Econometrics 1: Tutorial I

Heikki Korpela

November 4, 2024

Econometrics 1: Tutorial I

If you haven't yet downloaded RStudio and R (and you have a laptop with you), I suggest you download them now.

You can simply google RStudio, or go to https://posit.co/download/rstudio-desktop/.

Contents

Activities

Activity	Covers	Grading
Problem sets	Theory	No
Videos	Theory	No
Lectures	Problem sets, theory	No
Tutorials	Any questions	No
Homework	Empirical examples, some theory	Yes (40%)
Self-assessment and peer review	Feedback	Yes
Exercise groups	HW solutions	No
Exam	Theory	Yes (60%)

Homework contents

- Homeworks 1–3 will be mostly programming tasks
- Homework 4 is mostly a pen-and-paper task
- The exam will have no coding problems, but will be about understanding the theory and interpreting estimates given to you
- A few reasons for the programming tasks include:
  - Most of you will not end up doing econometric theory, but applied work, so you need to eventually learn some basics
  - The theory will be easier to understand by working with applied examples

Elements of a homework submission

Your submission should contain:

- The main numerical results
- A one or two sentence verbal interpretation for each of the results: what did you conclude, and why
- The relevant code used unless you're confident your results are correct; using R is safest
- You can use whatever editor/environment you like, or even pen and paper
- I would focus on making things easily readable, rather than fancy and pretty

Interpretation of estimates

- Thinking about structural questions (why is X associated with Y) is usually not required
- Safe keywords: 1 unit change in X is associated with a b unit change in Y
- Usually safe: X helps explain or predict Y
- Keywords that need robust justification: X causes Y to change

2. Some R examples to get you started

I will show a few things in R to get you started:

- How to do your first regression in R if you haven't yet
- How to test linear constraints on a model in R
- Any other questions you might have

3. Bias, consistency and precision – but why?

- In this course, we will repeatedly discuss questions like consistency, bias and precision (standard errors)
- Why are these concepts important? Why do we care?
- This part of the tutorial is not required material and you won't be asked questions about this in the exam
- I am trying to motivate why we discuss these concepts

These concepts help us bound our uncertainty

- Suppose we are asked "how much will an additional year of education increase wages on average?"
- Besides a single number – some β euros – we want to be able to assess how robust our answer is, or how much uncertainty there is
- I like to think that there are two layers of uncertainty: quantifiable and unquantifiable
- Given the assumptions of our model, we can perform tests and construct confidence intervals
- But at least some of the assumptions themselves cannot be fully tested (unquantifiable uncertainty)
- If the assumptions are wrong, the confidence interval may grossly overestimate what we know

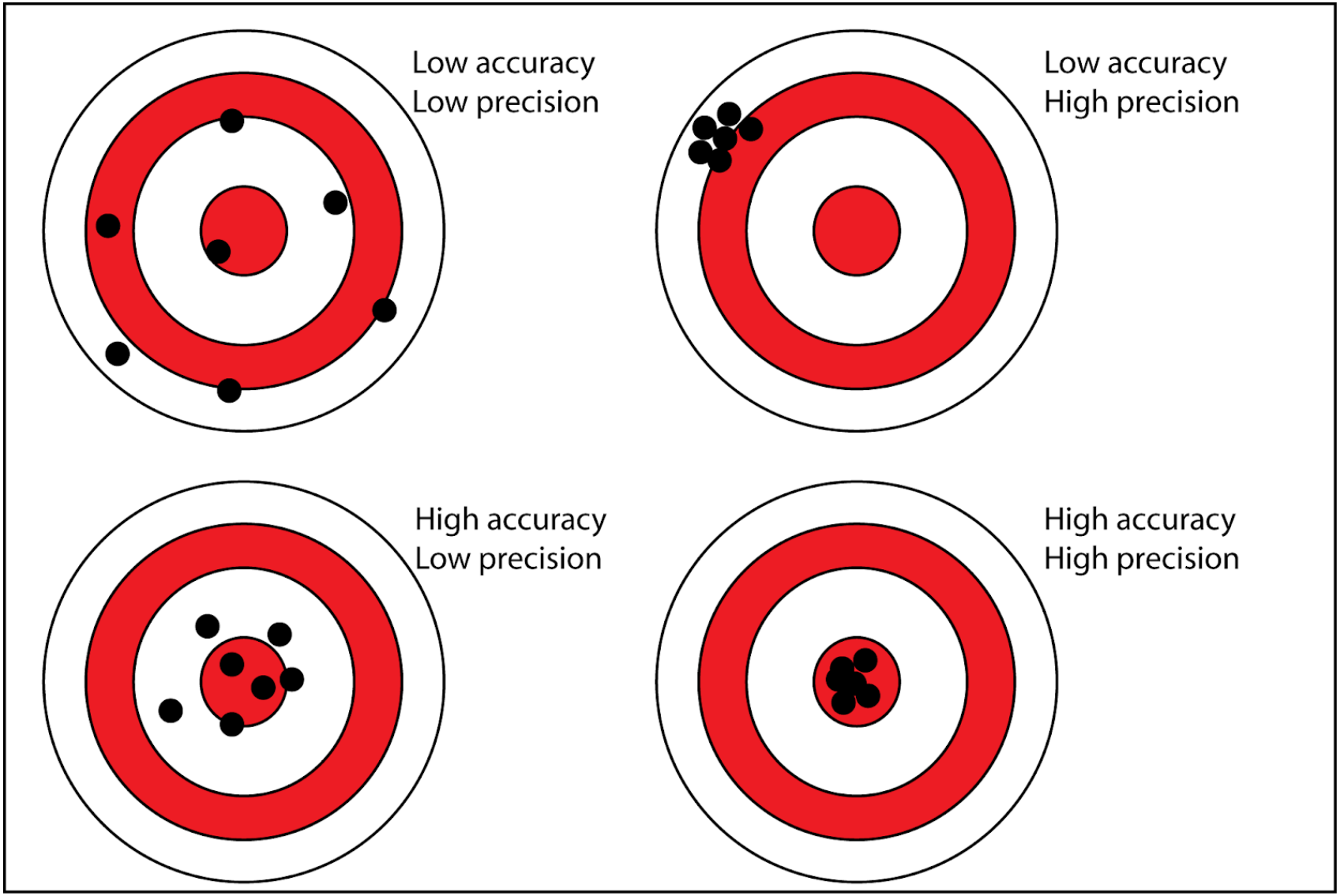
Simple example: election polls

- Election polls are conducted by random sampling
- We can calculate a statistical margin of error that is due to sampling (precision)
- However, the response rate in different surveys might be anything between 1% and 95%
- Suppose citizens who support the purple party systematically cannot be contacted or refuse to answer, but appear otherwise similar to the poll respondents
- Then polls would systematically underestimate the true support for the purple party (bias)

Simple example: election polls

- We cannot use data to estimate the bias in the context of a single poll, and the bias could be very large
- However, with polls we have an objective benchmark case: elections
  - (The election is actually not a perfect benchmark for the poll, although it is often viewed that way)
- That is, we can compare the typical range of differences between (the average of) polls and elections
- As a ballpark figure, the true margin of error in polling (close to an election) is usually roughly double the nominal margin

Accuracy and precision



Picture source: Nate Silver's substack.

Uncertainty about assumptions and quantifiable uncertainty

	Assumptions	Quantifiable uncertainty
Potential problem	Inferences systematically wrong across studies	Average across studies correct, but large variance
Scope can be assessed from data	No	Yes (relying on assumptions!)
Solving requires	Understanding the setting	Statistical theory and methods
Sample size	Does nothing	Improves precision
Controls	Often not helpful	Improves precision
Model search, "tweaking"	Often not helpful	Can improve precision
More plausible assumptions	Can solve the issue	Can decrease precision

Another example: homework in this course

- In 2022, scoring 1% higher on homework 1 was associated with scoring 0.68% (0.10) more points on the exam.
- In 2022, scoring 1% higher on homework overall was associated with scoring 0.66% (0.08) more points on the exam.
- Did doing better with the homework cause students to do better in the exam?
- Hopefully so, but the above does not tell us that: things like motivation, prior ability and available time drive both the homework and the exam results
- To disentangle the actual effect, we'd need something like a randomized controlled trial
- Luckily, studies with good identification strategies usually show flipped learning improves learning outcomes

An economic example: training programs

- Suppose we want to know whether training programs help the unemployed find and keep a job
- In a seminal paper, LaLonde (1986) compared experimental results to an observational assessment on this topic
- The experiments suggested that training was much more effective than the observational studies would find
- Potential source of bias: persons do not enter the training programs randomly

How to get bias: a hypothetical example

- Suppose we have two types of unemployed: 50% are "skilled" (educated, experienced, smart, healthy, motivated, ...), 50% are low-skilled
- The skilled are employed within 6 months with probability $\frac{10}{20}$, the low-skilled with $\frac{7}{20}$
- The training program improves this probability by $\frac{2}{20}$
- Suppose that a half of the low-skilled enter training, increasing their probability to $\frac{9}{20}$
- Now, the trained and untrained jobseekers have the same re-employment probability of $\frac{9}{20}$
- Simply comparing the trained and the untrained would yield a null/zero result (training has no effect), which is wrong

Connection to concepts

- Formally, let x_{i1} indicate participation in the training program, and y_i future employment
- In the specification $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, the unobserved skill resides in the error term ε_i
  - The error term ε_i captures everything in y_i not explained by the predictors X_i
- As only the low-skilled train, ε_i is correlated with x_{i1}
- The estimator $\hat{\beta}_1$ is inconsistent: increasing sample size N indefinitely does make the probability that $\hat{\beta}_{1,N} = \beta_1$ approach 1
- The estimator is biased: repeating the study will systematically (on average) yield the wrong estimate
- The model is misspecified: the error term and the predictor are correlated, and the model lacks an explicit variable for skill

Things that do not solve the problem

- Adding controls? This might extract some of the unobserved skill from the error term, but usually not all of it
  - The issue is that without a direct measure for skill we don't know how much of the bias remains
  - Even IQ is an imperfect measure for "skill"
- Tweaking the model?
  - Polynomials or other transforms of variables other than zskill: will not remove bias (except by chance)
  - Fancier distributional assumptions or estimators: will not remove bias (except by chance)

Things that do solve the problem

- A solid identification strategy (a correctly specified model)
  - An identification strategy is a set of solid assumptions that will yield a consistent estimator
  - Thus, any strategy still relies on assumptions, but those assumptions can be justified and might be much more convincing than the usual OLS assumptions
- Note: different writers may use ”model”, ”assumptions” and ”identification strategy” quite interchangeably
  - This usually becomes clear from the context
- A classic example is the randomized controlled trial (RCT). How does it remove the problem?

Why RCTs work

- The intuition: if training is randomly assigned, then the effects of the program cannot be due to selection
- Formally: if training x_{i1} is randomized properly, it is independent of everything else, including the error term
- RCTs in economics can be costly, impossible, or have their own problems
- Modern strategies usually seek to emulate RCTs in some way, by finding some way to consider x_{i1} as if randomly assigned, for example:
  - Instrumental variables (covered later in this course)
  - Differences-in-differences and regression discontinuity (covered in applied courses)

The identification issue and this course

- We will return to the identification issue in the context of instrumental variables
- In empirical work, one typically has to master both the high-level strategies and more technical aspects
- This course is relatively technical
- Some additional material on identification would include
  - The course Applied Microeconometrics I (not very technical)
  - The books Mostly Harmless and Mastering Metrics (latter is even less technical)
  - For alternative approaches to causality, see Imbens 2020: Potential Outcome and Directed Acyclic Graph Approaches to Causality

Identification and measures of model fit

- A great study can have poor measures of model fit, such as R^2
- Example: you have a randomized controlled trial that shows that an expensive program has no intended effects
- By virtue of the RCT, you only need one variable x_{i1} (the program assignment)
- Because x_{i1} is not correlated with the outcomes, your R^2 will be low, but the null result is extremely robust and valuable
- In applied econometric settings, R^2 is often quite low (and often uninteresting)

Identification and measures of model fit

- In practice, once you have come up with a solid identification strategy, you want the model that best fits your data
- Often, one will simply include all controls that aren't straight out so-called bad controls, and experiment a little with transforms and interactions
- If your identification strategy is sound, you can get an unbiased estimate with just a few predictor variables
- The controls are there to give you precision and one robustness check

Not everything is about causality

- Often, we cannot establish causality
- OLS can still be used to make predictions and to describe the data
- Machine learning can be much more useful here than with coming up with great causal identification
- In such cases, having the nitty-gritty details of the empirical model right becomes more important
- In the past (roughly before 1990's), empirical work was much less relaxed about credible identification
- To read these papers critically, one first needs to understand the limitations of "just running OLS"