Introduction

In this first introductory chapter, I will lay out the relations 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.

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** the other way. So, when we change Y, X will not necessarily change.

And identifying causal effects is what most applied econometric work nowadays is focused on. And we will focus on that as well in **?@sec-univariateregression**, **?@sec-modeling** and **?@sec-specification**. 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 theory can lead to **multiple** types of 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 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).

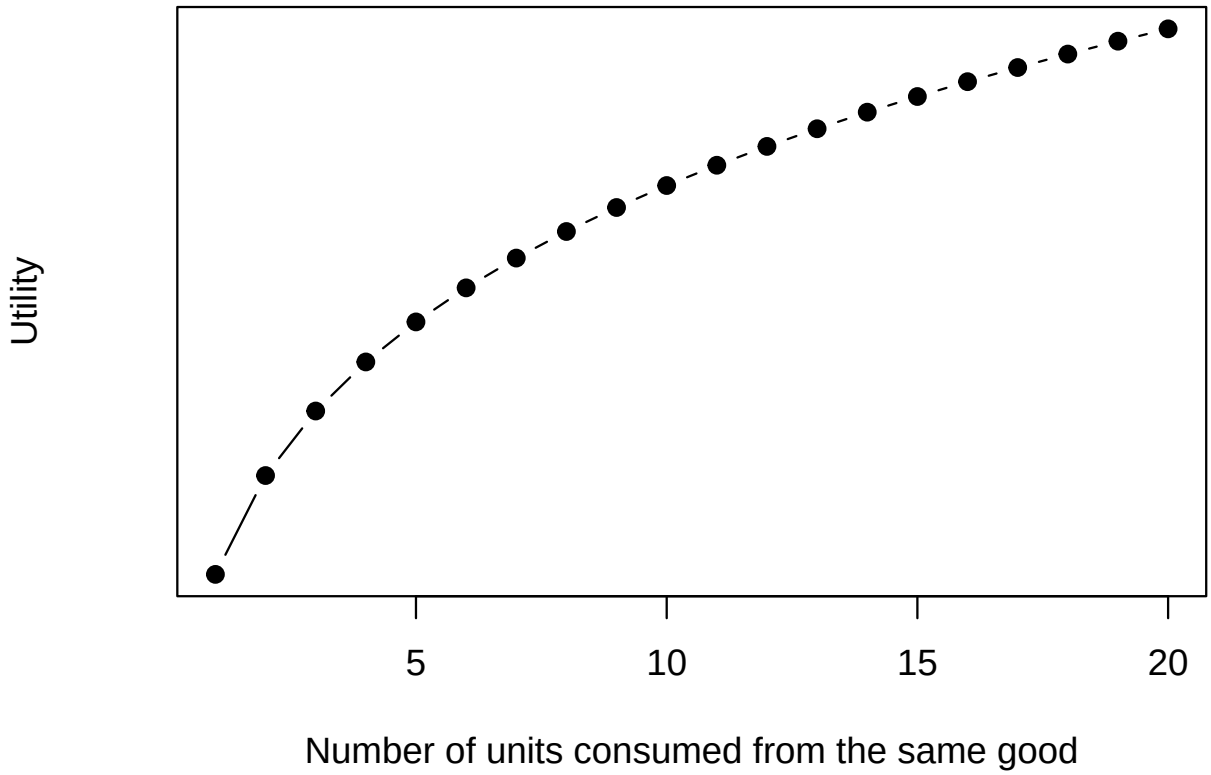

Figure 1: Law of Diminishing Marginal Utility

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?

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 **?@sec-univariateregression**). 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 (we will revisit its basic assumptions in **?@sec-univariateregression**). 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 2).

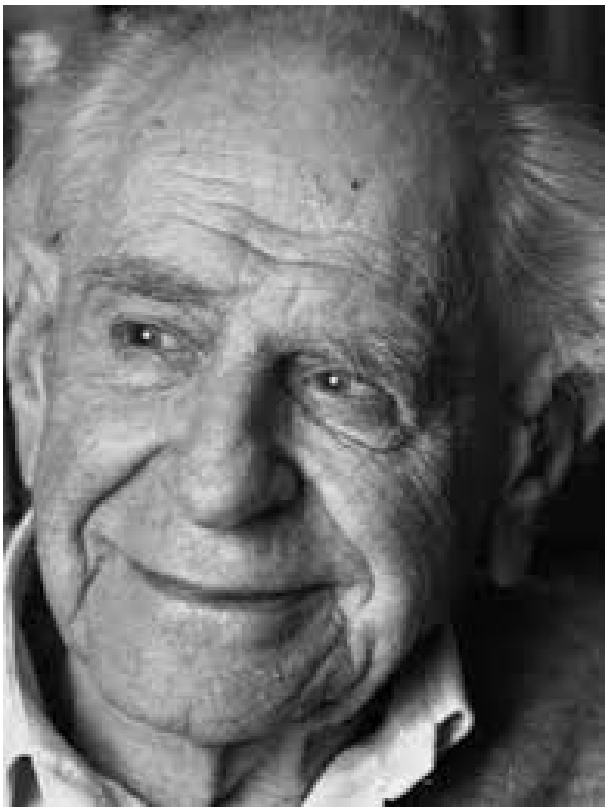

Figure 2: Karl Popper

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 intrinsically *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 **?@sec-univariateregression**)—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!

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.

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. **?@sec-specification** will spent additional time on working tidy and making it looking good.

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 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`).

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.

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

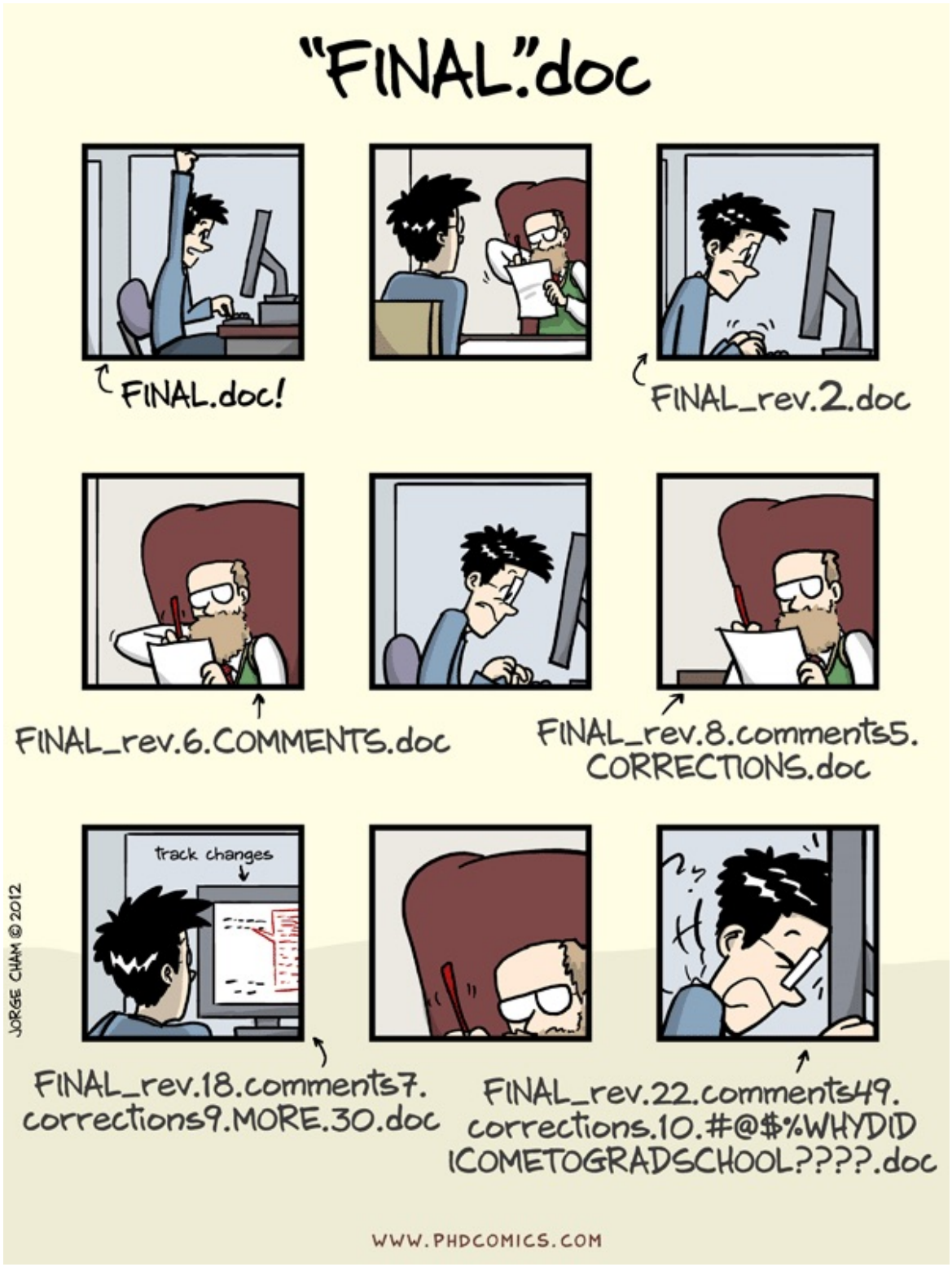

Figure 3: Version confusion

Statistical software

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**. Ideally, an open-source package is used (such as `R` or `Python`), but they have a steep-learning curve and do not work immediately out-of-the-box. The statistical software we use in this course is `STATA` and is more intuitive (compared to `R` or `Python` that is!) and, above all, all economists use it. So, the user base is large and that is important, because for each problem there is much material to be found on internet (including videos). Be aware though that there is one disadvantage in using `STATA` and that is that it is not open-source.

Reading Guide

This course will not concern itself with theory as such, but more with how to test that theory (the applied econometrics). **?@sec-univariateregression** introduces the basic concepts of applied econometrics in the form of univariate regression. **?@sec-modeling** 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. **?@sec-specification** 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.
Popper, Karl. 2005. *The Logic of Scientific Discovery.* Routledge.