Demystifying causal inference: ingredients of a recipe

Vikram Dayal
Anand Murugesan

2020

IEG Working Paper No. 393

सत्यमेव परमो धर्मः

Demystifying causal inference: ingredients of a recipe

Vikram Dayal*[1] and Anand Murugesan†[2]

[1]Institute of Economic Growth, New Delhi
[2]Central European University, Vienna

May 18, 2020

Abstract

In the last few decades, scholars have contributed to a flourishing literature on casual inference and the demand for its application in areas like programme evaluation has increased. Our suggestion is that the following ingredients are useful for demystifying causal inference in introductory courses: (1) using the potential outcomes and causal graph frameworks, (2) covering applications with real data that use key methods for causal inference: experiments, regression discontinuity etc., (3) using Monte Carlo simulation, and (4) using data graphs. The first two ingredients are components of the scholarship in causal inference, while the latter two are more general ingredients of statistical and econometric pedagogy. We discuss the case for these ingredients, drawing on the substantive and pedagogical literature, our experience, and student opinions.

Introduction

Causal inference is often of interest to researchers. As Pearl (2009, p.97) puts it, "The questions that motivate most studies in the health, social and behavioural sciences are not associational but causal in nature." Causal inference is interdisciplinary, but not

*vikday@iegindia.org
†murugesana@spp.ceu.edu

surprisingly, is understood and practiced differently in different disciplines. Causal inference has grown in importance in the last few decades; according to Gary King (blurb on the back cover of Morgan and Winship (2014)), 'More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history'. There has been a boom in the field of programme evaluation, which builds on causal inference (Abadie and Cattaneo 2018). Policy makers in several countries have explicitly begun relying on evidence-based research as a strict criterion for policy making. The demand for skills in causal inference extends to the private sector, (Angrist and Pischke 2017, p. 2): 'Google and Netflix post positions flagged 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.'

This demand for causal inference translates into a need for conveying the key ideas of this area. Causal inference is characterized by some subtle ideas. Since those encountering the subject for the first time may find it challenging, our paper is about demystifying causal inference. We suggest ingredients of a recipe, but not a recipe itself since in different contexts, different instructors and students are likely to use, be prepared for, or prefer different ingredients of the recipe. With a given amount of class time, the ingredients need to be balanced. The objective of an introductory course is to convey the key elements and applications of causal inference, so that it is enjoyable and intuitive understanding is promoted. We first draw on some scholarship on the subject of teaching econometrics.

More generally, in econometrics, scholars have reflected on the task of teaching the subject. In the Preface of their book, Johnston and Dinardo (1996) wrote that the applied econometrician often suffers from 'intellectual indigestion'; and this is also highly likely to be the case with students. Greene and Becker (2001) advocated greater use of computing technology as opposed to *chalk and talk*; this was at a time when computers were increasingly available to students of economics. The paucity of applications used in instruction made the subject dry and abstract. Since then, access to data has improved and econometric textbooks usually include many 'real' examples. As pointed out by Verbeek (2012), working with data is not not only interesting, it is substantively important. He quotes a seminar speaker: 'Econometrics is much easier without data.'

Kennedy (2009, p. 487) opened his article on econometric teaching with the follow-

ing: 'Contrary to the belief of most econometrics instructors, upon completion of introductory statistics courses, the vast majority of students do not understand the basic logic of classical statistics as captured in the sampling-distribution concept.' He emphasized the importance of teaching the sampling distribution concept and advocated explaining Monte Carlo studies to students. Craft (2003) and Briand and Hill (2013) provide specific guidance on using Monte Carlo simulations for teaching.

More recent commentaries on the teaching of econometrics reflect the need to include causal inference. Angrist and Pischke (2017) commented on econometric instruction. In their opinion, although empirical economics has changed greatly, econometric teaching has not. According to them, causal inference needs to be emphasized in teaching (they use the potential outcomes framework), and interesting applications should be used. Such methods as regression discontinuity and difference in difference should get sufficient class time. Angrist and Pischke have written two textbooks on causal inference. In their 2017 paper, they are generally critical of econometric textbooks, though they approve of the book by Stock and Watson, 'which comes closest to embracing the modern agenda.' Chen and Pearl (2013) reviewed six econometrics textbooks by posing what they feel were key questions (for example, 'Does the author present example problems that require causal reasoning?'). The textbook by Stock and Watson also fared well on most of their criteria.

Against this background, we propose some possible ingredients for a recipe for demystifying causal inference. Both the authors of this paper have some experience in teaching causal inference to those not exposed to it previously, and we draw on our experience here. We realize that different instructors may want to emphasize different aspects, and there is a given time constraint in any course.

The ingredients we propose are:

1. Using the potential outcomes framework and causal graphs.
2. Covering applications that use key methods for causal inference: experiments, regression discontinuity etc.
3. Using monte carlo simulation.
4. Using data graphs.

In a course on causal inference, it is natural to cover the key methods for causal inference that are currently being used in economics, ingredient 2, and this is also consistent with the view that real applications with real data should be used in courses.

Ingredient 1 is optional, but we feel that the potential outcomes framework and causal graphs help clarify issues, and in the next section we discuss the case for them. Several authors have advocated ingredient 3 as we have discussed in the introduction. We discuss the case for ingredient 4 in the next section, but we feel that graphs help connect the data to models; also, visualizing data is a skill in demand.

The rest of the paper discusses the case for the different ingredients, drawing on the literature, our own experience, and some opinions solicited from students. We asked graduate students at two locations for their opinions, one at TERI School of Advanced Studies, Delhi and the other at the Central European University, Vienna (CEU). The questions were administered via google forms. The students at CEU were enrolled in empirical economics courses that extensively reviewed methods such as Difference-in-Differences, Instrumental Variables and Regression Discontinuity Design. The students at TERI were studying a more standard econometrics course at the Master's Level.

We intersperse the responses from the students within our discussion of the case for the different ingredients for demystifying causal inference.

The case for the proposed ingredients

In this section we discuss the case for each of the different proposed ingredients, in turn and in the next section we see how these ingredients combine.

Starting point: causal questions

Before we begin with the ingredients, a starting point is clarifying what causal questions are (Chen and Pearl 2013). In their excellent text, Stock and Watson (2011, pp. 2 – 4) begin with four questions that economists examine:

1. Does reducing class size improve elementary school education?
2. Is there racial discrimination in the market for home loans?
3. How much do cigarette taxes reduce smoking?
4. What will be the rate of inflation be next year?

They point out that the first three questions are questions that require causal inference. Hernan, Hsu and Healy (2019) have stressed that it is important to classify data

science tasks into description, prediction, and causal inference (which can also be viewed as requiring counterfactual prediction).

Though most students did distinguish between descriptive, causal and predictive interpretations, several did not (Table 1). Also, among those students who had been taught a more conventional course (TERI), a higher proportion did not make the distinction. Since the students at CEU were often exposed to the distinction between correlation and causation within a regression framework, they may be better at recognizing the difference (even if only cursorily).

Table 1: When you run a regression, do you distinguish between (1) descriptive, (2) causal and (3) predictive interpretations?

	Number of students	
Response	TERI	CEU
Yes	13	18
No	7	3
Total	21	21

Source: survey of students conducted by us in 2020

Ingredient: potential outcomes and causal graph frameworks

The potential outcomes and causal graph frameworks make causal inference explicit. Is it important to have an explicit framework for causal inference? Yes, according to Pearl (2009, p. 100): 'Another ramification of the sharp distinction between associational and causal concepts is that any mathematical approach to causal analysis must acquire new notation for expressing causal relations—probability calculus is insufficient.'

We agree with the sentiment expressed above, and provide a more specific statement related to experiments. Causal inference is the underlying idea in evidence based policy. And many consider randomized experiment based inference as the gold standard, though some disagree. The potential outcomes and causal graph frameworks can help us unpack randomized experiments, help us see how they work, and how they relate to studies based on observational data, instead of only asserting that they are the 'gold standard'.

In an online presentation, the political methodologist Kosuke Imai (2016) wrote: 'I have been using potential outcomes in most of my research, but recently I have started using *DAGs*. Potential outcomes are useful when thinking about treatment assignment

mechanism, experiments, quasi-experiments. DAGs are useful when thinking about the entire causal structure, complex causal relationships. Both are better suited for causal inference than the standard regression framework.' Abadie and Catteneo (2018) use both the potential outcomes framework and causal graphs to explain econometric methods for program evaluation.

Imbens (2019) has recently discussed the potential outcomes and causal graph frameworks. His view is that both are complementary, but the potential outcomes framework is used more in economics. However, he sees expository value for the causal graph (also called *DAGs*, directed acyclic graphs) framework. He writes that 'The DAGs, like the path analyses that came before them, ... can be a powerful way of illustrating key assumptions in causal models. ... Ultimately some of this is a matter of taste and some researchers may prefer graphical versions to algebraic versions of the same assumptions and vice versa.' Causal graphs, or DAGS, are visual but are also mathematical objects. They tie in with structural causal models but can be related to the potential outcomes framework; but they only require modest formal training (Elwert 2013).

Our own view is that these two frameworks are important components of causal inference. We now discuss the substantive and pedagogical arguments for using the frameworks.

Substantive argument for the potential outcomes framework

- The potential outcomes framework is closely associated with the Neyman-Rubin causal model. It is a core element of causal inference ideas, and connects with methods of inference. It is useful for understanding the core of causal inference as well as developments in the subject that build on it.
- The potential outcomes framework can illustrate a simple decomposition of estimated effects into selection effects and true effects (Angrist and Pischke 2009).
- In the potential outcomes framework, there is a delineation of different types of treatment effects.
- The potential outcomes framework helps us understand experiments and one of the inferential methods often used with experiments, randomization inference (Abadie and Cattaneo 2018)

- Building on the potential outcomes model, Angrist, Imbens and Rubin (1996) developed insight into instrumental variables estimation used in experiments with imperfect compliance.
- Another extension by an econometrician, that of Manski bounds, builds on the potential outcomes model (see Manski 1995). Manski has pushed the frontier in terms of highlighting the role of assumptions in causal analysis.

Pedagogical argument for the potential outcomes framework

- The potential outcomes framework can be introduced simply with numbers in a table (Wikipedia).
- The potential outcomes framework can introduce the idea of a causal effect simply. Also, along with this, the idea of the counterfactual or the potential outcome that is not observed can be communicated.
- The idea of randomization inference can also be communicated simply as done by Rubin (2005) with a simple table of numbers and from first principles.
- Once students internalize the potential outcomes framework, and the idea of a counterfactual, it is easier to communicate the logic of experiments, regression discontinuity and difference-in-difference. In experiments we use randomization and the control group to fill in the missing potential outcome. In difference-in-difference, we use the 'if the treatment group had not got the treatment it would have changed like the control group' assumption to fill in the potential outcome.
- Useful pedagogical resources for the potential outcomes framework are: Rosenbaum's beautifully written book *Observation and Experiment*, Rubin's course notes (2005), the chapter on experiments in Angrist and Pischke (2015), the Wikipedia entry on the Rubin Causal Model.

Substantive argument for causal graphs

- Causal graphs, or more formally, causal directed acyclic graphs, are part of the structural causal model that is used by many practitioners of causal inference. Abadie and Cattaneo (2018) use it in their discussion.
- Causal graphs help make causal assumptions explicit.
- The literature using causal graphs clarified the issue of which covariates one should adjust for (Pearl and Mackenzie 2019). With a given causal graph

we can see which sets of covariates need to be adjusted for or not, given our interest in specific causal effects. Causal graphs are visual but there is a rigorous mathematical machinery underpinning them.

- A key contribution of causal graphs to causal inference thinking was the idea of collider bias where covariate adjustment was harmful. Elwert and Winship (2014) is a detailed explanation of the phenomenon with its applications.
- Causal graphs can be used to understand instrumental variables (see Abadie and Cattaneo 2018). Morgan and Winship (2015) use causal graphs to discuss valid instrumental variables and the tests for their validity. The same authors use causal graphs to discuss the charge that instrumental variables estimation may result in insufficiently deep explanations.

Pedagogical argument for causal graphs

- Causal graphs are intuitive, and reflect how we think of relationships. Applied researchers often use figures that are like causal graphs, with variables connected by arrows to convey their thinking (see, for example, Rodrik, Subramanian and Trebbi 2004).
- Causal graphs can be used both informally and more formally, as part of structural causal models. Their use can be adjusted to the audience.
- Since causal graphs are part of structural causal models, they represent equations, and can be useful along with simulation exercises. These can supplement traditional derivation based classes, where the key points of the derivation are highlighted while simulations with causal graphs provide intuition.
- We often say that measurement error or omitted variables are sources of 'endogeneity'. Left at that, it can be a bit mysterious. However, causal graphs are more specific in intuitively delineating possible data generating processes that lead to measurement error and omitted variables.
- Causal graphs are widely used in epidemiological pedagogy and practice.
- Useful pedagogical resources for causal graphs are the superbly written article by Elwert (2013), the *Book of Why* by Pearl and Mackenzie (2019). For causal graphs, two non-econometrics books that provide excellent introductory treatments are Shipley (2000) and Kaplan (2009), and both use causal graphs along with simulation. The brief introduction to the chapter on causal inference in

Dayal (2020), draws on the potential outcomes and causal graph frameworks.

Student opinions

Table 2 shows that students are unlikely to hear about the Rubin causal model by themselves (among TERI students who took a usual type of econometric class, only 2 out of 21 had heard about the Rubin causal model). Students at CEU are more familiar with the discussion of potential outcomes as it was fleetingly discussed in the course (and were assigned to read the introductory chapter in Angrist and Pischke (2015)).

Table 2: Have you heard of / read about the potential outcomes (also called counterfactual or Neyman-Rubin) approach?

	Number of students	
Response	TERI	CEU
Yes	2	12
No	19	9
Total	21	21

Source: survey of students conducted by us in 2020

The proportion of students who had read anything using causal graphs was very similar in both places, and they were in a minority (Table 3). Anand Murugesan uses causal graphs informally, i.e. diagrams/one-sided or two-sided arrows in one of his courses to illustrate relationships among dependent, independent, endogenous and exogenous variables.

Table 3: Have you read anything using causal graphs (also called directed acyclic graphs) or path diagrams?

	Number of students	
Response	TERI	CEU
Yes	6	7
No	15	14
Total	21	21

Source: survey of students conducted by us in 2020

Most students find econometrics somewhat mysterious and puzzling (Table 4). Whether they would find the potential outcomes and causal graphs approaches also somewhat

mysterious and puzzling is something we cannot say anything definite about, though we would hope that it helps clarify the nature of causal inference.

Table 4: Do you find econometrics mysterious...?

Response	Number of students TERI	CEU
Very mysterious / puzzling	1	4
Somewhat mysterious / puzzling	15	14
Not mysterious / puzzling at all	5	3
Total	21	21

Source: survey of students conducted by us in 2020

Most students find econometric discussions of exogenous explanatory variables somewhat mysterious in TERI though a substantial proportion are not mystified or puzzled by it (Table 5). At CEU, 4 out of 21 are very puzzled by such discussions. In our opinion, the potential outcomes and causal graph frameworks, help make the ideas of causal inference more explicit in comparison with discussions of exogenous explanatory variables typically found in textbooks.

Table 5: Do you find econometric discussions of exogenous explanatory variables... ($E(u \mid X) = 0$, in a regression of y on the vector of explanatory variables X)

Response	Number of students TERI	CEU
Very mysterious / puzzling	0	4
Somewhat mysterious / puzzling	13	7
Not mysterious / puzzling at all	8	10
Total	21	21

Source: survey of students conducted by us in 2020

Ingredient: Covering applications that use key methods for causal inference

Substantive reasons to teach key causal inference methods

There has been an 'exponential growth in economists use of quasi-experimental methods and randomized trials' (Angrist and Pischke 2017, p.2). Angrist and Pischke (2017) lament the 'failure to discuss modern empirical tools' by most econometric

textbooks that they surveyed. Others disagree with them, it is only fair to mention one contrary view, by Diebold of their earlier similar but more technical book, 'It's a novel treatment of that sub-sub-sub-area of applied econometrics, but pretending to be anything more is most definitely harmful, particularly to students, who have no way to recognize the charade as a charade.'

Pedagogical considerations in teaching key causal inference methods

- As we stated in the introduction, Becker and Greene (2001) recommended that applications be used in teaching econometrics.
- Angrist and Pischke (2015) cover what they consider the "Furious Five": random assignment, regression, instrumental variables, regression discontinuity, and difference-in-differences. Their book provides good introductory, example-based explanations of regression discontinuity and difference-in-differences.
- Textbooks are changing to meet courses targeting causal inference specifically, and more of the most used causal inference methods are appearing along with standard coverage. Stock and Watson (2011) have a chapter on experiments and quasi-experiments, and also a chapter on causal inference with time series. Bailey (2017) and the fifth edition of Hill et al. (2018) cover the "Furious Five"; and Imai's (2017) book on Quantitative Social Science covers four of the "Furious Five". Dayal (2020) covers these techniques, discussing relevant R packages and code, and in some cases drawing on examples in Angrist and Pischke (2015) and other leading texts, along with some papers, for which data are easily available.

Student opinions

In CEU, compared to TERI, a far larger proportion of students (18 out of 21) had studied an experiment (and regression discontinuity) as it was discussed extensively in the course (Table 6). In today's context we believe that it is vital to have studied experiments.

Source: survey of students conducted by us in 2020

The regression discontinuity experience is similar to that for experiments (Table 7).

Source: survey of students conducted by us in 2020

Table 6: Have you studied any specific experimental (RCT) study, or analysed experimental data?

	Number of students	
Response	TERI	CEU
Yes	10	18
No	11	3
Total	21	21

Table 7: Are you familiar with the regression discontinuity method?

	Number of students	
Response	TERI	CEU
Yes	3	18
No	18	3
Total	21	21

Ingredient: Using monte carlo simulation

Substantive reasons

- Classical statistics is about the sampling distribution, both when we estimate, and when we test. We quote Kennedy (2003, pp. 419 – 420): 'Using $\beta*$ to produce an estimate of β can be conceptualized as the econometrician shutting his or her eyes and obtaining an estimate of β by reaching blindly into the sampling distribution of $\beta*$ to obtain a single number. ... Hypothesis testing is undertaken by seeing if the value of a test statistic is unusual relative to the sampling distribution of that test statistic calculated assuming the null hypothesis is true.'
- Permutation or randomization tests are frequently used for statistical inference, especially with experiments (Gerber and Green 2012, Abadie and Cattaneo 2018).
- In applied work, researchers have to ponder the use of several possible alternative estimation methods. Monte carlo simulation is often used to compare different methods. For example, O'Neill et al. (2016) consider three alternatives to difference-in-differences estimation (synthetic control, lagged dependent variable, and matching on past outcomes). They write (p.1), 'We conduct the first Monte Carlo simulation study to contrast the relative performance of DiD compared to these alternative approaches. We consider scenarios where the parallel trends

assumption does, and does not hold. The simulation results show that DiD performs best under parallel trends, and when the parallel trends assumption is violated, the LDV approach reports the least biased, most efficient estimates.' While theory can tell us what the key assumption is, the Monte Carlo method can provide us with a sense of the relative performance of different estimators, given the input on different possible synthetic data generating processes.

- Deaton and Cartwright's (2018) well known paper titled 'Understanding and misunderstanding randomized controlled trials' uses Monte Carlo simulation to support one of its arguments.

Pedagogical considerations

- We once again quote Kennedy (2003, p.34), who made the case eloquently, 'understanding Monte Carlo studies is one of the most important elements of studying econometrics, not because a student may need actually to do a Monte Carlo study, but because an understanding of Monte Carlo studies guarantees an understanding of the concept of a sampling distribution and the uses to which it is put.' Carsey and Harden's (2014) text on Monte Carlo Simulation and Resampling Methods for Social Science is motivated by the idea that 'If you really want students to understand the properties of a model or the model's underlying assumptions, make them simulate a sample of data that has those properties.'

- Large-sample properties figure centrally in modern econometrics. This is where Monte Carlo simulation shines in providing an intuitive understanding. Four key econometrics texts that instructors may use in an introductory course in causal inference present simulation results, in varying degrees. Stock and Watson (2011, p. 47) present simulation results in a section titled 'Large-Sample Approximations to Sampling Distributions.' Angrist and Pischke (2015, p. 39) present simulation results in a section titled 'The t-statistic and the Central Limit Theorem.' Imai (2017) uses simulation for illustrating probability, confidence intervals and hypothesis tests. Hill, Griffiths and Lim (2018) have a number of Monte Carlo exercises presented in appendices, for example one on instrumental variables.

- It is possible to use Monte Carlo simulation to demonstrate from first principles randomization inference in the case of experiments. It is also possible to use sta-

tistical simulation with DAGS or causal graphs. Causal graphs are accompanied by structural equations, so we can use a synthetic data generating process that is consistent with the causal graph. We can thus verify the insights that are intuitively suggested by causal graphs, and have been mathematically verified by the formal, mathematical work on causal graphs.

- Instructors can opt to highlight key points of mathematical derivations, and let students use Monte Carlo simulations. Students could use already programmed functions, simply playing with the inputs to the functions and seeing what happens, or write their own functions.
- Gerber and Green's (2012) excellent book on Field Experiments relies on statistical simulation for hypothesis intervals and confidence intervals, which, they feel, makes the presentation more systematic and concise. Kaplan (2009) touches upon several of the concepts discussed in this section, including sampling distributions, randomization inference, and simulations used along with causal graphs. Dayal (2020)'s chapter on causal inference uses simulation extensively while discussing causal inference.

Student opinions

A minority of students had heard of Monte Carlo simulation in TERI and CEU (Table 7) and very few had done Monte Carlo simulation (Table 8). Simulation requires some time set aside, a fixed cost.

Table 8: Have you ever yourself done a statistical (Monte Carlo) simulation?

	Number of students	
Response	TERI	CEU
Yes	0	4
No	20	17
Total	21	21

Source: survey of students conducted by us in 2020

Ingredient: Using data graphs

Substantive considerations

- The statistician and polymath John Tukey (1962, p. 49) advocated the use of graphs of data: 'The simple graph has brought more information to the data

analyst's mind than any other device.' Being able to graph data is a key skill today, and often can be used in illuminating descriptions. Software like Stata and R, make it possible to make high quality graphs easily, though this too is a skill that needs to be developed.

- Rosenbaum (2010) uses boxplots to compare different groups through the entire book. He discusses the case of uncommon but dramatic effects in a separate chapter, i.e. where most subjects are not much affected by treatment, but a small fraction, are strongly affected. His starting point for the discussion is a boxplot of the observed differences in outcomes in paired subjects. Conventional statistical methods are geared to detecting typical treatment effects, but not dramatic effects for a few subjects. Chattopadhyay and Duflo (2004) conducted an experimental study of the effect of reserving positions of leadership in Village Councils in India on the kinds of projects undertaken by them. A subset of their data is presented in Dayal (2020) and boxplots show that a few of the treated villages had very high levels of water projects. Had we not seen the boxplots, we would have missed this feature in the sample.

- Data graphs play a role in making an analysis transparent, but as with statistical analysis in general, some skepticism is useful. Lee and Lemieux (2010) write: 'It has become standard to summarize RD analyses with a simple graph showing the relationship between the outcome and assignment variables. This has several advantages. The presentation of the "raw data" enhances the transparency of the research design. A graph can also give the reader a sense of whether the "jump" in the outcome variable at the cutoff is unusually large compared to the bumps in the regression curve away from the cutoff. ... The problem with graphical presentations, however, is that there is some room for the researcher to construct graphs making it seem as though there are effects when there are none, or hiding effects that truly exist.'

- Data graphs can help connect data to models, and thus, reveal problems in causal inference. Gelman (2013) discussed a paper by Chen et al. (2013) in which they based causal inference about the effect of air pollution on life expectancy in China on a regression discontinuity study, on his much read blog: 'Here's the key figure from the paper ... This is a beautiful graph. I love love love a plot that shows the model and the data together. One thing I like about this *particular* graph is that, just looking at it, you can see how odd the model is. Or, at least, how odd it looks to an outsider. A third-degree polynomial

indeed!' Later, Gelman (2018) wrote: 'The most obvious problem revealed by this graph is that the estimated effect at the discontinuity is entirely the result of the weird curving polynomial regression, which in turn is being driven by points on the edge of the dataset.'

Pedagogical considerations

From an expository perspective, data graphs can help make the subject less abstract. Stock and Watson (2011) and Bailey (2017) use scatterplots very effectively while introducing regression with panel data. Angrist and Pischke (2009) provide graphs in two different studies where the common trends assumption was consistent and inconsistent with past data. Gelman and Hill (2007), whose text includes chapters on causal inference, use graphs skilfully throughout, and have a very good appendix on statistical graphics. Dayal (2020) uses data graphs extensively. An example is his use of data graphs to provide a simplified step-by-step exposition of Manski and Pepper (2018), which uses Manski Bounds.

Student opinions

Very few students did not like data graphs (Table 9).

Coming to student opinions, relevant here, they are as follows:

Table 9: Do you like making graphs of data?

	Number of students	
Response	TERI	CEU
Like a lot	12	5
Ok with graphs	9	15
Don't like	0	1
Total	21	21

Source: survey of students conducted by us in 2020

Most students at TERI and CEU thought that graphs of data are helpful in data analysis, for communication, and substantively (Table 10).

Source: survey of students conducted by us in 2020

Table 10: Do you think graphs of data help?

Response	Number of students TERI	CEU
Communicate the results of data analysis?	2	4
Can be substantive tools that help analysis?	1	1
Both of the above	18	17
Total	21	21

Synergy of the ingredients and illustrations

Finally, we point out that there is a synergy between the ingredients. We provide specific illustrations of our use of the ingredients, drawing on the chapter on causal inference in Dayal (2020).

First illustration: Anchoring experiment

Ingredients used in this case: a causal inference method with real data, data graphs, simulation-based inference.

In this example we use Kahneman's (2011) well known anchoring experiment, that is conducted in class. This has a number of advantages in the context of such a course:

- It is a key idea of a Nobel Laureate in economics.
- It provides a feel for how the data of an experiment are generated.
- It illustrates how in a different setting the results may change (the issue of external validity).
- It is a simple example, easy to comprehend.
- It can be used to illustrate the use of randomization inference, both illustrating the use of computation in inference, and illustrating the randomization inference distribution.

The specific questions that students are asked are:

- We chose (by computer) a random number between 0 and 100.
- The number selected and assigned to you is X = ...
- Do you think the percentage of countries, among all those in the United Nations, that are in Africa is higher or lower than in X? ....

- Give your best estimate of the percentage of countries, among all those in the United Nations, that are in Africa ....

  Once the data are collected, it can be analysed.

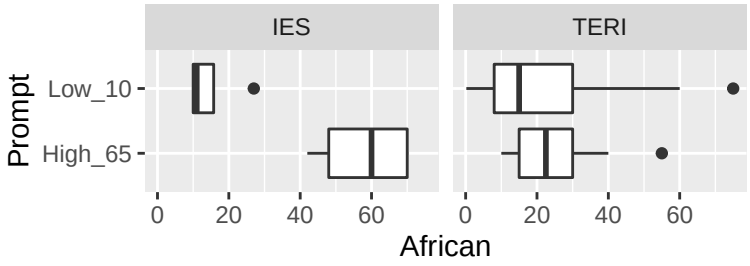

Figure 1: Boxplots comparing IES and TERI classes, best estimate of countries in Africa, anchoring experiment

Figure 1 shows a data graph, with the boxplots giving us a visual comparison. The graphs show us comparisons across two samples, two classes (IES and TERI), and within each, the comparison between treatment and control. The graph also shows that in different settings an experiment can give somewhat different results, although in both there is an anchoring effect.

The ri2 package is used for randomization inference; Figures 2 and 3 show the randomization inference for the samples.

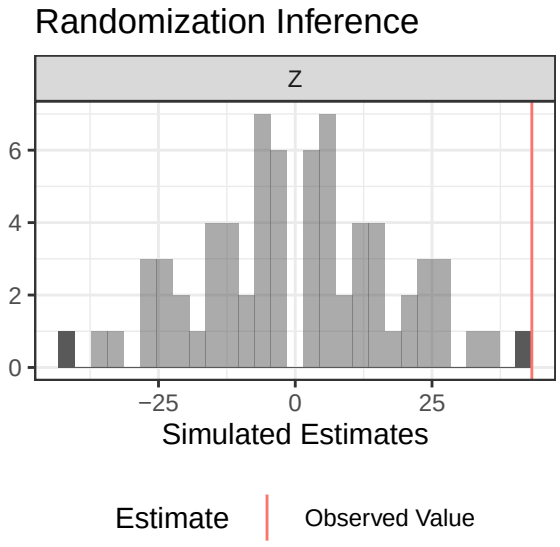

Figure 2: Randomization inference for IES

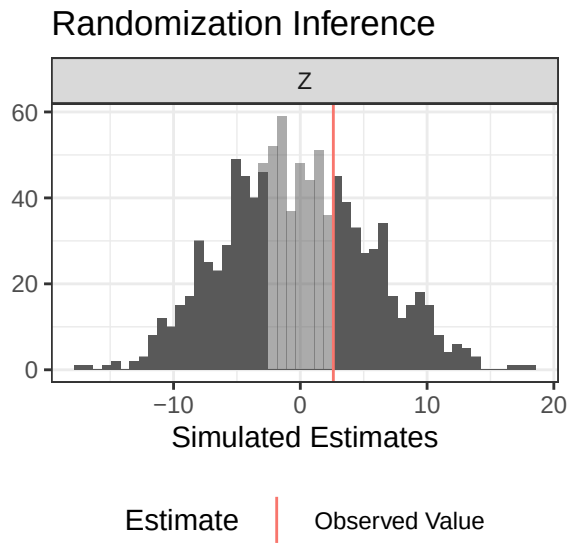

Figure 3: Randomization inference for TERI

Second illustration: regression discontinuity

Ingredients used: statistical simulation, data graphs, illustrating a causal inference method with real data.

In order to explain regression discontinuity, which can be intriguing, we believe it is useful to do a simple simulation, generate artificial data, and then estimate the true causal effect.

The synthetic data generating process broadly, is:

- The running variable, run, is drawn from a uniform distribution
- The treatment variable is 0 if run < 20, else it is 1.
- The outcome variable $= 10$ treat $- 0.4$ run $+$ noise.

Figure 4 shows the synthetically generated data. We see that there is a clear jump equal to the effect of the treatment at the cutoff point.

We then use real data in an example drawn from Angrist and Pischke (2015), about the Minimum Legal Drinking Age (MLDA). The r package rddtools is not only good for regression discontinuity analysis, it also produces graphs that help demystify. The package can be used for both parametric and non-parametric analysis. Below you can see the graph (Figure 5) accompanying the non-parametric analysis.

We can also illustrate the use of placebo tests visually (Figure 6). In a regression

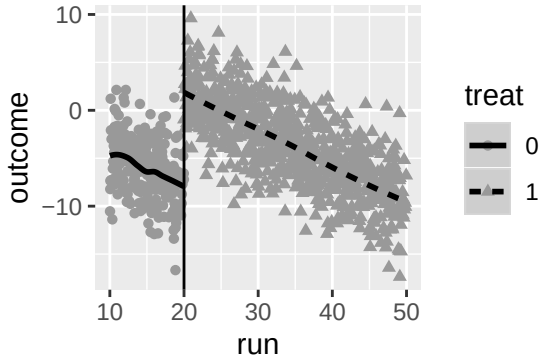

Figure 4: Visualising regression discontinuity with simulated data

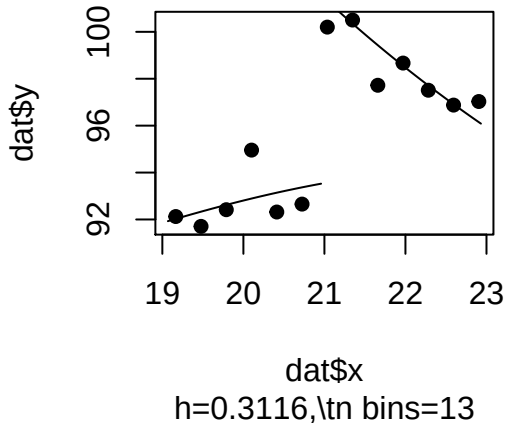

Figure 5: Non-parametric analysis of MLDA

discontinuity we should get an effect only at the cutpoint. Away from the cutpoint, the estimate should be zero. The placebo test tests whether this is zero. We think the following visual of the placebo test (Figure 6) is neat, confirming that we get an effect only at the cutpoint:

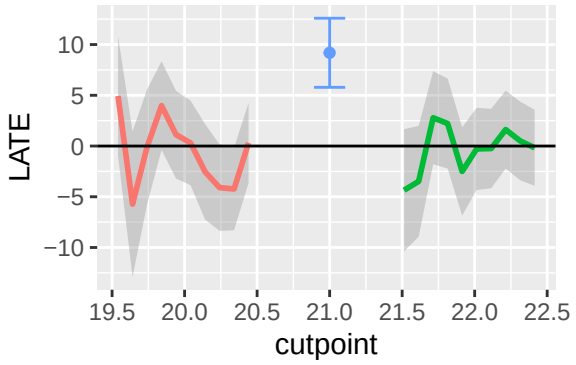

Figure 6: Placebo test

Similarly, the package does a sensitivity check with respect to the bandwidth used and can produce a plot (Figure 7).

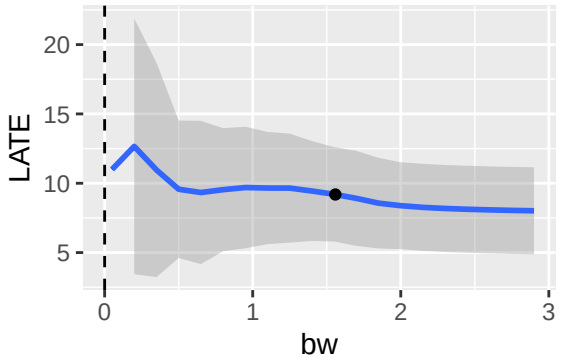

Figure 7: Sensitivity test

Third illustration: graphical approach to difference-in-difference

Ingredients used: method for causal inference illustrated with real data, crucial use of data graph, and application of the concept of potential outcomes

The difference-in-difference method is easy to use for answering policy questions; a key assumption is the parallel slopes assumption.

Manski and Pepper (2018) examined the effect of right to carry laws on crime. The state of Virginia allowed guns to be carried in 1989, while Maryland did not do so. Can a typical difference-in-difference analysis be used? However, in the following figure on murder rates in Virginia and Maryland, the parallel slopes assumption does not hold.

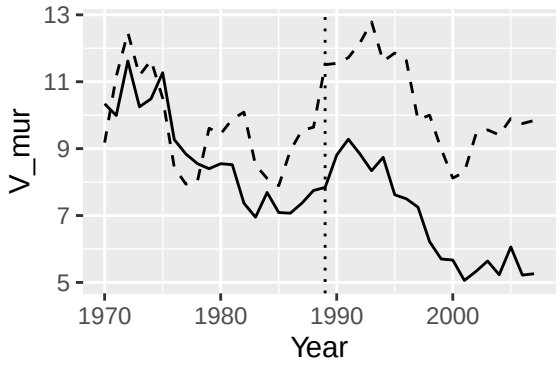

Figure 8: Murder rates in Virginia (solid line) and Maryland (dashed line). Virginia enacted a right to carry statute in 1989 (vertical dotted line).

The analysis by Manski and Pepper (2018) also brings home a practical application of the potential outcomes framework. Since the parallel slopes assumption fails, they use alternate assumptions and on that basis provide bounded estimates. Manski and Pepper's analysis can overwhelm initially because of the different approach, but through a careful step-by-step explanation that uses graphs like Figure 8, the simpler parts of their rich paper can be communicated.

Fourth illustration: instrumental variables

Ingredients used: causal graphs, simulation, data graphs

While presenting instrumental variables, we find it useful to introduce instrumental variables via a very simple example, using causal graphs. Abadie and Cattaneo 2018) use a similar causal graph in the beginning of their section on instrumental variables.

Consider a causal graph (part of a structural causal model as in Pearl et al. 2016) of four variables, U, Z, X and Y (Figure 9). We are interested in the causal effect of X on Y.

The individual causal links are:

- U causes X

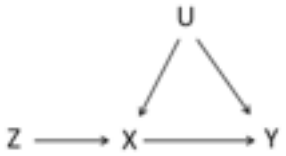

Figure 9: Causal graph

- U causes Y
- Z causes X
- X causes Y

The structural equations (we assume linear relationships) for the causal graph are:

$X = \gamma_0 + \gamma_Z Z + \gamma_U U + error_X$

$Y = \beta_0 + \beta_X X + \beta_U U + error_Y$.

We generate data according to the structural equations above, assuming numerical values for the parameters.

We now run the following regressions:

- Regress Y on X only, get coefficient of X, OLS1
- Regress Y on X and U, get coefficient of X, OLS2.
- Regress Y on X, using instrumental variables, with Z as and instrumental variable, and get coefficient of X, IV.

How do OLS1, OLS2 and IV relate to β_X, the structural equation parameter, or the "true" effect of X on Y? We generate data repeatedly and then repeatedly estimate OLS1, OLS2 and IV. We are then able to compare the sampling distributions of OLS1, OLS2 and IV.

We create a function in R, which allows us to carry out data generation and estimation repeatedly. A part of the R code is as follows:

sample_size = 300

coef_Z = 0.9

Z <- runif(sample_size, min = 1, max = 5) # generating Z

U <- runif(sample_size, min = 1, max = 5) # generating U

X <- U + rnorm(sample_size) + coef_Z *Z # generating X

Y <- U + X + rnorm(sample_size) # generating Y

OLS1 <- lm(Y ~ X)

OLS2 <- lm(Y ~ X + U)

IV <- ivreg(Y ~ X | Z)

Figure 10 displays the results; the true effect is 1. OLS2, which controls for U, is unbiased, but of course it requires knowledge of U. OLS1 is biased, but may have a smaller spread than IV, which uses the instrument Z. IV is consistent.

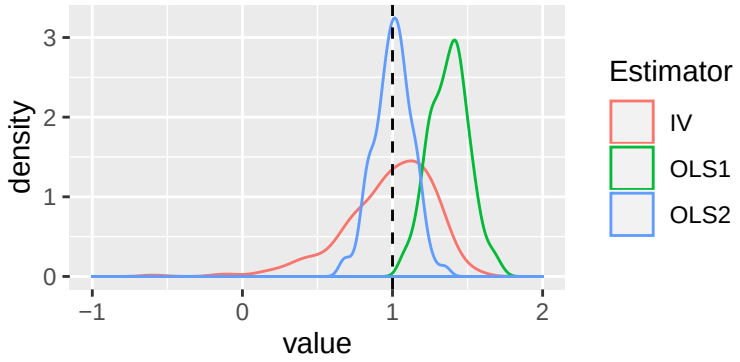

Figure 10: Simulation results. Sampling distributions of estimators. True effect of X on Y is 1. IV uses Z as instrument, OLS1 is Y regressed only on X, OLS2 is Y regressed on X and U

This helps illustrate:

- Both omitted variable bias and the use of instrumental variables.
- Instrumental variable estimates even with a good instrumental variable are consistent, but have a larger variance.

Conclusion

With developments in causal inference and computing we feel it is possible to convey the essence of this subject and promote active learning using the ingredients we have suggested. Different instructors can tailor the course differently depending on their students and emphasize different ingredients.

References

Abadie A, Cattaneo M D (2018) Econometric methods for program evaluation. Annual Review of Economics 10: 465-503.

Angrist J D, Imbens G, Rubin D B (1996) Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444-72.

Angrist J D, Pischke J-S (2009) Mostly Harmless Econometrics. Princeton, Princeton University Press.

Angrist J D, Pischke J-S (2015) Mastering Metrics. Princeton, Princeton University Press.

Angrist J D, Pischke J-S (2017) Undergraduate Econometrics Instruction: Through our Classes, Darkly. IZA Institute of Labor Economics. Discussion Paper Series. January 2017.

Bailey M A (2017) real econometrics: the right tools to answer important questions. New York, Oxford University Press.

Becker W E, Greene W H (2001) Teaching Statistics and Econometrics to Undergraduates. Journal of Economic Perspectives 15(4): 169—182

Briand G, Hill, R C (2013) Teaching basic econometric concepts using Monte Carlo simulations in Excel. International Review of Economics Education 12 (2013): 60—79

Carsey T M, Harden J J (2014) Monte Carlo Simulation and Resampling Methods for Social Science. New Delhi, Sage.

Chen Y, Ebenstein A, Greenstone M, Li H (2013) Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. PNAS. Url: https://doi.org/10.1073/pnas.1300018110

Chen B, Pearl J (2013) Regression and Causation: A Critical Examination of Six Econometrics Textbooks. Real-World Economics Review, 65: 2-20.

Craft R K (2003) Using Spreadsheets to Conduct Monte Carlo Experiments for Teaching Introductory Econometrics. Southern Economic Journal 69(3): 726-735

Chattopadhyay R and Duflo E (2004) Women as policy makers: Evidence from a randomized policy experiment in India. Econometrica 72(5): 1409-1443

Dayal V (2020) Quantitative Economics with R: A Data Science Approach. Singapore, Springer Nature.

Deaton A, Cartwright N (2018) Understanding and misunderstanding randomized controlled trials. Social Science and Medicine: 210, 2-21

Elwert F (2013) Graphical Causal Models. In Morgan S L (ed) Handbook of Causal Analysis for Social Research, pp.245-274. Springer

Elwert F, Winship C (2014) Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol. 2014 Jul; 40: 31–53.

Gelman A (2013) Evidence on the impact of sustained use of polynomial regression on causal inference (a claim that coal heating is reducing lifespan by 5 years for half a billion people). Url: https://statmodeling.stat.columbia.edu/2013/08/05/evidence-on-the-impact-of-sustained-use-of-polynomial-regression-on-causal-inference-a-claim-that-coal-heating-is-reducing-lifespan-by-5-years-for-half-a-billion-people/

Accessed: 24 April 2020

Gelman A (2018) China air pollution regression discontinuity update. Url: https://statmodeling.stat.columbia.edu/2018/08/02/38160/ Accessed: 24 April 2020

Gelman A, Hill J (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Gerber A S, Green D P (2012) Field Experiments – Design, Analysis, and Interpretation. W W Norton and Company.

Hernan M A, Hsu J, Healy B (2019) A Second Chance to get causal inference right: A classification of data science tasks. Chance 32:1, 42-49.

Hill R C, Griffiths W E, Lim G C (2018) Principles of Econometrics. Wiley

Imai K (2016) Causal Mediation Q&A: Use of Directed Acyclic Graphs (DAGs) and Potential Outcomes in Social Science Research. Url: https://imai.fas.harvard.edu/talk/files/StanfordGSB16.pdf

Accessed on 24 April 2020.

Imbens G W (2019) Potential outcome and directed acyclic graph approaches to causality: relevance for empirical practice in economics. NBER Working Paper No. 26104

Johnston J, Dinardo J (1996) Econometric Methods. McGraw-Hill Education.

Kahneman D (2011) Thinking, fast and slow. London, Penguin Books.

Kaplan D (2009) Statistical Modeling: A Fresh Approach. Ingram.

Kennedy P (2003) A Guide to Econometrics. Cambridge, The MIT Press.

Kennedy P E (2009) Teaching Undergraduate Econometrics: A Suggestion for Fundamental Change. AER 88(2), Papers and Proceedings, 487-492

Lee D S, Lemieux T (2010) Regression Discontinuity Designs in Economics. Journal of Economic Literature 48 (June 2010): 281–355

Manski C F (1995) Identification Problems in the Social Sciences. London, Harvard University Press.

Manski C F, Pepper J V (2018) How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded variation assumptions. Review of Economics and Statistics 100(2): 232-244

Morgan S L, Winship C (2015) Counterfactuals and causal inference: methods and principles for social research. New York, Cambridge University Press.

O'Neill s, Kreif N, Grieve R, Sutton M, Sekhon J S (2016) Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv Outcomes Res Methodol. 2016; 16: 1–21.

Pearl J (2009) Causal inference in statistics: an overview. Statistics Surveys 3: 96—146

Pearl J, Glymour M, Jewell N P (2016) Causal Inference in Statistics: A Primer. Wiley

Pearl J, Mackenzie D (2019) The Book of Why: The New Science of Cause and Effect. Penguin.

Rodrik D, Subramanian A, Trebbi F (2004) Institutions Rule: the primacy of institutions over geography and integration in economic development. Journal of Economic Growth 9, 131-165

Rosenbaum P R (2010) Design of Observational Studies. New York, Springer.

Rosenbaum P R (2017) Observation and Experiment: An Introduction to Causal Inference. Cambridge, Harvard University Press.

Rubin D B (2004) Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies. Journal of Educational and Behavioural Statistics 29(3) : 343-367

Shipley B (2000) Cause and Correlation in Biology: A user's guide to path analysis, structural equations and causal inference. Cambridge, Cambridge University Press.

Stock J H, Watson M W (2011) Introduction to econometrics, third edition. Boston, Addison-Wesley.

Tukey J (1962) The future of data analysis. The Annals of Mathematical Statistics. 33: 1-67

Verbeek M (2012) A guide to modern econometrics. Chichester, John Wiley and Sons.

Recent IEG Working Papers:

Dash, Minati (2020). 'OUR PLACE IN THE FUTURE': AN EXPLORATION OF CHALAKI AMONG YOUNG MEN DISPOSSESSED BY A MINING PROJECT, Working Paper Sr. No.: 392

Mishra, Ajit and Samuel, Andrew (2020). Does it matter who extorts? Extortion by competent and incompetent enforcers, Working Paper Sr. No.: 391

Sahoo, Pravakar, Ashwani (2020). COVID-19 AND INDIAN ECONOMY: Impact on Growth, Manufacturing, Trade and MSME sector, Working Paper Sr. No.: 390

Das, Saudamini, Jha, Prabhakar and Chatterjee, Archana (2020). Assessing Marine Plastic Pollution in India, Working Paper Sr. No.: 389

Murty, M. N., Panda, Manoj and Joe, William (2020). Estimating Social Time Preference Rate for India: Lower Discount Rates for Climate Change Mitigation and other Long Run Investment Projects* , Working Paper Sr. No.: 388

Nuthalapati, Rao, Chandra Sekhara, Bhatt, Yogesh and Beero, K., Susanto (2020). Is the Electronic Market the Way Forward to Overcome Market Failures in Agriculture? , Working Paper Sr. No.: 387

Ray, Saon and Kar, Sabyasachi (2020). Kuznets' tension in India: Two episodes , Working Paper Sr. No.: 386

IEG Working Paper No. 393

INSTITUTE OF ECONOMIC GROWTH

University Enclave, University of Delhi
(North Campus) Delhi 110007, India
Tel: 27667288/365/424
Email: system@iegindia.org