What's wrong with how we teach (and then practice) econometrics? What can we do about it?

Arnab Bhattacharjee (Heriot-Watt University & Nat Inst Econ Soc Res)

Mark E Schaffer (Heriot-Watt University, Edinburgh)

Developments in Economics Education Conference
Heriot-Watt University, Edinburgh
2-3 September 2021.

Outline: 3 problems, 3 solutions

Problem 1: Teaching statistical significance and "null hypothesis significance testing" (NHST).

Solution: Follow the statisticians and "embrace uncertainty". **Teach interval estimation and the coverage** as the key learning outcomes.

Problem 2: Teaching causality. When teaching we fail to distinguish sufficiently clearly between predictive inference and causal inference.

Solution: Teach prediction first, and then causal inference.

Problem 3: Disciplinary diversity in Big Data econometrics across the three relevant disciplines - computer science, economics and statistics - is not taught well.

Solution: Explain the different disciplinary approaches using examples. For example: what would happen to output if an economy is hit by a positive 10% demand shock, a negative 10% supply shock and a 5% monetary policy shock?

Problem 1: P-values and ‘Statistical Significance’

P-values, ‘statistical significance’, ‘null hypothesis significance testing’ (NHST)

- Much attention in the applied statistics literature in recent years, most of it critical.
- Economics profession just starting to pick up on this (e.g., JEP Summer 2021 symposium on statistical significance, with contributions from Imbens (2021), Kasy (2021), Miguel (2021).)
- American Statistical Association 2016 “Statement on Statistical Significance and P-Values”: “Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.” (Wasserstein and Lazar, 2016)
- 2019 *Nature* paper by Amrhein et al. (2019), cosigned by over 800 researchers (including one of us): researchers should “retire statistical significance” in favor of more nuanced interpretation.

A typical NHST example

A researcher estimates

$$y_i = x_i \beta + \varepsilon_i \tag{1}$$

usually with some ‘controls’, and then tests the null hypothesis

$$H_0 : \beta = 0 \tag{2}$$

based on the estimated $\hat{\beta}$ and its standard error. If the p-value is less than 5%, the researcher declares victory: β is ‘statistically significant’ and it’s time to write it up and send it off to a journal.

There is a long list of reasons why this is Bad Practice ... and we can’t cover them all.

A typical NHST example (continued)

NHST: a researcher tests

$$H_0 : \beta = 0 \tag{3}$$

based on the estimated $\hat{\beta}$ and its standard error, and if the p-value is less than 5%, the researcher declares victory: β is 'statistically significant'.

First on our list: it almost certainly doesn't help answer any question of economic interest.

As economists, we almost always want to know the answers to 'How big is the effect?' and 'How precisely is it estimated?' Testing $H_0 : \beta = 0$ helps answer neither of these questions.

Say the researcher rejects the null:

- What if $\hat{\beta}$ is extremely small but extremely precisely estimated?
- What if $\hat{\beta}$ is very large but the standard error is also huge?

(It is amazing that so many papers with this mistake still get circulated.)

NHST and teaching econometrics

Most of the wider debate has been about the problems of misuse of statistical significance, p-values, NHST etc. in the practice of research.

But the problem is rife in teaching, as a casual skim of econometrics textbooks will reveal.

(NB: Nostra culpa! Looking at our old teaching materials makes for uncomfortable reading in places.)

An example: gender wage differences

Source: Stock Watson, 4th edition - an excellent textbook.

"To investigate possible gender discrimination in a British firm, a sample of 120 men and 150 women with similar job descriptions are selected at random. [Data detail follows.]

1. What do these data suggest about wage differences in the firm? Do they represent statistically signficant evidence that the average wages of men and women are different?
2. Do these data suggest that the firm in guilty of gender discrimination in compensation policies? Explain."

What's wrong with this?

We report the estimated wage gap (good). But reporting just whether we can reject a zero gap or not is Not Enough.

What if we can reject that the gap is zero but we can't reject £10/year?

How precisely estimated is the gap?

Notes in passing

Imbens (2021) comes out strongly against misuse of statistical significance and p-values, but suggests that "cases commonly arise in economics" where "if a researcher is legitimately interested in assessing a null hypothesis versus an alternative hypothesis".

We think this is less common than he suggests. Of the examples he cites - sheepskin effects, CRS, the EMH, discrimination in the labour market and some others - all but the EMH are better served with an answer that includes "how precisely estimated is it?" (Arbitrage means even miniscule deviations from the EMH can be economically important.)

The CORE *Doing Economics* textbook: "In short: instead of asking students, ‘Is the difference in means statistically significant?’, we encourage students to ask themselves, ‘What do the p-values tell us about the difference in means?’" Same problem. The answer to the question "Is the difference in means different from zero?" is not very informative. Students should be taught to ask "How precisely estimated is the difference in means?" (Nostra culpa again - one of us contributed to the textbook.)

Interval estimation and coverage

What should be done instead?

$$y_i = x_i\beta + \varepsilon_i \tag{4}$$

Our recommendation: **Interval estimation and coverage should be the key teaching outcomes.**

- Report $[\hat{\beta}_{LL}, \hat{\beta}_{UL}]$ as the key estimand - **not** $\hat{\beta}_{OLS}$.
- Teaching interval estimation rather than point estimation as the key estimand automatically emphasises uncertainty.
- "Based on this sample, we estimate the firm's gender wage gap to be [15%, 21%]."
- "Interval estimation" means frequentist confidence intervals (unless it's a Bayesian course).

Teaching coverage

Compare:

- "Based on this sample, we estimate the firm's gender wage gap to be [15%, 21%] based on a 95% confidence interval."
- "Based on this sample, we estimate the firm's gender wage gap to be [1%, 35%] based on a 95% confidence interval."

It's easy to see, and to teach, the difference between these two results: the first estimate is obviously more precise, and the metric is easy to understand.

Compare the Imbens/CORE approach: "At the 5% significance level, we can reject the null hypothesis that there is no discrimination. The p-value is 0.00004% | 4%, respectively." What are students to make of this?

Teaching coverage

To interpret these intervals, students need to understand what "coverage" means. Teaching this concept is easier than it sounds (and **much** easier than teaching p-values).

Definition of coverage: "The probability that a confidence interval contains the true β.

Definition of a 95% confidence interval: An interval estimation method with 95% coverage. In repeated samples, 95% of the estimated intervals will contain the true β.

Need to emphasise to students: (1) **The interval (not the β) is random** - it's based on a sample dataset. (2) 95% coverage applies to the **method**.

Teaching this is easier than it sounds, because there are good analogies available:

- Mystery Ringtoss.
- Existential Pin-the-Tail-on-the-Donkey.
- ... and no doubt others.

Teaching coverage

Mystery Ringtoss:

- We are playing ringtoss, and we get to put the ring exactly where we want it, but we don't see where the stake is. The interval is the ring, and a dataset is a set of clues about where to throw the ring.
- 95% coverage: 95% of the time we play, the stake (the true β) will be inside the ring. But we never find out the result of any particular game.

Existential Pin-the-Tail-on-the-Donkey:

- We are playing Pin-the-Tail-on-the-Donkey. We're blindfolded as usual, but instead of tail with a pin, we're trying to place a ring where the tail goes. Our friends are yelling clues to us as we approach the donkey. The interval is the ring, and a dataset is the clues yelled by our friends.
- 95% coverage: 95% of the time we play, the donkey's tail location (the true β) will be inside the ring. But we never find out the result of any particular game.

Problem 2: Causality and Prediction

"Teaching causal inference is hard. Really hard. You just won't believe how vastly hugely mind-bogglingly hard it is. I mean, you may think it's hard to teach basic OLS, but that's just peanuts to teaching causal inference."

(With apologies and acknowledgements to Douglas Adams)

What's the problem?

- We teach OLS as an estimation procedure as if it automatically generates estimates of causal effects.
- ... but only if we lay on extra assumptions.
- These extra assumptions can be challenging to explain and motivate (zero conditional mean, conditional mean independence).
- Expositions often don't clearly distinguish between OLS as a prediction tool and OLS as a tool for estimation causal effects.
- Typically we teach causal effects first, and then later return to OLS as a tool for prediction and forecasting, where these extra assumptions aren't needed.

An example (time permitting)

Example quiz question: if one OLS coefficient is a biased estimator of a causal effect, what are the implications for the other OLS coefficients from the same estimation?

The way we teach things now, students will often struggle with this one. (Sometimes practicing researchers too.)

But it is a kind of trick question because the short answer is "there are no implications".

Indeed, the coefficients on control variables will often be biased estimates of causal effects, yet it is the inclusion of these controls that may well be what makes the OLS estimator of the causal effect of interest unbiased!

Problem 2: Causality and Prediction

Let's skip straight to our recommendation: **Teach prediction first.**

- The **theoretical assumptions** required for OLS predictive inference are **much easier to teach** ... and to satisfy in practice.
- The key assumption is about **out-of-sample prediction**: the observation for which we are going to predict $\hat{y}$ is drawn from the same sample as our dataset. Easy to motivate/justify.
- **Makes teaching the mechanics of OLS much easier** as well. (Why are we minimising the sum of squared residuals?)
- If students are already comfortable with OLS as a predictive tool, it will **help them understand the estimation of causal effects is hard**. Example: "hospital treatment predicts health status" vs "hospital treatment has a causal effect on health status".
- Teaching prediction first **facilitates introducing time series econometrics and forecasting**.
- Teaching prediction first **facilitates introducing machine learning**. (This will soon be a standard component of u/g courses.)

Problem 3: Disciplinary Diversity in Big Data Econometrics

- Teaching causal inference & structural models well half the job done
- But, how does Econometrics relate to other relevant disciplines?
  - What does Econometrics bring new to Big Data?
  - Computer Science – Discover patterns in (past) data
  - Statistics – Model selection and “true” DGP
  - Econometrics – Counterfactual scenarios potentially not seen yet
- Teaching with examples works well

Large amount of past data available. "What would happen to output if the economy is hit by a positive 10% demand shock, a negative 10% supply shock and a 5% monetary policy shock?"

- Is such a scenario covered in the data?
- Likely not exactly, but potentially approximately
- But, how do we know which periods?
- Need to look through lens of a structural model

Example: Big Data Disciplinary Diversity (Contd.)

- **Econometrics**: Suitable structural model implies a reduced form
- **Statistics**: Shocks can be interpreted as (fns of) reduced form errors
- **Statistics**: Thus, identify past periods “similar” to desired scenario, and quantify deviations (weights)
- **Computer Science**: Study patterns for “similar” periods
  – what would happen to output?

- Whither (structural) econometrics?
  - Wolpin (2013); McFadden (2021).
  - `https://twitter.com/s_delhommer/status/1430666261457899528`
- Whither statistics/inference?
  - For example: Hastie et al. (2019).
- Computer Science/ Machine Learning
  - For example: Chernozhukov et al. (2017).
  - *"Maximum Likelihood is old hat – Machine Learning is the New ML."*

Conclusions

- Traditional ways of teaching econometrics no longer fit for purpose
  - Socio-economic-business context has changed dramatically
  - Multi-disciplinary ethos – Need to speak not only to advances in econometrics, but also allied disciplines
- We provide 3 contexts/problems where this concern is important, and propose solutions
  - Null hypothesis significance testing & *p*-hacking
    – Embrace 'uncertainty', teach interval estimation and coverage
  - Teaching causality & structural models
    – Teach prediction first, then causal inference
  - Disciplinary diversity in Big Data econometrics
    – Use examples to highlight how computer science, statistics and econometrics have different objectives, and how they can be brought together

References I

G.W. Imbens. Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, (3):157–174, 2021.

G.W. Kasy. Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives*, (3):175–192, 2021.

E. Miguel. Evidence on research transparency in economics. *Journal of Economic Perspectives*, (3):193–214, 2021.

Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70 (2):129–133, 2016. doi: 10.1080/00031305.2016.1154108. URL `https://doi.org/10.1080/00031305.2016.1154108`.

Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, (567):305–307, 2019.

References II

K.I. Wolpin. *The Limits of Inference without Theory*. MIT Press, 1 edition, 2013. URL https://cowles.yale.edu/conferences/tjalling-c-koopmans-memorial-lecture/2010.

D.F. McFadden. Epilogue (annals issue: Structural econometrics honoring daniel mcfadden). *Journal of Econometrics*, (1):261–263, 2021.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC Press, 1 edition, 2019.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, (5):261–265, 2017.