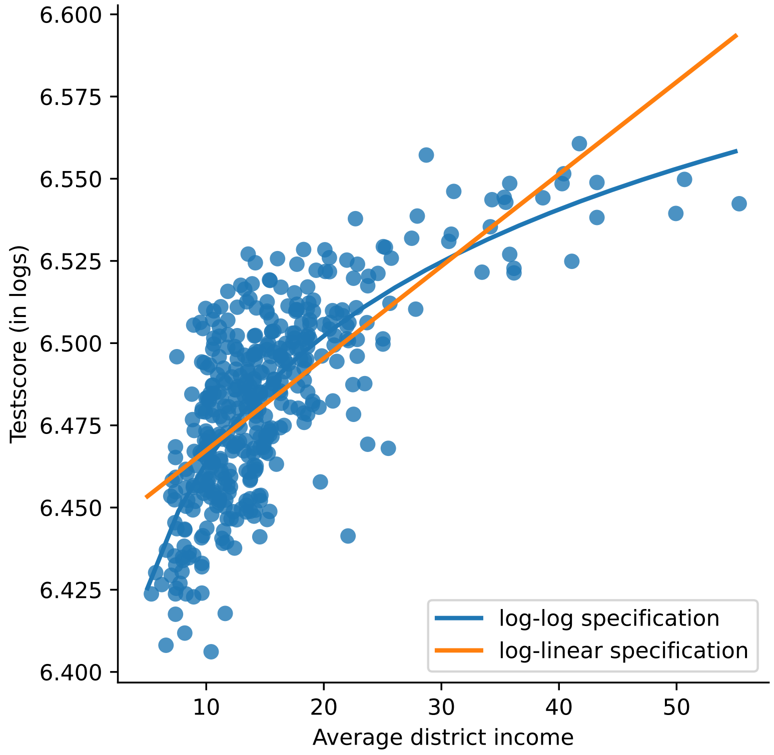


Figure 3.10: A non-linear relation

Note that the y-axis is on a logarithmic scale here, thus the log-linear specification is now a linear line.

3.3.3.4 Summary: logarithmic transformations

We have seen three different cases of logarithmic specification, differing in whether Y and/or X is transformed by taking logarithms. Now, the regression is linear in the new variable(s) $\ln(Y)$ and/or $\ln(X)$, and the coefficients can be estimated by OLS where hypothesis tests and confidence intervals are now implemented and interpreted 'as usual'. Only the interpretation of the coefficients differs from case to case and is directly related to percentage changes (growth) and elasticities. Oftentimes, the choice of specification, however, should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values. Sometimes, though, you have a structural economic model such as Eq. 2.1, which defines the type of specification you should use. Finally, see that in economics many models exists with decreasing or increasing return to scale and that these are very closely related with logarithmic specifications.

3.4 Using fixed effects in panel data

Multivariate regression is a powerful tool for controlling for the effect of variables for which we have data. But often we do not have data on what we suspect might be important—such as individual characteristics like ambition, intelligence, drive or stamina. Or regional of country data, where the type of soil, the ruggedness (hilliness), or population density determine to a large extent the behavior of people living on it. If we do not have this type of data, then it is not always the case that everything is lost. Especially, when we have repeated observations, so observations on the same entity throughout time. This is referred to as panel data and requires one additional subscript t as in X_{it} indicating the observation X on individual i made at time t. To understand why this sometimes works, we temporarily change to another dataset and that is the 'fatality' data collected by Levitt and Porter (2001) which deals with the relation between drunk driving and fatal accidents in the States of the US between 1982 and 1988. For this particular example we look at the impact of the 'beer tax', measured as the real tax in dollars on a case of beer, on 'fatality', measured as the number of annual traffic deaths per 10,000 people in the population of each state. For this we first read the data and manipulate the mortality variable.

```
df_fatality = pd.read_csv('/Users/tomba/projects/syllabus_python/data/fatality.csv')
# define the fatality rate
df_fatality['fatal_rate'] =  df_fatality['fatal'] / df_fatality['pop'] * 10000
```

and then run a simple regression:

```python
model_12 = smf.ols('fatal_rate ~ beertax', data = df_fatality).fit(cov_type = "HC0")
model_12.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:             fatal_rate   R-squared:                       0.093
Model:                            OLS   Adj. R-squared:                  0.091
Method:                 Least Squares   F-statistic:                     47.88
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           2.33e-11
Time:                        17:22:13   Log-Likelihood:                -271.04
No. Observations:                 336   AIC:                             546.1
Df Residuals:                     334   BIC:                             553.7
Df Model:                           1
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8533      0.047     39.441      0.000       1.761       1.945
beertax        0.3646      0.053      6.919      0.000       0.261       0.468
==============================================================================
Omnibus:                       66.653   Durbin-Watson:                   0.465
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              112.734
Skew:                           1.134   Prob(JB):                     3.31e-25
Kurtosis:                       4.707   Cond. No.                         2.76
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
"""
```

But these outcomes are very strange. For every dollar increase in tax, number of fatal accidents per 10,000 people increases with 0.36, which is statistically significantly different from 0. What is going on here. Most likely this effect is biased because of omitted variable bias. States in the US differ widely in terms of population density, environment, institutions, religion, poverty, and so on and so forth. And those state characteristics might influence both the variables beertax and fatality.

Fortunately, for each state we have yearly data. In fact for each state we have 7 observations. And we can make use of that by using so-called fixed effects, which is a very common technique in the social sciences—especially in economics. We model the use of fixed effects in this example as follows:

$$\text{fatality}_{it} = \beta_0 + \beta_1 \text{beertax}_{it} + \beta_3 S_1 + ... + \beta_5 1 S_{47} + u_{it}, \quad (3.64)$$

where S_i denote indicator (dummies) for each state which constitute the fixed effects. In total there are 48 states in this dataset, so we have 47 dummies (we have to leave out 1 dummy because of the *dummy trap*—Python does this automatically). Note that these fixed effects only depend on variation over states, not over years. So, essentially what these fixed effects capture is all state **specific** characteristics which are **constant** over time. And most of the characteristics' examples given above do not vary that much over time, so by using these state fixed effects we can **control** for them. In Python you can estimate this in a straightforward way as (but note all these additional coefficients for the US states):

```
model_13 = smf.ols('fatal_rate ~ beertax + C(state)', data = df_fatality).fit(cov_type = "HC0")
model_13.summary()
```

To be sure we use the C() syntax which ensures that the variable inside the parentheses is treated as a categorical variables (and thus creates all these kinds of dummy variables, one for each state).

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:             fatal_rate   R-squared:                       0.905
Model:                            OLS   Adj. R-squared:                  0.889
Method:                 Least Squares   F-statistic:                     147.8
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          1.67e-175
Time:                        17:22:13   Log-Likelihood:                 107.97
No. Observations:                 336   AIC:                            -117.9
Df Residuals:                     287   BIC:                             69.09
Df Model:                          48
Covariance Type:                  HC0
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          3.4776      0.324     10.727      0.000       2.842       4.113
```

```
C(state)[T.ar]    -0.6550      0.222     -2.950      0.003     -1.090     -0.220
C(state)[T.az]    -0.5677      0.273     -2.081      0.037     -1.102     -0.033
C(state)[T.ca]    -1.5095      0.307     -4.911      0.000     -2.112     -0.907
C(state)[T.co]    -1.4843      0.299     -4.957      0.000     -2.071     -0.897
C(state)[T.ct]    -1.8623      0.284     -6.565      0.000     -2.418     -1.306
C(state)[T.de]    -1.3076      0.308     -4.250      0.000     -1.911     -0.705
C(state)[T.fl]    -0.2681      0.134     -2.007      0.045     -0.530     -0.006
C(state)[T.ga]     0.5246      0.165      3.178      0.001      0.201      0.848
C(state)[T.ia]    -1.5439      0.269     -5.746      0.000     -2.071     -1.017
C(state)[T.id]    -0.6690      0.261     -2.566      0.010     -1.180     -0.158
C(state)[T.il]    -1.9616      0.295     -6.656      0.000     -2.539     -1.384
C(state)[T.in]    -1.4615      0.276     -5.303      0.000     -2.002     -0.921
C(state)[T.ks]    -1.2232      0.252     -4.856      0.000     -1.717     -0.729
C(state)[T.ky]    -1.2175      0.292     -4.165      0.000     -1.790     -0.645
C(state)[T.la]    -0.8471      0.207     -4.090      0.000     -1.253     -0.441
C(state)[T.ma]    -2.1097      0.278     -7.582      0.000     -2.655     -1.564
C(state)[T.md]    -1.7064      0.288     -5.925      0.000     -2.271     -1.142
C(state)[T.me]    -1.1079      0.205     -5.410      0.000     -1.509     -0.707
C(state)[T.mi]    -1.4845      0.240     -6.191      0.000     -1.955     -1.015
C(state)[T.mn]    -1.8972      0.267     -7.105      0.000     -2.421     -1.374
C(state)[T.mo]    -1.2963      0.273     -4.749      0.000     -1.831     -0.761
C(state)[T.ms]    -0.0291      0.152     -0.192      0.848     -0.327      0.269
C(state)[T.mt]    -0.3604      0.293     -1.231      0.218     -0.934      0.213
C(state)[T.nc]    -0.2905      0.119     -2.433      0.015     -0.524     -0.056
C(state)[T.nd]    -1.6234      0.286     -5.674      0.000     -2.184     -1.063
C(state)[T.ne]    -1.5222      0.256     -5.944      0.000     -2.024     -1.020
C(state)[T.nh]    -1.2545      0.220     -5.699      0.000     -1.686     -0.823
C(state)[T.nj]    -2.1057      0.311     -6.778      0.000     -2.715     -1.497
C(state)[T.nm]     0.4264      0.258      1.655      0.098     -0.078      0.931
C(state)[T.nv]    -0.6008      0.295     -2.040      0.041     -1.178     -0.023
C(state)[T.ny]    -2.1867      0.302     -7.247      0.000     -2.778     -1.595
C(state)[T.oh]    -1.6744      0.255     -6.570      0.000     -2.174     -1.175
C(state)[T.ok]    -0.5451      0.217     -2.517      0.012     -0.970     -0.121
C(state)[T.or]    -1.1680      0.293     -3.990      0.000     -1.742     -0.594
C(state)[T.pa]    -1.7675      0.279     -6.343      0.000     -2.314     -1.221
C(state)[T.ri]    -2.2651      0.300     -7.546      0.000     -2.853     -1.677
C(state)[T.sc]     0.5572      0.110      5.053      0.000      0.341      0.773
C(state)[T.sd]    -1.0037      0.230     -4.360      0.000     -1.455     -0.553
C(state)[T.tn]    -0.8757      0.274     -3.196      0.001     -1.413     -0.339
C(state)[T.tx]    -0.9175      0.263     -3.486      0.000     -1.433     -0.402
C(state)[T.ut]    -1.1640      0.196     -5.946      0.000     -1.548     -0.780
```

```
C(state)[T.va]     -1.2902      0.205     -6.292      0.000      -1.692      -0.888
C(state)[T.vt]     -0.9660      0.216     -4.467      0.000      -1.390      -0.542
C(state)[T.wa]     -1.6595      0.286     -5.802      0.000      -2.220      -1.099
C(state)[T.wi]     -1.7593      0.297     -5.924      0.000      -2.341      -1.177
C(state)[T.wv]     -0.8968      0.249     -3.596      0.000      -1.385      -0.408
C(state)[T.wy]     -0.2285      0.347     -0.659      0.510      -0.908       0.451
beertax            -0.6559      0.188     -3.491      0.000      -1.024      -0.288
==============================================================================
Omnibus:                       53.045   Durbin-Watson:                   1.517
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              219.863
Skew:                           0.585   Prob(JB):                     1.81e-48
Kurtosis:                       6.786   Cond. No.                         187.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
"""
```

Now, see what happens with the coefficient of the beer tax variable. It changes sign! So from positive it becomes negative. That is how **disruptive** omitted variable bias can be. Also see that by including all these state fixed effects, the $\bar{R}^2$ now increases enormously to 89%, which does make sense because the states explain the variation in fatality rate to a large extent (e.g., compare Georgia (`T.ga`) with New York (`T.ny`)).

This is just a snapshot of the use of fixed effects in panel data, but for now this is enough. The main message you should take home with it that the use of fixed effects can go a long way in addressing omitted variable bias.

3.5 Conclusion and discussion

This chapter dealt with multivariate regression analysis in which I argue that from an applied econometrics perspective the main reason to control for additional variables is to control for omitted variable bias. In the social sciences, this bias is almost always an issue. That is because in the socio-economic domain most phenomena are correlated with each other.

Humans make decisions and base their behavior on decisions and behavior of other humans creating many forms of feedback loops.

Now, this is not to say that we need to include every variable that we have data on. This is only to say that we have think *a priori* about our selection of control variables, preferably in some kind of causal framework.

Moreover, this chapter dealt as well with non-linear transformations of variables, so that more realistic specification can ben estimated. Especially, in economics with its many forms of decreasing (sometimes increasing) returns to scale, it is important to account for concave (the slope keeps positive but gets smaller and smaller) relations.

Finally, I dealt very shortly with the use of fixed effects—in this case still a set of additional dummies to account for omitted variable bias. Fixed effects are *very* often used in economics. Every empirical paper addresses them and in more advanced courses more time and attention will be spent on them.

4 Specification and Assessment Issues

This concluding chapter deals with an epistemological question , namely how to convey your key results, and an ontological question , how do you know whether what you convey is true (e.g., unbiased). To do so, we first look at specification issues in Section 4.1. Which variables should you include? Thereafter, Section 4.2 discusses how to present your results. And, subsequently, Section 4.3 deals with the question about the validity of your results. When do you know that you really obtained estimated an **unbiased** parameter. The final section concludes.

Epistemology says something about what knowledge is and how we communicate that knowledge.

Ontology is more about the nature of being and our own 'beliefs' in that.

4.1 Specification of your model

A long-standing but simple question is how to decide which variables to include in a regression model. Unfortunately, the answer to this question is rather complex. A straightforward but naive approach would be to include them all! So, throw every variable in that is in your database. This, however, leads to "causal salad" (a term coined by McElreath 2020) as displayed in Figure Figure 4.1 and can actually lead to a biased estimator. One reason for this is that if you include a variable that is related to the error term then all other parameters are biased as well.

So, for the final time we return to our Californian school district data and now try to devise a specification that *mimimizes* the changes upon biased estimator. So, our focus is to get an unbiased estimate of the effect on test scores of changing class size, holding constant student and school characteristics but

Figure 4.1: Causal salad

not necessarily holding constant the budget (we do not want to control for budget as this actually governs class sizes).

To do this we need to think about what variables to include and what regressions to run—and we should do this before we actually sit down at the computer. Think beforehand about your model specification and try to avoid throwing everything in (your causal salad).

In practice, and especially economics, most follow the following general approach to variable selection and model specification:

1. First you specify a base or benchmark model. In this case that is the univariate regression of test scores on class size.
2. Then you specify a range of plausible alternative models, which include additional candidate variables.
3. Then you assess whether a candidate variable changes the coefficient of interest ($\hat{\beta}_1$)? You keep focusing on the effect of class size!
4. You assess whether a candidate variable is statistically significantly different from zero; so whether it has an impact of not.

    - Use judgment, not a mechanical recipe, meaning that variable statistically insignificant different from zero should not automatically be thrown out.
    - In all cases, do not just try to maximize $\bar{R}^2$. You focus on identifying a causal effect, not on prediction.

Considering the last point, it is easy to fall into the trap of maximizing the $\bar{R}^2$—but this loses sight of our real objective, an unbiased estimator of the class size effect. Recall that a high $\bar{R}^2$ means that the regressors explain the variation in Y. It does **not** mean

- that you have eliminated omitted variable bias;
- that you have an unbiased estimator of a causal effect (β_1);
- that the included variables are statistically significant.

So, in this case, what variables would you want—ideally—to include to estimate the effect on test scores of STR using school district data? There is a whole set of potential relevant variables in the California class size data set, being:

- student-teacher ratio (STR)—the variable we focus on
- percent English learners in the district ($PctEL$)—as a proxy for large migrant communities
- school expenditures per pupil—largely correlated with student-teacher ratio
- name of the district (so we could look up average rainfall, for example)
- percent eligible for subsidized/free lunch—proxies district income
- percent on public income assistance—proxies district income
- average district income—a measure for district affuency

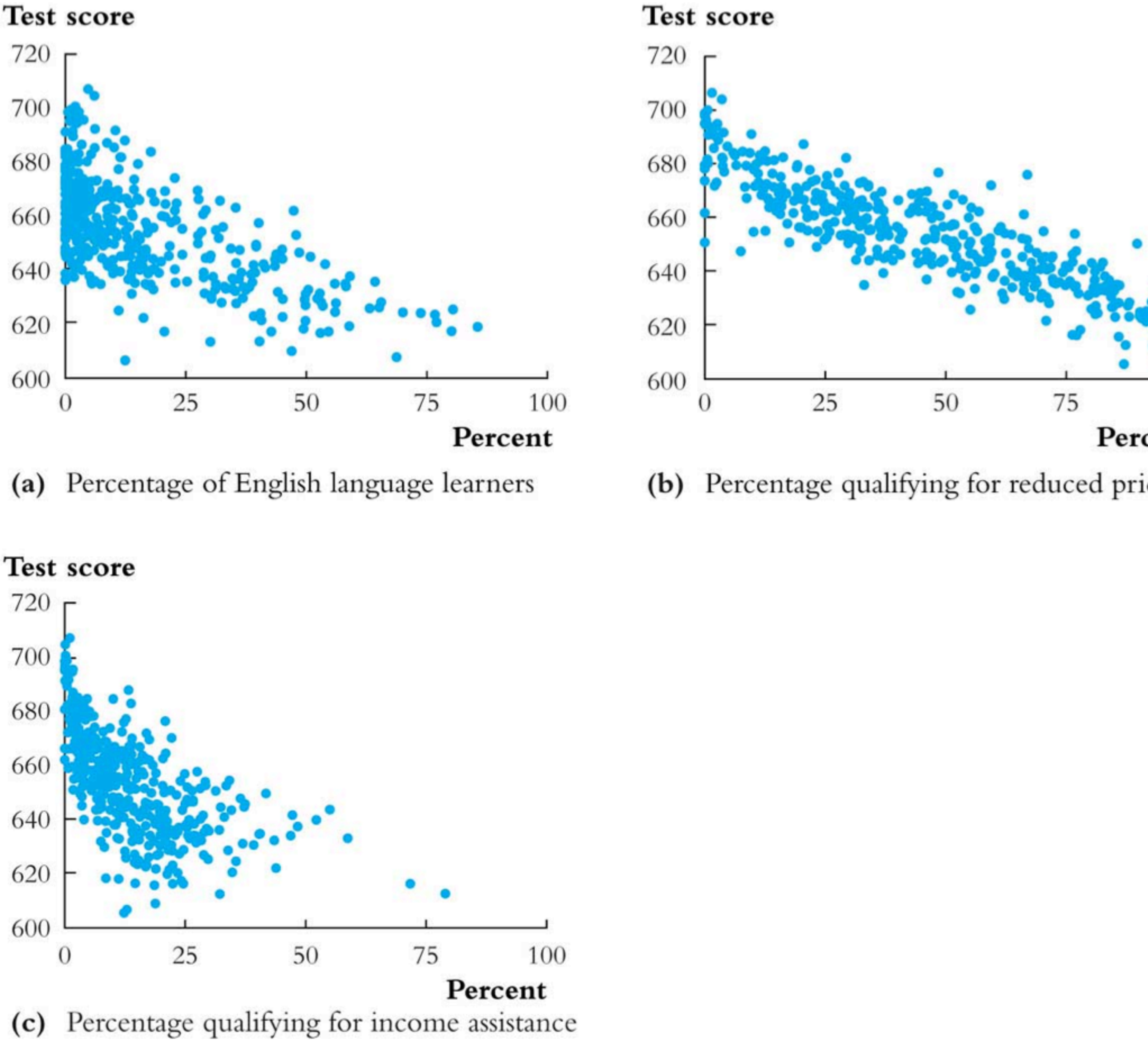
So, which of these variables would you want to include?

Looking at Figure 4.2, all three percentage variables (English learners, subsidized lunch, and income assistance) behave in a similar manner. But interestingly, the strongest relation is between subsidized lunch and test scores and that is at least the variable that we would like to include.

4.2 Presentation of results

So, we have a number of regressions (also called specifications) and we want to report them. Often, it is awkward and difficult to read regressions written out in equation form, so instead it is conventional to report them in a table. Note that reading regression estimates from computer output is even more difficult. On top of that it is ugly and contains way too much information. Try to avoid statistical computer output as much as possible—at least in your thesis. Now, regression tables should include a couple of elements:

- The estimated regression coefficients.
- The standard errors or the t-statistics. Having both of them is too much. Do not report p-values, because often they are not informative (as they often are reported as $p = 0.000$).



(a) Percentage of English language learners

(b) Percentage qualifying for reduced pri

(c) Percentage qualifying for income assistance

Figure 4.2: Test scores versus various independent variables

- Some measures of fit (usually just the $\bar{R}^2$ would do).
- The number of observations.
- Some relevant F-statistics, if any. Usually they are not included.
- Any other pertinent information but typically there is none.

You can find most of this information in the final estimation Table @ref(fig:catable) as presented in Stock, Watson, et al. (2003).

So, here the variable of interest (student-teacher ratio) is the first variable on top. And the table keeps focusing on that one. Moreover, specification (3) and (5) seems to be preferred as they have the highest $\bar{R}^2$, although that is perhaps of lesser importance. What we can infer from this is that the estimate for student-teacher ratio remains robust around -1 and is significantly different from 0. Does this now mean that this effect is **unbiased**? Most likely not, but that is something that the next section will discuss.

4.3 Potential sources of bias

No to include we would like to answer the question whether there is a systematic way to assess regression studies? We already have seen that multivariate regression models have some key virtues:

1. They provide an estimate of the marginal effect of the variable of interest X on Y.
2. They resolve the problem of omitted variable bias, if an omitted variable can be measured and included.
3. They can handle nonlinear relations (effects that vary with the X's) and therefore resolve the problem of misspecification bias.

Still, OLS might yield a **biased** estimator of the true causal effect. In other words, it might not yield valid inferences. That what you want to measure is not what you actually measure. In general there is two ways to assess statistical studies: threats to internal and threats to external validity.

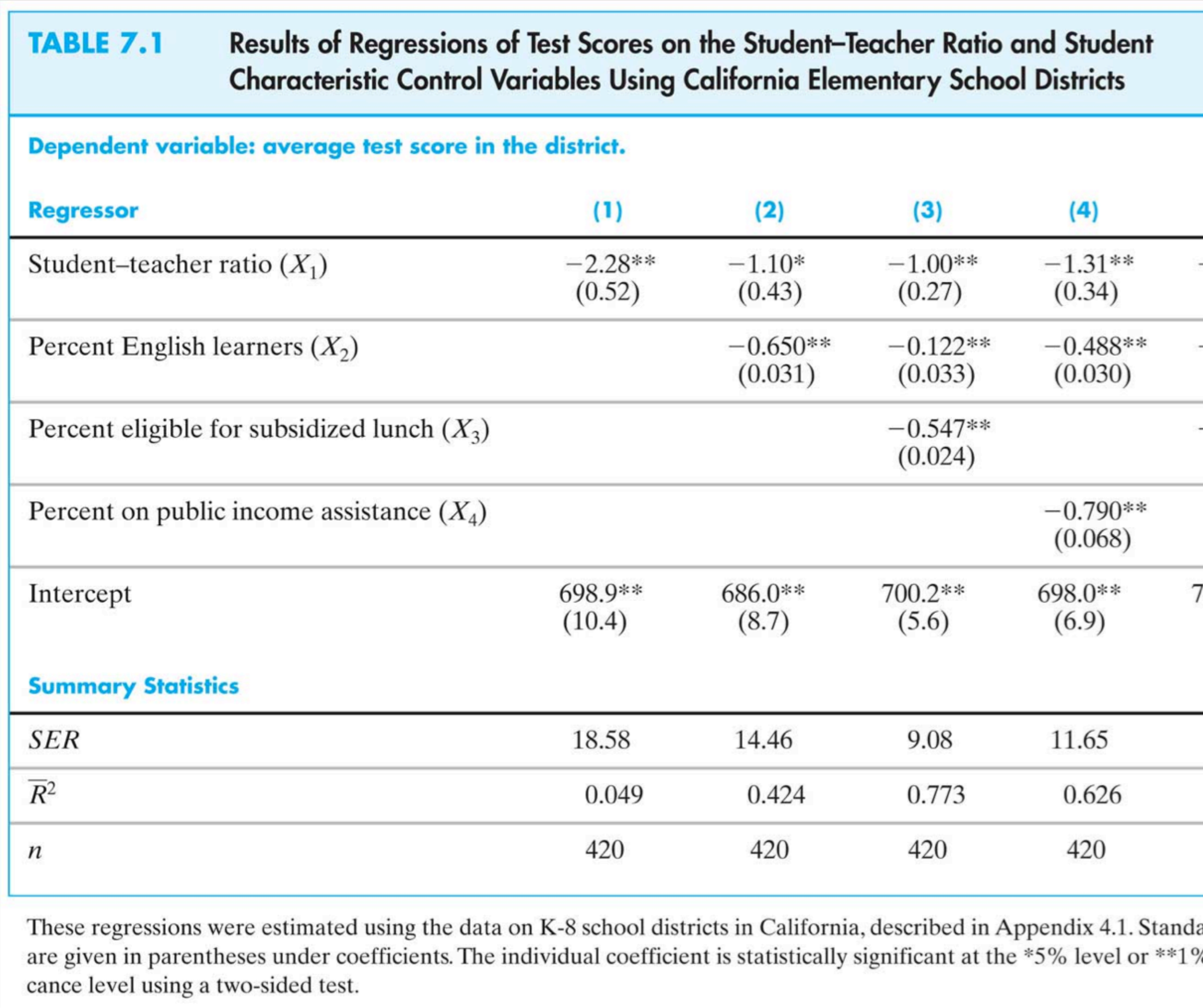


TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)	
Percent on public income assistance (X_4)				−0.790** (0.068)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)
Summary Statistics				
SER	18.58	14.46	9.08	11.65
$\overline{R}^2$	0.049	0.424	0.773	0.626
n	420	420	420	420

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standa are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% cance level using a two-sided test.

Figure 4.3: Various specifications of test score models

- **Internal validity**: the statistical inferences about causal effects are valid for the population being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and setting.

4.3.1 Threats to external validity

So, above we came to a (tentative) conclsusion about the impact of class size on test scores. But we have done so in the context of Californian school districts in the year 2005. Can we extend this finding and generalize class size results from California school districts to other population, for example to that of Massachusetts or Mexico in 2005? And can we do so for differences in institutional settings as there are different legal requirements concerning special education, different treatment of bilingual education, and differences in teacher characteristics across regions and countries.

We therefore always to be careful to transfer our finding to that of other settings. Note that this as well a special case of omitted variable bias but now outside the scope of our study (our population).

4.3.2 Threats to internal validity

In applied econometrics, the following five threats to the internal validity of regression studies are usually given (in statistics there is a different framework for this, but in most cases they come down to the same thing)

1. Omitted variable bias
2. Wrong functional form
3. Errors-in-variables bias or measurement error
4. Sample selection bias
5. Simultaneous causality bias

All of these imply that $E(u_i|X_{1i}, \ldots, X_{ki}) \neq 0$, in which case the OLS estimates are therefore **biased**.

4.3.2.1 Omitted variable bias

Omitted variable bias arises if an omitted variable is both a determinant of Y and a determinant of at least one included regressor. We first discussed omitted variable bias in regression with a single X, but Omitted variable bias will arise when there are multiple X's as well, if the omitted variable satisfies the two conditions above. Fortunately, there are potential solutions to omitted variable bias

- If the variable can be measured, include it as an additional regressor in multiple regression;
- Possibly, use panel data in which each entity (individual) is observed more than once;
- If the variable cannot be measured, use instrumental variables regression (for later courses);
- Run a randomized controlled experiment.

4.3.2.2 Wrong functional form

This threat to internal validity arises if the functional form is incorrect. For example, if an interaction term is incorrectly omitted, then inferences on causal effects will be biased. There is a potential solution to functional form misspecification and that is to use the appropriate nonlinear specifications in X (logarithms, interactions, etc.). Sometimes this is not possible and then one has to resort to direct non-linear estimation techniques.

4.3.2.3 Errors-in-variables bias or measurement error

The third threat is measurement error or sometimes know as Errors-in-variables bias. So far we have assumed that X is measured without error. In reality, (economic) data is often measured with error have measurement error. Especially surveys are prone to measurement error. For example recollection errors that arise with questions as "which month did you start your current job?". Or ambiguous questions problems as "what was your income last year?" What is meant with latter: monthly or yearly income, gross or net income? Also respondents sometimes have an incentive not to answer honestly (intentionally false response problems) with questions as "What is the current value of your financial assets?" or "How often do you drink and drive?". There are potential solutions to errors-in-variables bias, such as

- Obtain better data, but that is bit easy
- Develop a specific model of the measurement error process. This is only possible if a lot is known about the nature of the measurement error—for example a subsample of the data are cross-checked using administrative records and the discrepancies are analyzed and modeled.
- Instrumental variables regression.

4.3.2.4 Sample selection bias

So far we have assumed simple random sampling of the population. In some cases, simple random sampling is thwarted because the sample, in effect, **selects itself**. Now, then sample selection bias arises when a selection process both influences the availability of data and if that process is related to the dependent variable. To illustrate this, I will adopt a hypothetical example given by McElreath (2020). Here we want to look at the relation between trustworthy science and newsworthy science. This example is motivated by the fact that newsworthy science (clickbait in the social media) oftentimes turns out not to be true. To given a reason why this might, we first simulate an artificial database of 400 observations of both newsworthy and trustworthy. Both variable are constructed such that they are *i.i.d.* and standard normally distributed. So, there is no relation whatsoever and, indeed, Figure @ref(fig:Rplot) shows a rather random cloud plot.

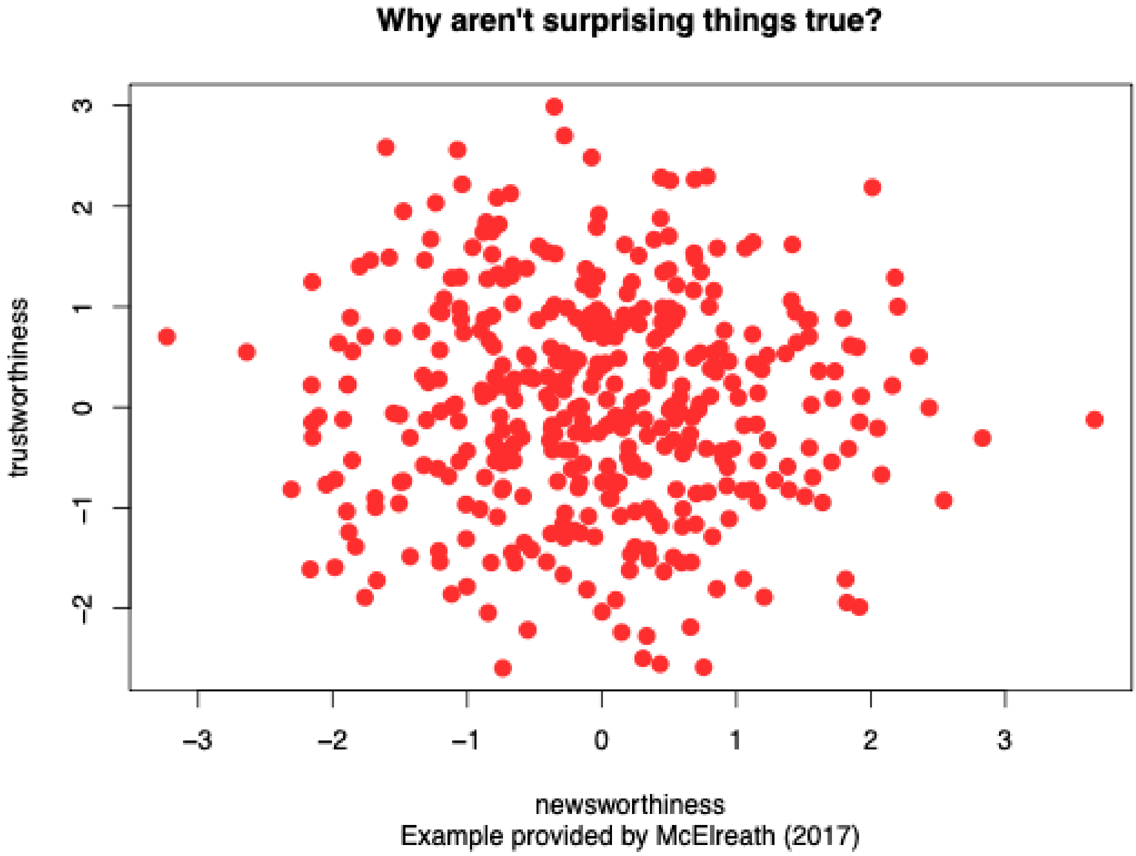


Figure 4.4: Random observations of newsworthy and trustworthy

But what if editors on social media have a decision rule: scientific output should be either thrustworthy or newsworthy, and preferably both. So, as a rule of thumb the select only the top 10% scientific outcomes , so the ones that score in the top 10% when both scores are added up (trustworthy + newsworthy). If we now depict the selected ones in grey in Figure @ref(fig:Rplot01), then clearly suddenly a negative relation emerges between newsworthiness and thrustworthiness. And that negative relation is caused by the selection (external) editors make. So, if there is a selection somewhere in the process, estimates of what you want to estimates can quickly become biased.

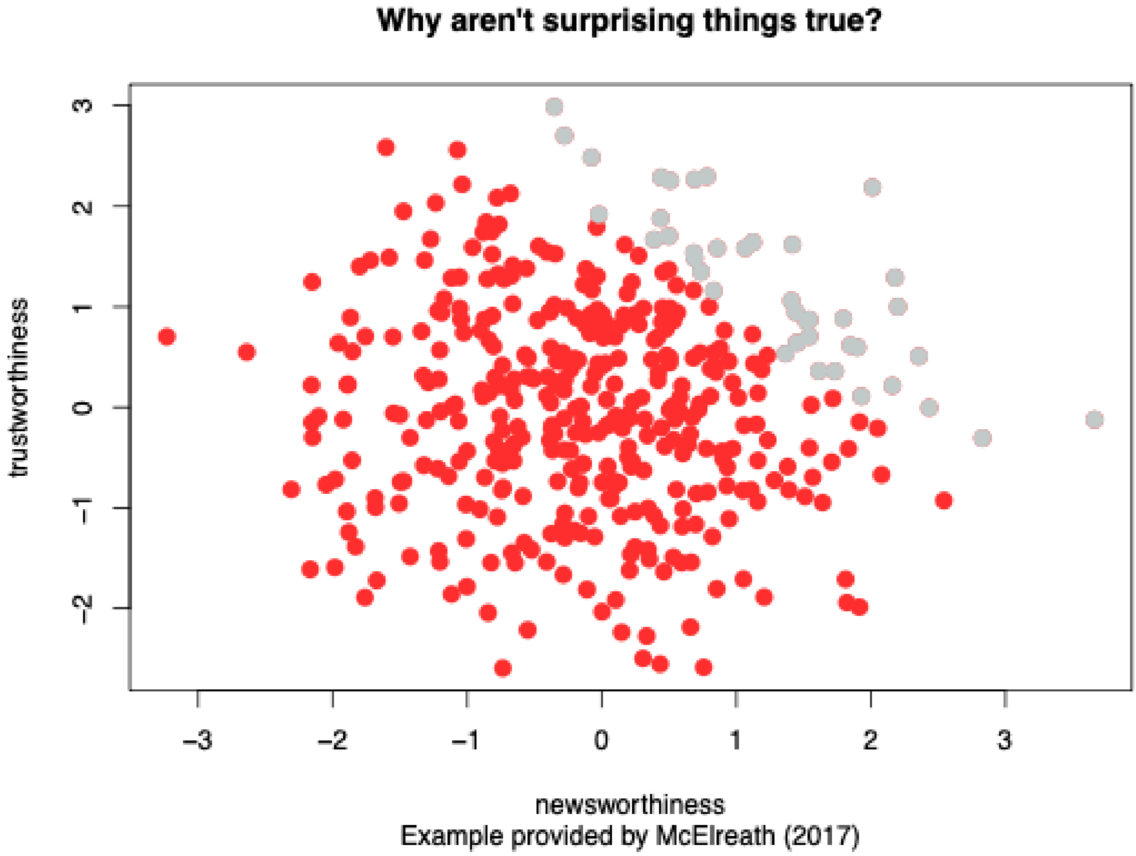


Figure 4.5: Negative relation amongst the selected points

This process occurs more often than you might think. Consider the two following examples:

1. **Aircraft noise externality**. Here the question is to what extent people "value" aircraft noise (that is in a negative sense)? To aim for an answer we adopt the following empirical strategy; we collect housing prices close to Schiphol airport (say Zwanenburg) and compare them with identical houses further away (say Schagen). We have data for individual housing prices (including characteristics) since 1985. As an estimator we assess then the average mean difference between the Zwanenburg and Schagen location. Now the question is whether there is

sample selection bias. And indeed there is and that is caused by the fact that humans react on their own situation based upon their preferences. In this case, they react by means of moving residence. So, those who have strong negative preference regarding aircraft noise are the first to move out Zwanenburg (if possible). So the population in both locations is not identical, instead they sorted spatially.

2. **Returns to education**. The question here is rather straightforward and involved the monetary returns to an additional year of education. As empirical strategy we collect data of all employed workers in the Netherlands (actually this data exists and is called micro-data), including worker characteristics, years of education, and hourly wages. Our approach is here to regress $\ln(Earnings)$ on $YearsEducation$ and a large set of other characteristics. Now, ignore issues of omitted variable bias and measurement error, then the question is: is there sample selection bias? And, indeed, there is again, as you only sample those people who are employed and not the unemployed (they have no current wage). And this leads to a different population than you wanted in the first place.

In there there are some potential solutions to sample selection bias and most of them deal with data issues. For example, you might want to collect the sample in a way that avoids sample selection. For example you might want to focus on those people who moved between Schagen and Zwanenburg or you include the unemployed as well in the returns to education example.

4.3.2.5 Simultaneous causality bias in equations

Finally, our last threat to causality is simultaneous or reverse causality bias. This means that the causal effect might go either way as in the following system

- Causal effect on Y of X: $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Causal effect on X of Y: $X_i = \gamma_0 + \gamma_1 Y_i + v_i$

Where a large u_i means a large Y_i, which implies large X_i (if $\gamma_1 > 0$) and therefore, by definition, $corr(X_i, u_i) \neq 0$. Thus, $\widehat{\beta}_1$ is biased and inconsistent. In our Californian school district example it might as well be that a district with particularly bad test scores given the STR (negative u_i) receives extra resources, thereby lowering its STR; so STR_i and u_i are then correlated

There are some potential solutions to simultaneous causality bias

The first and always the best one is to conduct a randomized controlled experiment. Because,if X_i is chosen at random by the experimenter, there is no feedback from the outcome variable to Y_i (assuming perfect compliance). Secondly, you can develop and estimate a complete model of both directions of causality. This is the idea behind many large macro models (e.g. those of the Federal Reserve Bank in the US). This is difficult in practice. Finally, you can use instrumental variables regression again to estimate the causal effect of interest (effect of X on Y, ignoring effect of Y on X). But that is not for this course.

4.4 Concluding remarks

5 In conclusion

5.1 So, what was this all about?

This syllabus gave an introduction to applied econometrics. As such, it follows in structure closely the first chapters of introductory econometric textbooks such as Stock, Watson, et al. (2003). However, it is not only much more concise, it as well focuses on what I (and colleagues) find important. Thus, it is not that the focus is on learning formula's. Statistical software will know what to do. It focuses much more on thinking about underlying causal mechanisms and on the interpretation of the findings. Moreover, writing my own syllabus also allows me to highlight what I find important in doing empirical research. And one of the most important elements I find is the close connection between a theoretical framework and empirical research. And note though that previous literature can as well provide a theoretical framework.

Also be fully aware of the focus of applied econometrics. It is really about the **identification** of a **causal** effect and not about prediction. In fact, some of the tools in applied econometrics (such as the use of fixed effects) hamper prediction. But know that for good prediction you need to establish a good causal framework as well.

Actually, you as well need a good causal framework for descriptive statistical work.

5.2 And now what?

The course is named *applied* econometrics and I really would like to emphasize the 'appliedness' of it. In the end you get intuition for this but only after doing this multiple times. And not only in this course, but also in other courses and for writing theses. Empirical work is becoming more and more important—not only in economics but also in the other social sciences.

After finishing this course you should have a good understanding of the basics of applied econometrics. More advanced courses always go back go this and especially to the least squares assumptions as defined in Section 2.4. All more advanced techniques are only used because one of these assumptions (typically the first one) are violated. Finally, understanding this also allows you to read and understand most empirical findings as displayed in empirical economic articles.

References

Adams, Douglas. 1995. *Hitchhiker's Guide to the Galaxy (Book 1)*. New York: Del Rey Bonks.

Ahrens, Sönke. 2022. *How to Take Smart Notes: One Simple Technique to Boost Writing, Learning and Thinking*. Sönke Ahrens.

Amrhein, Valentin, Sander Greenland, and Blake McShane. 2019. "Scientists Rise up Against Statistical Significance." Nature Publishing Group.

Galton, Francis. 1886. "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–63.

Levitt, Steven D, and Jack Porter. 2001. "How Dangerous Are Drinking Drivers?" *Journal of Political Economy* 109 (6): 1198–1237.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Popper, Karl. 2005. *The Logic of Scientific Discovery*. Routledge.

Senn, Stephen. 2011. "Francis Galton and Regression to the Mean." *Significance* 8 (3): 124–26.

Stock, James H, Mark W Watson, et al. 2003. *Introduction to Econometrics*. Vol. 104. Addison Wesley Boston.

A Reviewing probability and statistics

```python
# Python packages we need for this chapter
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

This appendix very briefly reviews the statistical knowledge you need for applied econometrics. We assume you had an introductory course in statistics already and will go over this material very quickly.

A.1 Reviewing probability

A.1.1 Probability

As a definition of probability we use the concept of *empirical probability* which is the proportion of time that something (a specific outcome or event X) occurs in the long-run of total events. Usually it is give by:

$$\text{probability} = p = \frac{\text{Number of times specific event } X \text{ happens}}{\text{Total amount of events that can happen}} \tag{A.1}$$

Now probabilities are *defined* by a set of definations (axioms). These are:

1) Probabilities, p, are always between 0 and 1. So, $0 \leq p \leq 1$

2) If something does not happen, then $p = 0$
3) If something always happens, then $p = 1$
4) Probabilities for the total amount of events always add up to 1. So, if the probability that something happens is p, then the probabilities that it will not happen is $1 - p$ (see that $p + 1 - p = 1$)

A.1.2 Population & random variables

In general we see a population as the the group or collection of all possible entities of interest (school districts, inhabitants of the Netherlands, homeowners) and we will think of populations as infinitely large (∞). From this population we then *sample* specific observations. This sample contains then a **random** variable Y, which denotes a characteristics of the entity (district average test score, prices of houses, prices of meat). An important feature is that the specific contents of the sample is unknown, that is before measurement (y), after measurement the sample is know and is called data.

So, a random variable (also called a stochastic variable) is a mathematical formalization of something that depends on **random** outcomes. Unfortunately, randomness is not clearly defined and depends on specific scientific philosophical schools. The scientific philosophical school we implicitly assume in this course—and, in fact, in most statistical courses—is that of frequentist statistics. Here we assume that all things we measure are *intrinsically* random. In fact, this is an *ontological* argument—in other words, what are our beliefs in the state of the world. Because all things we measure are random, every time we measure something our measurements are (slightly) different. However, the more we measure, the more precise we *know* something. But there is still randomness.

In general, there are two types of random variables. First, there are *discrete* random variables, where outcomes can be counted, such as $0, 1, 2, 3, ...$ and *continuous* random variables, where outcomes can be any real number.[1]

[1] There is slightly more to this as fractions such as $\frac{1}{2}$ can in fact be counted as well, and continuous outcomes can be as well complex numbers. But for now we typically see integer numbers as discrete, and real numbers

A.1.3 Distribution functions

Random variables are governed by *distribution functions* which are mathematical functions that provides the probabilities of occurrence of all different possible outcomes of a specific *experiment*[2]: e.g. for a discrete distribution, $f(x) = \Pr(Y = y)$ $\forall y$. Or, in other words, the distribution function maps discrete outcomes to probabilities. For continuous distribution function, this is not possible as there an infinite number of possible outcomes, so that means that for each specific y must yield $\Pr(Y = y) = 0$. Therefore, with continuous distributions, often the *cumulative distribution function* is used, which is defined as $F(x) = \Pr(Y \leq y)$. This is why we always use the surface of areas under the *normal* distribution function.

Distribution functions have characteristics of which the most important are:

- The mean, also known as the expected value (or expectation) of Y. It is usually denoted as $E(Y) = \mu_Y$ and can as well be interpreted as the long-run average value of Y over repeated realizations of Y: $\frac{1}{n}\sum_{i=1}^{n} y_i$
- The variance, which is denoted as $E(Y - \mu_Y)^2$. Usually it is associated with the symbol σ_Y^2 and provides a measure of the squared spread of the distribution. If we take the square root then we have the standard deviation ($= \sqrt{\text{variance}} = \sigma_Y$). For a symmetrical normal distribution, it is useful to know that the mean plus or minus 1 time the standard deviation governs about 2/3 of all probability while the mean plus or minus 2 times the standard deviation governs about 95% of all probability associated with that random variable.

Now, in statistics we are usually related in relations between random variables, and luckily most entities in real life are related. To capture that relation we need two concepts, joint distributions and covariance. If we assume that that random

as continuous.

[2] This could be the throw of a dice but as well the measurement of 10,000 house prices.

variables X and Z have a joint distribution then the covariance between X and Z is:

$$cov(X,Z) = E[(X-\mu_X)(Z-\mu_Z)] = \sigma_{XZ} \tag{A.2}$$

Note that this covariance is a measure of the *linear* association between X and Z and that its units are units of X times units of Z. $cov(X,Z) > 0$ means a positive relation between X and Z, and finally if X and Z are independently distributed, then $cov(X,Z) = 0$. Note that the covariance of a random variable with itself is just its variance:

$$cov(X,X) = E[(X-\mu_X)(X-\mu_X)] = E[(X-\mu_X)^2] = \sigma_X^2 \tag{A.3}$$

However, the covariance is still measured in the units of X and Z. To correct for that, we often use the correlation coefficient, defined by:

$$corr(X,Z) = \frac{cov(X,Z)}{\sqrt{var(X)var(Z)}} = \frac{\sigma_{XZ}}{\sigma_X\sigma_Z} = r_{XZ} \tag{A.4}$$

where $-1 \leq corr(X,Z) \leq 1$, a $corr(X,Z) = 1$ means perfect positive linear association, a $corr(X,Z) = -1$ means perfect negative linear association, and a $corr(X,Z) = 0$ denotes no linear association.

It is very important to notice that a correlation coefficient measures **linear** association. So, $corr(X,Z) = 0$ does not mean that there is no relation, there is only no linear correlation. This is illustrated by Figure A.1. In panel (a) there is clearly a positive relation, and panel (b) shows a negative relation, but what about panel (d)? Here, the correlation coefficient is 0, just as in panel (c), but obviously there is a clear **non-linear** relation.

A.1.4 Conditional distributions and conditional means

An important notion in applied statistics (and in applied econometrics) is that of the conditional distribution, that is the distribution of Y, given value(s) of some other random variable, X. For example, in our California school example, we might want

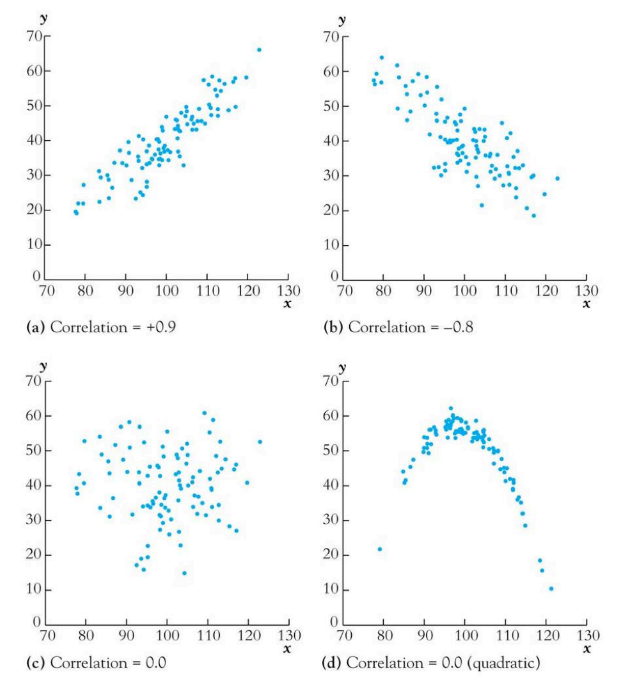


Figure A.1: The correlation coefficient and the relation between observed x and y

to know something about the distribution of test scores, **given** that $STR < 20$. Therefore, we use the concept of conditional mean, which is defined as the mean of a conditional distribution $= E(Y \mid X = x)$. Note here the | symbol—it means the expected value of Y **given** that a random variable X is measured with x. As an example: $E(Testscores \mid STR < 20)$ which denotes the mean of test scores among districts with small class sizes. We also denote this with the *conditional* mean.

Now, if we want to know the difference in means, then we can denote that with

$$\Delta = E(Testscores \mid STR < 20) - E(Testscores \mid STR \geq 20), \tag{A.5}$$

which is a very important concept in applied economics as it resembles two groups of which one received *treatment* and the other one not. Other examples of the use of conditional means: difference in wages among gender (in the possible case of a glass ceiling for females) and mortality rate differences between those who are treated and those who are. Now if $E(X \mid Z)$ is constant, then $corr(X, Z) = 0$. We then say that X and Z are independent.

A.2 Sampling in frequentist statistics

So, we mentioned above that we sample from the population which is assumed to be infinitely large. Now, how does this sampling then carry over to statistics. For that we need a statistical framework based on *random sampling*. First, choose an individual, i, (or district, firm, etc.) at random from the population. Now, prior to sample selection, the value of what we want to know Y_i is random because the individual is **randomly** selected. Once the individual is selected and the value of Y is observed, then Y is just a number—not random anymore but data. And then we say it has the value y. Hence the notation $\Pr(Y = y)$.

If we sample multiple entities, then we can construct a data set that looks like $(y_1, y_2, \ldots, y_n)$, where y_i = value of y for the i^{th} individual (district, entity) sampled. Again the lower case here denotes a realisation—the dataset. Now, we want to know what the distribution of the random variables $Y_1, \ldots, Y_n$ is under simple random sampling. Note that because entities (say individuals) #1 and #2 are selected at random, the value of Y_1 has no information content for Y_2. Thus: Y_1 and Y_2 are independently distributed. And if Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are identically distributed, then we say that, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (*i.i.d.*). More generally, under simple random sampling, Y_i, $i = 1, \ldots, n$, are *i.i.d*—this term always come back in all sorts of statistics.

This simple framework already allows rigorous statistical inferences about, e.g., *the mean* $\bar{Y}$ of population distributions using a sample of data from that population. The next subsection does this because the mean is not only an important statistic, but because the results can be immediately transferred to the regression context as well.

A.2.1 The sampling distribution of $\bar{Y}$

Now because $\bar{Y}$ is formed by a sample of $\{Y_i\}'s$ it is as well a random variable, and its properties are determined by the *sampling distribution* of $\bar{Y}$. Again, we assume that the elements in the sample are drawn at random, that thus the values of $(Y_1, \ldots, Y_n)$ are random, and that thus functions of $(Y_1, \ldots, Y_n)$, such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value. Finally, the distribution of $\bar{Y}$ over different possible samples of size n is called the sampling distribution of $\bar{Y}$, which underpins all of *frequentists* statistics.

A.2.2 Example: simple binomial random variables

So how does this work. Let's take the easiest statistical example: coin flipping, where the coin is this case is notoriously biased. Suppose the random variable Y takes on 0 (head) or 1 (tails) with the following probability distribution, $\Pr[Y = 0] = 0.22$, $\Pr(Y = 1) = 0.78$. Then the mean and variance are given by:

$$\begin{aligned}
\mu_Y &= p \times 1 + (1-p) \times 0 = p = 0.78 \\
\sigma_Y^2 &= E[Y - \mu_Y]^2 = p(1-p) \\
&= 0.78 \times 0.22 = 0.17
\end{aligned} \tag{A.6}$$

But this is only one throw ($throw = 1$). We would like to have multiple observations to derive at our sampling distribution of $\bar{Y}$, which we assume to depend on the number of throws, n.

Consider therefore first the case of $throw = 2$. The sampling distribution of $\bar{Y}$ is,

$$\begin{aligned}
\Pr(\bar{Y} = 0) &= 0.22^2 &= 0.05 \\
\Pr(\bar{Y} = 1/2) &= 2 \times 0.22 \times 0.78 &= 0.34 \\
\Pr(\bar{Y} = 1) &= 0.78^2 &= 0.61.
\end{aligned} \tag{A.7}$$

but this start to become boring as the number of throws increases. Therefore, we turn to `Python`. Let's first check for $throw = 2$.

```
throw = 2
reps = 10000

np.random.seed(42)

# perform random sampling
sample_means = [] # empty array
# fill array with mean of coin throws
for i in range(reps):
    coin_2 = np.random.binomial(throw, 0.78)
    sample_means.append(np.mean(coin_2) / throw)

# Create histogram
plt.hist(sample_means)
# Add vertical line denoting theoretical mean
plt.axvline(x = .78, color = 'red', label = 'Probability tails')
plt.legend()
plt.show()
```

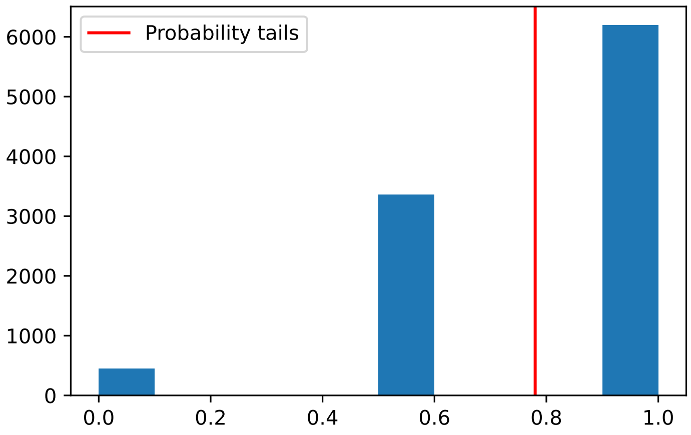


Figure A.2: Sampling distribution when you throw a coin 2 times

The first two lines of this code sets the number of throws (`throws`) and how often I do this (`reps`). So, I throw a coin twice, for 10000 times in a row. The third line sets a random seed for the random number generator, so that every time that I run this code, I get the same answer. The fourth line of code generates an empty array. The for loop calculate the number of heads, which in this case are no heads (0), head once (1), or two heads (2). To arrive at probabilities I divide by the number of throws again (2) again. Finally, the last four lines gives a histogram of the density with a legend and a vertical line showing the probability to throw a tail.

Concerning the number of 42, there is of course nothing special with that number; any number will do. Except that this gives *"the answer to the ultimate question of life, the universe, and everything"* (Adams 1995).

But what if I do this a 1,000 times, so $throw = 1000$?

```
throw = 1000
reps = 10000

# perform random sampling
sample_means = []
for i in range(reps):
    coin_100 = np.random.binomial(throw, 0.78)
    sample_means.append(np.mean(coin_100) / throw)
```

```python
# Create histogram
plt.hist(sample_means, bins = 20)
plt.axvline(x = .78, color = 'red', label = 'Probability tails')
plt.legend()
plt.show()
```

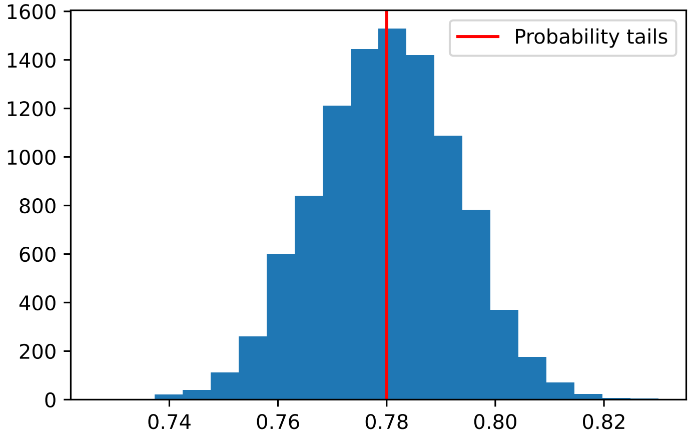


Figure A.3: Sampling distribution when you throw a coin a 1,000 times

The histogram can be seen now in Figure A.3.

But isn't this strange. We can now observe a couple of things. First, the average of the distribution of Figure A.3 is very close to 0.78, which is the actual probability that our biased coin provides tails. But, more importantly the distribution starts to look like a symmetric normal distribution. And we started with a binomial distribution!

This is the result of two amazing statistical theorems:

1. **The law of large numbers**: the average of the results obtained from a large number of trials should be close to the expected value and tends to become closer to the expected value as more trials are performed. That is, if there a no biases in the experiment itself. It also means

that with more experiments the precision become better, or the variance decreases. In general this implies that:

- $\bar{Y}$ is an *unbiased* estimator of μ_Y (that is, $E(\bar{Y}) = \mu_Y$)
- $\text{var}(\bar{Y})$ is *inversely proportional* to n
- The standard error associated with $\bar{Y}$ is $\sqrt{\frac{\sigma_Y^2}{n}}$ (that means that with larger samples there is less uncertainty but see the square-root law)

2. **The Central Limit Theorem**: when independent random variables are summed up[3], their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. So $\bar{Y}$ is approximately distributed $N(\mu_Y, \frac{\sigma_Y^2}{n})$

    - When working with standardized variables then $\bar{Y} = \frac{\bar{Y}-\mu_Y}{\sigma_Y/\sqrt{n}}$ is approximately distributed as $N(0, 1)$

    - The larger is n, the better is the approximation. And this already holds for $n \geq 50$.[4] So with a reasonable amount of observations, the mean of *i.i.d.* variables is normally distributed

[3] Taking the mean is as well a sum but then divided by a constant.

[4] All applied econometrics assumes the number of observations to be larger than 50.