Causal Inference with Observational Data[*]

Justin Esarey[†]

June 16, 2015

Abstract

Much of our interest in studying the world stems from the desire to change that world through political and technological intervention. If we wish to change social outcomes, we must understand how these changes are precipitated by factors under our control. But a central aphorism of social science is that correlation is not necessarily causation: when we observe that two things are associated, it does not logically imply that changes in one of them cause changes in the other. There are many alternative explanations for an observed relationship that must be ruled out before causality can be inferred. Experimental methods (covered elsewhere in this volume) allow us to rule out many of these alternatives via random assignment of the treatment and strict control of other factors in the environment. The subject of this chapter, and in large measure the subject of most research into quantitative analysis and research design, is what we can do to rule out these alternatives when we cannot conduct an experiment.

1 What is "causal inference?"

As Sekhon[1] notes in his essay on causal inference, every student of statistics quickly learns the aphorism that "correlation is not causation." Yet a great deal of research in the social sciences is aimed precisely at inferring causal relationships from empirical observation. Through clever research design and careful statistical modeling, we try to identify the hidden causal relationships embedded in the correlations that we can see. In this sense, a great deal of the work of policy scholars and those in allied

[*] Chapter 4 in *Analytics, Policy, and Governance*, eds. Jennifer Bachner, Kathyrn Wagner Hill, and Benjamin Ginsberg. New Haven: Yale University Press. Forthcoming in 2015. I thank Katie Dessi, Carolina Tchintian, Matt Blackwell, and Ahra Wu for helpful comments and corrections.

[†] Assistant Professor of Political Science, Rice University. E-mail: justin@justinesarey.com.

fields (political science, economics, sociology, etc.) could be considered part of a grand tradition in causal inference with many approaches and variations.

At present, *causal inference* tends to have a much more specific meaning. The term is now strongly linked to what is referred to[2] as the "credibility revolution" in empirical social science, a movement designed to improve the robustness and overall believability of quantitative research. As Angrist and Pischke describe, the sensitivity of non-experimental findings to model assumptions and disagreement between non-experimental studies and experimental studies of the same phenomena led many social scientists to re-examine their methodologies with an eye toward producing more robust inferences from non-experimental data.

The result of this "credibility revolution" in social science is a renewed interest in particular epistemological frameworks, research designs, and statistical models that are aimed at producing reliable causal inferences. The process begins with a question: what effect does imposing some treatment T have on an outcome of interest y? A policy scholar answering the question using a *causal inference* approach must first specify the conditions required to infer what changes in y are caused by T. These conditions are usually derived from a definition of causality and associated epistemological framework, such as those proposed by Rubin[3] or Pearl.[4] The person moves on to propose a research design and analytical strategy under which the necessary conditions for causal inference can be met; this proposal is collectively referred to as an *identification strategy*.[5] Generally speaking, the strategy is to mimic the conditions of an ideal experiment to maximize the validity of a causal inference. Where possible, the strategies explicitly include experimental techniques like random assignment to a treatment and strict control over external influences. When this is not possible, the strategies attempt to reconstruct the features of an ideal experiment by using instrumental variables, matching procedures, or other strategies to imitate experimental conditions.

While causal inference is of interest to all social scientists, it has a special value for policy-oriented academics and practitioners. Policy work is fundamentally about using targeted interventions to produce better social and economic outcomes. This goal leads naturally to an interest in determining and measuring how much these interventions actually change outcomes. The driving question of policy research is often, "how much should I expect a new program or policy to change an outcome of interest?"[6] The causal inference framework is designed to answer just this sort of question.

In this essay, I will explain the counterfactual reasoning framework that underlies causal inference according to the Rubin causal model and link this framework to the known advantages of experiments for causal inference. I describe three common procedures for causal inference in observational (viz., non-experimental) data: matching methods, regression models with controls, and instrumental variable models. The goal of the essay is to compile and summarize some of the basic ideas of causal inference as they are presented in a variety of sources and to relate these ideas to the problems of policy analysis.[7]

Throughout my exposition, I tie the discussion of causal inference to a policy-relevant question: how much does a high school education increase income for those who receive it? This question is interesting from the perspective of causal inference because income and education are difficult to disentangle. Although education increases one's own income, many confounding factors (such as innate ability) probably cause both.[8] Using U.S. Census data originally collected by Angrist and Krueger,[9] I show how matching, regression, and instrumental variables approaches can be used to recover the causal effect of a high school education on (log) weekly wage.

2 Counterfactual reasoning and average treatment effects

The Rubin causal model[10] is built on the idea of *counterfactual* reasoning: the causal effect of a treatment T is equivalent to the difference between what an outcome y

would be in the presence of T compared to the outcome in its absence. Under some conditions, we can obtain a reliable estimate of this effect with empirical observation.

2.1 The Rubin causal model[11]

For a particular individual observation i, the causal effect of T on y is:

$$y_i\left(T=1, X\right)-y_i\left(T=0, X\right) \tag{1}$$

where X includes any influences on y other than the treatment T. This equation says that the causal effect of a treatment on some unit is the difference between what we observe for that unit when the treatment is applied, and what we observe for that same unit when the treatment is *not* applied.

The fundamental problem of causal inference tells us that we can never observe this causal effect directly because we cannot simultaneously observe the same unit with and without the treatment condition.[12] Even if we observe the same unit at different times, influences on the unit that are included in X (including its past history with the treatment condition) have changed in the interval between the two observations; it is not a comparison of identical units.

We may, however, be able to observe a collection of units that are on average the same except for the presence or absence of the treatment. This will give us the average treatment effect (ATE) on the population that is represented by this sample. Consider[13] taking expectations over i for equation 1:

$$\begin{aligned} ATE &= E\left[y_i\left(T=1, X\right)-y_i\left(T=0, X\right)\right] \\ &= E\left[y_i\left(T=1, X\right)\right]-E\left[y_i\left(T=0, X\right)\right] \end{aligned}$$

and then taking expectations over X:

$$\begin{aligned} ATE &= E\left[E\left[y_i\left(T=1,X\right)|X\right]\right]-E\left[E\left[y_i\left(T=0,X\right)|X\right]\right] \\ &= E\left[y|T=1\right]-E\left[y|T=0\right] \end{aligned} \tag{2}$$

Note that I drop the individual indexing for conciseness in places where it can be inferred. This tells us that we might be able to estimate the average causal effect of a treatment in some population by comparing the average outcome of treated units to the average outcome of non-treated units in a sample from that population. Although we cannot simultaneously observe a single unit with and without the treatment, we *can* simultaneously observe a group with the treatment that is functionally identical to one without the treatment.

Under certain conditions, both of the components of equation 2 can be estimated in principle (unlike equation 1, which is conceptually impossible to observe). For example, we can draw a random sample of N observations out of the population of interest, randomly assign half of this sample to be exposed to the treatment, record their value for y, and then average over the $n = N/2$ observations to estimate the first term of equation 2:

$$\hat{E}\left[y|T=1\right]=\frac{1}{n}\sum_{i=1}^{n}y_i(T=1,X=X_i) \tag{3}$$

If we average y for the n observations randomly assigned *not* to be exposed to the treatment, we get an estimate of the second term of equation 2:

$$\hat{E}\left[y|T=0\right]=\frac{1}{n}\sum_{i=1}^{n}y_i(T=0,X=X_i) \tag{4}$$

This is the venerable logic of experimental design.[14]

When will the difference between (3) and (4) be equal to the ATE? We require two assumptions. First, we must make the Stable Unit Value Treatment Assumption (SUTVA); quoting Rubin (p. 961):[15]

> SUTVA is simply the a priori assumption that the value of y for unit i when exposed to treatment T will be the same no matter what mechanism is used to assign treatment T to unit i and no matter what treatments the other units receive, and this holds for all $i = 1, ..., N$ and all $T = 1, ..., T_k$.[16]

This assumption makes clear that the underlying estimand (Equation 2) exists and is not subject to influences from other units. Without this assumption, y_i could be a function not just of T and X but also of the other units' treatment assignments, T_{-i}. Thus our expectation $E[y|T = k]$ would have to be taken not only over i, but also over all possible combinations of other units' treatment assignments, in order to accurately reflect the ATE. This requirement would make most experiments prohibitively complex and data-intensive.

The second assumption we require is that assignment to the treatment is "strongly ignorable,"[17] which is defined by two conditions:

$$\begin{aligned} \{y\left(T=1,X\right), y\left(T=0,X\right)\} \perp T|X \\ \Pr(T=1|X) \in (0,1) \end{aligned}$$

This assumption tells us that a unit's assignment to a treatment condition is not a function of that unit's potential outcomes; it rules out, for example, the possibility that certain people are selected into a treatment because a particular outcome is expected.[18]

An integral property of the experimental environment ensures strong ignorability: random assignment to the treatment.[19] It also allows us to assume that the empirical distribution of X in the treatment cases matches the empirical treatment of X in the control cases, and furthermore that both of these empirical distributions match the distribution of X in the population from which the sample was drawn. In an experiment, equations (3) and (4) are estimators of $E\left[y(T = k)\right]$ for the treatment and control groups respectively.[20]

The assumption of strong ignorability is most transparently met under experimental conditions. Subjects in an experiment are selected randomly out of a population of interest, and then randomly assigned to the treatment or control condition. Any external influences on subjects are held constant by the carefully regulated environment of the laboratory. Consequently, the only difference (on average) between the subjects in the treatment and control conditions is the presence or absence of the treatment itself. Any observed difference in outcome between these two groups must therefore be attributable to the treatment. Furthermore, because the subjects are selected randomly out of the population of interest, the ATE we calculate from the experiment can be inferred to apply to that population.[21]

2.2 Non-experimental data

Outside the laboratory, where treatment conditions are generally *not* randomly assigned, we usually cannot directly compare the average outcomes of treated and untreated units in order to calculate an ATE. Very often even the fact of choosing to be treated is itself an important influence on outcomes. Returning to the example of high school education, it is plausible that students who stand to gain the most from a high school education are the students most likely to choose to attain it. It is therefore problematic to simply compare the expected earnings y for high school graduates to the expected earnings of non-graduates when estimating the effect of a high school education on income.

If we designate T as the treatment of receiving a high school education ($= 1$) or not ($= 0$) and S as being selected (or choosing oneself) to finish high school ($= 1$) or not (= 0), then this simple comparison of expected values[22] works out to:

$$E[y|T = 1, S = 1] - E[y|T = 0, S = 0] \tag{5}$$

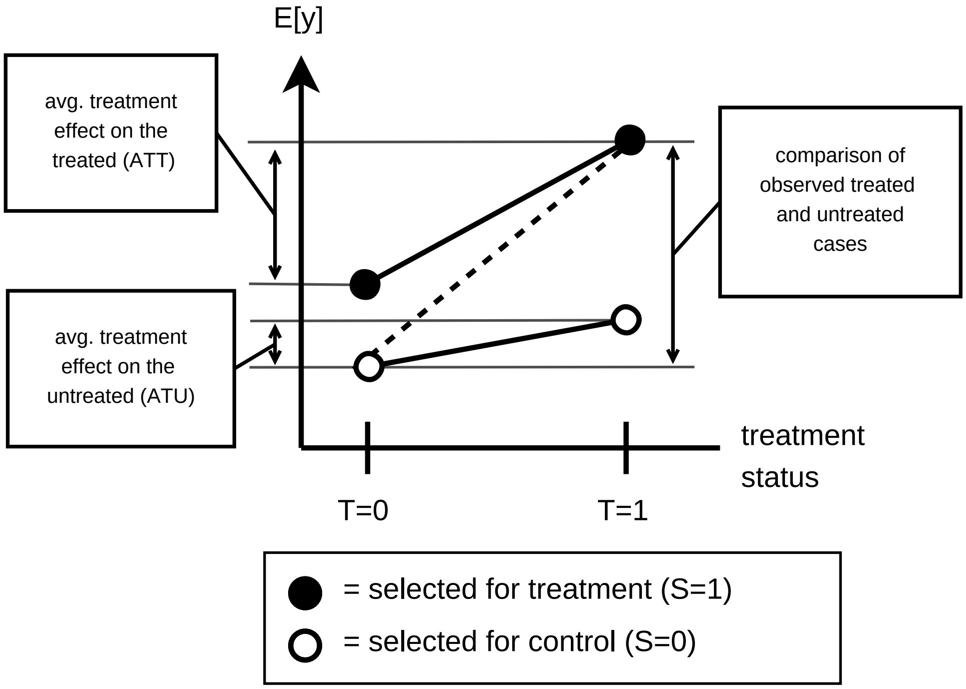
but the average treatment effect[23] (presuming no other confounding influences) is:

$$\begin{aligned} ATE = p\,(E[y|T = 1, S = 1] - E[y|T = 0, S = 1]) \\ + (1 - p)\,(E[y|T = 1, S = 0] - E[y|T = 0, S = 0]) \end{aligned} \tag{6}$$

where p is the fraction of the population selecting into the treatment condition. If assignment to receive a high school education is random, as in an experiment, then equations (5) and (6) are equivalent because $E[y|T = 1, S = 1] = E[y|T = 1, S = 0]$ and $E[y|T = 0, S = 1] = E[y|T = 1, S = 0]$; this is why the ATE is so easy to calculate in an experimental setting. But it is rare that assignment is truly random outside the laboratory, and in this case $E[y|T = 1, S = 1] \neq E[y|T = 1, S = 0]$ and $E[y|T = 0, S = 1] \neq E[y|T = 0, S = 0]$. That is, we expect the effect of the treatment on the sample selected to receive it to differ systematically from the treatment effect on the sample selected *not* to receive it.[24]

The problem is illustrated in Figure 1. I depict two cases: observations that have been assigned to the treatment ($S = 1$, shown as black circles) and observations that are assigned to the control ($S = 0$, shown as white circles). Each circle shows the expected value of y for each type of case when the treatment is present ($T = 1$) and absent ($T = 0$). The effect of the treatment on $S = 1$ cases is shown by the line connecting the black circles; this effect is larger than the treatment effect on $S = 0$ cases, which is shown by the line connecting the white circles. The simple comparison of equation (5) is equivalent to the dashed line connecting observed untreated cases ($T = 0, S = 0$) to observed treated cases ($T = 1, S = 1$) cases. Note that this dashed line is not equal to the treatment effect for cases selected into the treatment ($S = 1$), cases selected into the control ($S = 0$), or the average between the two.

Figure 1: An example of biased causal inference caused by simple comparison of treated and untreated cases


E[y]
avg. treatment effect on the treated (ATT)
avg. treatment effect on the untreated (ATU)
comparison of observed treated and untreated cases
treatment status
T=0
T=1
● = selected for treatment (S=1)
○ = selected for control (S=0)

This problem leads us to differentiate between estimating the treatment effect on units selected to receive the treatment and the treatment effect on units selected to receive the control. The average treatment effect on the treated population, or ATT,[25] is:

$$ATT = E[y|T = 1, S = 1] - E[y|T = 0, S = 1]$$

This is the degree to which a treatment impacts those who are selected to receive that treatment in a natural setting (the difference between the two black circles in Figure 1. To continue our education example, we might assume that the ATT of a high school education on income would be larger than the average treatment effect on the untreated, or ATU:

$$ATU = E[y|T = 1, S = 0] - E[y|T = 0, S = 0]$$

which corresponds to the difference between the two white circles in Figure 1. We might expect that $ATT > ATU$ in this case because rational, self-interested people are more likely to incur the effort and opportunity costs of an education when they expect to receive larger benefits from that education.

The distinction between ATT, ATU, and ATE is important for those interested in estimating the effect of a policy intervention. It is likely that any policy change will have heterogeneous effects on the population to which it is applied; some will be more greatly affected, others less so. The change in outcomes that we observe from any policy initiative will be a complex function of who receives the treatment that wasn't receiving it before, and how large an effect that the treatment has on that group.

In the example of high school education, suppose that the government chose to make it legally compulsory to complete high school (removing the option for older students to drop out). We might expect this change to have an impact on students' life outcomes. However, the observed change would likely be much closer to the ATU than to the ATE or ATT. Prior to the policy change, those who self-select into receiving a high school education are probably the people who most strongly benefit from that education.[26] These people received a high school education before the policy intervention, and will continue to receive it afterward. It is those who do *not* already self-select into receiving an education who will be most impacted by the policy.

2.3 A word on endogeneity

It is possible that increased education causes increased income, but that increases in one's income also cause better access to education even net of the effects of other variables (e.g., individual ability level or parental household income). That is, one may argue that education and income are *endogenous*, such that the two variables are simultaneously determined. Endogenously related variables are well-known to social scientists; to take a famous example, every student of microeconomics learns that the price and quantity of a good are simultaneously determined by the intersection of a supply and demand function. A simple regression predicting the quantity of a good using its price in a functioning market will not necessarily uncover the law of demand ($d\text{quantity}/d\text{price} < 0$) because the countervailing law of supply ($d\text{quantity}/d\text{price} > 0$) is simultaneously in operation.

The Rubin causal model does not explicitly contemplate simultaneous relationships of this type. This fact is evident from the definition of a causal effect embodied in equation (1), which writes y_i as a function of T and X. When y and T are endogenously related, their levels are *jointly* determined by the solution of a set of simultaneous equations. We can also see this in the assumption of strong ignorability, which presumes that potential outcomes are independent of the value of the treatment conditional on observable covariates. When y and T are endogenously related, then the value of T is a direct function of the value of y.

To operate inside of the framework of the Rubin causal model, we must be able to recast any apparent simultaneity between y and T as the product of an external factor X which determines y and T, thus restoring the assumption of strong ignorability. In the case of education and income, an apparent endogeneity between education levels and one's own income *might* be explainable as the product of innate ability, which determines both. But if y and T really are simultaneously determined and we wish to isolate the causal impact of T on y, we will have to take a different approach; a structural interpretation of instrumental variable models can serve this function.[27]

3 Applied causal inference: statistical procedures

The Rubin causal model provides an epistemological and ontological framework for drawing a causal inference. But actually drawing such an inference outside the lab requires us to specify how we will estimate an particular treatment effect in practice. I describe three approaches to observational data in this section: matching methods, regression, and instrumental variable models.

3.1 Regression with matched data

Recognizing the ATT and ATU as distinct estimands suggests a solution for estimating causal effects in a non-experimental setting. When estimating the ATT, for example, we might compare a unit that was selected to receive the treatment and did ($T = 1, S = 1$) to a unit that was selected to receive the treatment but didn't ($T = 0, S = 1$). If we match every selected and treated case in our sample to a selected but untreated case, we might be able to get an estimate of the sample ATT.[28]

Indeed, as demonstrated by Rosenbaum and Rubin,[29] a treatment effect can still be estimated on observational data as long as (1) SUTVA holds, (2) the treatment is strongly ignorable conditional on observable covariates X, and (3) observations can be compared according to their similarity on X or on a measure of the propensity to be assigned to the treatment $\Pr(S = 1)$ (which is typically a function of X).

In describing using regression with matching methods, I follow recent methodological literature in focusing on estimation of the sample ATT.[30] Methods to estimate the ATU and ATE using matching are extensions of these ideas; a variety of sources give helpful details on how to implement these procedures.[31]

3.1.1 Setting up a matching procedure to estimate a sample ATT

When estimating the sample ATT, all matching methods aim to pair cases in the treated group with cases in the untreated group that are the same in every observable respect *except* for the fact that they are untreated. Once this is done, the result is assessed for *balance*, the degree to which the observable characteristics of the treated sample match that of the newly constructed control sample; some procedures aim to achieve balance automatically without extensive post-matching assessment.[32]

There are many possible ways to set up a matching procedure,[33] and this essay is too short to give a thorough description of these choices or analyze the tradeoffs embedded in them. I opt instead to describe one approach in detail: *coarsened exact matching* (or CEM) as developed by Iacus, King, and Porro.[34]

CEM approaches the matching problem by "coarsening" multivalued covariates into a small number of discrete bins (or strata) that cover a range of values for each covariate; these ranges can be automatically determined by the program or defined by the user. Observations are matched when they fall into a stratum containing a non-zero number of treated and control cases, meaning that these cases have close to the same value (inside of the stratum range) for all covariates; other observations are discarded. Inside of each stratum, treatment and control cases are weighted so that each receives equal weight inside of the stratum. The within-stratum weights are:

$$\omega_i^s = \begin{cases} 1 & \text{if } T_i = 1 \\ \frac{m_{T=1}^s}{m_{T=0}^s} & \text{if } T_i = 0 \end{cases}$$

where i indexes observations inside of stratum s, T_i gives the treatment status of observation i, and $m_{T=k}^s$ is the number of matched observations in stratum s with treatment status k. Let $m_{T=k}$ be the total number of matched observations in the data set with treatment status k. Using these within-stratum weights, $\sum_{i \in \mathbb{S}_{T=1}} \omega_i^s = m_{T=1}$ (as appropriate) but $\sum_{i \in \mathbb{S}_{T=0}} \omega_i^s$ also $= m_{T=1}$, which means that the sum of these weights would not equal the total matched sample size.[35] Therefore every observation i in the matched data set is instead assigned a normalized weight equal to:

$$\omega_i = \begin{cases} 1 & \text{if } i \in \mathbb{S}_{T=1} \\ \frac{m^s_{T=1}/m_{T=1}}{m^s_{T=0}/m_{T=0}} & \text{if } i \in \mathbb{S}_{T=0} \end{cases} \tag{7}$$

which correspond to the weights given by Iacus, King, and Porro.[36]

If there are observations in a stratum that does not include both treatment and control cases, then these cases are dropped from the analysis. Dropping these cases means that our causal estimand is the *local sample* ATT, "the treatment effect averaged over only the subset of treated units for which good matches exist among available controls".[37]

3.1.2 Analysis of the matched data

After matching has been performed, the matched sample can be used to estimate a quantity of interest. If we are aiming to estimate the sample ATT, the matched sample ideally includes all treatment cases paired with appropriately selected and weighted control cases. We might then simply compare the average value of the outcome variable y between the two groups to estimate the sample ATT; this is equivalent to specifying the following regression model on the matched and weighted sample:

$$y_i = \beta_T * T_i + \beta_C * (1 - T_i) + \varepsilon_i$$

where i indexes observations in the matched sample, $T \in \{0, 1\}$ is a binary indicator of being exposed to the treatment, and the β_T and β_C coefficients correspond to estimates of the predicted value of y for the treatment and control cases respectively. The sample ATT is $\beta_T - \beta_C$, and its regression estimate is $\hat{\beta}_T - \hat{\beta}_C$.[38] More simply, we can estimate:

$$y_i = \beta_0 + \beta_{TE} * T_i + \varepsilon_i$$

The estimated coefficient $\hat{\beta}_{TE}$ is an estimator of the sample ATT.[39]

Alternatively, one might include the K many elements of X that were previously used to create the matched sample as control variables in the regression:

$$y_i = \beta_T * T_i + \beta_C * (1 - T_i) + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \tag{8}$$

An estimate of the sample ATT is $\frac{1}{n}\sum_{i \in \mathbb{S}_{T=1}} [y_i - \hat{y}_i(T_i = 0, X_i)]$ from this modified regression for the set of n treated cases in the sample $\mathbb{S}_{T=1}$.[40] The y_i cases are simply the observed outcomes for the treated cases, while the $\hat{y}_i$ are constructed by computing the regression prediction for these same cases for $T = 0$ (and with all other covariates X left the same) using equation (8). The sample ATT can also be estimated[41] using $\hat{\beta}_{TE}$ in the regression:

$$y_i = \beta_0 + \beta_{TE} * T_i + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \tag{9}$$

These procedures have the advantage of being "doubly robust" in that they will yield a correct estimate of the sample ATT if either the matching procedure achieves appropriate balance *or* the regression model in equations (8) and (9) is a mathematically accurate description of the true functional relationship between y, T, and X.[42]

3.2 Regression analysis of unmatched data

We can also run the regression in equation (9) *without* performing a matching procedure first; this is certainly a common practice in the social scientific literature. In this case, $\hat{\beta}_{TE}$ will have a causal interpretation[43] in the event that the regression estimates are unbiased. It is sufficient that:

1. The model is correctly specified (that is, equation (9) accurately describes the data generating process);

2. The ε noise term is uncorrelated with the regressors T and X; and

3. The matrix of all covariates, including the treatment indicators, has full rank (i.e., the number of observations exceeds the number of covariates).

These assumptions are standard in the econometric literature for unbiased estimation of regression coefficients.[44]

Under these conditions, $\hat{\beta}_{TE}$ is an estimate of the ATT, ATU, and ATE of the treatment. We see this via the derivative of equation (9) with respect to T:

$$\frac{dy_i}{dT} = \beta_{TE}$$

Under the assumption of correct specification, this relationship is constant for all treated and control cases, and thus $ATT = ATU = ATE$.

The cost of this greatly simplified procedure lies in the stronger assumptions we must make about the data generating process in order to derive a causal inference in the sense of the Rubin causal model. The Gauss-Markov theorem tells us that the regression-only approach will be the most efficient linear unbiased estimate of causal estimands possible, meaning that the uncertainty around our answers will be smaller than any other (linear) unbiased estimator–but only when additional assumptions hold.[45]

3.3 Instrumental Variables

Instrumental variables (or IV) techniques provide a third avenue for drawing causal inferences. The procedure attempts to replace the random assignment to a treatment that takes place in a laboratory with a model for treatment assignment that is known to be unrelated to the outcome of interest except through its effect on $\Pr(T = 1)$, the probability of receiving the treatment. The idea hinges upon being able to find a set of eponymous *instrumental variables*, variables that are correlated with $\Pr(T = 1)$ but not correlated with y except through $\Pr(T = 1)$. We thus substitute the controlled assignment to treatment of a laboratory for assignment to treatment by random observable factors.

If we find a binary instrumental variable (call it z), with $z \in \{0, 1\}$, then the Wald IV estimator[46] for the relationship between T and y in a set of observations $i \in 1...N$ is:

$$\hat{\rho}_w = \frac{\hat{E}[y_i|z_i = 1] - \hat{E}[y_i|z_i = 0]}{\hat{E}[T_i|z_i = 1] - \hat{E}[T_i|z_i = 0]} \tag{10}$$

More generally, we can estimate the impact of treatments T on an outcome y using instruments Z with instrumental variables regression:[47]

$$\hat{\beta}_{IV} = (X'P_V X)^{-1} X'P_V y$$

where P_V is the projection matrix onto the space defined by V:

$$P_V = V(V'V)^{-1}V'$$

where V is an $\{N \times L\}$ matrix of the N observations in rows, and the K control variables and $L - K$ instruments in the columns. X is the $\{N \times (K + 1)\}$ matrix of the treatment variable and the controls; thus, $V = [\, {\scriptstyle Z\ X_{-T}} \,]$ where X_{-T} is the X matrix deleting the column corresponding to the treatment variable. The relationship between T and y, β_T, is estimated by the element of the $\hat{\beta}_{IV}$ vector corresponding to the treatment variable T. Control variables need not be included, but can be in order to improve the efficiency of results; T can be continuous or binary.

3.3.1 A causal interpretation for IV: the local average treatment effect

Under some assumptions, $\hat{\rho}_w$ and $\hat{\beta}_T$ have a causal interpretation under the Rubin causal model. Consider the Wald IV estimator first; I follow the presentation of Angrist, Imbens, and Rubin.[48] Distinguish between T_i and z_i, unit i's value for the treatment and instrument respectively, and $\mathbf{T}$ and $\mathbf{z}$, the $N \times 1$ vectors of treatment and instrument values for the entire set of units. Define the following terms:

- $T_i(\mathbf{z})$ is unit i's treatment status contingent on the configuration of all units' instrument value $\mathbf{z}$;
- $y_i(T_i(\mathbf{z}), \mathbf{z})$ is unit i's outcome value contingent on the unit's own treatment status $T_i(\mathbf{z})$ and all units' instrument vector $\mathbf{z}$;
- $y_i(\mathbf{T}, \mathbf{z})$ is unit i's outcome value contingent on all units' treatment status and instrument value; and
- $\mathbf{y}(\mathbf{T}, \mathbf{z})$, the vector of all units' outcome values contingent on all units' treatment status and instrument value.

A causal interpretation of ρ_w requires:

1. The stable unit treatment value assumption (SUTVA) as defined earlier, which can now be formally stated as:
   (a) $z_i = z'_i \rightarrow T_i(\mathbf{z}) = T_i(\mathbf{z}')$
   (b) $z_i = z'_i$ and $T_i = T'_i \rightarrow y_i(\mathbf{T}, \mathbf{z}) = y_i(\mathbf{T}', \mathbf{z}')$
2. Random assignment[49] of the instrument: $\Pr(\mathbf{z} = \mathbf{c}) = \Pr(\mathbf{z} = \mathbf{c}')$ for any two $N \times 1$ instrument vectors $\mathbf{c}$ and $\mathbf{c}'$;
3. The *exclusion restriction*: z affects y only through its effect on $\Pr(T = 1)$, or $\mathbf{y}(\mathbf{T}, \mathbf{z}) = \mathbf{y}(\mathbf{T}, \mathbf{z}')$ for all values of $\mathbf{z}$, $\mathbf{z}'$, and $\mathbf{T}$; and
4. *Strong monotonicity*: the instrumental variable moves the probability of assignment to treatment in one direction, or $T_i(z_i = 1) \geq T_i(z_i = 0)$ for all $i \in 1...N$ and $\exists j : T_j(z_j = 1) > T_j(z_j = 0)$

Under these conditions, proposition 1 in Angrist, Imbens, and Rubin[50] shows that $\hat{\rho}_w$ is an estimate of the *Local Average Treatment Effect* or LATE:

$$LATE = \frac{E[y_i(T_i(z_i = 1), z_i = 1) - y_i(T_i(z_i = 0), z_i = 0)]}{E[T_i(z_i = 1) - T_i(z_i = 0)]} \tag{11}$$

Given our earlier assumptions, the LATE has an interesting interpretation: it is the effect of the treatment on the outcome for that subgroup of people who would take the treatment when exposed to the instrument ($T_i(z_i = 1) = 1$) but would *not* take the treatment when not exposed to the instrument ($T_i(z_i = 0) = 0$). Consider Table 1, a duplication of Table 1 in Angrist, Imbens, and Rubin.[51] The table shows that there are four types of observations in the $z = 1$ and $z = 0$ pools: never-takers, always-takers, defiers, and compliers. The never-takers and always-takers have the same treatment status regardless of z, compliers take the treatment when $z = 1$ and do not when $z = 0$, and defiers take the treatment when $z = 0$ but do not when $z = 1$. The existence of defiers is ruled out by the strong monotonicity assumption. The exclusion restriction tells us that the never- and always-takers have the same value for y regardless of z because their treatment status does not change. Thus, the only difference in the treated and un-treated samples is the response of compliers.

Table 1: Unit types and treatment assignment
(duplicates Table 1 in Angrist, Imbens, and Rubin 1996)

		$T_i(0)$	
		0	1
$T_i(1)$	0	Never-takers	Defiers
	1	Compliers	Always-takers

Ergo, $\hat{E}[y_i|z_i = 1] - \hat{E}[y_i|z_i = 0]$ is a weighted average of the effect of the treatment on compliers, and zero effect (for the never- and always-takers).[52] Once we multiply by the estimated inverse proportion of compliers in the population, $\hat{E}[T_i|z_i = 1] - \hat{E}[T_i|z_i = 0]$, we obtain an estimate of the treatment effect on compliers (the LATE).

When 2SLS is used to estimate β_T, the resulting estimand is still a local average treatment effect, albeit averaged over some characteristic of the sample; quoting Angrist and Pischke (p. 173):[53]

> 2SLS with multiple instruments produces a causal effect that averages IV estimands using the instruments one at a time; 2SLS with covariates produces

> an average of covariate-specific LATEs; 2SLS with variable or continuous treatment intensity produces a weighted average derivative along the length of a possible non-linear causal response function. These results provide a simple causal interpretation for 2SLS in most empirically relevant settings.

3.3.2 The LATE and its connection to policy scholarship

The LATE is of particular interest to those studying policy because of the close conceptual connection between instruments and policy interventions. As the exposition above indicates, an instrument is an influence which has an isolated impact on exposure to the treatment; that is, it impacts receipt of the treatment without impacting (or being impacted by) anything else. Under ideal circumstances, the instrument is actually randomly assigned just as a treatment would be in an experiment. But the instrument is not the treatment itself.

Similarly, most policy interventions are not identical to the treatment. Policy interventions are instruments that we hope will change behavior, which in turn changes an outcome of interest. The change in behavior is the actual treatment. Scholars of policy are thus intrinsically interested in the causal impact of a behavioral change on outcomes through the mechanism of instruments.

Furthermore, not everyone who is exposed to the instrument will receive the treatment, and not everyone who is not exposed to the instrument will not receive the treatment. This is analogous to the effect of a policy intervention. For example, imposing cigarette taxes (the instrument) will cause some people to stop smoking (the treatment), but some people will smoke whether the tax exists or not while others will remain non-smokers regardless of the tax. The relevant information for a policy maker is how much the change in taxes will change outcomes (like life expectancy or health care costs) through its impact on the behavior of those who do respond to the instrument. This is precisely the definition of the LATE.[54]

Returning to our example of education, an IV estimator can be used to tell us the effect of attaining a high school education (the treatment) on income (the outcome of interest) for the subset of people whose school attendance choices are actually changed by a policy initiative. An ideal instrument would be created by a government initiative to randomly select a subset of 1,000 students in a particular cohort and guarantee to pay them a cash bounty for completing high school. The instrument is ideal because (a) we strongly expect the instrument to be correlated with the decision to complete high school, and (b) we do not expect weekly earning potential to be affected by the receipt of a one-time cash bounty except through its effect on whether the student completes high school. Furthermore, we know the instrument is not spuriously correlated with other influences on income (e.g., through individual ability or parental education levels) because it has been randomly assigned: participants cannot preferentially select themselves into participation in the program.

Inside of the group, some set of "always-takers" will complete high school with or without the bounty; another set of "never-takers" will not complete a high school education even with the bounty. However, some set of individuals will complete high school with the bounty when they would not have done so without it; this set of "compliers" has their behavior changed by the program. By comparing the expected earnings of the set of people in the program to the set of people not in the program using equation (10), we can calculate the impact of high school completion on this compliant population.

Note, however, that any particular estimated LATE is not necessarily a measure of the impact of an arbitrary policy change.[55] It is not, for example, when "compliers" with an instrument are not the same as the "compliers" with a policy change. Quoting Carneiro, Heckman, and Vytlacil[56] (p. 16):

> ...if the instrumental variable we use is exactly the policy we want to evaluate, then the IV estimand and the policy relevant parameter coincide. But whenever that is not the case, the IV estimand does not identify the effect

> of the policy when returns vary among people and they make choices of treatment based on those returns. For example, if the policy we want to consider is a tuition subsidy directed toward the very poor within the pool [of people who would not ordinarily attend schooling], then an instrumental variable estimate based on compulsory schooling will not be the relevant return to estimate the policy.

4 Application: the effect of high school education on earnings

To illustrate the application of causal inference techniques to an important policy problem, I reanalyze data collected by Angrist and Krueger[57] using standard regression techniques, regression with coarsened exact matching, and two stage least squares regression using an instrumental variable. Angrist and Krueger originally used the public use file of the 1970 U.S. Census to study the relationship between income and total years of education received. I will use their replication data to estimate the causal impact of receiving a high school education on income; the original paper briefly discusses this relationship (pp. 1004-1005) but does not directly study it.

4.1 Regression with control variables

The easiest and most straightforward approach to estimating the causal effect of a high school education on income is to simply regress one on the other in a large observational data set of individual respondents and control for likely sources of spurious correlation. The Angrist and Krueger data set is derived from the public sample of the 1970 U.S. Census; I analyze the cohort of respondents born between 1920 and 1929.[58] It contains the following variables:

- **log Weekly Earnings**

- **Education** completed in years
- **Race** (1 = black, 0 = white)
- **Married** (1 = yes, 0 = no)
- **SMSA** (1 = respondent lives in the central city of a metropolitan statistical area)
- Census **region** indicators (9 binary dummy variables)
- **Year of birth** (10 binary dummy variables)

I recode the years of education variable to a high school completion variable (= 1 if the respondent has ≥ 12 years of education, and = 0 otherwise); this is the treatment variable, T.

I use this data to estimate the following regression model:

$$y_i = \beta_0 + \beta_{TE} * T_i + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \tag{12}$$

The control variables of race, marriage status, SMSA, census region, and year of birth are included as elements $k \in 1...K$. The results are shown in Table 2, Column 1.

The regression coefficient for attaining high school is 0.346, indicating that high school graduates earn $\approx 35\%$ more than non-graduates on average. If equation 12 is an accurate specification of the data generating process, then on average receipt of a high school education causes a 35% increase in weekly wages for both the populations that received the treatment (high school graduates) and those who did not (non-graduates).

4.2 Regression with coarsened exact matching

One weakness of the regression procedure is that it relies on an accurate specification of the data generating process. By contrast, matching procedures do not require this condition. We still, however, require SUTVA and strong ignorability of $\Pr(T = 1)$ contingent on our set of control variables. Thus, the validity of the procedure re-

Table 2: How much does completing high school change future income?

	(1) Regression	(2) CEM w/ Reg	(3) IV 2SLS
At least 12 yrs. schooling (1 = yes)	0.346^{***}	0.339^{***}	0.388^{**}
	(139.39)	(140.94)	(3.21)
Race (1 = black)	-0.367^{***}	-0.357^{***}	-0.355^{***}
	(-82.02)	(-61.49)	(-10.73)
Married (1 = yes)	0.306^{***}	0.305^{***}	0.304^{***}
	(78.84)	(76.21)	(45.64)
Metro Area (1 = center city)	-0.156^{***}	-0.143^{***}	-0.153^{***}
	(-58.65)	(-53.57)	(-20.50)
N	247199	247090	247199

t statistics in parentheses
dependent variable: log weekly wage in \$. Data from the 1920-1929 Cohort of the 1970 U.S. Census.
All models also include year of birth dummy variables, region dummies, and a constant.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

quires (among other things) that we use a complete set of confounding variables in the matching process.

As noted above, there are many possible choices that one can make as a part of a matching procedure. For this demonstration, I elect to use *coarsened exact matching* (or CEM) as developed by Iacus, King, and Porro[59] and implemented in Stata by Blackwell et al.[60]

I use the `cem` command in Stata to match the Angrist and Krueger data on all control covariates (including year of birth and region of residence dummy variables); the process discards 89 control and 20 treatment cases that fall into a stratum with no matching observations and assigns the rest weights given by equation (7). I then repeat the regression of equation (12) on the matched and weighted data; the principle of "double robustness" ensures an accurate estimate of the local sample ATT if either the matching procedure is adequate *or* the regression is accurately specified.

The matching procedure is successful at improving balance between the observable characteristics of the treated and control samples. Balance is calculated in the `cem` software using the $\mathcal{L}_1$ statistic:

$$\mathcal{L}_1(f,g) = \frac{1}{2}\sum_{\ell_1 \ldots \ell_K} |f_{\ell_1 \ldots \ell_K} - g_{\ell_1 \ldots \ell_K}|$$

where K indexes the number of (coarsened) independent variables, $\ell_1 \ldots \ell_K$ is the coordinate of a matching stratum, $f_{\ell_1 \ldots \ell_K}$ is the frequency of treated cases in the stratum, and $g_{\ell_1 \ldots \ell_K}$ is the frequency of control cases in the stratum. Perfect balance is indicated by an $\mathcal{L}_1 = 0$.[61] Before performing matching on the Angrist and Krueger data set, $\mathcal{L}_1 = .159$; after matching, $\mathcal{L}_1 = 2.29 * 10^{-14}$.

The results are shown in Column 2 in Table 2. The coefficient for high school attainment is 0.339, indicating that earning a high school diploma causes a 33.9% increase in weekly wage for those in the sample who received the treatment (i.e., this is the local sample ATT). This is very similar to the estimate of a 34.6% increase that we achieved with a regression analysis alone.

4.3 Instrumental variables models

Finally, we consider the possibility of using the 2SLS IV procedure to determine the causal effect of high school education on earnings. The first and most important question is: what instrumental variable is correlated with the decision to complete high school, but not correlated with weekly income except through its effect on high school completion?

Angrist and Krueger argue that an individual's quarter of birth (that is, the quarter of the year in which a student was born) is a good choice of instrument. Their reasoning is that compulsory school attendance laws in the United States (which require students to achieve a certain age before they may discontinue their education) create a link between educational attainment and quarter of birth. They argue (on p. 982):

> Students who are born early in the calendar year are typically older when they enter school that children born late in the year. ...Because children born in the first quarter of the year enter school at an older age, they attain the legal dropout age after having attended school for a shorter period of time than those born near the end of the year. Hence, if a fixed fraction of students is constrained by the compulsory attendance law, those born in the beginning of the year will have less schooling, on average, than those born near the end of the year.

4.3.1 2SLS estimates

In their original article, Angrist and Krueger interact three separate dummies for quarter of birth (one for each quarter, omitting the fourth as a base category) with ten years of birth dummies (one for each year between 1920-1929) for a total of 30 instruments. We can use this larger set of instruments, along with the full set of control variables, to create a 2SLS IV estimate of the effect of completing high school on log weekly earnings. The F-test of a regression of all instruments on high school completion rejects the null that the full set of instruments is not related to completing high school ($F_{(30,247148)} = 3.48$, $p < 0.001$). The R^2 of this regression $= 0.0465$, which is relatively weak.

The presence of multiple instruments allows me to perform a Sargan test for overidentifying restrictions in the set of instruments.[62] The Sargan test checks a key assumption of the IV procedure:

$$E[Z'\varepsilon] = 0$$

That is, the $N \times (L-K)$ matrix of instrumental variables Z should be uncorrelated with the second stage 2SLS errors. In the context of the 2SLS model, this is a restatement of the exclusion restriction: $\mathbf{y}(\mathbf{T}, \mathbf{Z}) = \mathbf{y}(\mathbf{T}, \mathbf{Z}')$ for all values of $\mathbf{Z}$, $\mathbf{Z}'$, and $\mathbf{T}$. When the

2SLS model of equation (11) is just identified (one instrument with a single treatment condition), we cannot perform this test. The abundance of instruments allows us to implicitly compare the results of single-instrument models to one another; if the instruments are valid, "the estimates should differ only as a result of sampling error."[63] The test is equivalent to regressing the fitted error terms of the 2SLS regression $\hat{\varepsilon}$ on the set of instruments Z; under the null hypothesis that all instruments are valid, N times the R^2 from this regression is distributed $\chi^2_{L-(K+1)}$.[64]

For the quarter of birth × year of birth dummy instruments, the Sargan test yields a p-value of 0.0885. This is a marginal result, close to rejecting the validity of the instruments but not quite doing so at the 0.05 level. Thus, we cautiously proceed to interpret the 2SLS estimates of the relationship between high school education and log weekly earnings as a measurement of the local ATE.

The 2SLS results are shown in column 4 of Table 2. The coefficient of 0.388 indicates a LATE of a 38.8% increase in weekly earnings associated with attainment of a high school education. Thus, the model indicates that a high school graduate who completed high school as a result of birth timing (when they would not have otherwise) receives 38.8% more weekly earnings as a result of completing high school. The uncertainty associated with these results is somewhat higher than in the other models due to the addition of the second stage of the 2SLS estimator and the comparative weakness of the instruments; this is reflected in smaller t-statistics. However, the magnitude and direction of the 2SLS results are extremely similar to those for our plain regression and regression-with-matching estimates.

5 Conclusion: what can policy makers and scholars learn from the causal inference movement?

As I hope this chapter has communicated, it is challenging to use observational data from outside the laboratory to draw a causal inference. Observing an association between the value of a treatment T and an outcome y is not sufficient to conclude that a change in the treatment will cause a change in expected outcome $E[y]$. At a minimum, we must rule out the possibility that confounding factors could be simultaneously causing y and T; confounding can create an observed association beteween y and T where no causal association exists.

Even if changes in T *do* cause changes in y, the strength of the observed association is unlikely to correspond to the magnitude of the causal impact of an change in T initiated by a policy intervention. In observational data, people choose whether to be exposed to a treatment, and thus those who stand to derive the greatest benefit from a treatment are often the most likely to choose to receive the treatment. Thus, simply comparing the outcomes of treated and untreated units from observational data is likely to give a highly misleading indication of how much outcomes would change if a change in treatment status was *imposed* by an external event, such as a legal mandate.

In my judgment, the most important lesson that policy-makers and scholars can draw from the causal inference literature is that describing (let alone predicting) the practical impact of a policy change using observational data is complicated. Scholarship on causal inference alerts us to the practical problem of confounding, but perhaps even more importantly it reminds us that causal relationships are heterogeneous and that policy interventions will not cause uniform changes across a target population.

Consider the results of our inquiry into the causal impact of high school completion on income. Which, if any, of the results in Table 2 is most informative about the potential increase in earnings that would result if we legally compelled all students to complete high school? Our matching estimator of the sample ATT is probably not the right estimand; this is the effect of high school completion on those who already receive it, not on the population of those with less education whose behavior would be changed.

The 2SLS IV estimator of the LATE is perhaps more informative; this, at least, tells us the response of people whose status was changed as a result of an accident of birth. But there is still a substantial set of "never-compliers" whose decision to drop out of high school was not changed by time of birth, and these non-compliers would be a significant subset of the group affected by the policy change. Our LATE estimate is not guaranteed to describe the change in earnings for this subset.[65]

Moreover, there is the possibility that our 2SLS estimates are subject to "coarsening bias."[66] Increased years of schooling have an impact on weekly earnings even if high school is not completed, and the quarter of birth instrument impacts years of schooling aside from increasing the chance of completing high school; this creates a pathway between the instrument and the dependent variable that does not pass through the treatment (high school completion). Because high school completion necessarily entails attaining more years of schooling, it may be challenging to causally separate these effects with an instrumental variable.

All of these approaches depend directly on the stable unit treatment value assumption (SUTVA), and there is a strong reason to believe that SUTVA would not hold in the presence of our proposed policy change. Weekly earnings are determined in part by the market forces of supply and demand, and increasing the supply of high school graduates via the proposed policy without changing demand is likely to drive down the wage for a person with a high school diploma. In terms of SUTVA, the relationship between treatment and outcome for person i depends on the treatment status of person j; a high school diploma is more valuable if fewer people have one. This is particularly true if education serves a signal of underlying quality for employers rather than a source of practical skills.[67] Thus, we have reason to doubt that *any* of our results in Table 2 are truly reflective of the probable impact of our policy intervention on income.[68]

None of this is to say that it is impossible to derive a causal inference from observational data, or that quantitative tools cannot determine the causal impact of a policy intervention. Instead, it is important to precisely determine the desired causal estimand before performing an empirical analysis, ensure that our empirical design gives us the best chance of obtaining an accurate estimate, and be aware of the sensitivity of any causal inference we draw to the assumptions we had to make in order to justify that inference. For example, the estimates in Table 2 are more defensible as estimates of the causal impact of a smaller-scale policy intervention to increase high school graduation that is unlikely to influence the market as a whole; this is true because such a program is unlikely to entail a violation of SUTVA. The causal inference framework is well-suited to reminding us of these limitations such as these and preventing us from drawing inferences beyond the support of our evidence.

Notes

1. Sekhon, "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods."

2. Angrist and Pischke, "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics."

3. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies"; Holland, "Statistics and Causal Inference."

4. Pearl, *Causality: models, reasoning and inference*.

5. Angrist and Card, "Empirical Strategies in Labor Economics," p. 1284; Angrist and Pischke, "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," pp. 16-17

6. Angrist and Card, "Empirical Strategies in Labor Economics," p. 1282-1283.

7. There are other such summaries which share similar themes and structures but are somewhat different in their emphases and depth of topical coverage. See, for example, Angrist and Card, "Empirical Strategies in Labor Economics," Sekhon, "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods," and Stuart, "Matching Methods for Causal Inference: A Review and a Look Forward."

8. See Willis, "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings

Functions" for a review of ideas and findings from relevant literature.

9. Angrist and Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?"

10. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies"; Holland, "Statistics and Causal Inference."

11. This subsection loosely follows the presentation of Holland, "Statistics and Causal Inference."

12. Ibid., p. 947.

13. The following illustration is given in Titiunik, "ATE, ATT and Potential Outcomes," particularly on p. 8.

14. See the section titled "Independence" in Holland, "Statistics and Causal Inference," pp. 948-949 and Sekhon, "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods," pp. 272-275.

15. Rubin, "Comment: Which Ifs Have Causal Answers," p. 961.

16. Note that I have changed Rubin's notation in this quote to match the notation used in the rest of this essay.

17. Rosenbaum and Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," p. 43.

18. Imbens, "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review" calls these two conditions "unconfoundedness" and "overlap" respectively (pp. 7-8).

19. Holland, "Statistics and Causal Inference," p. 948.

20. Imbens, "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," p. 8.

21. Holland, "Statistics and Causal Inference," pp. 948-949.

22. The notation and basic ideas of this section are similar to those presented in Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Section 2.2 and Chapter 4.

23. ibid., p. 103.

24. See ibid., pp. 46-48 and Holland, "Statistics and Causal Inference" p. 948. I assume that there is no possibility for those selected to receive the treatment not to receive it ($T = 0, S = 1$) or for those not selected to receive the treatment to receive it ($T = 1, S = 0$). The possibility of the former leads to estimation of an Intention to Treat Effect (ITT), $ITT = E[y|S = 1] - E[y|S = 0] = q(E[y|T = 1, S = 1]) + (1 - q)E[y|T = 0, S = 1] - E[y|T = 0, S = 0]$, which averages the results of all those selected to receive the treatment (where $(1 - q)$ is the proportion of "noncompliers" who do not take the treatment). For more information on the ITT, see Angrist, Imbens, and Rubin, "Identification of Causal Effects Using Instrumental Variables" and Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, pp. 163-164.

25. See Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, p. 42.

26. See Willis, "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions," section 2.4 (pp. 534-535). Evidence for this phenomenon in the context of college education is given by Carneiro, Heckman, and Vytlacil, "Estimating Marginal Returns to Education."

27. Davidson and MacKinnon, *Econometric Theory and Methods*, Chapter 8.

28. These ideas are covered in Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Chapter 4.

29. Rosenbaum and Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," see especially p. 46, Theorem 4.

30. See, e.g., Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, pp. 105-116, Ho et al., "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," Iacus, King, and Porro, "Causal inference without balance checking: Coarsened exact matching," Sekhon, "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods," pp. 275-276, and Blackwell et al., "cem: Coarsened exact matching in Stata."

31. See, e.g., Imbens, "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review" and Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, pp. 98-105.

32. Iacus, King, and Porro, "Causal inference without balance checking: Coarsened exact matching."

33. To see these choices laid out in detail, consider the options available in the `MatchIt` package for using matching methods in the R statistical environment of Ho et al., "Matchit." See also Section 4.4.1, pp. 107-109 of Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*.

34. Iacus, King, and Porro, "Causal inference without balance checking: Coarsened exact matching."

35. King et al., "Comparative effectiveness of matching methods for causal inference," p. 4.

36. Iacus, King, and Porro, "Causal inference without balance checking: Coarsened exact matching," p. 8.

37. Ibid., p. 5.

38. The estimated sample *ATT* can be assessed for statistical significance using the relevant t-statistic for the estimated quantities:

$$t = \frac{\hat{\beta}_T - \hat{\beta}_C}{\hat{\text{se}}_{\text{dif}}}$$
$$\hat{\text{se}}_{\text{dif}} = \sqrt{\text{Var}(\hat{\beta}_T) + \text{Var}(\hat{\beta}_C) - 2\text{Cov}(\hat{\beta}_T, \hat{\beta}_C)}$$

The variances can be obtained from the variance-covariance matrix normally produced in a regression analysis, and will be appropriate in the case that we treat the matching procedure as non-stochastic and the usual Gauss-Markov assumptions for OLS are met in this case; see pp.

207-208, 223-224 of Ho et al., “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference.”

39. See p. 537 of Blackwell et al., “cem: Coarsened exact matching in Stata”; see also Imbens, “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” p. 19 for a procedure with slightly different weights applied to estimating an ATE.

40. See Rubin, “The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies,” Rubin, “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies,” Imbens, “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” pp. 12-13, 19, and Ho et al., “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference,” pp. 207-208 and 223-224.

41. See p. 537-538 of Blackwell et al., “cem: Coarsened exact matching in Stata.”

42. Ho et al., “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference,” p. 215.

43. See Angrist and Card, “Empirical Strategies in Labor Economics,” pp. 1284-1293 and Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Chapter 5 for an in-depth examination of causally interpreting regression estimates.

44. See, e.g., Davidson and MacKinnon, *Econometric Theory and Methods*, Chapters 1-3, especially Section 3.2.

45. Ibid., p. 106.

46. Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, pp. 127-128.

47. Davidson and MacKinnon, *Econometric Theory and Methods*, p. 321, eq. 8.29.

48. Angrist, Imbens, and Rubin, “Identification of Causal Effects Using Instrumental Variables,” pp. 446-447; see also Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Chapter 4, Sections 4.4-4.5.

49. It is also sufficient to assume that assignment to z is strongly ignorable; see Angrist, Imbens, and Rubin, “Identification of Causal Effects Using Instrumental Variables,” p. 446. I follow their original proof in assuming the stronger condition of random assignment.

50. Angrist, Imbens, and Rubin, “Identification of Causal Effects Using Instrumental Variables,” p. 448.

51. Ibid.

52. A related point is made in Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, pp. 163-165.

53. Ibid., p. 173; see also pp. 174-186.

54. For a more thorough explanation of the policy relevance of the LATE, see Angrist and Card, “Empirical Strategies in Labor Economics,” pp. 1320-1326.

55. Carneiro, Heckman, and Vytlacil, "Estimating Marginal Returns to Education."

56. Carneiro, Heckman, and Vytlacil, "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education," p. 16.

57. Angrist and Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?"

58. Quoting from Angrist and Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," pp. 1010-1011:

    > Our extract [of the 1970 Census Data] combines data from three separate public-use files: the State, County group, and Neighborhood files. Each file contains a self-weighting, mutually exclusive sample of 1 percent of the population (as of April 1, 1970), yielding a total sample of 3 percent of the population. The data sets we use are based on the questionnaire that was administered to 15 percent of the population. The sample consists of white and black men born between 1920-1929 in the United States. Birth year was derived from reported age and quarter of birth. In addition, we excluded any man whose age, sex, race, veteran status, weeks worked, highest grade completed or salary was allocated by the Census Bureau. Finally, the sample is limited to men with positive wage and salary earnings and positive weeks worked in 1969.

59. Iacus, King, and Porro, "Causal inference without balance checking: Coarsened exact matching."

60. Blackwell et al., "cem: Coarsened exact matching in Stata."

61. ibid., p. 530; Iacus, King, and Porro, "Causal inference without balance checking: Coarsened exact matching," p. 7.

62. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, p. 122-123 refers to this as a Hausman test after Hausman, "Specification Tests in Econometrics," but a standard reference (e.g., in the Stata 13 help file for the test) is Sargan, "The Estimation of Economic Relationships using Instrumental Variables."

63. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, p. 123.

64. Ibid.

65. Carneiro, Heckman, and Vytlacil, "Estimating Marginal Returns to Education."

66. Marshall, "Coarsening bias: How instrumenting for coarsening treatments upwardly biases instrumental variable estimates."

67. Spence, "Job Market Signaling."

68. A similar argument, applied to the effectiveness of Catholic schools and job training programs, is made by Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, pp. 38-40.

References

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

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91, no. 434 (1996): 444–455.

Angrist, Joshua D., and Alan B. Krueger. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (1991): 979–1014.

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

———. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *The Journal of Economic Perspectives* 24, no. 2 (2010): 3–30.

Blackwell, M., S. Iacus, G. King, and G. Porro. "cem: Coarsened exact matching in Stata." *Stata Journal* 9, no. 4 (2009): 524–546.

Carneiro, Pedro, James J. Heckman, and Edward Vytlacil. "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education." URL: `http://goo.gl/PWQMcA`, July 2003.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. "Estimating Marginal Returns to Education." *American Economic Review* 101 (2011): 2754–2781.

Davidson, Russell, and James G. MacKinnon. *Econometric Theory and Methods.* New York: Oxford University Press, 2003.

Hausman, J.A. "Specification Tests in Econometrics." *Econometrica* 46 (1978): 1251–1271.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. "Matchit: nonparametric preprocessing for parametric casual inference." *Journal of Statistical Software* 42 (2011): 1–28.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political Analysis* 15, no. 3 (2007): 199–236.

Holland, Paul W. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, no. 396 (1986): 945–960.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "Causal inference without balance checking: Coarsened exact matching." *Political Analysis* 20, no. 1 (2012): 1–24.

Imbens, Guido W. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86, no. 1 (2004): 4–29.

King, Gary, Richard Nielsen, Carter Coberley, James E. Pope, and Aaron Wells. "Comparative effectiveness of matching methods for causal inference." URL: `http://gking.harvard.edu/files/psparadox.pdf`. 2011.

Marshall, John. "Coarsening bias: How instrumenting for coarsening treatments upwardly biases instrumental variable estimates." URL: `http://goo.gl/YA5Sjo`, 2014.

Morgan, Stephen L., and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press, 2007.

Pearl, Judea. *Causality: models, reasoning and inference.* Cambridge Univ Press, 2000.

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70, no. 1 (1983): 41–55.

Rubin, Donald B. "Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81, no. 396 (1986): 961–962.

———. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66, no. 5 (1974): 688–701.

———. "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29 (1973): 185–203.

———. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74, no. 366 (1979): 318–328.

Sargan, J.D. "The Estimation of Economic Relationships using Instrumental Variables." *Econometrica* 26 (1958): 393–415.

Sekhon, Jasjeet S. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press, 2008.

Spence, Michael. "Job Market Signaling." *The Quarterly Journal of Economics* 87, no. 3 (1973): 355–374.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25, no. 1 (2010): 1–21.

Titiunik, Rocio. "ATE, ATT and Potential Outcomes." Online, URL: `http://goo.gl/ZKprMS` accessed 6/8/2015. 2007.

Willis, Robert J. "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions." Chap. 10 in *Handbook of Labor Economics, Vol. 1*, edited by Orley Ashenfelter and Richard Layard, 525–602. Elsevier, 1986.

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, 2002.