The Stata Journal (2009)
9, Number 2, pp. 315–320

A statistician's perspective on "Mostly Harmless Econometrics: An Empiricist's Companion", by Joshua D. Angrist and Jörn-Steffen Pischke

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University
gelman@stat.columbia.edu

Abstract. This article reviews *Mostly Harmless Econometrics: An Empiricist's Companion*, by Joshua D. Angrist and Jörn-Steffen Pischke.

Keywords: gn0046, experimentation, causal inference, ordinary least-squares regression, instrumental variables, difference in differences, fixed effects, discontinuity analyses, mhe

1 Introduction

Joshua D. Angrist and Jörn-Steffen Pischke, two leading researchers on the practical use of instrumental variables in econometrics, have come out with a new book, suitable for students, applied researchers, and researchers such as myself who work in adjacent disciplines such as political science and statistics. The book's eight chapters begin with a conceptual overview of experimentation and causal inference, then cover ordinary least-squares regression, instrumental variables, and a few models such as difference in differences, fixed effects, and discontinuity analyses, which can be viewed as special cases of regression adapted to particular forms of natural experiments. The book closes with some more-technical discussion of quantile regression and calculations of standard errors.

Despite its broad title, the book's coverage is limited to one corner of econometrics. The data structures considered are cross-sectional, with no time series and virtually no panel or time-series cross-sectional data. And, in the world of Angrist and Pischke, the problems of interest are all causal: there is no forecasting, no descriptive modeling, and no theory testing. Finally, when it comes to methods, the book is all about least-squares estimation of regression coefficients: there is little to nothing about nonparametric methods, Bayesian inference, or models other than linear regression.

 gn0046

That said, the corner of econometrics covered by Angrist and Pischke's book is important and perhaps deserves more emphasis, given that much of the theoretical literature in econometrics and statistics focuses on issues of standard errors, robustness, and asymptotics. It can be debated whether *Mostly Harmless Econometrics* is indeed mostly harmless, but it is certainly well adapted to the econometric issues that researchers in labor economics, program evaluation, and political science are concerned with in their applied work.

The book is focused, which has got to be a good thing: it is only 300 pages, and the pages themselves are pretty small. It is written in a conversational style (except for the theorems; more on that below) and is informative and pleasant to read.

I will now give a statistician's impression of the book. My own research is on statistical methods with applications in political science and public health, and I have written books on Bayesian data analysis and applied regression/multilevel modeling. I have done some work on causal inference (including estimating the advantage of incumbency in congressional elections, an example discussed in Angrist and Pischke's book), and I hope this review will be of interest in revealing some of the similarities and differences between statisticians' and econometricians' attitudes.

2 Nothing on model building

The book's perspective is as follows: you want to do a causal inference, you already have an outcome measure and a treatment indicator—in their examples, the outcome is almost always continuous and the treatment is almost always binary—and you also probably have some pretreatment measures. And then you run a regression. Angrist and Pischke explain how direct regression works, and then they discuss how various methods such as instrumental variables and discontinuity analysis can help you make use of any quasi-experimental structure in your data. My main criticism of the book is that, in keeping their sharp focus, the authors spend almost no time discussing model building. They discuss the idea of the conditional expectation function, $E(y \mid x)$, right away, but they seem to imply that the model is prespecified or that the researcher has designed the study so well ahead of time that he or she can simply take all available variables and do nothing to them but throw them in. Realistically, in the many examples of regression analyses I have seen, there can be a lot of transformation and combination of variables.

In particular, the book does not mention interactions at all (except in the special case of their use to create a saturated model with discrete predictors), a point to which I shall return below. From my perspective, these omissions are not a problem, because model building and interactions are standard material in any statistical textbook on regression methods. But I am a little worried that students and researchers in economics might be misled by Angrist and Pischke's conversational yet authoritative tone into thinking that the model and predictors come to the researchers fully formed.

Also here are some things that are *not* in the book, or, at least, words that are entirely missing from the index:

- forecasting
- time series
- Bayes
- mediation
- multilevel (or hierarchical)
- sampling

Again this is no criticism—but I would have liked to have seen a few sentences near the beginning or the end of the book bounding their topics, so that students would have a better sense of what else is out there.

To defend Angrist and Pischke here, I might say that statisticians such as myself are all too concerned about modeling the data, enough so that they (we) shortchange the ultimately more important goal of causal inference. Angrist and Pischke are keeping the focus where it counts. With its focus on practical tips and understanding of models, Angrist and Pischke's book is much closer to my sensibility than Wooldridge (2002), for example, which is of very high quality but is structured much more around theoretical results. And in many ways, there is something refreshing to me about the economists' (and econometricians') focus on estimating one "beta" using all means necessary. On the other hand, I think there is a big gap in practice when there is no discussion of how to set up the model, an implicit assumption that variables are just dumped raw into the regression.

Maybe students of econometrics could read a statistics text (that is, a book more focused on model building) as a supplement.

3 Treatment interactions

In keeping with the econometric and statistical literature on causal inference, Angrist and Pischke spend a bit of time discussing concepts such as the local average treatment effect. The idea is that, in any experiment or observational study, the inference about the treatment effect applies only to the people who could have had the treatment or the control done to them, with various designs and estimation strategies corresponding to different estimands (the effect of the treatment on the treated, and so forth).

All this is important, but it seems funny to me for it to be considered in isolation of the model being fit. In particular, were the treatment effect truly constant across all units, we could just speak of "the treatment effect" without having to specify which cases we are averaging over. Any discussion of particular average treatment effects is relevant because treatment effects vary; that is, the treatment interacts with pretreatment variables.

Given that, I think it can be important to model such interactions. This is done in econometrics (see, for example, Dehejia [2005]), so I am not proposing anything revolutionary here. What I am saying is that in a 300-page book on econometrics, where there is much discussion of average treatment effects, I would like to see some discussion and examples of models with treatment interactions. This is one area where statisticians' more open-ended philosophy of "modeling the data" might have some advantage over econometricians' often laser-like focus on average treatment effects.

4 Which comes first, the causal question or the natural experiment?

The other day, I was discussing Angrist and Pischke's book with a colleague, and I mentioned my struggle with instrumental variables: where do they come from, and doesn't it seem awkward when you see someone studying a causal question and looking around for an instrument?

And my colleague said: No, it goes the other way. What Angrist and his colleagues do is to find the instrument first, and then they go from there. They might see something in the newspaper or hear something on the radio and think, Hey—there is a natural experiment—it could make a good instrument! And then they go from there.

This sounded wrong at first, but now that I think about it, I actually prefer this to the usual presentation of instrumental variables. The "find the instrumental variables first" approach is cleaner: in this story, all causation flows from the instrumental variable, which has various consequences. So if you have a few key researchers such as Angrist keeping their ears open, hearing of instrumental variables, then you will learn some things.

This approach also fits in with a more general approach in which you focus on the direct effect of the instrument. Suppose z is your instrument, T is your treatment, and y is your outcome. So the causal model is $z \to T \to y$. The trick is to think of (T, y) as a joint outcome and to think of the effect of z on each. For example, an increase of 1 in z is associated with an increase of 0.8 in T and an increase of 10 in y. The usual "instrumental variables" summary is just to say the estimated effect of T on y is $10/0.8 = 12.5$, but I would rather just keep it separate and report the effects on T and y separately.

Sometimes the approach of leading with the natural experiment can lead to missteps, as illustrated by Angrist and Pischke's overinterpretation of David Lee's work on incumbency in elections. (In my opinion, Lee is estimating the "incumbent party advantage" rather than the advantage of individual incumbency.) But, generally, it seems like the way to go, much better than the standard approach of starting with a causal goal of interest and then looking around for an instrumental variable. My only complaint with Angrist and Pischke on this point is that they frame the identification problem as one of looking for a good instrument or discontinuity, even though it might make more sense to consider the natural experiment as coming first.

5 Other impressions

Reading this enjoyable book provoked various additional thoughts:

Psychological experimentation: The authors discuss Milgram's famous experiment on obedience to authority but then suggest that it would have been "better left on the drawing board". Why do they say this—because somebody said that the participants of that study might have been upset by it? My impression is that the Milgram experiment was a great contribution and that, as a society, we are better off that it was done. I can see that some people might be skeptical about Milgram's result, but as an empirical researcher, I appreciate those people like Milgram who go to the trouble to get the data that the rest of us analyze.

Thinking in terms of interventions: In introducing the foundations of experimentation and causal inference, I wish Angrist and Pischke would discuss potential interventions as a way to understand causality. For example, they give an example of a "fundamentally unidentified question" that I would actually call a fundamentally *undefined* question. In their example, why not just consider potential interventions such as sending a given kid to first grade at age 5, 6, or 7? The causal inference all flows from this.

Matching and regression: The authors provide a useful discussion of matching and regression and the essential unity of these methods. In a second edition, I recommend that they more clearly distinguish between two different (but related) goals of matching: balance and overlap (for example, see chapter 10 of Gelman and Hill [2007] for some graphs that illustrate these concepts). Also they should make more clear the two-step process: do matching to get comparable groups, and then do regression for further adjustment and for modeling interactions. A casual reader of the book might be left with the unfortunate impression that matching is a competitor to regression rather than a tool for making regression more effective.

Weighting: When discussing weighted regression, it would have been good if Angrist and Pischke had pointed out that if you include as regression predictors the variables that affect the treatment assignment, then there is no need to weight for them in the regression. Weighting is intended to correct for variables that have not been included in the model. Making this point would have unified the book's presentation of weighting, instead of presenting it more as a matter of taste as they do here.

Mathematical style: As a statistician, I trace my descent from R. A. Fisher, who wrote books (notably, *Statistical Methods for Research Workers*) with methods and examples and discussions but no theorems. Statistics books such as Fisher's (and, I hope, mine) have a logical flow but are not in the theorem/proof style. In contrast, Angrist and Pischke, despite their conversational style, here and there slip in pages of theorems and formulas. It does not make a lot of sense to me, but there you go. I suppose it is how econometricians communicate.

Writing style: Throughout, the authors use too many abbreviations for my taste (and, I suspect, for many students' tastes as well). It is a well-written book, but the constant barrage of acronyms disrupts the flow of the book. One or two abbreviations (for example, OLS and IV) are okay, but it gets out of control when they start with the more obscure acronyms such as HIE, CEF, LDV, CIA, OVB, MD, CQF, DD, LIML, ACR, and so forth.

Presentation of results: I like their use of graphs. I would have liked some scatterplots of raw data, but I know that is not economists' style. Overall, I feel the presentation is excellent.

6 Summary

I like *Mostly Harmless Econometrics* a lot and hope it is widely read, not just by economists but also by statisticians, political scientists, and other social researchers. If the book does do any harm, then it would be in misleading readers into thinking that causal inference is a bit too cut-and-dried, a misunderstanding that will be avoided as long as a more traditional book on statistical modeling is read at the same time. Much will be learned by students who go back and forth between the two books and ask their hapless instructors to explain the differences.

7 Acknowledgments

I thank the National Science Foundation for partial support of this research and Jennifer Hill and several others for helpful suggestions.

8 References

Dehejia, R. H. 2005. Program evaluation as a decision problem. *Journal of Econometrics* 125: 141–173.

Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

About the author

Andrew Gelman is a professor of statistics and political science at Columbia University. His books include *Bayesian Data Analysis* (with John B. Carlin, Hal S. Stern, and Donald B. Rubin), *Teaching Statistics: A Bag of Tricks* (with Deborah Nolan), *Data Analysis Using Regression and Multilevel/Hierarchical Models* (with Jennifer Hill), and, most recently, *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do* (with David Park, Boris Shor, Joseph Bafumi, and Jeronimo Cortina).