Econometrics 1 (ECON 4003)
Suggested Solutions - Tutorial 1

Max Schroder

October 4, 2019

Abstract

This guide is supposed to be complementary to the official solutions supplied by the lecturer. All errors are my own.

Question 1

This question is supposed to get you started to think about how to test economic hypotheses using data. In particular we want to test the hypothesis that training (call it x) makes workers more productive (call it y) *on average.*[1] A bit more generally, we are looking for a *relationship* of the form $y = f(x)$ and/or an *effect* of the form $\frac{\partial y}{\partial x}$. Note also that the data that we have available is firm level data, and in particular describes averages of the productivities and training hours over all the workers in the establishment.[2]

a)

Ceteris Paribus tends to be translated as *"other things being (held) equal",* and is supposed to emulate the ideal type of experiment available to the physical sciences. The question can be put as "if two firms A and B were identical in all respects except for the amount of training they provided to their workers, what would be the difference in measured productivity between A and B?"[3]

[1] Most of the time economic/metric methods are concerned with average outcomes. The real world is messy and sometimes things don't pan out like they do in our neat models, so it's good to remind ourselves that we are mostly looking what happens on average.

[2] Question: Do you think we could do a better job if we had the same data on a worker per worker level?

[3] Note that this formulation falls a little short of the actual idealized experimental standard that supposes that if A and B were **the same firm** except that they differed in their training provision. This is sometimes called a *counterfactual experiment*, but apart from some philosophical thought experiments I have never heard of anyone pulling off a true counterfactual experiment.

b)

This is probably going to depend on the type of training provided - a lot of worker training is related to security (fire drills etc.) which is (hopefully) provided to all workers **independently** of their characteristics - but then again we shouldn't think that fire drills add to worker output, or should we? Generally, however we should expect (meaningful) training provision to vary with some characteristics of the workers in question. Here the question asks you to distinguish between **measurable** and **unmeasurable** characteristics, but I feel that this is not a very good distinction. With the right scale virtually *everything* is measurable (indeed a main job of economists is to quantify all manner of stuff) - take for example ability which is suggested here as an unmeasurable characteristic and think about what an IQ test does. Rather I'd like to suggest that you think of characteristics in terms of **observed** or **unobserved.**[4]

c)

This is obviously a huge problem: if the variation in productivity due to other factors is large, compared to the variation in training, it will be difficult to figure out what effect training has. Remind yourself of the measure of the output available to us (number of nondefect items per worker per hour) and try to imagine that we compare a pencil factory to a company producing heavy duty machines. Surely the output of the pencil factory will be many orders of magnitude larger than that of the capital goods producer, irrespective of the hours of training put in in either firm, simply due to the nature of the product they produce! A way of overcoming this might be to change the measure of productivity (e.g. value added per worker), or compare like with like (i.e. pencil factories with other pencil factories) in an attempt to restore ceteris paribus. However, even under these conditions there might be other factors that account for productivity differentials, such as available technology, machinery, location, etc.

d)

Showing that the data really shows that there exists an effect of the sort that you have hypothesized, is the hardest job of any econometrician.[5] There are many reasons why something that looks like a straightforward case turns out to be very complicated. Some of these reasons are generally know as a class of **endogeneity** problems which come in many varieties. For example it could be the case, that very productive workers simply pick firms where they also receive a lot of training, even though the training doesn't improve their productivity

[4] It is important to point out that these designations refer to what *you,* the econometrician see or don't see. Usually what causes the biggest issues for our job, is if there are characteristics, that are unobserved by the econometrician (i.e. they are not in your data set), but are observed by the actors (i.e. the individuals you are observing them and base their decisions on this information.

[5] In the lingo this is called "identification".

as such (self selection). Or firms with the latest up to date technology also like to put on a lot of training sessions even if they have no added benefit (unobserved characteristics). Or being a productive worker makes individuals seek out training (reverse causality).

There are many more of these issues and if you continue to study econometrics you will doubtlessly come across more of these and (hopefully) find creative ways around them. For now, I encourage you to keep thinking about possible (even outlandish) ways of how it could be the case that what you see in the data is caused by a mechanism that is different from the one you have in mind. You might think it's obvious, but many a seminar participant or referee might disagree.

Question 2

This question is a little more open ended, instead of given parameters you are supposed to think about what you might want to do in order to answer your research question. This might be a good place to point out, that historically econometric analysis was mainly limited to data sources that were available, because they were collected for some specific (administrative) reason, such as census or sales data. As such these early data sets were rarely custom built and usually didn't contain exactly the kinds of variables that we were looking for in order to test our theories. Only more recently economists have branched out into the territory of "field" and "laboratory" experiments (*randomized controlled trials*) where they design and execute purpose built studies in much the same way that drug companies perform medical trials. Even though there are some huge advantages to this approach, there are some drawbacks, regarding the scope and applicability of the experimental methodology (as well as the substantial monetary cost involved).[6] Even more recently, the rise of "*big data*" and associated advances in *machine learning* have opened up new challenges and opportunities, even though most econometricians are still trying to differentiate themselves from (catch up to) the computer sciences.

a)

In short, a perfect experiment is one that as closely as possible reproduces the *ceteris paribus* conditions. So in order to perform the perfect experiment, we should pick a 4th grader (say Jimmy), and clone[7] him say 1,000,000 times and distribute all these Jimmys randomly into classes of different sizes. Then we are also going to take all of Jimmy's teachers and make copies of them to make sure that every class gets taught in the exact same way. Then we are also going to make copies of Jimmy's parents, siblings, friends, neighbors and even strangers

[6] Some malicious tounges suggest that the rise of the RCT has turned economists away from answering "economically meaningful" questions to questions "that can be answered using an RCT"...

[7] More accurately, we should make perfect copies of Jimmy, using a hypothesized duplication machine...

that he might accidentally run into on the streets, just to make absolutely sure that there is no outside influence that might affect one of the different Jimmys in any way that might mess with our ceteris paribus. Indeed we are going to cover each of our model city copies with one of those giant glass bolwes out of the Simpson's movie and use our advanced technology to keep atmospheric conditions equal across all cities down to the sub atomic level. I could go on...

By now you have probably realized that it is often impossible to truly avoid all the potential pitfalls, but luckily most of the time this is not strictly necessary. Many times taking a few steps towards addressing endogeneity issues is enough to get a satisfying answer to your question. Furthermore, sometimes insisting on crystal clear identification can get in the way of answering a relevant economic question. As in the example above whilst this experiment would be doubtlessly praised in the academic community should you pull it off, you can imagine policymakers wondering what they are supposed to do with results that have been obtained in such restrictive circumstances (as well as with the army of Jimmys running rampant).

b)

The reasons for an observed negative correlation between class size and performance might have several origins, and it is good to try and order them in your mind. To get the first out of the way, it might be simply a *spurious correlation:* sometimes random data looks like there is a relationship when there simply isn't one. Indeed a lot of the tools and techniques that you will be learning in this course will help you distinguish between *true* and *spurious* relationships.[8]

The second type are presumably the type of mechanisms that you have in mind and that constitute the hypothesis that you are looking to test. There *is something about* class size that directly affects student performance through for example noise, or limited attention from teachers, etc.

Finally, there are the kind of tricky cases that we have already mentioned. Namely cases, where there appears to be a negative correlation between class size and performance, but instead of being directly caused by class size, the causal channel is something unexpected. For example better teachers might prefer to teach smaller classes, or communities with lower educational budgets might have both larger classes and worse teachers/materials.

c)

So why do we care about all these confounding factors? If you are a little philosophically minded, you might want to say: "It doesn't matter, what the exact causal mechanism is. If the data shows that smaller class sizes are regularly associated with better performance, then that's all there is to causation." However, think about why a policymaker might be interested in figuring out exactly what causal channel affects this result? Think about what your recommendation would be if class size really was the thing that caused students to perform

[8] Or more accurately *significant* and *insignificant* correlations.

better or worse. How would that recommendation change, if you were reasonably certain, that class size is just an indicator for whether a school has a lot of money?

Question 3

Famously there is no such thing as the plural of data, but for an econometrician there is hardly ever enough. The most basic building block of data is the *unit of observation,* which is an object (person, company, country, etc.) from which different measurments are taken. These measures are generally called variables, and each unit of observation can have many different variable values attached to it. For example, if your unit of observation is a person, different variables could be the persons height, age, income, political beliefs, etc. Commonly we can think of data being defined by its informational content across two dimensions: the *cross section* (or ensemble) that describes the observations made on different units of observation (people, countries) at roughly the same point in time. And the *time series* dimension, which describes the behaviour of a single unit of observation across time (e.g. quarterly GDP data). If our data set exhibits both a cross sectional and a time dimension, we speak of a *panel*. Panel data can come in different flavours, but generally we can distinguish the "repeated cross section" - where the units of observation are different at different points in time; and the "longitudinal panel" - where the units of observation are the same at every point in time.

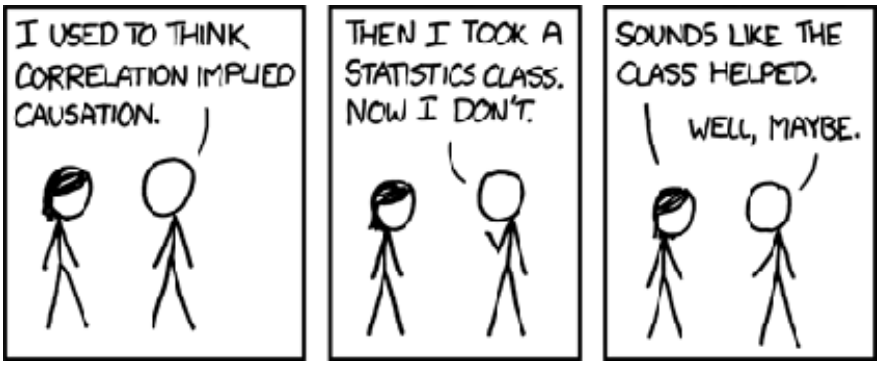

Figure 1: Source: https://xkcd.com/