Joshua D. Angrist and Jorn-Steffen Pischke
Undergraduate econometrics instruction: through our classes, darkly

Article (Published version)
(Refereed)

Undergraduate Econometrics Instruction: Through Our Classes, Darkly

Joshua D. Angrist and Jörn-Steffen Pischke

As the Stones' Age gave way to the computer age, applied econometrics was mostly concerned with estimating the parameters governing broadly targeted theoretical description of the economy. Canonical examples include multi-equation macro models describing economy-wide variables like unemployment and output, and micro models characterizing the choices of individual agents or market-level equilibria. The empirical framework of the 1960s and 1970s typically sought to explain economic outcomes with the aid of a long and diverse list of explanatory variables, but no single variable of special interest.

Much of the contemporary empirical agenda looks to answer specific questions, rather than provide a general understanding of, say, GDP growth. This agenda targets the causal effects of a single factor, such as the effects of immigration on wages or the effects of democracy on GDP growth, often focusing on policy questions like the employment effects of subsidies for small business or the effects of monetary policy. Applied researchers today look for credible strategies to answer such questions.

Empirical economics has changed markedly in recent decades, but, as we document below, econometric instruction has changed little. Market-leading econometrics texts still focus on assumptions and concerns motivated by a model-driven approach to regression, aimed at helping students produce a statistically precise

■ *Joshua D. Angrist is Ford Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Jörn-Steffen Pischke is Professor of Economics, London School of Economics, London, United Kingdom. Their email addresses are angrist@mit.edu and s.pischke@lse.ac.uk.*

account of the processes generating economic outcomes. Much of this material prioritizes technical concerns over conceptual matters. We still see, for example, extended textbook discussions of functional form, distributional assumptions, and how to correct for serial correlation and heteroskedasticity. Yet this instructional edifice is not of primary importance for the modern empirical agenda. At the same time, newer and widely-used tools for causal analysis, like differences-in-differences and regression discontinuity methods, get cursory textbook treatment if they're mentioned at all.

How should changes in our *use* of econometrics change the way we *teach* econometrics?

Our take on this is simple. We start with empirical strategies based on randomized trials and quasi-experimental methods because they provide a template that reveals the challenges of causal inference, and the manner in which econometric tools meet these challenges. We call this framework the *design-based* approach to econometrics because the skills and strategies required to use it successfully are related to research design. This viewpoint leads to our first concrete prescription for instructional change: a revision in the manner in which we teach regression.

Regression should be taught the way it is now most often used: as a tool to control for confounding factors. This approach abandons the traditional regression framework in which all regressors are treated equally. The pedagogical emphasis on statistical efficiency and functional form, along with the sophomoric narrative that sets students off in search of "true models" as defined by a seemingly precise statistical fit, is ready for retirement. Instead, the focus should be on the set of control variables needed to insure that the regression-estimated effect of the economic variable of interest has a causal interpretation.

In addition to a radical revision of regression pedagogy, the exponential growth in economists' use of quasi-experimental methods and randomized trials in pursuit of causal effects should move these tools to center stage in the classroom. The design-based approach emphasizes single-equation instrumental variables estimators, regression-discontinuity methods, and variations on differences-in-differences strategies, while focusing on specific threats to a causal interpretation of the estimates generated by these fundamental tools.

Finally, real empirical work plays a central role in our classes. Econometrics is better taught by example than abstraction.

Causal questions and research design are not the only sort of econometric work that remains relevant. But our experience as teachers and researchers leads us to emphasize these skills in the classroom. For one thing, such skills are now much in demand: Google and Netflix post positions flagged by keywords like causal inference, experimental design, and advertising effectiveness; Facebook's data science team focuses on randomized controlled trials and causal inference; Amazon offers prospective employees a reduced form/causal/program evaluation *track*.[1]

---

[1] See also the descriptions of modern private sector econometric work in Ayres (2007), Brynjolfsson and McAfee (2011), Christian (2012), and Kohavi (2015).

Of course, there's econometrics to be done beyond the applied micro applications of interest to Silicon Valley and the empirical labor economics with which we're personally most engaged. But the tools we favor are foundational for almost any empirical agenda. Professional discussions of signal economic events like the Great Recession and important telecommunications mergers are almost always arguments over causal effects. Likewise, Janet Yellen and the hundreds of researchers who support her at the Fed crave reliable evidence on whether *X* causes *Y*. Purely descriptive research remains important, and there's a role for data-driven forecasting. Applied econometricians have long been engaged in these areas, but these valuable skills are the bread-and-butter of disciplines like statistics and, increasingly, computer science. These endeavors are not where our comparative advantage as economists lies. Econometrics at its best is distinguished from other data sciences by clear causal thinking. This sort of thinking is therefore what we emphasize in our classes.

Following a brief description of the shift toward design-based empirical work, we flesh out the argument for change by considering the foundations of econometric instruction, focusing on old and new approaches to regression. We then look at a collection of classic and contemporary textbooks, and a sample of contemporary reading lists and course outlines. Reading lists in our sample are more likely to cover modern empirical methods than are today's market-leading books. But most courses remain bogged down in boring and obsolete technical material.

Good Times, Bad Times

The exponential growth in economists' use of quasi-experimental methods and randomized trials is documented in Panhans and Singleton (forthcoming). Angrist and Krueger (1999) described an earlier empirical trend for labor economics, but this trend is now seen in applied microeconomic fields more broadly. In an essay on changing empirical work (Angrist and Pischke 2010), we complained about the modern macro research agenda, so we're happy to see recent design-based inroads even in empirical macroeconomics (as described in Fuchs-Schündeln and Hassan 2016). Bowen, Frésard, and Taillard (forthcoming) report on the accelerating adoption of quasi-experimental methods in empirical corporate finance.

Design-based empirical analysis naturally focuses the analyst's attention on the econometric tools featured in this work. A less obvious intellectual consequence of the shift towards design-driven research is a change in the way we use our linear regression workhorse.

Yesterday's Papers (and Today's)

The changed interpretation of regression estimates is exemplified in the contrast between two studies of education production, Summers and Wolfe (1977) and Dale and Krueger (2002). Both papers are concerned with the role of schools in generating human capital: Summers and Wolfe with the effects of elementary

school characteristics on student achievement; Dale and Krueger with the effects of college characteristics on post-graduates' earnings. These questions are similar in nature, but the analyses in the two papers differ sharply.

Summers and Wolfe (1977) interpret their mission to be one of modeling the complex process that generates student achievement. They begin with a general model of education production that includes unspecified student characteristics, teacher characteristics, school inputs, and peer composition. The model is loosely motivated by an appeal to the theory of human capital, but the authors acknowledge that the specifics of how achievement is produced remain mysterious. What stands out in this framework is lack of specificity: the Summers and Wolfe regression puts the change in test scores from 3rd to 6th grade on the left-hand side, with a list of 29 student and school characteristics on the right. This list includes family income, student IQ, sex, and race; the quality of the college attended by the teacher and teacher experience; class size and school enrollment; and measures of peer composition and behavior.

The Summers and Wolfe (1977) paper is true to the 1970s empirical mission, the search for a true model with a large number of explanatory variables:

> We are confident that the coefficients describe in a reasonable way the relationship between achieving and GSES [genetic endowment and socioeconomic status], TQ [teacher quality], SQ [non-teacher school quality], and PG [peer group characteristics], for this collection of 627 elementary school students.

In the spirit of the wide-ranging regression analyses of their times, Summers and Wolfe offer no pride of place to any particular set of variables. At the same time, their narrative interprets regression estimates as capturing causal effects. They draw policy conclusions from empirical results, suggesting, for example, that schools not use the National Teacher Exam score to guide hiring decisions.

This interpretation of regression is in the spirit of Stones' Age econometrics, which typically begins with a linear regression equation meant to describe an economic process, what some would call a "structural relation." Many authors of this Age go on to say that in order to obtain unbiased or consistent estimates, the analyst must assume that regression errors are mean-independent of regressors. But since *all* regressions produce a residual with this orthogonality property, for *any* regressor included in the model, it's hard to see how this statement promotes clear thinking about causal effects.

The Dale and Krueger (2002) investigation likewise begins with a question about schools, asking whether students who attend a more selective college earn more as a result, and, like Summers and Wolfe (1977), uses ordinary least squares regression methods to construct an answer. Yet the analysis here differs in three important ways. The first is a focus on specific causal effects: there's no effort to "explain wages." The Dale and Krueger study compares students who attend more- and less-selective colleges. College quality (measured by schools' average SAT score) is but one factor that might change wages, surely minor in an R^2 sense. This highly

focused inquiry is justified by the fact that the analysis aspires to answer a causal question of concern to students, parents, and policymakers.

The second distinguishing feature is a research strategy meant to eliminate selection bias: Graduates of elite schools undoubtedly earn more (on average) than those who went elsewhere. Given that elite schools select their students carefully, however, it's clear that this difference may reflect selection bias. The Dale and Krueger (2002) paper outlines a selection-on-observables research strategy meant to overcome this central problem.

The Dale and Krueger (2002) research design compares individuals who sent applications to the same set of colleges and received the same admission decisions. Within groups defined by application and admission decisions, students who attend different sorts of schools are far more similar than they would be in an unrestricted sample. The Dale and Krueger study argues that any remaining within-group variation in the selectivity of the school attended is essentially serendipitous—as good as randomly assigned—and therefore unrelated to ability, motivation, family background, and other factors related to intrinsic earnings potential. This argument constitutes the most important *econometric content* of the Dale and Krueger paper.

A third important characteristic of the Dale and Krueger (2002) study is a clear distinction between causes and controls on the right hand side of the regressions at the heart of their study. In the modern paradigm, regressors are not all created equal. Rather, only one variable at a time is seen as having causal effects. All others are controls included in service of this focused causal agenda.[2]

In education production, for example, coefficients on demographic variables and other student characteristics are unlikely to have a clear economic interpretation. For example, what should we make of the coefficient on IQ in the earlier Summers–Wolfe regression? This coefficient reveals only that two measures of intellectual ability—IQ and the dependent variable—are positively correlated after regression-adjusting for other factors. On the other hand, features of the school environment, like class sizes, can sometimes be changed by school administrators. We might indeed want to consider the implications of class size coefficients for education policy.

The modern distinction between causal and control variables on the right-hand side of a regression equation requires more nuanced assumptions than the blanket statement of regressor-error orthogonality that's emblematic of the traditional econometric presentation of regression. This difference in roles between right-hand variables that might be causal and those that are just controls should emerge clearly in the regression stories we tell our students.

Out of Control

The modern econometric paradigm exemplified by Dale and Krueger (2002) treats regression as an empirical control strategy designed to capture causal effects. Specifically, regression is an automated matchmaker that produces within-group

---

[2] We say "one variable at a time," because some of the Dale and Krueger (2002) models replace college selectivity with tuition as the causal variable of interest.

comparisons: there's a single causal variable of interest, while other regressors measure conditions and circumstances that we would like to hold fixed when studying the effects of this cause. By holding the control variables fixed—that is, by including them in a multivariate regression model—we hope to give the regression coefficient on the causal variable a *ceteris paribus*, apples-to-apples interpretation. We tell this story to undergraduates without elaborate mathematics, but the ideas are subtle and our students find them challenging. Detailed empirical examples showing how regression can be used to generate interesting, useful, and surprising causal conclusions help make these ideas clear.

Our instructional version of the Dale and Krueger (2002) application asks whether it pays to attend a private university, Duke, say, instead of a state school like the University of North Carolina. This converts college selectivity into a simpler, binary treatment, so that we can cast the effects of interest as generated by simple on/off comparisons. Specifically, we ask whether the money spent on private college tuition is justified by future earnings gains. This leads to the question of how to use regression to estimate the causal effect of private college attendance on earnings.

For starters, we use notation that distinguishes between cause and control. In this case, the causal regressor is P_i, a dummy variable that indicates attendance at a private college for individual i. Control variables are denoted by X_i, or given other names when specific controls are noteworthy, but in all cases distinct from the privileged causal variable, P_i. The outcome of interest, Y_i, is a measure of earnings roughly 20 years post-enrollment.

The causal relationship between private college attendance and earnings is described in terms of potential outcomes: Y_{1i}, representing the earnings of individual i were he or she to go private ($P_i = 1$), and Y_{0i}, representing i's earnings after a public education ($P_i = 0$). The causal effect of attending a private college for individual i is the difference, $Y_{1i} - Y_{0i}$. This difference can never be seen; rather, we see only Y_{1i} or Y_{0i}, depending on the value of P_i. The analyst's goal is therefore to measure an average causal effect, like $E(Y_{1i} - Y_{0i})$.

At MIT (where we have both taught), we ask our private-college econometrics students to consider their personal counterfactual had they made a public-school choice instead of coming to MIT. Some of our students are seniors who have lined up jobs with the likes of Google and Goldman. Many of the people they work with at these firms—perhaps the majority—will have gone to state schools. In view of this fact, we ask our students to consider whether MIT-style private colleges really make a difference when it comes to career success.

The first contribution of a causal framework based on potential outcomes is to explain why naive comparisons of public and private college graduates are likely to be misleading. The second is to explain how an appropriately constructed regression strategy leads us to something better.

Naive comparisons between alumni of private and public universities will confound the average causal effect of private attendance with selection bias. The selection bias here reflects the fact that students who go to private colleges are, on average, from stronger family backgrounds and probably more motivated and better

prepared for college. These characteristics are reflected in their *potential* earnings, that is, in how much they could earn without the benefit of a private college degree. If those who end up attending private schools had instead attended public schools, they probably would have had higher incomes anyway. This reflects the fact that public and private students have different Y_{0i}'s, on average.

To us, the most natural and useful presentation of regression is as a model of potential outcomes. Write potential earnings in the public college scenario as $Y_{0i} = \alpha + \eta_i$, where α is the mean of Y_{0i}, and η_i is the difference between this potential outcome and its mean. Suppose further that the difference in potential outcomes is a constant, β, so we can write $\beta = Y_{1i} - Y_{0i}$. Putting the pieces together gives a causal model for observed earnings

$$Y_i = \alpha + \beta P_i + \eta_i.$$

Selection bias amounts to the statement that Y_{0i} (potential earnings after going to a public college) and hence η_i depends (in a statistical sense) on P_i, that is, on the type of school one chooses.

The road to a regression-based solution to the problem of selection bias begins with the claim that the analyst has information that can be used to eliminate selection bias, that is, to purge Y_{0i} of its correlation with P_i. In particular, the modern regression modeler postulates a control variable X_i (or perhaps a set of controls). Conditional on this control variable, the private and public earnings comparison is apples-to-apples, at least on average, so those being compared have the same average Y_{0i}'s or η_i's. This *ceteris paribus*-type claim is embodied in the *conditional independence assumption* that ultimately gives regression estimates a causal interpretation:

$$\mathrm{E}(\eta_i | P_i, X_i) = \mathrm{E}(\eta_i | X_i).$$

Notice that this is a weaker and more focused assumption than the traditional presentation, which says that the error term is mean-independent of *all* regressors, that is, $\mathrm{E}(\eta_i | P_i, X_i) = 0$.

In the Dale and Krueger (2002) study, the variable X_i identifies the schools to which the college graduates in the sample had applied and were admitted. The conditional independence assumption says that, having applied to Duke and UNC and having been admitted to both, those who chose to attend Duke have the same earnings potential as those who went to the state school. Although such conditioning does not turn college attendance into a randomized trial, it provides a compelling source of control for the major forces confounding causal inference. Applicants target schools in view of their ambition and willingness to do the required work; admissions offices look carefully at applicant ability.

We close the loop linking causal inference with linear regression by introducing a functional form hypothesis, specifically that the conditional mean of potential earnings when attending a public school is a linear function of X_i. This can be written formally as $\mathrm{E}(\eta_i | X_i) = \gamma X_i$. Econometrics texts fret at length about linearity

and its limitations, but we see such hand-wringing as misplaced. In the Dale and Krueger research design, the controls are a large set of dummies for all possible applicant groups. The key controls in this case come in the form of a saturated model, that is, an exhaustive set of dummies for all possible values of the conditioning variable. Such models are inherently linear. In other cases, we can come as close as we like to the underlying conditional mean function by adding polynomial terms and interactions. When samples are small, we happily use linearity to interpolate, thereby using the data at hand more efficiently. In some of the Dale and Krueger models, for example, dummies for groups of schools are replaced by a linear control for the schools' average selectivity (that is, the average SAT scores of their students).

Combining these three ingredients, constant causal effects, conditional independence, and a linear model for potential outcomes conditional on controls, produces the regression model

$$Y_i = \alpha + \beta P_i + \gamma X_i + e_i,$$

which can be used to construct unbiased and consistent estimates of the causal effect of private school attendance, β. The causal story that brings us to this point reveals what we mean by β and why we're using regression to estimate it.

This final equation looks like many seen in market-leading texts. But this apparent similarity is less helpful than a source of confusion. In our experience, to present this equation and recite assumptions about the correlation of regressors and e_i clouds more than clarifies the basis for causal inference. As far as the control variables go, regressor-residual orthogonality is *assured* rather than assumed; that is, regression algebra makes this happen. At the same time, while the controls are surely uncorrelated with the residuals, it's unlikely that the regression coefficients multiplying the controls have a causal interpretation. We don't imagine that the controls are as good as randomly assigned and we needn't care whether they are. The controls have a job to do: *they are the foundation for the conditional independence claim* that's central to the modern regression framework. Provided the controls make this claim plausible, the coefficient β can be seen as a causal effect.

The modern regression paradigm turns on the notion that the analyst has data on control variables that generate apples-to-apples comparisons for the variable of interest. Dale and Krueger (2002) explain what this means in their study:

> If, conditional on gaining admission, students choose to attend schools for reasons that are independent of [unobserved determinants of earnings] then students who were accepted and rejected by the same set of schools would have the same expected value of [these determinants, the error term in their model]. Consequently, our proposed solution to the school selection problem is to include an unrestricted set of dummy variables indicating groups of students who received the same admissions decisions (i.e., the same combination of acceptances and rejections) from the same set of colleges.

In our analysis of the Dale and Krueger data (reported in Chapter 2 of Angrist and Pischke 2015), estimates from a regression with no controls show a large private school effect of 13.5 log points. This effect shrinks to 8.6 log points after controlling for the student's own SAT scores, his or her family income, and a few more demographic variables. But controlling for the schools to which a student applied and was admitted (using many dummy variables) yields a small and statistically insignificant private school effect of less than 1 percent.

Comparing regression results with increasing numbers of controls in this way—that is, comparing uncontrolled results, results with crude controls, and results with a control variable that more plausibly addresses the issue of selection bias—offers powerful insights. These insights help students understand why the last model is more likely to have a causal interpretation than the first two.

First, we note in discussing these results that the large uncontrolled private differential in wages is apparently driven by selection bias. We learn this from the fact that the raw effect vanishes after controlling for students' precollege attributes, in this case, ambition and ability as reflected in the set of schools a student applies to and qualifies for. Of course, there may still be selection bias in the private–public contrast conditional on these controls. But because the controls are coded from application and admissions decisions that predate college enrollment decisions, they cannot themselves be a consequence of private school attendance. They must be associated with differences in Y_{0i} that generate selection bias. Eliminating these differences, that is, comparing students with similar Y_{0i}'s, is therefore likely to generate private school effects that are less misleading than simpler models omitting these controls.

We also show our students that after conditioning on the application and admissions variables, ability and family background variables in the form of SAT scores and family income are uncorrelated with private school attendance. The finding of a zero private-school return is therefore remarkably insensitive to further control beyond a core set. This argument uses the omitted variables bias formula, which we see as a kind of golden rule for the modern regression practitioner. Our regression estimates reveal robustness to further control that we'd expect to see in a well-run randomized trial.

Using a similar omitted-variables-type argument, we note that even if there are other confounders that we haven't controlled for, those that are positively correlated with private school attendance are likely to be positively correlated with earnings as well. Even if these variables remain omitted, their omission leads the estimates computed with the variables at hand to *overestimate* the private school premium, small as it already is.

Empirical applications like this demonstrate the modern approach to regression, highlighting the nuanced assumptions needed for a causal interpretation of regression parameters.[3] If the conditional independence assumption is violated,

---

[3] In a recent publication, Arcidiacono, Aucejo, and Hotz (2016) use the Dale and Krueger conditioning strategy to estimate causal effects of enrolling at different University of California campuses on graduation and college major.

regression methods fail to uncover causal effects and are likely to be misleading. Otherwise, there's hope for causal inference. Alas, the regression topics that dominate econometrics teaching, including extensive discussions of classical regression assumptions, functional form, multicollinearity, and matters related to statistical inference and efficiency, pale in importance next to this live-or-die fact about regression-based research designs.

Which is not to say that causal inference using regression methods has now been made easy. The question of what makes a good control variable is one of the most challenging in empirical practice. Candidate control variables should be judged by whether they make the conditional independence assumption more plausible, and it's often hard to tell. We therefore discuss many regression examples with our students, all interesting, but some more convincing than others. A particular worry is that not all controls are good controls, even if they're related to both P_i and Y_i. Specific examples and discussion questions—"Should you control for occupation in a wage equation meant to measure the economic returns to schooling?"—illuminate the bad-control issue and therefore warrant time in the classroom (and in our books, Angrist and Pischke 2009, 2015).

Take It or Leave It: Classical Regression Concerns

It is easiest to use the conditional independence assumption to derive a causal regression model when the causal effect is the same for everyone, as assumed above. While this is an attractive simplification for expository purposes, the key result is remarkably general. As long as the regression function is suitably flexible, the regression parameter capturing the causal effect of interest is a weighted average of underlying covariate-specific causal effects. In fact, with discrete controls, regression can be viewed as a matching estimator that automates the estimation of many possibly heterogeneous covariate-specific treatment effects, producing a single weighted average in one easy step.

More generally, linearity of the regression function is best seen as a convenient approximation to possibly nonlinear functional forms. This claim is supported by pioneering theoretical studies such as White (1980a) and Chamberlain (1982). To the best of our knowledge, the first textbook to highlight these properties is Goldberger (1991), a graduate text never in wide use and one rarely seen in undergraduate courses. Angrist (1998), Angrist and Krueger (1999), and our graduate text (Angrist and Pishke 2009) develop the theoretical argument that regression is a matching estimator for average treatment effects (see also Yitzhaki 1996).

An important consequence of this approximation and matchmaking view of regression is that the assumptions behind the textbook linear regression model are both implausible and irrelevant. Heteroskedasticity arises naturally as a result of variation in the closeness between a regression fit and the underlying conditional mean function it approximates. But the fact that the quality of the fit may vary does not obviate the value of regression as a summarizer of economically meaningful causal relationships.

Classical regression assumptions are helpful for the derivation of regression standard errors. They simplify the math and the resulting formula reveals the features of the data that determine statistical precision. This derivation takes little of our class time, however. We don't dwell on statistical tests for the validity of classical assumptions or on generalized least squares fix-ups for their failures. It seems to us that most of what is usually taught on inference in an introductory undergraduate class can be replaced with the phrase "use robust standard errors." With a caution about blind reliance on asymptotic approximations, we suggest our students follow current research practice. As noted by White (1980b) and others, the robust formula addresses the statistical consequences of heteroskedasticity and nonlinearity in cross-sectional data. Autocorrelation in time-series data can similarly be handled by Newey and West (1987) standard errors, while cluster methods address correlation across cross-sectional units or in panel data (Moulton 1986; Arellano 1987; Bertrand, Duflo, and Mullainathan 2004).

In Another Land: Econometrics Texts and Teaching

Traditional econometrics textbooks are thin on empirical examples. In Johnston's (1972) classic text, the first empirical application is a bivariate regression linking road casualties to the number of licensed vehicles. This example focuses on computation, an understandable concern at the time, but Johnston doesn't explain why the relationship between casualties and licenses is interesting or what the estimates might mean. Gujarati's (1978) first empirical example is more substantive, a Cobb–Douglas production function estimated with a few annual observations. Production functions, implicitly causal relationships, are a fundamental building block of economic theory. Gujarati's discussion helpfully interprets magnitudes and considers whether the estimates might be consistent with constant returns to scale. But this application doesn't appear until page 107.

Decades later, real empirical work was still sparse in the leading texts, and the presentation of empirical examples often remained focused on mathematical and statistical technicalities. In an essay published 16 years ago in this journal, Becker and Greene (2001) surveyed econometrics texts and teaching at the turn of the millennium:

> Econometrics and statistics are often taught as branches of mathematics, even when taught in business schools ... the focus in the textbooks and teaching materials is on presenting and explaining theory and technical details with secondary attention given to applications, which are often manufactured to fit the procedure at hand ... applications are rarely based on events reported in financial newspapers, business magazines or scholarly journals in economics.

Following a broader trend towards empiricism in economic research (documented in Hammermesh 2013 and Angrist, Azoulay, Ellison, Hill, and Lu

forthcoming), today's texts are more empirical than those they've replaced. In particular, modern econometrics texts are more likely than those described by Becker and Greene to integrate empirical examples throughout, and often come with access to websites where students can find real economic data for problem sets and practice.

But the news on the textbook front is not all good. Many of today's textbook examples are still contrived or poorly motivated. More disappointing to us than the uneven quality of empirical applications in the contemporary econometrics library is the failure to discuss modern empirical tools. Other than Stock and Watson (2015), which comes closest to embracing the modern agenda, none of the modern undergraduate econometrics texts surveyed below mentions regression-discontinuity methods, for example. Likewise, we see little or no discussion of the threats to validity that might confound differences-in-differences–style policy analysis, even though empirical work of this sort is now ubiquitous. Econometrics texts remain focused on material that's increasingly irrelevant to empirical practice.

To put these and other claims about textbook content on a firmer empirical foundation, we classified the content of 12 books (listed in online Appendix Table A1), six from the 1970s and six currently in wide use. Our list of classics was constructed by identifying 1970s-era editions of the volumes included in Table 1 of Becker and Green (2001), which lists undergraduate textbooks in wide use when they wrote their essay. We bought copies of these older first or second edition books. Our list of classic texts contains Kmenta (1971), Johnston (1972), Pindyck and Rubinfeld (1976), Gujarati (1978), Intriligator (1978), and Kennedy (1979). The divide between graduate and undergraduate books was murkier in the 1970s: unlike today's undergraduate books, some of these older texts use linear algebra. Intriligator (1978), Johnston (1972), and Kmenta (1971) are noticeably more advanced than the other three. We therefore summarize 1970s book content with and without these three included.

Our contemporary texts are the six most often listed books on reading lists found on the Open Syllabus Project website (http://opensyllabusproject.org/). Specifically, our modern market leaders are those found at the top of a list generated by filtering the Project's "syllabus explorer" search engine for "Economics" and then searching for "Econometrics." The resulting list consists of Kennedy (2008), Gujarati and Porter (2009), Stock and Watson (2015), Wooldridge (2016), Dougherty (2016), and Studenmund (2017).[4]

Recognizing that such an endeavor will always be imperfect, we classified book content into the categories shown in Table 1. This scheme covers the vast majority of the material in the books on our list, as well as in many others we've used or read. Our classification scheme also covers three of the tools for which growth in usage appears most impressive in the bibliometric data tabulated by Panhans and Singleton (forthcoming), specifically, instrumental variables, regression-discontinuity methods,

---

[4] These books are also ranked highly in Amazon's econometrics category and (at one edition removed) are market leaders in sales data from Nielsen for 2013 and 2014. Dougherty (2016) is number eight on the list yielded by Open Syllabus, but the sixth book, Hayashi (2000), is clearly a graduate text, and the seventh, Maddala (1977), is not particularly recent.

Table 1

Topic Descriptions

Topic	Which includes …
Bivariate regression	Basic exposition of the bivariate regression model, interpretation of bivariate model parameters
Regression properties	Derivation of estimators, classical linear regression assumptions, mathematical properties of regression estimators like unbiasedness and regression anatomy, the Gauss–Markov Theorem
Regression inference	Derivation of standard errors for coefficients and predicted values, hypothesis testing and confidence intervals, R^2, analysis of variance, discussion and illustration of inferential reasoning
Multivariate regression	General discussion of the multivariate regression model, interpretation of multivariate parameters
Omitted variables bias	Omitted variables bias in regression models
Assumption failures and fix-ups	Discussion of classical assumption failures including heteroskedasticity, serial correlation, non-normality, and stochastic regressors; multicollinearity, inclusion of irrelevant variables, generalized least squares (GLS) fix-ups
Functional form	Discussion of functional form and model parametrization issues including the use of dummy variables, logs on the left and right, limited dependent variable models, other nonlinear regression models
Instrumental variables	Instrumental variables (IV), two-stage least squares (2SLS), and other single equation IV-estimators like limited information maximum likelihood (LIML) and k-class estimators, the use of IV for omitted variables and errors-in-variables problems
Simultaneous equations models	Discussion of multi-equation models and estimators, including identification of simultaneous equation systems and system estimators like seemingly unrelated regressions (SUR) and three-stage least squares (3SLS)
Panel data	Panel techniques and topics, including the definition and estimation of models with fixed and random effects, pooling time series and cross section data, and grouped data
Time series	Time series issues, including distributed lag models, stochastic processes, autoregressive integrated moving average (ARIMA) modeling, vector autoregressions, and unit root tests. This category omits narrow discussions of serial correlation as a violation of classical assumptions
Causal effects	Discussion of causal effects and the causal interpretation of econometric estimates, the purpose and interpretation of randomized experiments, and threats to a causal interpretation of econometric estimates including sample selection issues
Differences-in-differences	Differences-in-differences assumptions and estimators
Regression discontinuity methods	Sharp and fuzzy regression discontinuity designs and estimators

and differences–in–differences estimators.[5] Our classification strategy counts pages devoted to each topic, omitting material in appendices and exercises, and omitting remedial material on mathematics and statistics. Independently, we also counted pages devoted to real empirical examples, that is, presentations of econometric results computed using genuine economic data. This scheme for counting examples omits the many textbook illustrations that use made-up numbers.

Not Fade Away

For the most part, legacy texts have a uniform structure: they begin by introducing a linear model for an economic outcome variable, followed closely by stating that the error term is assumed to be either mean-independent of, or uncorrelated with, regressors. The purpose of this model—whether it is a causal relationship in the sense of describing the consequences of regressor manipulation, a statistical forecasting tool, or a parameterized conditional expectation function—is usually unclear.

The textbook introduction of a linear model with orthogonal or mean-independent errors is typically followed by a list of technical assumptions like homoskedasticity, variable (yet nonstochastic!) regressors, and lack of multicollinearity. These assumptions are used to derive the good statistical properties of the ordinary least squares estimator in the classical linear model: unbiasedness, simple formulas for standard errors, and the Gauss–Markov Theorem, (in which ordinary least squares is shown to be a best linear unbiased estimator, or BLUE). As we report in Table 2, this initial discussion of *Regression properties* consumes an average of 11 to 12 percent of the classic textbooks. *Regression inference*, which usually comes next, gets an average of roughly 13 percent of page space in these traditional books.

The most deeply covered topic in our taxonomy, accounting for about 20 percent of material in the classic textbooks, is *Assumption failures and fix-ups*. This includes diagnostics and first aid for problems like autocorrelation, heteroskedasticity, and multicollinearity. Relief for most of these maladies comes in the form of generalized least squares. Another important topic in legacy texts is *Simultaneous equations models*, consuming 14 percent of page space in the more elementary texts. The percentage given over to orthodox simultaneous equations models rises to 18 percent when the sample includes more advanced texts. Ironically, perhaps, *Assumption failures and fix-ups* claims an even larger share of the classics when more advanced books are excluded. These older books also devote considerable space to *Time series*, while *Panel data* get little attention across the board.

A striking feature of Table 2 is how similar the distribution of topic coverage in contemporary market leading econometrics texts is to the distribution in the classics. As in the Stones' Age, well over half of the material in contemporary texts is concerned with *Regression properties*, *Regression inference*, *Functional form*, and

---

[5] Panhans and Singleton (forthcoming) also document growth in the number of articles using the terms "natural experiment" and "randomized control trial."

Table 2

Topics Coverage in Econometrics Texts, Classic and Contemporary

(page counts as percentage)

Topic	1970s (1)	1970s excluding more-advanced texts (2)	Contemporary (3)
Bivariate regression	2.5	3.6	2.8
Regression properties	10.9	11.9	9.9
Regression inference	13.2	13.3	14.6
Multivariate regression	3.7	3.7	6.4
Omitted variables bias	0.6	0.5	1.8
Assumption failures and fix-ups	18.4	22.2	16.0
Functional form	10.2	9.3	15.0
Instrumental variables	7.4	5.1	6.2
Simultaneous equations models	17.5	13.9	3.6
Panel data	2.7	0.7	4.4
Time series	12.3	15.2	15.6
Causal effects	0.7	0.7	3.0
Differences-in-differences	–	–	0.5
Regression discontinuity methods	–	–	0.1
Empirical examples	14.0	15.0	24.4

Note: We classified the content of 12 econometrics texts, six from the 1970s and six currently in wide use (see text for details): Our classic texts are Kmenta (1971), Johnston (1972), Pindyck and Rubinfeld (1976), Gujarati (1978), Intriligator (1978), and Kennedy (1979). Our contemporary texts are Kennedy (2008), Gujarati and Porter (2009), Stock and Watson (2015), Wooldridge (2016), Dougherty (2016), and Studenmund (2017). We report percentages of page counts by topic. All topics sum to 100 percent. Empirical examples are as a percentage of the whole book. Column 2 excludes Kmenta (1971), Johnston (1972), and Intriligator (1978), the more advanced classic econometrics texts. Dashes indicate no coverage.

Assumption failures and fix-ups. The clearest change across book generations is the reduced space allocated to *Simultaneous equations models.* This presumably reflects declining use of an orthodox multi-equation framework, especially in macroeconomics. The reduced coverage of *Simultaneous equations* has made space for modest attention to *Panel data* and *Causal effects,* but the biggest single expansion has been in the coverage of *Functional form* (mostly discrete choice and limited dependent variable models).

Some of the volumes on our current book list have been through many editions, with first editions published in the Stones' Age. It's perhaps unsurprising that the topic distribution in Gujarati and Porter (2009) looks a lot like that in Gujarati (1978). But more recent entrants to the textbook market also deviate little from the classic template. On the positive side, recent market entrants are more likely to at least mention modern topics.

The bottom row of Table 2 reveals the moderate use of empirical examples in the Stones' Age: about 15 percent of pages in the classics are devoted to illustrations

involving real data. This average conceals a fair bit of variation, ranging from zero (no examples at all) to more than one-third of page space covering applications. Remarkably, the most empirically oriented textbook in our 12-book sample remains Pindyck and Rubinfeld (1976), one of the classics. Although the field has moved to an average empirical content of over 24 percent, no contemporary text on this list quite matches their coverage of examples.[6]

BLUE Turns to Grey: Econometrics Course Coverage

Many econometrics instructors rely heavily on their lecture notes, using textbooks only as a supplement or a source of exercises. We might therefore see more of the modern empirical paradigm in course outlines and reading lists than we see in textbooks. To explore this possibility, we collected syllabuses and lecture schedules for undergraduate econometrics courses from a wide variety of colleges and universities.[7]

Our sampling frame for the syllabus study covers the ten largest campuses in each of eight types of institutions. The eight groups are research universities (very high activity), research universities (high activity), doctoral/research universities, and baccalaureate colleges, with each of these four split into public and private schools. The resulting sample includes diverse institutions like Ohio State University, New York University, Harvard University, East Carolina University, American University, US Military Academy, Texas Christian University, Calvin College, and Hope College. We managed to collect syllabuses from 38 of these 80 schools. Each of the eight types of schools we targeted is represented in the sample, but larger and more prestigious institutions are overrepresented. Most syllabuses are for courses taught since 2014, but the oldest is from 2009. A few schools contribute more than one syllabus, but these are averaged so each school contributes only one observation to our tabulations. The appendix available with this paper at http://e-jep.org lists the 38 schools included in the syllabus dataset.

For each school contributing course information, we recorded whether the topics listed in Table 1 are covered. A subset of schools also provided detailed lecture-by-lecture schedules that show the time devoted to each topic. It's worth noting that the amount of information that can be gleaned from reading lists and course schedules varies across courses. For example, most syllabuses cover material we've classified as *Multivariate regression*, but some don't list *Regression inference* separately, presumably covering inference as part of the regression module without spelling this out on the reading list. As a result, broader topics appear to get more coverage.

With this caveat in mind, the first column of Table 3 suggests a distribution of econometric lecture time that has much in common with the topic distribution in textbooks. In particular, well over half of class time goes to lectures on *Regression*

---

[6]The average is pulled down by the fact that one book on the list has no empirical content. Our view of how a contemporary undergraduate econometrics text can be structured around empirical examples is reflected in our book, Angrist and Pischke (2015).
[7]Our thanks to Enrico Moretti for suggesting a syllabus inquiry.

Table 3
Course Coverage

	Lecture time (percent)	Courses covering topic (percent)
Topic		
Bivariate regression	11.7	100.0
Regression properties	8.7	43.4
Regression inference	12.4	92.1
Multivariate regression	10.5	94.7
Omitted variables bias	1.9	28.5
Assumption failures and fix-ups	20.2	73.7
Functional form	15.7	92.1
Instrumental variables	3.9	51.8
Simultaneous equations models	0.4	19.3
Panel data	3.6	36.8
Time series	5.0	45.6
Causal effects	2.5	25.4
Differences-in-differences	2.0	27.2
Regression discontinuity methods	1.4	16.7
Number of institutions	15	38

Notes: The first column reports the percentage of class time devoted to each topic listed at left for the 15 schools for which we obtained a detailed schedule. This column sums to 100 percent. Column 2 reports the percentage of courses covering particular topics for the 38 schools for which we obtained a reading list.

properties, *Regression inference*, *Assumption failures and fix-ups*, and *Functional form*. Consistent with this distribution, the second column in the table reveals that, except for *Regression properties*, these topics are covered by most reading lists. The *Regression properties* topic is very likely covered under other regression headings.

Also paralleling the textbook material described in Table 2, our tabulation of lecture time shows that just under 6 percent of course schedules is devoted to coverage of topics related to *Causal effects*, *Differences-in-differences*, and *Regression discontinuity methods*. This is only a modest step beyond the modern textbook average of 3.6 percent for this set of topics. Single-equation *Instrumental variables* methods get only 3.9 percent of lecture time, less than we see in the average for textbooks, both old and new.

Always looking on the bright side of life, we happily note that Table 3 shows that over a quarter of our sampled instructors allocate at least some lecture time to *Causal effects* and *Differences-in-differences*. A healthy minority (nearly 17 percent) also find time for at least some discussion of *Regression discontinuity methods*. This suggests that econometric instructors are ahead of the econometrics book market. Many younger instructors will have used modern empirical methods in their PhD work, so they probably want to share this material with their students. Textbook authors are probably older, on average, than instructors, and therefore less likely to have personal experience with tools emphasized by the modern causal agenda.

Out of Time

Undergraduate econometrics instructions is overdue for a paradigm shift in three directions. One is a focus on causal questions and empirical examples, rather than models and math. Another is a revision of the anachronistic classical regression framework away from the multivariate modeling of economic processes and towards controlled statistical comparisons. The third is an emphasis on modern quasi-experimental tools.

We recognize that change is hard. Our own reading lists of a decade or so ago look much like those we've summarized here. But our approach to instruction has evolved as we've confronted the disturbing gap between what we do and what we teach. The econometrics we use in our research is interesting, relevant, and satisfying.

Why shouldn't our students get some satisfaction too?

References

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66(2): 249–88.

Angrist, Joshua D., Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Lu. Forthcoming. "Economic Research Evolves: Citations Fields and Styles." *American Economic Review.*

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." Chap. 23 in *Handbook of Labor Economics*, vol. 3, edited by Orley Ashenfelter and David Card, 1277–1366. Elsevier.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2): 3–30.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton University Press.

Arcidiacono, Peter, Esteban M. Aucejo, and V. Joseph Hotz. 2016. "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California." *American Economic Review* 106(3): 525–62.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics* 49(4): 431–34.

Ayres, Ian. 2007. *Super Crunchers.* Bantam Books.

Becker, William E., and William H. Greene. 2001. "Teaching Statistics and Econometrics to Undergraduates." *Journal of Economic Perspectives* 15(4): 169–82.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly*

Journal of Economics 119(1): 249–75.

Bowen, Donald E., III, Laurent Frésard, and Jérôme P. Taillard. *Forthcoming.* "What's Your Identification Strategy? Innovation in Corporate Finance Research." *Management Science.*

Brynjolfsson, Erik, and Andrew McAfee. 2011. "The Big Data Boom Is the Innovation Story of Our Time." *The Atlantic,* November 21.

Chamberlain, Gary. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18(1): 5–46.

Christian, Brian. 2012. "The A/B Test: Inside the Technology That's Changing the Rules of Business." *Wired,* April 25.

Dale, Stacy Berg, and Alan B. Krueger. 2002. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." *Quarterly Journal of Economics* 117(4): 1491–1527.

Dougherty, Christopher. 2016. *Introduction to Econometrics.* 5th edition. Oxford University Press.

Fuchs-Schündeln, Nicola, and Tarek A. Hassan. 2016. "Natural Experiments in Macroeconomics." Chap. 12 in *Handbook of Macroeconomics,* vol. 2, edited by John B. Taylor and Harald Uhlig, 923–1012. Elsevier.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Harvard University Press.

Gujarati, Damodar. 1978. *Basic Econometrics.* New York: McGraw-Hill.

Gujarati, Damodar N., and Dawn C. Porter. 2009. *Basic Econometrics.* 5th Edition. Boston: McGraw-Hill.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51(1): 162–72.

Hayashi, Fumio. 2000. *Econometrics.* Princeton University Press.

Intriligator, Michael D. 1978. *Econometric Models, Techniques, and Applications.* Englewood Cliffs, NJ: Prentice Hall.

Johnston, J. 1972. *Econometric Methods,* 2nd Edition. New York: McGraw-Hill.

Kennedy, Peter. 1979. *A Guide to Econometrics.* Cambridge, MA: The MIT Press.

Kennedy, Peter. 2008. *A Guide to Econometrics.* 6th Edition, Malden, MA: Blackwell Publishing.

Kmenta, Jan. 1971. *Elements of Econometrics.* New York: The Macmillan Company.

Kohavi, Ron. 2015. "Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM.

Maddala, G. S. 1977. *Econometrics.* McGraw-Hill.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3): 385–97.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55(3): 703–08.

Panhans, Matthew T., and John D. Singleton. Forthcoming. "The Empirical Economist's Toolkit: From Models to Methods." *History of Political Economy.*

Pindyck, Robert S., and Daniel L. Rubinfeld. 1976. *Econometric Models and Economic Forecasts.* New York: McGraw-Hill.

Stock, James H., and Mark M. Watson. 2015. *Introduction to Econometrics.* 3rd Edition. Boston: Pearson.

Studenmund, A. H. 2017. *Using Econometrics: A Practical Guide.* 7th Edition, Boston: Pearson.

Summers, Anita A., and Barbara L. Wolfe. 1977. "Do Schools Make a Difference?" *American Economic Review* 67(4): 639–52.

White, Halbert. 1980a. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21(1): 149–70.

White, Halbert. 1980b. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4): 817–38.

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics: A Modern Approach.* 6th edition. Boston: Cengage Learning.

Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics* 14(4): 478–86.