Impact Evaluation and Analysis of Development Interventions III

Introduction to Causal Models

1 Introduction to Causality and Randomized Trials

Causal effects are of interest to economists (and other social scientists) because we would often like to know what the effects of manipulating a particular program or policy are. Take, for example, the return to schooling – possibly the most heavily analyzed quantity in labor economics (maybe even in all applied microeconomics!). Using survey data in the United States, it is easy to estimate the relationship between schooling and earnings – we can, for example, use linear regression to approximate the expected value of earnings conditional upon years of schooling. However, this only reveals to us how these two variables covary in the US population. It does not, in general, reveal what effects a policy manipulation that increased schooling by one year for each student in the US might have on earnings. Using the terms that you are familiar with from other econometrics courses, years of schooling is "endogenously" determined by individual students and their parents. If you are interested in predicting how earnings change when you draw a different individual with a higher level of education from the US population, then it is perfectly reasonable to apply the regression coefficient. The policy manipulation, however, refers to an "exogenous" change in years of schooling. There is therefore no reason that the regression coefficient – estimated using data in which schooling is endogenously determined – should correspond to the effect of an exogenous change in years of schooling. Note that it is not the case that one quantity is "right" and the other is "wrong" – which one is "correct" depends on what question you are trying to answer. Rather, it's simply the case that the two quantities are different, and one cannot be substituted for the other.

Despite the central role that causality plays in answering policy-relevant questions (since a policy intervention implies, almost by definition, some sort of external manipulation), many econometrics courses do not formally present or discuss a model of causality. Instead,

they often begin by presenting a structural model of some economic phenomenon – which is implied to have an underlying causal interpretation – and then proceed to discuss the cases in which linear regression (or some other estimator) will estimate this model. This presentation, however, sometimes leaves students thinking that a regression is inherently "wrong" or useless if it doesn't provide unbiased or consistent estimates of an underlying structural model (which, as we will see later, is certainly not the case). It is also true that in some (many?) cases the structural parameters themselves do not correspond to any meaningful causal effect without further transformations or assumptions.

The causal model we discuss today has come to be known as the *Rubin Causal Model* (RCM), in reference to Rubin (1974) and subsequent publications. The RCM relies heavily upon the notion of *potential outcomes* – that is to say, possible outcomes under different values of a variable we shall refer to as the *treatment* – and it is useful for two reasons. First, it is useful when understanding many common estimation techniques, such as instrumental variables, regression discontinuity design, propensity score matching, etc. More importantly, however, it can be useful in framing or understanding what question you are trying to answer or what effect you are trying to estimate. If the quantity cannot be conceptualized as arising from an experimental manipulation of some type of treatment, then it cannot be estimated from a randomized trial, and the techniques that we learn which simulate randomized experiments will be inappropriate.[1]

1.1 The Rubin Causal Model

Suppose that we have N units, $i = 1, ..., N$, drawn randomly from a large population. We are interested in the effect of some binary treatment variable, D_i, on an outcome, Y_i. We refer to $D_i = 1$ as the *treatment condition* and $D_i = 0$ as the *control condition*. Given these two possiblities – treatment and control – we postulate the existence of two potential outcomes for each unit: $Y_i(0)$ under the control condition and $Y_i(1)$ under the treatment

[1] Of course, the question may still be of interest, but you will have to find a different (possibly easier!) way to answer it, and you should understand that the answer will not correspond to the effect of a policy intervention.

condition.[2] The key here is that, although we will never observe both $Y_i(0)$ and $Y_i(1)$ (we will observe at most one or the other, but never both), it is theoretically possible that we could observe either. In Holland's terminology, every unit must be *potentially exposable* to every value of the treatment variable. If you cannot conceptualize both $Y_i(0)$ and $Y_i(1)$ for the same unit, then D does not correspond to a treatment that is potentially manipulable and we cannot talk about the causal effect of manipulating D without further defining the problem. Holland, for example, argues that race is not something to which each unit is potentially exposable – we do not in general think of race as being something that we can experimentally manipulate, and it is unclear what it would mean to ask what my potential outcomes would be if I changed my race to be, for example, African-American.

Using the notation above, we define the *causal effect* of treatment $D = 1$ on outcome Y for unit i as:

$$Y_i(1) - Y_i(0) = \tau_i$$

Alternatively, we often refer to τ_i as the treatment effect for unit i. Several things are important to note here. First, the effect of a treatment is always defined in a relative sense – in this case it is the effect of the treatment condition $D = 1$ relative to the potential outcome that would have occurred under the control condition $D = 0$. In medicine, $D = 1$ might correspond to giving a drug (e.g., Lipitor) to a patient, while $D = 0$ corresponds to giving a placebo to the patient. In development, $D = 1$ might correspond to implementing a conditional cash transfer program in Senegal, while $D = 0$ corresponds to not doing so.[3] Second, the effect of the treatment need not be constant across different units, as indicated by the fact that τ is indexed by i – many (probably most) treatments have heterogeneous

[2]Note that the notation here is slightly different than in the excellent Holland (1986) article. In Holland's article, the subscript of $Y_t(i)$ corresponds to treatment/control while the argument inside the parentheses corresponds to the unit number ($1, ..., N$ in our case). In our notation, the subscript corresponds to the unit number while the argument inside the parentheses corresponds to treatment/control. We do this because our notation corresponds to the notation used in seminal articles such as Angrist, Imbens, and Rubin (1996).

[3]If the treatment variable can take on more than two values (e.g., 0, 1, or 2), then multiple treatment effects exist for each unit (e.g., $Y_i(1) - Y_i(0)$ and $Y_i(2) - Y_i(1)$), and these effects need not be equal, just as the relationship between a dependent variable and an explanatory variable need not be linear.

effects. Finally, we will never observe both $Y_i(1)$ and $Y_i(0)$ for any given unit. This is because, although it is not evident in the notation, treatments also involve a time dimension. When we write $D = 1$ and $D = 0$, we implicitly mean that we are applying the treatment or control condition at a specific point in time. In the medical example, if we administer Lipitor to a patient on his 55th birthday, we cannot simultaneously not administer Lipitor to him at the exact same moment. In the development policy example, if we implement a conditional cash transfer program for the 2014 fiscal year in Senegal, we cannot simultaneously not implement that program in Senegal during the same fiscal year. We might choose not to implement the program in 2013 or 2015 – just as we might choose not to administer Lipitor to the patient on his 54th or 56th birthdays – but since other factors affecting the unit can change during the interim period, we are not guaranteed of observing the outcome that would have occurred had we implemented the control condition in 2014 (or on the 55th birthday).

This inability to observe both $Y_i(0)$ and $Y_i(1)$ for any given unit leads to the following theorem:

Fundamental Problem of Causal Inference: It is impossible to observe the value of $Y_i(0)$ and $Y_i(1)$ on the same unit i and, therefore, it is impossible to observe τ_i, the effect for unit i of the treatment on Y_i. (Holland 1986)

The Fundamental Problem of Causal Inference would appear to rule out any precise estimation of τ_i, and, at the unit level, it is true that we can never observe the exact treatment effect. However, all is not lost. We are often interested in relationships that hold "on average," or in expectation. In this context, it is possible to estimate quantities of interest. We define the *average causal effect* or *average treatment effect* (ATE) of the treatment relative to the control as the expected value of the difference $Y_i(1) - Y_i(0)$, or

$$\bar{\tau} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

With the appropriate research design, it is possible to estimate ATE.

1.2 Estimation of Treatment Effects: The Randomized Controlled Trial

For each unit i, there *exist* the quantities ($Y_i(0)$, $Y_i(1)$, D_i). However, we only *observe* (Y_i, D_i), where

$$Y_i = (1 - D_i)Y_i(0) + D_iY_i(1)$$

The distinction between what exists conceptually and what we can actually observe is subtle but tremendously important. Although we can only observe $Y_i(0)$ for untreated units and $Y_i(1)$ for treated units, we can conceive of the counterfactual quantities $Y_i(1)$ for untreated units (i.e., the outcome that control unit i would have realized under the treatment condition) and $Y_i(0)$ for treated units (i.e., the outcome that treated unit i would have realized under the control condition). Understanding the distinction between the observed Y_i and the unobserved-but-still-existent counterfactual quantities ($Y_i(0)$ or $Y_i(1)$) will be crucial in subsequent derivations in this course.

By definition,

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1]$$

$$E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0]$$

Note that in general $E[Y_i(0)|D_i = 0] \neq E[Y_i(0)|D_i = 1]$ (and $E[Y_i(1)|D_i = 1] \neq E[Y_i(1)|D_i = 0]$). That is to say, people who select into the control condition generally have different outcomes under the control condition ($Y_i(0)$) than people who do not select into the control condition. Thus, the average control outcome for the control unit $E[Y_i(0)|D_i = 0]$ need not equal the average control outcome for all units $E[Y_i(0)]$, which is a combination of both control and treated units. The fact that we do not observe control outcomes ($Y_i(0)$) for any of the treated units, however, does not prevent us from imagining the existence of these counterfactual outcomes. In the context of our medical example, Y is cholesterol level and

D represents treatment with Lipitor. Patients who choose to take Lipitor ($D_i = 1$) are likely to have high cholesterol levels in the absence of Lipitor (i.e., $Y_i(0)$ is high, though we do not observe $Y_i(0)$ for them). Patients who choose not to take Lipitor ($D_i = 0$) are likely to have low cholesterol levels in the absence of Lipitor (i.e., $Y_i(0)$ is low, and for these patients we observe $Y_i(0)$ since $Y_i = (1 - D_i)Y_i(0) + D_iY_i(1) = Y_i(0)$). The average untreated cholesterol level for patients not taking Lipitor, $E[Y_i(0)|D_i = 0]$, is therefore less than both the average untreated cholesterol level for treated patients, $E[Y_i(0)|D_i = 1]$, and the average untreated cholesterol level for all patients, $E[Y_i(0)]$.

There is, however, an important case in which $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] = E[Y_i(0)]$ (and $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] = E[Y_i(1)]$). Suppose that the treatment assignment, D, is *randomly assigned*. In that case, D is *independent* of both $Y(0)$ and $Y(1)$. The conditional distribution of $Y_i(0)$ (and $Y_i(1)$) given D_i is therefore equal to the unconditional distribution, and it must be the case that

$$E[Y_i(0)|D_i = 0] = E[Y_i(0)]$$

$$E[Y_i(1)|D_i = 1] = E[Y_i(1)]$$

The average causal effect, $\bar{\tau}$, is thus

$$\bar{\tau} = E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

We can easily estimate $\bar{\tau}$ by taking the difference between the average value of Y_i in the treatment group and the average value of Y_i in the control group. Because it allows estimation of ATE, the randomized controlled trial is considered the "gold standard" of evidence in medicine, and in many areas of social science as well.

In some instances we may be willing to assume that $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] = E[Y_i(0)]$ but not that $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] = E[Y_i(1)]$. In other words, we may be willing to assume that the untreated potential outcomes are mean-independent of the

treatment assignment, but not that the treated potential outcomes are mean-independent of the treatment assignment. This is equivalent to saying that there is no selection into treatment based on the level of untreated outcomes, but there is selection into treatment based on the potential gains of being treated. You could probably write down an economic model that would give this result, but to be honest I doubt it would be a palatable assumption in most empirical settings. Regardless, under this slightly weaker assumption, you can still identify

$$\bar{\tau}_{TOT} = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

This quantity is commonly referred to as "the effect of the treatment on the treated," or TOT (treatment-on-treated) or ATOT (average treatment-on-treated) or some other strange permutation of those letters. It is the causal effect of the treatment on those who select into treatment.

1.3 The Stable Unit Value Treatment Assumption: SUTVA

Beyond the assumption of random assignment of D, there is an implicit assumption embedded in the previous section that is known rather awkwardly as the *stable unit treatment value assumption*, or *SUTVA*. Let $\mathbf{D}$ be a $N \times 1$ column vector that contains the treatment values for all N units. Formally, SUTVA states that

$$\text{If } D_i = D_i', \text{ then } Y_i(\mathbf{D}) = Y_i(\mathbf{D}').$$

We have not yet defined what $Y_i(\mathbf{D})$ is, but it is exactly analogous to our definition of $Y_i(D_i)$ (i.e., $Y_i(0)$ and $Y_i(1)$). That is to say, $Y_i(\mathbf{D})$ is the potential outcome for unit i under treatment regime $\mathbf{D}$. Now, instead of just specifying whether unit i is receiving the treatment or the control, we are specifying values of D_i for all units in the sample. For this reason, SUTVA is often referred to as the "no interference" assumption, since it states that unit i's potential outcomes are unaffected by whether unit j ($j \neq i$) is treated or untreated. A

classic example of SUTVA *not* holding is the case of vaccines. If D_i represents inoculation of unit i with the measles vaccine, and Y_i represents whether unit i gets measles, clearly $Y_i(D_i)$ depends on the values of the entire vector $\mathbf{D}$. In particular, if $D_j = 1$ for all $j \neq i$, then $Y_i(0)$ will likely be 0 despite the fact that unit i is unprotected, because there are no other unprotected units to spread the disease to unit i. If $D_j = 0$ for all $j \neq i$, however, then $Y_i(0)$ might change to 1. Another example of SUTVA not holding is if the treatment is a carbon tax intended to address global warming. Suppose that such a tax is enacted in the European Union (EU). If $Y_e(0)$ is the average temperature in the EU (in the year 2100) in the absence of an EU carbon tax, it should be clear that $Y_e(0)$ depends on whether other regions implement carbon taxes. If SUTVA does not hold, then there is not just one treatment effect, τ_i, per unit but rather a multitude of treatment effects (one for each different permutation of $\mathbf{D}$). More importantly, it may be impossible to estimate the treatment effect relative to the "no intervention" scenario (i.e., the scenario in which $D_i = 0$ for all units i), because as soon as one unit is treated, all are potentially affected (so it is impossible to construct an unbiased estimate of $E[Y_i(0)]$ with the data).

Rubin (1986) discusses SUTVA in the context of poorly defined treatments. That is to say, he focuses on cases in which, even if $\mathbf{D} = \mathbf{D}'$, it is still the case that $Y_i(\mathbf{D}) \neq Y_i(\mathbf{D}')$. This occurs because the treatment, and in particular the assignment mechanism, is not precisely defined, so even though $\mathbf{D} = \mathbf{D}'$, it's really not the same treatment (we will discuss some examples shortly). In subsequent years, however, interest has focused on the "no interference" aspect of SUTVA – in many cases, treating one unit indirectly affects other units, and SUTVA does not hold.

1.4 Applications and Discussion

1.4.1 Poorly Defined Treatments

When does it make sense to talk about D as a cause and when does it not? Holland (1986) has a nice discussion on pp. 954-955 that I urge you to read, as does Rubin's comment to

that article. In Holland's example, there are three hypothetical scenarios:

(1) She scored highly on the exam because she is female.

(2) She scored highly on the exam because she studied.

(3) She scored highly on the exam because her teacher tutored her.

In scenario (3), it is clear that the treatment is well-defined: the teacher tutors her. We can easily conceive of manipulating whether or not this tutoring occurs. In scenario (1), Holland argues that the student's gender cannot be considered a "cause" because we cannot manipulate it. It is certainly the case that the treatment is not well-defined in this case and cannot fit within the causal framework, although Rubin points out that further refinements could allow the scenario to fit within the RCM. For example, if we said, "She scored highly on the exam because she received sex reassignment surgery," then we would have a clearly defined treatment. Scenario (2) is the most problematic because it involves a voluntary activity that the student can choose to do. Although we could certainly conceive of an intervention that might *prevent* the student from studying (anesthesia, for example, would be a pretty good bet), it is hard to imagine a manipulation that would *force* the student to study (or at least force her to study as well as she would if she voluntarily studied). Since we cannot manipulate this attribute (studying for the exam), we cannot think of it as cause, at least not within the potential outcomes framework. Hence Holland's phrase, "No Causation Without Manipulation." It should also be clear from this discussion why there is such a close linkage between the potential outcomes framework and policy relevance. If you can't conceive of manipulating a particular attribute, then by definition you cannot design a policy that would manipulate that attribute!

1.4.2 Effects of Causes vs. Causes of Effects

Holland writes that "an emphasis on the effects of causes rather on the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation and directly opposes more traditional analyses of causation." This distinction

between effects of causes and causes of effects may seem somewhat pedantic. It is not.

A concrete example should help clarify the distinction. Consider the obesity "epidemic," an issue of great importance to both agricultural economists (on the input side) and health economists (on the output side). Researchers in different fields cite a myriad of "causes" of this epidemic: increased consumption of snack foods, larger portions at restaurants, more frequent consumption at restaurants, more sedentary jobs, the introduction of high fructose corn syrup, etc. Under these explanations, however, it is rarely clear what the counterfactual is. Take, for example, the increased consumption of snack foods over the last 30 years. One possible counterfactual is what would have happened if people had not increased their consumption of snack foods while everything else remained unchanged. It is unclear, however, what policy manipulation could enforce that counterfactual scenario. Although it is easy to imagine limiting snack food consumption (through a quota or a tax, for example), it is impossible to imagine doing so while simultaneously preventing individuals from compensating in any other manner. Another possible counterfactual is to imagine a world in which the complex set of technologies and changes in consumer preferences that led to the increase in snacking had been inhibited from developing. Even if it were possible to imagine this sort of manipulation, however, there is no guarantee that total caloric consumption would fall by exactly the amount that snack food consumption has increased (in fact, it almost surely would not). It becomes clear that the answer to the question of what has "caused" the obesity epidemic is every single thing about the world that affects weight and has changed since 1970. But even this turns out to be an incomplete answer because the distribution of weight in 1970 is itself a cause of the distribution of weight today, so the obesity epidemic is in fact "caused" by everything in the history of the world that has ever had an effect on weight. As Holland (1986) notes, there is really no definable answer to this type of question.

In contrast to the causes of effects – which is effectively an unlimited exercise in accounting and description – the effects of causes are clearly defined under the RCM. Even if we cannot measure them with existing data, we can at least conceive of what they are.

1.4.3 Quasi-experimental Methods

For reasons that should be obvious, randomized controlled trials (RCTs) have become popular in the program evaluation literature. However, in many cases a RCT is too expensive, infeasible, or potentially unethical. In those cases we have no choice but to use observational data (i.e., data in which the treatment is not randomly assigned) if we wish to estimate a treatment effect. The modern program evaluation toolkit contains many different techniques to estimate treatment effects: regression, matching, propensity score methods, differences-in-differences, instrumental variables, regression discontinuity, and more. However, the underlying motivation for *all* of these techniques is that they are attempting to recreate or simulate a randomized experiment within the observational data. That is to say, they are attempting to isolate some subset of variation in the treatment D_i that can be considered to be "as good as randomly assigned."

One implication of this is that if you cannot conceptualize of an idealized RCT to estimate the treatment effect that you are interested in, then you likewise won't be able to use any of the program evaluation techniques to estimate that treatment effect. That simply follows from the fact that the program evaluation techniques are attempting to recreate the idealized RCT, so they can never do better than that RCT! Before developing a research design for estimating a specific treatment effect, I thus often find it useful to ask the question, "What is the idealized experiment that I would run to estimate this effect if I had unlimited funding and no data constraints?" If you cannot suitably answer this question, then the effect you're hoping to estimate is not compatible with the techniques we're learning here! Furthermore, by answering this question you may gain insight in developing the research design that you apply to the observational data.

2 Regression: What does it do?

Linear regression is the bread and butter of econometrics, so it is worth doing a quick review of it. But this review is unlikely to resemble anything you saw in other econometrics courses,

primarily because it focuses on what linear regression *is*, rather than on what you would like it to be. That is to say, as long as you satisfy certain trivial conditions (e.g., your matrix of regressors is full rank), you can *always* run a linear regression. And there is absolutely nothing wrong with doing that – regardless of problems involving "endogeneity" or "omitted variables" or "measurement error" – as long as you interpret the results appropriately. In this section we will focus on what I refer to as "agnostic regression."

2.1 The Conditional Expectation Function

Consider a dependent variable, y_i, and a vector of explanatory variables, x_i. We are interested in the relationship between the dependent variable and the explanatory variables. There are several possible reasons that we may be interested in this relationship, including:

1. Description – What is the observed relationship between y and x?
2. Prediction – Can we use x to create a good forecast of y?
3. Causality – What happens to y if we experimentally manipulate x?

It is generally the last item that generates comments about "exogeneity conditions" and so forth. We will ignore all that negative energy for the moment and, instead of worrying about what you can't infer, focus on what you can infer. Think positive!

Of course, few real-world relationships are deterministic. Recognizing this fact, we focus on relationships that hold "on average," or "in expectation." Given our variables y and x, we may be interested in the *conditional expectation* of y given x. That is to say, given a particular value of x, where is the distribution of y centered? This relationship is given by the *Conditional Expectation Function*, or the CEF.

$$E[y_i|x_i] = h(x_i)$$

We define the CEF residual as:

$$\varepsilon_i = y_i - h(x_i) \text{ where}$$

$$E[\varepsilon_i|x_i] = 0$$

Note that, because ε_i is the CEF residual, $E[\varepsilon_i|x_i] = 0$ holds by definition – we do not require any exogeneity assumptions regarding x_i.

Proof.

$$E[\varepsilon_i|x_i] = E[y_i - h(x_i)|x_i] = E[y_i|x_i] - E[h(x_i)|x_i]$$

$$= E[y_i|x_i] - h(x_i) = E[y_i|x_i] - E[y_i|x_i] = 0$$

To recap, the CEF residual always has zero conditional expectation. By definition. No assumptions necessary. Always.

Theorem. CEF residuals are *mean-independent* of the arguments in the CEF, x_i (see above). They are therefore orthogonal to any function of the conditioning variables.

Proof. Iterated expectations.

$$E[\varepsilon_i \cdot f(x_i)] = E[E[\varepsilon_i \cdot f(x_i)|x_i]] = E[E[\varepsilon_i|x_i]f(x_i)] = E[0] = 0$$

More importantly, the CEF is the "best" function of x that exists for predicting y (where "best" is defined in terms of expected squared loss).

Theorem. $E[y_i|x_i] = argmin_g\ E[(y_i - g(x_i))^2]$ In other words, the CEF is the function that minimizes the expected squared deviations from y_i. We say that "the CEF is the *minimum mean-square error* (MMSE) predictor for y_i given x_i."

Proof.

$$E[(y_i - g(x_i))^2] = E[((y_i - E[y_i|x_i]) + (E[y_i|x_i] - g(x_i)))^2] =$$

$$E[(y_i - E[y_i|x_i])^2 + 2(y_i - E[y_i|x_i])(E[y_i|x_i] - g(x_i)) + (E[y_i|x_i] - g(x_i))^2] =$$

$$E[E[(y_i - E[y_i|x_i])^2 + 2(y_i - E[y_i|x_i])(E[y_i|x_i] - g(x_i)) + (E[y_i|x_i] - g(x_i))^2|x_i]] =$$

$$E[E[(y_i - E[y_i|x_i])^2|x_i] + 2(E[y_i|x_i] - E[y_i|x_i])(E[y_i|x_i] - g(x_i)) + (E[y_i|x_i] - g(x_i))^2] =$$

$$E[E[(y_i - E[y_i|x_i])^2|x_i]] + E[(E[y_i|x_i] - g(x_i))^2]$$

It should be clear that choosing $g(x_i)$ such that $g(x_i) = E[y_i|x_i]$ minimizes the second term in the last line. The first term in the last line does not contain $g(x_i)$ and is therefore unaffected by our choice of $g(x_i)$. The CEF, $E[y_i|x_i]$, therefore solves $min_g\ E[(y_i - g(x_i))^2]$.

2.2 Regression and the CEF: Why We Regress

Clearly the CEF has some desirable properties in terms of summarizing the relationship between x_i and y_i and making predictions about y_i given x_i. In particular, we have seen that it is the MMSE predictor of y_i. But what does this have to do with linear regression, and why might we want to use linear regression?

2.2.1 Reason the First: Regression-CEF Theorem

Theorem. If the CEF is linear, then the regression of y_i on x_i estimates the CEF. Formally, if $E[y_i|x_i] = x_i\gamma$, then $\gamma = E[x_i'x_i]^{-1}E[x_i'y_i]$ (which is what the regression coefficient converges to).

Proof.

$$E[x_i'x_i]^{-1}E[x_i'y_i] = E[x_i'x_i]^{-1}E[E[x_i'y_i|x_i]] = E[x_i'x_i]^{-1}E[x_i'E[y_i|x_i]] =$$

$$E[x_i'x_i]^{-1}E[x_i'x_i\gamma] = E[x_i'x_i]^{-1}E[x_i'x_i]\gamma = \gamma$$

Of course, there is no reason the CEF has to be linear. Two of the most common sufficient conditions for a linear CEF are: (1) joint normality of x_i and y_i or (2) a *saturated model* for discrete regressors. A saturated model is one in which you estimate a separate parameter for each point in the support of x_i (e.g., you have a separate dummy variable for each unique

value of the vector x_i in your data set). This is more common in empirical work than joint normality.

In most cases, however, the CEF is not linear. But we still run regressions anyway. Why do we do this? One reason is that it is computationally tractable and that we understand its properties both when it is correctly specified and under misspecification (or, at least, we understand its properties under misspecification better than we understand the properties of other estimators). Nevertheless, there are good theoretical reasons to regress as well.

2.2.2 Reason the Second: BLP Theorem

Theorem. If you want to predict y_i, and you limit yourself to linear functions of x_i, then $x_i\beta = x_i E[x_i'x_i]^{-1}E[x_i'y_i]$ is the best linear predictor (BLP) of y_i in a MMSE sense. Formally, $\beta = E[x_i'x_i]^{-1}E[x_i'y_i] = argmin_b\ E[(y_i - x_ib)^2]$.

Proof.

$$\partial E[(y_i - x_ib)^2]/\partial b = 2E[x_i'(y_i - x_ib)] = 0$$

$$E[x_i'y_i] - E[x_i'x_i]b = 0$$

$$b = E[x_i'x_i]^{-1}E[x_i'y_i] = \beta$$

If you're limiting yourself to linear combination of x_i, then linear regression gives you the best predictor of y_i. Of course, this isn't a big surprise given that the OLS estimator is derived by minimizing the sample analog of $E[(y_i - x_ib)^2]$. Regardless, this property is nice if you're in the business of forecasting, but it's not as useful if your interest is in estimating the CEF as a summary of the underlying relationship between y_i and x_i. Which brings us to our third reason to regress (arguably the best reason).

2.2.3 Reason the Third: Regression Approximation Theorem

Theorem. The MMSE linear approximation to the CEF is $\beta = E[x_i'x_i]^{-1}E[x_i'y_i]$. Formally, $\beta = E[x_i'x_i]^{-1}E[x_i'y_i] = argmin_b\ E[(E[y_i|x_i] - x_ib)^2]$.

Proof.

$$\partial E[(E[y_i|x_i] - x_ib)^2]/\partial b = 2E[x_i'(E[y_i|x_i] - x_ib)] = 0$$

$$E[E[x_i'y_i|x_i]] - E[x_i'x_i]b = 0$$

$$b = E[x_i'x_i]^{-1}E[x_i'y_i] = \beta$$

So regression provides the best linear approximation to the CEF, even when the CEF is non-linear. Regression can therefore give you a pretty decent approximation of the CEF as long as you don't try to extrapolate beyond the support of x_i.

2.3 Discussion

If your object of interest is the CEF, then linear regression is a good tool for estimating it. Specifically, it is the best linear predictor in terms of minimizing the mean squared error from the CEF. More importantly, this result depends on absolutely nothing. In particular, it does not depend on:

1. Whether your data are i.i.d.
2. Whether you treat your regressors as random variables or fixed quantities.
3. Whether your regressors are correlated with the CEF residuals (by definition, they are not, since the residuals are mean-independent of any function of the conditioning variables).
4. Whether the CEF is linear or not.

5. Whether your dependent variable is continuous, discrete, non-negative, or anything else.

Regression is therefore remarkably robust as an estimation tool, provided that you interpret it for what it actually is – an approximation of the conditional expectation function – rather than what you might like it to be (an estimate of a causal relationship). So if you're only interested in description or prediction, we can probably end the class right here.

2.4 Application: Predicting College Success

Geiser and Santelices (2007) use high school GPA, standardized test scores (SAT), and other covariates to predict college performance (college GPA) using linear regression for University of California (UC) freshman entering between Fall 1996 and Fall 1999. The results from this exercise are listed in Table 4 of their article, reproduced below. They find that, in this sample, high school grade point average (GPA) is a more effective predictor of college GPA than any other measure. In particular, it is much more effective than SAT I (a standardized test that most college applicants take). This can be seen in at least two ways. First, in comparing Model 1 – which uses high school GPA as a predictor – and Model 2 – which uses SAT I as a predictor – we see that Model 1 has a much higher R^2; in other words, high school GPA is explaining much more of the variation in college GPA than SAT I score is (Note: This is probably the *only* time in this course that you will hear reference to R^2. In general it is *not* an interesting statistic in answering policy-relevant questions.) We also see, in Model 7, that the standardized coefficient on high school GPA is substantially larger than the standardized coefficient(s) on SAT I (the standardized coefficient is a normal regression coefficient that has been rescaled to indicate how many standard deviations y changes with a one standard deviation change in x). Moving up one standard deviation in the high school GPA distribution is therefore much more beneficial for college GPA (in a predictive sense) than moving up one standard deviation in the SAT I score distribution.

Does this relationship answer any interesting, policy-relevant questions? Arguably, yes.

I. Validity of Admissions Factors in Predicting Cumulative Fourth-Year GPA

We begin with findings on the relative contribution of admissions factors in predicting cumulative four-year college GPA. Table 4 shows the percentage of explained variance in cumulative fourth-year GPA that is accounted for by HSGPA, SAT I verbal and math scores, and SAT II Writing, Mathematics and Third Test scores. The estimated effects of these admissions factors on cumulative fourth-year GPA were analyzed both singly and in combination. Parents' education, family income and school API rank were also included in all of the regression models in order to control for the "proxy" effects, noted above, of socioeconomic status on standardized test scores and other admissions variables.

Table 4

Relative Contribution of Admissions Factors in Predicting Cumulative Fourth-Year GPA

	Standardized Regression Coefficients										
	High School GPA	SAT I Verbal	SAT I Math	SAT II Writing	SAT II Math	SAT II 3rd Test	Parents' Education	Family Income	School API Rank	Number	% Explained Variance
Model 1	**0.41**	x	x	x	x	x	**0.12**	**0.03**	**0.08**	59,637	20.4%
Model 2	x	**0.28**	**0.10**	x	x	x	**0.03**	**0.02**	0.01	59,420	13.4%
Model 3	x	x	x	**0.30**	**0.04**	**0.12**	**0.05**	**0.02**	**-0.01**	58,879	16.9%
Model 4	**0.36**	**0.23**	0.00	x	x	x	**0.05**	**0.02**	**0.05**	59,321	24.7%
Model 5	**0.33**	x	x	**0.24**	**-0.05**	**0.10**	**0.06**	**0.02**	**0.04**	58,791	26.3%
Model 6	x	**0.06**	-0.01	**0.26**	**0.04**	**0.12**	**0.04**	**0.02**	**-0.01**	58,627	17.0%
Model 7	**0.34**	**0.08**	-0.02	**0.19**	**-0.04**	**0.09**	**0.05**	**0.02**	**0.04**	58,539	26.5%

Boldface indicates coefficients are statistically significant at 99% confidence level.
Source: UC Corporate Student System data on first-time freshmen entering between Fall 1996 and Fall 1999.

If you are a UC admission officer, and you are tasked with reducing acceptance rates due to state budget cuts (sadly, this scenario is likely to occur), then you can use the regression results to predict which students are least likely to succeed. We know from the previous theorems that the CEF provides the MMSE prediction of y (college GPA) and that regression provides the MMSE linear approximation to the CEF. So in a predictive sense you are likely to do well (at least relative to alternative choices), and in this case what you care about is prediction.

These results also have policy relevance in that the University of California would like to maintain a diverse student body but is not allowed to give any weight to ethnicity as an admission criterion. UC administrators are aware, however, that weighting SATs more heavily (as is traditionally done) tends to favor Caucasians (and possibly Asians?), while weighting high school GPA more heavily tends to favor African Americans and Latinos (in a relative sense). But will putting more weight on high school GPA and less weight on SAT scores result in a lower quality student body? The results from Table 4 indicate that it will not; in fact, if anything, it may result in a higher quality student body.

Are the estimated relationships causal? Highly unlikely. Even after controlling for parental education and income, there are probably unobserved individual, family, neighborhood, and peer characteristics that affect college success and are correlated with high school GPA and SAT scores.[4] The regression results, however, are still useful for prediction and have interesting applications in policy-relevant questions.

One can still take issue with the results along multiple dimensions. For example, should some adjustment be done to GPA to reflect the student's choice of major?[5] Might there be other variables collected from the applicants that could improve the predictive power of the

[4]In fact, it doesn't even make sense to talk about an experimental manipulation of high school GPA or SAT scores. The effects on college success will almost surely depend on whether the treatment entails raising these attributes through cram sessions or through mentoring programs or through intensive intervention earlier in life. The treatment is better defined as the actual intervention than as raising GPA by one point or increasing SAT scores by 100 points.

[5]I would strongly recommend against getting into this debate – it will be a great way to alienate a lot of colleagues very quickly.

model? Nevertheless, the fact remains that the results are useful and interesting despite the fact that the coefficients do not have causal interpretations. This example makes appropriate use of a descriptive relationship estimated via linear regression, which is probably more than can be said for the vast majority of empirical applications in economics.

3 Selection on Observables Designs

We now begin our discussion of what we refer to as "selection on observables" designs. The key assumption underlying these designs is that the treatment assignment is "ignorable" – which you can interpret as "as good as randomly assigned" – after you condition on a set of observable factors. There are a variety of estimation techniques available in this scenario: standard linear regression, flexible nonparametric regression, matching estimators, and propensity score estimators. The underlying (untestable) assumption of all of these estimators, however, is that you observe all of the factors that affect treatment assignment *and* are correlated with the potential outcomes. In other words, to the extent that there is systematic selection into treatment, this selection is only a function of the observable variables. Hence, if you can "control" for the effects of these variables on the probability of selection, then you can produce consistent estimates of causal effects. The flip side is that, if you don't observe all the determinants of selection, then these methods do not, in general, produce estimates with a causal interpretation. This important fact is often overlooked by applied practitioners who focus on the sophistication of the estimation technique (matching and propensity score techniques are somewhat en vogue these days). In my opinion, the underlying selection on observables assumption is too strong to hold in most cases, *so these methods are probably applied more often than they should be.* Nevertheless, in some cases the assumption is palatable (or at least defensible), and in those cases these techniques can be quite helpful.[6]

[6]For example, consider a case in which individuals apply for some program or job, and then are assigned to different areas/departments/treatments/whatever based upon the data in their applications. In this scenario, the researcher can observe all of the non-random factors that affected selection (i.e. the data in the applications), and the selection on observables assumption clearly makes sense.

Suppose we wish to estimate the effect of a treatment, D_i, on an outcome, Y_i. The key underlying assumption motivating all selection on observables designs is that the treatment is independent of the potential outcomes (particularly the untreated potential outcomes) after conditioning on a set of observable covariates, X_i. We write this assumption as:

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

This assumption is referred to as the "unconfoundedness assumption," the "selection on observables assumption," or the "conditional independence assumption." When combined with an assumption about overlap, $0 < P(D_i = 1|X_i) < 1$, it is referred to as "strongly ignorable treatment assignment."

In this case the covariates X_i are potentially confounding variables – i.e., variables that are correlated both with D_i and Y_i – that we must control for if we wish to estimate the causal effect of D on Y. The selection on observables assumption implies that, if you hold fixed the covariates X_i, then the treatment D_i is "as good as randomly assigned." In other words, if two units – one treated and one untreated – have identical values of X_i, then you can compare these two units as if they were randomly assigned. However, to take advantage of the selection on observables assumption, you need to find a way to "hold X constant" while comparing treated and untreated units.

One method of "holding X constant" that you are all familiar with is linear regression. Specifically, consider a regression of Y_i on D_i that controls for X_i:

$$Y_i = \alpha + \beta D_i + X_i \delta + u_i$$

Loosely speaking, this regression "controls" for the relationship between X_i and Y_i and thus estimates the causal effect of D_i on Y_i. However, as you are likely aware from previous econometrics courses, whether this regression is sufficient to control for the potentially complex relationship between X_i and Y_i depends on whether the conditional expectation function $E[Y_i|D_i, X_i]$ is linear. If this CEF is non-linear, then we know from our previous

discussion that the linear regression will give us the "best" linear approximation to the non-linear CEF. However, if the CEF is highly non-linear, then this linear approximation may be very poor, particularly in cases in which the regression must engage in substantial extrapolation because the treated units have very different values of X than the control units. This fact has motivated the development of alternatives to regression that are "non-parametric" in the sense that they are less dependent on functional form assumptions. We briefly discuss two of these alternatives – matching and propensity score methods – below.

3.1 Matching

The idea behind matching is very simple. If the selection on observables assumptions holds, i.e., $Y_i(0), Y_i(1) \perp D_i|X_i$, then we can estimate $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ because the treatment is effectively randomly assigned after conditioning on X_i. The idea behind matching is to compare treated units ($D_i = 1$) to control units ($D_i = 0$) that have similar values of X_i. This guarantees that every treatment-control comparison is performed on units with identical (or close to identical) values of X_i, so we are literally conditioning on $X_i = x$. Given the selection on observables assumption, we know that D_i is as good as randomly assigned after conditioning on X_i, so we should get causal estimates.

For every treated unit, i.e. every unit with $D_i = 1$, the goal of the matching estimator is to find a comparison unit among the controls that has similar values of observable characteristics X_i. It is important to note, however, that this "comparison unit" need not be a single unit – rather, it can be a composite (i.e., a weighted average) of several different control units that have similar values of X_i.[7] Assume that there are N_T treated units and N_C control units. Define N_T sets of weights, with N_C weights in each set: $w_i(j)$ $(i = 1, ..., N_T,\ j = 1, ..., N_C)$. For each set of weights, let $\sum_j w_i(j) = 1$. Then the generic matching estimator is:

$$\hat{\tau}_M = \frac{1}{N_T} \sum_{i \in \{D=1\}} [y_i - \sum_{j \in \{D=0\}} w_i(j) y_j]$$

[7] There are obvious efficiency gains to doing this, particularly if there are more control units than treated units.

In other words, we are simply computing the average difference between the treated units and the composite comparison units. The key to this estimator is how you calculate the weights used to construct the composite comparison units, $w_i(j)$. For example, you could set $w_i(j) = \frac{1}{N_C}$. In that case, $\sum_j w_i(j)y_i$ simply equals the control group mean, $\bar{y}_C$, for all i, and $\hat{\tau}_M$ is just the difference in means between the treated and control groups. Obviously that is not very exciting estimator and it does not solve the selection on observables problems.

In general, we want to choose $w_i(j)$ so that it measures the "nearness" of X_j to X_i – $w_i(j)$ is what I will call the distance measure. If X is discrete, then in principle you could choose $w_i(j)$ such that it equals one if $X_i = X_j$ and zero otherwise (you would of course have to rescale $w_i(j)$ by $\sum_j w_i(j)$ so that it summed to one for each i). If X is continuous, then that particular measure won't work, but there are several common choices of distance measures. The most popular is probably "nearest-neighbor" matching. With nearest-neighbor matching, $w_i(j)$ is a function of the Euclidean distance between X_i and X_j. Specifically, $w_i(j)$ equals one for the control unit with the closest X_j to X_i – where closeness is measured by Euclidean distance ($(X_i - X_j)'(X_i - X_j)$) – and zero otherwise. Thus, $w_i(j)$ selects the "nearest (control) neighbor" j to treated unit i, and $\hat{\tau}_M$ computes the mean difference between each treated unit and its nearest control neighbor. This procedure should produce valid causal estimates under the selection on observables assumption, *assuming that there is sufficient overlap between the treated and control groups* (we will return to this issue shortly).

Of course, the choice of units for each component of X_i is arbitrary, so it may not make sense to weight each component equally when computing the distance between two points, as the Euclidean distance metric does. A popular alternative to the Euclidean metric is thus the Mahalanobis distance metric, $(X_i - X_j)'\Sigma_x^{-1}(X_i - X_j)$, where Σ_x is the covariance matrix of X – note the parallel to GLS. Effectively, you are normalizing the components of $(X_i - X_j)$ by the root of the inverse covariance matrix.[8]

[8]The odd thing about Mahalanobis distance is that, depending on the covariance structure, you can end up in situations in which (10, 10) is closer to (0, 0) than (8, 2). I believe this is because the weight in the inverted covariance matrix can become negative.

The main problem with matching is something known as the "Curse of Dimensionality." The problem is that the sparsity of the data rapidly increases with the dimension of X. Thus, the more variables you have in X, the less likely you are to find a comparison control unit lying close to any given treatment unit – there are simply too many dimensions to match along. For example, supposed that X contains age, gender, race, education, income, height, weight, and city of residence. Suppose that individual i is a 32 year old black female with 16 years of education, an income of \$42,000, 1.63 meters in height, weighing 60 kg, and living in a small town outside of London. It is unlikely that you will find another individual with those identical attributes in your data unless you have a very, very large data set. From a technical standpoint, this problem represents a failure of what we call the "overlap assumption." The overlap assumption posits that for any value of X_i that occurs in the data, we observe both treated and control units with that value of X_i. Loosely speaking, the overlap assumption implies that for every treated unit, we can find a "good" control unit, where the control unit is "good" in the sense that it shares the exact same values of X_i as the treated unit.

So what can be done? To address the "Curse of Dimensionality," statisticians developed propensity score matching.

3.2 Propensity Score Methods

Assume that we have unconfoundedness: $(Y_i(0), Y_i(1)) \perp D_i|X_i$. Also assume that the overlap assumption holds: $0 < P(D_i = 1|X_i) < 1$. Combining these two assumptions, we say that the treatment assignment is "strongly ignorable." We know that if we condition on X_i, then we can get a consistent estimate of ATE by simply comparing the difference in means between treated and control units. In practice, however, it is hard to condition on X_i if X_i is high dimensional. Note that this is effectively because the overlap assumption fails in finite samples – for most observations, it is impossible to find a comparison unit with the opposite treatment assignment and the same value of X.

An important result is that, under strongly ignorable treatment assignment, it is sufficient

to condition simply on $p(X_i) = E[D_i|X_i]$, also known as the *propensity score*. Formally, if we assume $(Y_i(0), Y_i(1)) \perp D_i|X_i$, then

$(Y_i(0), Y_i(1)) \perp D_i \mid p(X_i)$

Proof

We will show that $P(D_i = 1 \mid Y_i(0), Y_i(1), p(X_i)) = P(D_i = 1 \mid p(X_i)) = p(X_i)$. This implies independence of D_i and $(Y_i(0), Y_i(1))$ after conditioning on $p(X_i)$.

$$P(D_i = 1 \mid Y_i(0), Y_i(1), p(X_i)) = E[D_i \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[E[D_i \mid Y_i(0), Y_i(1), p(X_i), X_i] \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[E[D_i \mid Y_i(0), Y_i(1), X_i] \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[E[D_i \mid X_i] \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[p(X_i) \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= p(X_i)$$

For completeness, note that:

$$P(D_i = 1 \mid p(X_i)) = E[D_i \mid p(X_i)]$$

$$= E[E[D_i \mid p(X_i), X_i] \mid p(X_i)]$$

$$= E[E[D_i \mid X_i] \mid p(X_i)]$$

$$= E[p(X_i) \mid p(X_i)]$$

$$= p(X_i)$$

So $P(D_i = 1 \mid Y_i(0), Y_i(1), p(X_i)) = p(X_i) = P(D_i = 1 \mid p(X_i))$. Since D_i is binary, this implies independence of D_i and $(Y_i(0), Y_i(1))$ after conditioning on $p(X_i)$. In other words, it is sufficient to merely condition on $p(X_i)$ – we don't have to condition on X_i.

Why is it sufficient to condition on the propensity score? Our concern is that units selecting into treatment differ in some meaningful way from units that do not select into treatment, and that this difference is consistently related to the probability of entering treatment. If, however, we only compare units with the exact same probability of treatment, then it is impossible for the differences to be consistently related to the probability of treatment.[9] After conditioning on the propensity score, the units are "as good as randomly assigned."

3.2.1 Estimating the Propensity Score

Before you can condition on the propensity score, $p(X_i) = E[D_i|X_i]$, you have to estimate it. There are several ways to do this – it's not clear that one method is uniformly superior, so your choice may be context dependent. The easiest way is to use a flexible logit specification (flexible in the sense that there are interactions between the various components of X_i). The general rule of thumb here is to err on the side of being more flexible, i.e. including more higher-order terms and interactions. There are other methods as well, but in general the results are not highly sensitive to the specification of the propensity score.

3.2.2 Blocking on the Propensity Score

Once you've estimated the propensity score, the next question is what to do with it, i.e. how to condition on it. A popular way to use the propensity score is to "block," or stratify, on the propensity score. That is to say, divide the range of the propensity score into K blocks (Dehejia and Wahba use 20 blocks of width 0.05) and place observations in each

[9] If they were, then we would be using them to estimate the propensity score, or so our unconfoundedness assumption claims.

block according to their estimated propensity scores, $\hat{p}(X_i)$. Within each block k, compute, $\hat{\tau}_k$, the difference in means between treated and untreated observations. Finally, combine all K treatment effect estimates as follows:

$$\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k \cdot \frac{N_{1k} + N_{0k}}{N}$$

In other words, the average treatment effect is a weighted sum of the block-level treatment effects, with the each block's weight equal to the number of observations contained in that block. Choosing the number of blocks is at the researcher's discretion. One popular algorithm is to start with a given number of blocks (e.g., 10), and check whether the covariates are balanced within each block. If they are not, then split the blocks and check again. Continue until the covariates are balanced.[10] If the covariates remain unbalanced within blocks even when the propensity score is balanced, then you may need to estimate the propensity score more flexibly.

The overlap assumption becomes prominent when blocking on the score. When a block contains either zero treated units or zero control units, no estimate of the treatment effect exists for that block, and it must be discarded. Furthermore, because the logit specification forces $0 < \hat{p}(X_i) < 1$, it may appear that the overlap assumption is satisfied for all units when in fact it is not. To be safe, one should discard all control units with $\hat{p}(X_i)$ less than the minimum $\hat{p}(X_i)$ in the treated group and all treated units with a $\hat{p}(X_i)$ greater than the maximum $\hat{p}(X_i)$ in the control group.[11]

Note that blocking on the score is analogous to matching on the score in that you are only comparing observations with propensity scores that are close to one another. One could formally implement a matching estimator, however, using one of the methods discussed in

[10] Note that if you have many covariates and many blocks, you should not expect 100% of the covariates to have no significant relationship to the treatment status in every block - some coefficients should be significant simply by chance. A more realistic target would be, for example, to find that only 10% of the covariates are significantly related to treatment status at the 10% level.

[11] I am assuming that the minimum $\hat{p}(X_i)$ occurs in the control group and the maximum $\hat{p}(X_i)$ occurs in the treated group. If not, perform the trimming so that the minimum $\hat{p}(X_i)$ is (virtually) the same for both groups and the maximum $\hat{p}(X_i)$ is (virtually) the same for both groups.

Section 3.1. Dehejia and Wahba, for example, use nearest-neighbor matching as an alternative estimator to blocking on the propensity score (both estimators give similar results in most, but not all, cases).

3.2.3 Overlap

If you recall from the initial discussion on regression adjustment, we noted that "if the CEF is highly non-linear, then [the regression's] linear approximation may be very poor, particularly in cases in which the regression must engage in substantial extrapolation because the treated units have very different values of X than the control units." This is another way of stating that the "overlap assumption" does not hold, because for some values of X_i we only observe treated units or we only observe control units (but not both). We have also seen that overlap is important for matching (i.e., if you don't have overlap of the covariates, then you can't find matches) and for the propensity score (if $p(X_i)$ gets close to zero or one, then it becomes hard to find matches if you are blocking or matching). The dominant theme is thus that the estimation technique itself is probably not as important as:

1. Whether the unconfoundedness assumption holds.
2. Whether there is overlap in the treatment and control distributions of the covariates.

4 Real-world Evidence on Selection Bias

We have seen so far that:

1. it is almost always valid to run a regression as long as you interpret it correctly,
2. randomized experiments are generally the preferred method for estimating causal effects under the potential outcomes framework, and
3. absent an RCT, one might try to use a selection on observables design – regression, matching, or propensity scores – to estimate a causal effect.

Under what conditions will a linear regression – the bread and butter of econometrics – or another selection on observables design approximate a randomized experiment? Loosely speaking, we need it to be the case that D_i is "as good as randomly assigned" after conditioning on the available covariates X_i. How often does it hold up in practice? Not as often as we would like.

4.1 LaLonde (1986): The NSW

LaLonde (1986) analyzes a randomized experiment evaluating a job training program, the National Supported Work Demonstration (NSW). The NSW, operated by Manpower Demonstration Research Corporation (MDRC), "admitted into the program AFDC women, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes."[12] (LaLonde 1986, p. 605) While the NSW is shown to increase post-training earnings by \$800-\$900 (1982 dollars), that is not the main focus of the article. Instead, LaLonde uses the experimental estimates as a "benchmark" to test whether typical econometric techniques can reproduce the same results. The short answer is that they cannot.

LaLonde needs a simulated control group in order to conduct his exercise – if he applies any sensible estimator to the experimental data (treated and control groups), he will get reasonable estimates because the treatment is randomly assigned. He therefore constructs a series of simulated control groups using data from the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS) (merged with Social Security Administration data). This is somewhat unusual in that the treated individuals and the control individuals are drawn from two entirely separate data sets, but it is not unreasonable for his purposes. LaLonde begins the benchmarking exercise by applying a series of differences-in-differences type estimators. The basic model is:

$$(1)\ y_{i,1979} - y_{i,1975} = \delta D_i + (\varepsilon_{i,1979} - \varepsilon_{i,1975})$$

[12]It is unclear from LaLonde's description how the MDRC administrators chose which applicants would enter the experiment. Given that there were only 6,616 trainees distributed between 10 cities, there was presumably a scarcity of slots relative to potential applicants.

This specification differences out any unobserved individual effects that are constant over time – it is equivalent to including individual fixed effects in the cross-sectional regression. Identification comes from comparing the change in earnings for those that participated in training to the change in earnings for those that did not participate. LaLonde also supplements this model with a regression that, instead of differencing, controls for pre-treatment earnings. This specification is more flexible in that it does not restrict the coefficient on pre-treatment earnings to be one.

$$(2)\ y_{i,1979} = \delta D_i + \beta y_{i,1975} + X_i\gamma + \varepsilon_{i,1979}$$

Table 1 presents estimates from these two specifications. The "Pre-Treatment" column presents differences between the income of treatment and control groups (or simulated control groups) in 1975, before the training program starts. If the treatment is randomly assigned, this difference should be close to zero, and for the true controls it is. LaLonde presents eight simulated control groups – for brevity I present the two control groups per gender that were closest to the experimental sample in terms of pre-treatment income. Table 1 is therefore more favorable to the nonexperimental estimates than LaLonde's equivalent tables. Nevertheless, we observe large variations between the experimental estimates and the nonexperimental estimates.

The "Diffs-in-Diffs" column presents estimates using the first differences specification presented above (equation 1). The experimental benchmarks are $833 for females and $847 for males. The nonexperimental estimates range from -$1,637 to $3,145, and in only one case does the nonexperimental confidence interval contain the experimental point estimate. The last two columns of Table 1 apply the model presented in equation 2 (first without additional covariates, and then with additional covariates). These models perform slightly better – three of the eight point estimates get reasonably close to the experimental benchmarks (one of them gets quite close). Nevertheless, the pre-treatment differences are actually the worst for the samples that produce the closest results, so there is no reason to believe that an objective econometrician would reliably choose point estimates close to the experimental benchmarks.

Table 1: One-Stage Estimates

Estimator:	Pre-Treatment Differences	Diffs-in-Diffs	Controlling For Previous Earnings	Fully Adjusted
Females				
Controls	-17	833	843	854
	(122)	(323)	(308)	(312)
PSID-3	-77	3,145	3,070	2,919
	(202)	(557)	(531)	(592)
CPS-4	-1,189	2,126	1,222	827
	(249)	(654)	(637)	(814)
Males				
Controls	39	847	897	662
	(383)	(560)	(467)	(506)
PSID-3	455	242	629	397
	(539)	(884)	(757)	(1,103)
CPS-3	337	-1,637	-1,396	1,466
	(343)	(631)	(582)	(984)

Notes: Standard errors in parentheses. Source: LaLonde (1986).

LaLonde then considers the performance of more advanced two-stage estimators. In particular, he applies the Heckman selection correction model from Heckman (1978). The Heckman selection correction models two equations separately: the participation equation (the first stage, a non-linear probit model) and the earnings equation (the second stage). In this sense it is not unlike two-stage least squares, but there are a couple key differences. First, it uses a "control function" approach to solve the endogeneity problem. Specifically, it uses estimates from the first stage equation as a regressor to control for the expected value of the earnings residual conditional on participation and the determinants of participation. Second, because it specifies the participation (treatment) dummy as a non-linear function of the covariates, it is possible to identify the training coefficient without any instruments (i.e., exclusion restrictions). Nevertheless, LaLonde experiments with several (questionable) instruments to see how well this model performs.

The results from the Heckman two-step estimator are reported in Table 2. The Heckman correction allows you to test the exogeneity of the treatment indicator by testing whether the

Table 2: Two-Stage Estimates

	Females		Males	
	Training	Participation	Training	Participation
Controls	861	284	889	-876
	(318)	(2,385)	(840)	(2,601)
5 Sketchy IVs	1,102	-606	-22	-1,437
	(323)	(480)	(584)	(449)
3 Sketchy IVs	1,256	-823		
	(405)	(410)		
2 Sketchy IVs	1,564	-552	13	-1,484
	(604)	(569)	(584)	(450)
No Instrument	1,747	-526	213	-1,364
	(620)	(568)	(588)	(452)

Notes: Standard errors in parentheses. Source: LaLonde (1986).

coefficient on the selection correction term in the second stage is significantly different than zero. For brevity I present only results for the samples that displayed the least evidence of selection into the treatment. The two-step estimators perform somewhat better than the one-step estimators, but the results are still not encouraging. On the positive side, the confidence intervals for all but one of the nonexperimental estimates contain the experimental point estimates. But the standard errors are so large that much of this "encouraging" performance is primarily due to the fact that the confidence intervals are huge. Male nonexperimental estimates are particularly bad, ranging from -$1,333 to $213 (see LaLonde's Table 6 for the full set of results). It seems likely that if additional data were available, the nonexperimental estimates would converge to different values than the experimental estimates.

When LaLonde's paper was published in 1986, it caused significant consternation among applied researchers trying to estimate causal effects. It is probably not an understatement to say that it sparked the pursuit of clean, transparent research designs that continues to this day.

4.2 Arceneaux, Gerber, and Green: More recent evidence

Arceneaux, Gerber, and Green (2006) (henceforth AGG) perform an exercise similar to LaLonde's exercise using data from a large-scale voter mobilization effort (this type of effort is often referred to as a "Get Out the Vote" campaign). In this effort, households are randomly called and encouraged to vote. Although the calling assignment is random, whether a household is actually contacted is non-random – people often do not answer their phones. Regressing a household's voting behavior on whether or not that household was contacted can thus give biased estimates of the causal effect of encouragement on voting. One way to correct for this bias is to use the original random calling assignment as an instrument for actual contact/encouragement. This estimator will consistently estimate the causal effect of encouragement on voting for households for whom the original calling assignment changed whether or not they were contacted (i.e., households that actually got contacted). In this case, that means that the instrumental variables estimator will estimate TOT, the effect of the treatment (being contacted and encouraged) on the treated (those who were contacted and encouraged). I refer to these estimates as the "experimental estimates."

Alternatively, however, we could try to use a matching estimator to condition on the observed covariates and, in that manner, estimate TOT. Specifically, we could find a match for every treated unit, and compare the difference in voter participation for the treated units and their matched pairs. We could then benchmark the results of this matching estimator, which should be valid under the selection on observables assumption, against the experimental estimates. This is exactly what AGG do.

The AGG data have at least one important advantage over the NSW data: the AGG sample size is massive. There are approximately 60,000 treated individuals and almost two million control individuals. All individuals (treated and control) were taken from voter registration lists, which contain detailed information on voting histories and demographic characteristics. Once included in the study, individuals were randomly assigned to treatment or control groups. Obviously, most people (97%) were assigned to the control group.

The first column of Table 3 reports experimental benchmark estimates (i.e., IV estimates). These estimates suggest that voter encouragement raises the probability of voting by approximately 0.3 to 0.5 percentage points. These estimates are precisely estimated and are not significantly different than zero – voter encouragement appears to have no appreciable effect on voting behavior, at least for this population.

Table 3: "Effect" of Voter Encouragement on Voting

	Experimental	OLS	Matching
Sample w/o Unlisted No.	0.5	2.7	2.8
	(0.4)	(0.3)	(0.3)
N	1,905,320	1,905,320	22,711
Sample w/ Unlisted No.	0.3	4.4	4.4
	(0.5)	(0.3)	(0.3)
N	2,474,927	2,474,927	23,467

Source: Arceneaux, Gerber, and Green (2006). Parentheses contain standard errors.

The second column of Table 3 reports OLS estimates that regress voting behavior on whether an individual was contacted, conditioning on a variety of covariates. These covariates include age, household size, gender, contest indicators, county indicators, and two years of previous voting behavior.[13] This is roughly comparable to LaLonde (1986), who has age, education, marital status, gender, race, and two years of prior earnings. In particular, both studies include two years worth of pre-treatment outcomes. The OLS estimates range from 2.7 to 4.4 and are highly significant (t-statistics of 9 to 14). These estimates imply that voter encouragement raises the probability of voting by 2.7 to 4.4 percentage points. Clearly the OLS estimates are biased – does this bias occur because of a lack of overlap in the covariates between the treated and untreated groups?

The third column of Table 3 reports matching estimates that find an *exact* match in the control group for each treated unit (this is possible because all covariates are discrete and

[13]Though AGG do not show the covariate balance across treated versus untreated individuals, there must be substantial imbalance (i.e., covariates can predict whether an individual answers her phone) because controlling for covariates has a strong effect on the OLS estimates.

the control group is enormous). They are able to match about 91% of observations using exact matches and 99.9% of observations using slightly less exact matches (e.g., coding age in 3 year intervals and dropping some geographic indicators). This means that close matches can be found for virtually all observations, enabling an estimator that is almost entirely non-parametric. Nevertheless, matching estimates range from 2.8 to 4.4 and are highly significant. Matching therefore does not appear to solve the selection bias problem, even with excellent overlap in the covariate distributions of the treated and control observations.

4.3 Shadish, Clark, and Steiner: Some Balance

Shadish, Clark, Steiner (2008) raise a somewhat different, but important, issue with LaLonde's NSW paper (and others like it). Their complaint is that LaLonde's NSW exercise confounds the assignment mechanism (random assignment versus observational data) with other factors – for example, sites, times, variable measurements, missing outcome data, etc. The core of Shadish, Clark, and Steiner's (henceforth SCS) argument is that the randomly assigned "experimental" control units come from one data set – the NSW data – while the non-randomly assigned "observational" control units come from other data sets – CPS and PSID. Thus it is possible that factors specific to the different data sets may be contributing to the observed differences in the estimates generated by using the CPS/PSID controls instead of using the experimental NSW controls. Simply put, LaLonde's study of confounding in the context of observational data may itself by confounded by other factors that correlate with the assignment mechanism.

SCS's response is, naturally, to randomly assign the assignment mechanism. Their study proceeds in three basic steps. First, they recruit students as test subjects. Next, they randomly assign students to either have a treatment randomly assigned to them or to be given a choice about which treatment they would like. In the random assignment arm, students are randomly assigned to either math training or vocabulary training. In the choice arm, students either choose math training or vocabulary training. In all cases students are tested on math and vocabulary after completing the training. Finally, SCS estimate the "effects" of training

in the choice arm using various selection on observable techniques (regression adjustment and several propensity score methods). These estimates adjust for a rich set of covariates that SCS collect from the students: sex, age, marital status, race, pretreatment vocabulary and math scores, number of math courses taken, stated preferences regarding math and literature, a personality measure, parental education, math intensity of major, ACT (standardized test) scores, and GPAs. The outcome of interest is the difference between a student's post-training math performance and her post-training vocabulary performance (or vice versa). SCS compare these estimates to the "true" effects of training that they estimate using the data from the random assignment arm. This is similar in spirit to LaLonde's exercise, but SCS can be confident that no other factors are correlated with whether a student ends up in the random assignment arm or the choice arm.

SCS present their results in Table 1 of their paper (p. 1338). I summarize the notable patterns from their results here. First, students randomly assigned to math training do better at math relative to vocabulary, and students randomly assigned to vocabulary training do better at vocabulary relative to math. Second, selection bias turns out to be modest in this experiment. The unadjusted estimate in the non-randomly assigned data is only 25% higher than the true treatment effect (from the randomly assigned data) for math. It is only 9% higher than the true treatment effect for vocabulary. Third, of the various selection on observables designs that SCS implement, OLS (aka "ANCOVA") does as well or better than anything else. Regression adjustment (OLS) achieves an 84% bias reduction in the math estimate using non-randomly assigned data and a 94% bias reduction in the vocabulary estimate using non-randomly assigned data. Among the propensity score methods, SCS use blocking, including the propensity score as a regressor, and weighting. They also implement some "doubly robust" methods. Blocking works well for both math and vocabulary, but weighting only works well in the vocabulary case. Including the propensity score as a regressor results in less bias reduction than just controlling directly for the covariates that go into the score.

The most interesting part of SCS's study is that they also try propensity score blocking

using only "predictors of convenience" – sex, age, marital status, and race. This purposely omits a lot of rich covariates that they have at their disposal, and it simulates a scenario in which a researcher has limited controls available in the data. Propensity score methods that only use predictors of convenience do much worse than propensity score methods (or simple linear regressions) that use the full set of predictors – bias reduction is in the range of only 0% to 40%. Thus SCS provide evidence that *if* you have a rich set of relevant covariates, you may get decent estimates using non-experimental data. However, in most cases it will be difficult to say when you have a sufficiently rich set of covariates!

4.4 Summary

The LaLonde NSW study demonstrates that conventional econometric techniques may be insufficient to solve the selection bias problem in a typical program evaluation scenario (when not using randomized treatment assignments). The AGG experiment further demonstrates that even in cases that seem well-suited to the selection on observables design, estimates can be biased pretty badly. The main problem that AGG face (or would face, if they didn't have the experimental estimates) is that it's hard to make the case that they observe all of the important factors in determining whether a caller makes contact with an individual or whether an individual turns out to vote. Of course, *one could say the same thing about almost any dataset with observational (i.e., not randomly assigned) data*, so it's difficult for the real-world econometrician to determine when the selection on observables design does or does not hold.[14] The SCS conclusions are somewhat more optimistic than AGG, but again it's hard to know when you have observed "enough" selection factors. An alternative to the selection on observables design is, of course, the selection on unobservables design. The next section of the course focuses on this category of research designs.

[14]My personal belief is that it's most plausible when the selection was performed by an individual or body that has observes the same data that is available to the researcher – e.g., a college admissions officer who does not conduct interviews or read long personal essays. In cases in which units are self-selecting (which to be honest is most cases), it's far less plausible.

4.5 Freedman (1991): A Natural Experiment

Freedman (1991) offers a critique of linear regression applications along with an example of an historical natural experiment as an alternative research design. Freedman begins with four possible views of regression, progressing from the most optimistic to the most pessimistic:

> (1) Regression usually works, although it is (like anything else) imperfect and may sometimes go wrong.
>
> (2) Regression sometimes works in the hands of skillful practitioners, but it isn't suitable for routine use.
>
> (3) Regression might work, but hasn't yet.
>
> (4) Regression can't work.

Source: Freedman (1991), p. 292.

Freedman professes that his own view falls between (2) and (3). I'm not sure exactly what (3) entails – the properties of linear regression are pretty well-established, so if it were going to work, I would think it would have done so by now. But, like Freedman, I agree that "good examples [of causal estimates from regression] are quite hard to find."

In contrast to regression models (and more sophisticated models, like matching or propensity score methods), Freedman presents the work of John Snow on cholera in the 1850s (that is to say, Snow conducted the work during the 1850s, on cholera at that time). Snow postulated that unsanitary water caused cholera outbreaks (at the time it was believed that cholera arose from poisonous particles in the air). Snow had several pieces of circumstantial evidence to support his position, but in order to prove his hypothesis he observed that water distribution in London gave rise to a natural experiment.

In the area that Snow was studying, two water supply companies, Southwark and Vauxhall Company and Lambeth Company, competed for customers. One company (Lambeth) drew water upstream of the sewage discharge points in the River Thames, while the other

(Southwark and Vauxhall) drew water downstream of the discharge points. Both companies had pipes running down virtually every street and alley, and which houses chose which company appeared to be virtually random. Snow wrote, "Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies." In today's terminology, Snow would say that the observable attributes (covariates) were balanced across the two companies. Having convinced himself that the choice of water company was nearly random, he examined the cholera death rate for customers of both companies.

The cholera results, presented in Table 4, are striking. Death rates for the downstream company are over eight times higher than death rates for the upstream company. Given the sample size, and the fact that the customers of both companies are spatially intermixed, it is clear that these results are highly significant despite the absence of standard errors. As Freedman writes (p. 298):

> As a piece of statistical technology, Table [4] is by no means remarkable. But the story it tells is very persuasive. The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data.

Table 4: Snow's Table IX

	Number of Houses	Deaths from Cholera	Deaths per 10,000 Houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Notes: Source: Freedman (1991).

Freedman's emphasis is that the findings' credibility is due to the persuasive research design in conjunction with an impressive data set, rather than the sophistication of the statistical modeling technique. The implication is that, if you can't get data that have some sort of clean (i.e., as good as randomly assigned) variation in the treatment of interest, then

you can't convincingly identify a causal effect, no matter how fancy an estimation technique or theoretical model you apply. My *personal* view is in line with Freedman's, but it is certainly a matter considered open for debate within economics/econometrics. Regardless, as a piece of empirical evidence, Snow's 150-year-old study is clearly more credible than the vast majority of articles published today in economics (or other social sciences).

With this motivation in mind, we begin our discussion of "selection on unobservables" research designs. Virtually all of these techniques have historical roots that precede the LaLonde (1986) paper by years, if not decades. Nevertheless, their increased popularity is likely in part a reaction to LaLonde's study.

5 Panel Data and Differences-in-Differences

Panel, or *longitudinal*, data sets consist of repeated observations for the same political jurisdictions, firms, individuals, or other economic agents. Typically the observations are at different points in time. The most common research design for policy analysis with panel data is the differences-in-differences model. In its simplest incarnation, the diffs-in-diffs model entails identifying two cross-sectional units (states, cities, countries, etc.), one of which was exposed to a policy change (or some other treatment) and the other of which was not. With longitudinal data, we collect information on the two units both before the policy change and after the policy change. To estimate the effect of the policy on a given outcome, we simply compare the change in the outcome for the treated unit to the change in the outcome for the control unit.

5.1 Differences-in-Differences

Suppose that we observe two states, $s = 0$ and $s = 1$, one of which is affected by a policy change and the other of which was not. Further suppose that we observe these states for two time periods, $t = 0$ (pre-policy change) and $t = 1$ (post-policy change). Formally, for some outcome Y_{ist} that we observe at the individual level, the differences-in-differences estimator

is

$$(\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00})$$

where $\overline{Y}_{st} = \frac{1}{N_{st}} \sum_i Y_{ist}$. To examine the strengths and weaknesses of this estimator, write $Y_{ist} = \overline{Y} + \tau D_{st} + \gamma_s + \delta_t + \varepsilon_{st} + u_{ist}$. Note that the inclusion of ε_{st} guarantees that $\overline{u}_{st} = 0$.

$$(\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00}) =$$

$$[(\overline{Y} + \tau + \gamma_1 + \delta_1 + \varepsilon_{11}) - (\overline{Y} + \gamma_1 + \delta_0 + \varepsilon_{10})] - [(\overline{Y} + \gamma_0 + \delta_1 + \varepsilon_{01}) - (\overline{Y} + \gamma_0 + \delta_0 + \varepsilon_{00})] =$$

$$(\tau + \delta_1 - \delta_0 + \varepsilon_{11} - \varepsilon_{10}) - (\delta_1 - \delta_0 + \varepsilon_{01} - \varepsilon_{00}) =$$

$$\tau + (\varepsilon_{11} - \varepsilon_{10}) - (\varepsilon_{01} - \varepsilon_{00})$$

The key assumption for identifying τ will therefore be $E[\varepsilon_{11} - \varepsilon_{10}] = E[\varepsilon_{01} - \varepsilon_{00}]$. In other words, the outcomes for the two states must have similar trajectories over the two time periods absent any treatment effect. Any factor that is specific to state s but does not change over time, or changes over time but changes in equal amount for both states, is netted out in the diffs-in-diffs estimator.

It is important to note, however, that the condition above only guarantees that we will identify τ *in expectation*. Because we only observe a single observation for each of the ε_{st} terms in the expression above, there is no guarantee that the noise from ε_{st} will not swamp our estimate of the treatment effect, τ – we cannot appeal to the law of large numbers as we do when we have N independent observations.

The diffs-in-diffs estimator can also be easily implemented within a regression framework. Consider running the regression:

$$Y_{ist} = \alpha + \tau D_{st} + \gamma \mathbf{1}(s = 1) + \delta \mathbf{1}(t = 1) + \varepsilon_{st} + u_{ist}$$

In other words, simply regress Y on a treatment indicator, a state dummy, and a time dummy. The state dummy controls for between-state differences in Y that are constant over time, and the time dummy controls for between-time period differences in Y that are identical across states. Identification of τ again comes from the assumption that ε_{st} is uncorrelated with the treatment indicator (which is equal to the interaction between the state dummy and the time dummy) conditional on the state dummy and the time dummy. Note that in the regression format, it is easy to control for individual-level covariates. You can also see the standard errors issue in this framework. If we use the typical OLS standard errors that assume independence across all observations, we are effectively claiming that the only error in our estimator is sampling error that arises because we do not observe the entire population of each state. However, if $\sigma_{\varepsilon}^2 \neq 0$, i.e. there are state-specific shocks that vary over time, then this independence assumption is violated, and our standard errors will be wrong.

The key identifying assumption in differences-in-differences is that $E[\varepsilon_{11} - \varepsilon_{10}] = E[\varepsilon_{01} - \varepsilon_{00}]$. This is often referred to as the "parallel trends assumption," because it implies that the outcome followed similar (parallel) trends in both states prior to the policy intervention, and more importantly that, absent the policy intervention, it would have continued to follow the same trends following the intervention date. Returning to the notation of the Rubin Causal Model, the parallel trends assumption implies that $E[Y_{s1}(0) - Y_{s0}(0)]$ is identical for both states; the potential outcome under no treatment changes similarly in both states.

Like all identifying assumptions, the parallel trends assumption is impossible to definitively test. Nevertheless, we can often provide suggestive evidence that it holds by plotting the outcome for treated and control states prior to the policy intervention. If the parallel trends assumption holds, then the outcome's time series for the control state should follow a similar trend to the outcome's time series for the treated state prior to the intervention. If the trends are not similar – for example, if the outcome is moving upwards for the treated state but moving downwards for the control state – then this implies that the control state does not provide a good counterfactual for the treated state, and the differences-in-differences design is suspect.

5.2 Triple Differences

A diffs-in-diffs research design can sometimes be made more compelling by adding another layer of differencing to the estimator, resulting in a triple-diffs estimator. For example, consider a policy change in state 1 in time period 1 that only affects persons 65 years and older. In that case, we might use individuals aged 55-64 as an additional "control" group. In practice, we would implement this with a triple differences estimator. Let $\overline{Y}_{sta}$ be defined as above, but with $a = 0$ signifying persons of age 55-64 and $a = 1$ signifying persons of age 65 and older. Then the triple differences estimator is:

$$[(\overline{Y}_{111} - \overline{Y}_{110}) - (\overline{Y}_{101} - \overline{Y}_{100})] - [(\overline{Y}_{011} - \overline{Y}_{010}) - (\overline{Y}_{001} - \overline{Y}_{000})]$$

In other words, we compare the evolution of the gap between 65+ year olds and 55-64 year olds in the treated state to the evolution of the gap between 65+ year olds and 55-64 year olds in the control state. The advantage of this triple-diffs structure is that it allows us to relax our assumptions on ε_{st}. We no longer need to assume that outcomes for both states would evolve similarly in expectation – we now need only assume that, to the extent that outcomes evolve differently in state $s = 1$ than state $s = 0$, the differences affect age groups $a = 1$ and $a = 0$ similarly.

We can easily implement this triple-diffs estimator within the regression framework. The key is to put in an indicator for every main effect or interaction up to, but not including, the level at which the treatment varies. Thus we include main effects for age, state, and time, as well as all possible two-way interactions between each of those indicators. The regression looks like:

$$Y_{ista} = \alpha + \tau D_{sta} + \gamma_1 \mathbf{1}(s = 1) + \gamma_2 \mathbf{1}(t = 1) + \gamma_3 \mathbf{1}(a = 1) + \gamma_4 \mathbf{1}(s = 1)\mathbf{1}(t = 1)$$

$$+\gamma_5 \mathbf{1}(s = 1)\mathbf{1}(a = 1) + \gamma_6 \mathbf{1}(t = 1)\mathbf{1}(a = 1) + \varepsilon_{sta} + u_{ista}$$

5.3 Applications: Card (1990), Ashenfelter & Greenstone (2004), and Kellogg & Wolff (2008)

A canonical example of a diffs-in-diffs papers is Card's (1990) study of the Mariel Boatlift.[15] The Mariel Boatlift occurred from May to September of 1980 when Cuba allowed any citizen wishing to emigrate to the United States free passage from the port of Mariel. Approximately 125,000 Cuban immigrants arrived in Miami during this time period, increasing the local labor force by about 7%.

Card examines wage and employment outcomes for various groups of natives, particularly blacks and lower-skilled workers – the latter group is more likely to be in direct competition with the newly arrived immigrants (who were relatively low-skilled). He compares the evolution of these outcomes over the 1979 to 1981 period in Miami to their evolution in four comparison cities: Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. For blacks, the difference in log wages between Miami and comparison cities changes from -0.15 in 1979 to -0.11 in 1981, so the diffs-in-diffs estimate for log wages is 0.04 (with a standard error of about the same size). The difference in the employment-to-population ratio between Miami and comparison cities changes from 0.00 in 1979 to 0.02 in 1981, so the diffs-in-diffs estimate for employment is 0.02. Estimates for unemployment rates and low-skilled blacks show similar patterns. Overall, there is no evidence that immigrants harm natives' labor market outcomes.[16]

Ashenfelter and Greenstone (AG) use a combination of diffs-in-diffs and triple differences designs to estimate the effect of speed limits on highway fatalities. They leverage a "natural

[15]The term "differences-in-differences" is thrown around often, but to my knowledge there is no formal definition for what classifies as a "diffs-in-diffs" paper. Arguably many panel data papers that control for both individual-specific effects and aggregate time effects are using some form of double differencing estimator, but that doesn't mean that we'd necessarily refer to them as diffs-in-diffs papers. In my mind, a diffs-in-diffs paper generally uses some sort of variation in the treatment that occurs at an aggregated level, e.g. the city or state level. We therefore tend not to worry so much about individuals selecting into the treatment (it's unlikely that most will move in response to just one shock), but rather we worry that the treatment was implemented in one area rather than another for some non-random reason (e.g., legislative endogeneity).

[16]Interestingly, for Americans of Cuban descent, Card finds evidence of some increase in unemployment rates, but no effect on wages.

experiment" that occurred when the Federal government permitted states to raise rural (but not urban) Interstate highway speed limits to 65 mph (from 55 mph) in 1987. The ultimate goal of the paper is to estimate the value of a statistical life (VSL) based on the idea that states that adopted higher speed limits must value the time savings more than the lives lost from higher speeds. However, here we focus mainly on the fatality effects of higher speed limits.

As a response to the energy crisis in the early 1970s, the US Federal government enacted a 55 mph limit in 1974. The intent of this law was to reduce gasoline consumption (cars are more efficient at 55 mph than 65 mph), but an unintended effect of this law was an apparent decline in traffic fatalities. In 1987, the 55 mph speed limit was partially rescinded, and states were allowed 65 mph *on rural Interstates only*. Not all states chose to raise speed limits, however, which enables a diffs-in-diffs strategy: AG compare changes in fatalities for states that increased their speed limits in 1987 with changes in fatalities for states that did not.

However, one problem when examining the map of states that changed speed limits is that there is strong geographic clustering in the policy variable –almost everyone changed their speed limits except Northeastern states. Thus we may confound changes in fatalities from speed limits with differing geographic trends in fatalities across regions. AG thus augment their design with a triple-diffs strategy. Urban Interstates are unaffected by the speed limit change, even in treated states. We can thus compare how the rural/urban fatality rate difference changes for treated states versus untreated states.

In practice, AG estimate rural and urban Interstate effects separately. Write the rural effect estimate as $(Y_{T1R} - Y_{T0R}) - (Y_{C1R} - Y_{C0R})$ and the urban effect estimate as $(Y_{T1U} - Y_{T0U}) - (Y_{C1U} - Y_{C0U})$. This strategy allows AG to test for differing geographic trends in fatalities, as long as those differing trends affect both the rural fatality rate and the urban fatality rate in an equal manner. They also examine rural arterial roads, which are another group of roads on which the speed limit should not change, even in the treated states.

Using a diffs-in-diffs strategy, AG find an average effect of speed limit increases on rural Interstates fatalities of +36% ($t = 4$). The same strategy estimates an average effect of speed limit increases on urban Interstate fatalities of -6% ($t = 1$), and an average effect on rural arterial fatalities of +8% ($t = 2$). We expect an effect in the first case but not in the latter two, because speed limits do not change on urban Interstates or rural arterial roads. The fact that the estimates are much smaller and insignificant (or close to insignificant) for urban Interstates and rural arterial roads is thus reassuring. If we found large effects on these roads, it would imply that the parallel trends assumption is violated.

Kellogg and Wolff (2008) provide another nice example of a triple differences research design. Their interest is in estimating the effect of Daylight Savings Time (DST) on electricity usage. DST may reduce energy usage because, for example, it aligns the hours at which people are awake with the hours at which the sun is up, thus reducing lighting needs. On the other hand, it may increase energy usage because people wake up when the sun rises (as opposed to after it has risen) and need to heat their homes during this time.

Kellogg and Wolff leverage an extension to DST in Australia that was put in place for the Summer 2000 Olympics. Some Australian states, including New South Wales (where the Olympics were held) and Victoria, extended DST beyond the date at which normally terminates. Other states, including South Australia, did not. They compare the change in electricity usage for Victoria (the "treated" state) to the change in electricity usage for South Australia (the "control" state). They are concerned, however, that electricity usage might be trending differently in these two states for reasons unrelated to the DST extension.

To address this concern, they observe that DST should not affect electricity usage during the middle of the day, when the sun is always in the sky regardless of whether you are on DST or standard time. The midday hours thus provide an extra "control" group that should be unaffected by DST. This allows them to implement a triple differences estimator. Specifically, they define the treated portion of the day as the hours from 0:00 to 12:00 and 14:30 to 24:00. They define the control portion of the day as the hours from 12:00 to 14:30.

Using a simple differences-in-differences estimator with electricity usage during the treated portion of the day as the outcome, they find that electricity usage fell by 0.4% in Victoria as compared to South Australia. However, they also find that electricity usage during the control portion of the day fell by 0.2% in Victoria as compared to South Australia. The triple differences estimator takes the difference between these two double differences estimators; thus their final estimate is that DST reduced electricity usage by 0.2% (with a standard error of 1.5%). The identifying assumption here is that, if Victoria and South Australia are trending differently from each other, these differential trends still have the same proportional effect on electricity usage from 12:00-14:30 and electricity usage from 0:00-12:00/14:30-24:00.

To increase the precision of their estimates, they also implement the triple differences estimator in a regression framework. The regression framework allows them to control for other determinants of electricity usage (e.g., day of week, weather, etc.). This reduces the unexplained variation in the outcome and thus reduces their standard errors. An observation in this regression is the half-hour-by-day-by-state. They regress electricity usage on the treatment variable (one if DST is in effect and it is before 12:00 or after 14:30, zero otherwise) and day-by-state indicators (which basically correspond to the state-by-time interactions in Section 5.2), hour-by-state indicators (which basically correspond to the state-by-age interactions in Section 5.2), hour-by-year indicators (which basically correspond to the time-by-age interactions in Section 5.2), and other control variables.

In the regression framework, they find that DST *increases* energy usage by 0.02% (if they impose a homogeneous effect across all treated hours) or 0.09% (if they allow for heterogeneous effects of DST across different times of day). The standard error drops to 0.4%, so they are able to rule out substantial electricity savings from DST – savings of 0.5% or higher, for example, are unlikely.

6 Instrumental Variables

Instrumental variables (IV) methods are a cornerstone of econometrics – these methods date back to the work of Tinbergen and Haavelmo in the 1930s and 1940s. Our understanding of IV methods advanced significantly during the 1990s, however, with seminal work on IV in the context of treatment effect heterogeneity and IV methods in the case of a large number of weak instruments. For the purposes of these notes, I will use the phrase "IV methods" to refer generally to methods using instrumental variables, including IV, two stage least squares (2SLS), and limited information maximum likelihood (LIML). Also, several other popular estimators – in particular, regression discontinuity (RD) – are in fact special cases of IV.

6.1 Basic IV

Consider a model of the form

$$y_i = \beta_0 + \beta_1 d_i + \varepsilon_i \tag{1}$$

I will sometimes refer to this equation as the "structural equation." At this point we are not assuming that d_i is binary – it may have more than two points of support, or it may be continuous. The standard condition that we need for a linear regression of y_i on d_i to consistently estimate β_1 is $\text{Cov}(d_i, \varepsilon_i) = 0$. This will be true if d_i is randomly assigned, and it could be true in other situations as well. In general, however, it will not be true.

An alternative way to estimate β_1 is via instrumental variables. The goal in IV is to find some subset of the variation in d_i, call it z_i, that is uncorrelated with ε_i (i.e., as good as randomly assigned). Formally, our goal is to find an instrument z_i, not in equation (1), that satisfies the following two properties:

1. $\text{Cov}(z_i, d_i) \neq 0$
2. $\text{Cov}(z_i, \varepsilon_i) = 0$

The first assumption ensures that z_i actually captures some of the variation in d_i. If it doesn't, then it will be of no use to us in estimating the effect of d_i on y_i. The second assumption ensures that z_i is uncorrelated with ε_i (obviously). This assumption is often referred to as the "exclusion restriction" because it implies that the instrument, z_i, can be excluded from equation (1). If z_i were correlated with ε_i, we would want to include it as a covariate (given that it's also correlated with d_i by our first assumption). This would violate our condition that z_i not be in equation (1).

To fix ideas, let us consider an application from Angrist (1990). Suppose that we would like to estimate the causal effect of military service on earnings. One way to estimate this effect would be to regress earnings on a military service indicator and a bunch of other covariates. Is it plausible that the variation in service is uncorrelated with everything else that affects earnings, such as unobserved ability? Probably not, even after conditioning on covariates. We generally think that on average people who go into the army are different in fundamental ways from people who do not, and specifically we believe that the two groups are likely different in ways that affect earnings.

An alternative way to estimate the effect of military service on earnings is to find an instrument for service that satisfies the criteria above. Angrist uses the Vietnam draft lottery as an instrument for military service. During the Vietnam War the military needed a "fair" way to determine which young men got drafted, so they held a lottery in which men were assigned to be drafted in a certain order based on their birthdays. The lottery determined men with certain birthdays to be "draft eligible," and other men to be draft ineligible. Draft eligible men *might* be drafted (depending on manpower needs), but draft ineligible men would not be.

If we let z_i be a dummy variable that is 1 if a man is draft eligible (based on the lottery results) and 0 otherwise, then z_i is a promising instrument for service. We know that the first condition, $\text{Cov}(z_i, d_i) \neq 0$, will be satisfied because men who are draft eligible will be more likely to serve than those who don't, inducing a positive correlation between z_i and d_i. We also know that the second condition for a good instrument, $\text{Cov}(z_i, \varepsilon_i) = 0$, is likely to

be satisfied because draft eligibility was randomly determined by the lottery. By definition, no characteristic, other than those characteristics directly affected by draft eligibility, can possibly be correlated with eligibility.

The IV estimator is:

$$\hat{\beta}_{IV} = (Z'D)^{-1}(Z'Y)$$

In the general case, Z could contain not only the instrument z_i (draft eligibility), but also predetermined covariates x_i (gender, race, parental education, etc.). D would then contain both the treatment of interest, d_i, and the predetermined covariates x_i. In the case in which there are no covariates, we can write $\hat{\beta}_{IV} = \text{Cov}(z_i, y_i)/\text{Cov}(z_i, d_i)$.

It is straightforward to show that $\hat{\beta}_{IV}$ is a consistent estimator of β_1 given the assumptions above.

$$\text{plim}(\hat{\beta}_{IV}) = \text{plim}[(Z'D)^{-1}(Z'Y)]$$

$$= \text{plim}[(Z'D)^{-1}(Z'D\beta + Z'\varepsilon)]$$

$$= \text{plim}[(Z'D)^{-1}(Z'D)\beta] + \text{plim}[\frac{1}{N}(Z'D)^{-1}]\text{plim}[\frac{1}{N}(Z'\varepsilon)]$$

$$= \beta$$

This formal derivation, however, gives limited intuition regarding why or how IV operates. For intuition, we will turn to alternative methods of implementing the IV estimator.

6.2 The Reduced Forms and 2SLS

The most popular way to implement the IV estimator is via a two stage procedure known as two stage least squares (2SLS). If we have one instrument and one variable that we want

to instrument for, 2SLS and IV are the exact same thing (in this case we would say that we are "exactly identified"). IV is thus a special case of 2SLS – you can always use 2SLS in any scenario in which you can use IV, though the reverse is not true. We begin by writing out the two stages of 2SLS, and then consider what is going on:

1. *First Stage* We first estimate a regression of d_i (the variable that we want to instrument for – e.g., military service, in our hypothetical example) on the instrument, z_i (e.g., the lottery number), and all of the predetermined covariates, x_i. This regression looks like:

$$d_i = \gamma_1 z_i + x_i \gamma_2 + u_i$$

where z_i and γ_1 are scalars, x_i is a $1 \times K + 1$ vector that includes all covariates and a 1, and u_i is a residual term. Take the predicted values of d_i (e.g., predicted military service, $\hat{d}_i = \hat{\gamma_1} z_i + x_i \hat{\gamma_2}$) from this regression and use them in place of the actual values of d_i in the second stage.

2. *Second Stage* In the second stage, we run the regression that we originally wanted to estimate, but instead of including the variable that we want to instrument for (d_i), we include its predicted values from the first stage ($\hat{d}_i$). In our example, instead of running earnings on service and other covariates, we would run earnings on predicted service (from the first stage) and other covariates. Thus the regression looks like:

$$y_i = \beta_0 + \beta_1 \hat{d}_i + x_i \beta_2 + \varepsilon_i$$

The estimate of β_1 from this regression will be consistent.

Note that both the first and second stages always contain the same set of covariates (you can't exclude certain covariates from the first stage and then include them in the second stage, and you can't exclude covariates from the second stage and include them in the first stage, unless you intend to use them as instruments). In matrices, define Z to

be a matrix that includes the instrument (z_i) and the predetermined covariates (x_i). D is a matrix that includes the treatment (d_i) and the predetermined covariates (x_i). Then $\hat{\beta}_{2SLS} = (D'P_Z D)^{-1}(D'P_Z Y)$, where $P_Z = Z(Z'Z)^{-1}Z'$.

Now that we have introduced the first stage and the second stage, we are almost done, but before we move on to the next section I will introduce the reduced form equation. Technically the term "reduced form" refers to any regression which regresses an endogenous variable (i.e., a not-exogenous variable; in our case y_i and d_i are our two endogenous variables) on all of the exogenous variables (z_i and x_i). So, if you consider the two regressions that we estimated above, you will see that the first stage is in fact a reduced form equation. However, in general I will use the term "reduced form" to refer specifically to the reduced form equation that regresses y_i on all of the exogenous variables (z_i and x_i). So the reduced form in our example is:

$$y_i = \pi_1 z_i + x_i \pi_2 + v_i$$

What does the reduced form measure? The reduced form measures the causal effect of the instrument (z_i) on the outcome variable (y_i). In our example, the coefficient that we get from running the reduced form gives us an estimate of the effect on earnings of winning the scholarship lottery. Note that if z_i is a good instrument, then the causal effect should run only through the variable that is being instrumented for (d_i). In our example, that means that becoming draft eligible should affect your income only because it makes it more likely that you'll serve, not for some other reason (e.g., because the draft eligible individuals flee to Canada to avoid going to Vietnam).

So there are three equations we want to keep in mind:

1. The first stage, which regresses the variable we're instrumenting for on the instrument(s) and the other exogenous variables. This predicts how the variable we're instrumenting for changes as our instrument changes.

2. The second stage, which regresses y_i on the predicted values from the first stage and the other exogenous variables. This gives us our IV estimate of β_1.

3. The reduced form, which regresses y_i on the instrument and the other exogenous variables. This measures how y_i changes as we change our instrument z_i. Note that we never have to run the reduced form in the 2SLS procedure, but as you will see in the next section, it is a useful concept to keep in mind.

6.3 IV Intuition

At this point, some might ask, "Why not just run the reduced form? Why bother with IV (2SLS) at all? After all, the reduced form gives unbiased predictions, and it's much less complex than this two stage procedure." In other words, why not simply replace the variable we want to instrument for (d_i) with the instrument (z_i)? Actually, this isn't necessarily a bad idea. As Josh Angrist (a former advisor of mine, and an author of the *Mostly Harmless Econometrics* textbook) says, "Many papers would do well to stop with the reduced form." The reduced form makes explicit exactly where the identification in the research design is coming from, and it does not suffer from some of the "weak instruments" issues that we will discuss later. Any time you are dealing with a single instrument, it's a good idea to estimate the reduced form and check whether it conforms to your expectations, even if you don't put it into the paper.

The answer to the question above, however, is that we usually are not interested in measuring the effect of z_i on y_i, which is what the reduced form gives us. Instead, we are interested in measuring the effect of d_i on y_i. That is what IV gives us. In our example, we are interested in measuring the effect of military service on wages, so we run IV. If we just ran the reduced form, we would get the effect of becoming draft eligible on wages. While that may be of some policy interest in evaluating the draft lottery, it is not what we are looking for.

From a linear algebra perspective, 2SLS/IV estimates β by first projecting all of the data

onto the subspace spanned by Z – all of the exogenous variables in the regression (i.e., the instrument and the predetermined covariates) – and then running the regression of y_i on d_i and x_i after they have been projected onto this subspace. In this sense it should be clear that we are only using the "good" variation in d_i (i.e., the variation in d_i that comes from z_i) to estimate β_1. However, in the case in which you have a single instrument (which is all we have discussed so far), there is an even cleaner interpretation.[17]

In the case of one treatment and one instrument, the estimate of β_1 that we get from IV equals the reduced form coefficient rescaled by the first stage coefficient. That is to say:

$$\hat{\beta}_{1IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$$

What this shows is that the IV estimate is very closely related to the reduced form estimate – in fact, it's exactly proportional to the reduced form estimate. Why is this a useful formulation? Well, consider what each coefficient means.

In our example, the reduced form coefficient ($\hat{\pi}_1$) measures the effect of becoming draft eligible on earnings. But that is not what we want; what we want is the effect of military service on earnings. Because the draft lottery only affects earnings due to its effect on increasing the probability of service (or so we're assuming), the reduced form coefficient represents the effect of an unknown change in the probability of service. The problem is that we don't have the units right. If we knew that everyone who was draft eligible served, and that everyone who was not draft eligible did not serve, then we could interpret the reduced form coefficient as the causal effect of military service on earnings. Why? Well, remember that because the instrument z_i is randomly assigned (i.e. draft eligibility is randomly picked), the eligible and non-eligible are on average comparable in every way except that the eligible serve and the non-eligible do not. So any difference in earnings between the eligible (i.e. those with $z_i = 1$) and the non-eligible (i.e. those with $z_i = 0$) must be due to service. Thus

[17] If you are "overidentified," i.e. you have more instruments than you need, then this interpretation does not hold anymore, though it is still conceptually useful. In our example, we are "just identified" (one variable to instrument for, i.e. service, and one instrument, i.e. the draft lottery), so you can apply the interpretation that I'm about to give.

the coefficient on z_i in the reduced form ($\hat{\pi}_1$) is the effect of military service on earnings.

In general, however, it is unlikely that the draft eligibility is the only determinant of military service. Some people volunteer to serve, for example. So how do we rescale the reduced form coefficient so that we get the units right? The answer is that we divide through by the first stage coefficient, $\hat{\gamma}_1$. Why does this work? Consider our specific example. The first stage estimates the effect of draft eligibility on the probability of military service. So the first stage coefficient, $\hat{\gamma}_1$, tells you how much, on average, your chance of serving increases if you become draft eligible. So suppose that $\hat{\gamma}_1 = 0.1$, i.e. that those who are draft eligible are 10 percentage points more likely to serve than those who are not eligible. Also suppose that $\hat{\pi}_1 = -500$, i.e., those who are draft eligible earn \$500 less per year on average than those who do are not draft eligible. Then we know that the people who are draft eligible are earning \$500 less because they are more likely to serve (this comes from the reduced form). And we know that they are on average 10 percentage points more likely to serve. So what is the effect of service? It is $-\$500/0.1$ (the change in earnings divided by the change in the probability of service), or \$5000. In other words, our estimate of the effect of service on earnings is $\hat{\beta}_{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$. So the reduced form coefficient represents the causal effect of some additional probability of service on earnings (how much additional probability is unknown until we see the first stage), and the first stage coefficient rescales that coefficient appropriately to reflect the change in the probability of service that the instrument (draft eligibility) generates.

So far you have taken it on faith that the formula $\hat{\beta}_{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$ is actually true. But it is actually simple to prove. Recall that $\hat{\beta}_{IV} = (Z'D)^{-1}(Z'Y)$. If you accept that we can apply partitioned regression to IV just like we can with OLS, then it is trivial to transform the formula for $\hat{\beta}_{IV}$ into one in which Z and D are always vectors.[18] If Z contains covariates X, simply redefine Z such that $\tilde{Z} = M_X Z_1$, where Z_1 is a column vector containing only the instrument and M_X is the orthogonal projection matrix for the covariates, $M_X = I -$

[18]That partitioned regression works for the 2SLS procedure should be fairly obvious. It is less self-evident that partitioning must work for the IV formula as well.

$X(X'X)^{-1}X'$ (X is an $N \times K+1$ matrix containing all covariates and a column of ones).[19] Thus we can always write $\hat{\beta}_{1IV}$ as

$$\hat{\beta}_{1IV} = (\tilde{Z}_1' D_1)^{-1}(\tilde{Z}_1' Y) = \text{Cov}(\tilde{z}_i, y_i)/\text{Cov}(\tilde{z}_i, d_i)$$

Now consider $\hat{\pi}_1$ and $\hat{\gamma}_1$. The former comes from a regression of y_i on $\tilde{z}_i$, so $\hat{\pi}_1 = \text{Cov}(\tilde{z}_i, y_i)/\text{Cov}(\tilde{z}_i, z_i)$. The latter comes from a regression of d_i on $\tilde{z}_i$, so $\hat{\gamma}_1 = \text{Cov}(\tilde{z}_i, d_i)/\text{Cov}(\tilde{z}_i, z_i)$. Thus

$$\hat{\pi}_1/\hat{\gamma}_1 = \frac{\text{Cov}(\tilde{z}_i, y_i)/\text{Cov}(\tilde{z}_i, z_i)}{\text{Cov}(\tilde{z}_i, d_i)/\text{Cov}(\tilde{z}_i, z_i)} = \hat{\beta}_{1IV}$$

The takeaway of all of this is that, when working with an IV estimator, the entire experiment is in the reduced form. The reduced form measures the causal impact of the instrument on the outcome – the first stage exists only to rescale that estimate and "get the units right." Thus, when applying IV, you should always consider the underlying reduced form that you are running and ascertain whether it makes sense and whether it is identifying the causal effect in the manner that you originally imagined.

6.4 Multiple Instruments

It's often very difficult to find one good instrument, let alone two or more good instruments. Nevertheless, in some cases a single conceptual instrument will be parameterized though multiple variables (we will see an example of this in the next section). In those cases, we say that the equation is "overidentified," in the sense that we have more instruments than we need. It's impossible to incorporate more than one instrument into the IV estimator because $\hat{\beta}_{IV} = (Z'D)^{-1}(Z'Y)$; Z and D must have the same number of columns, or else the first half of $\hat{\beta}_{IV}$ won't be conformable with the second half. One option would be to simply pick one instrument and discard the rest, but this seems undesirable from an efficiency standpoint

[19] Also define the column vector D_1 such that D_1 contains only the treatment, d_i.

because you're throwing away valid information for estimating β. An attractive alternative then is to use 2SLS, which can trivially accommodate more than one instrument.

In the two stage procedure, simply include all instruments in the first stage when you predict the value of d_i. For example, if you have two instruments, z_{1i} and z_{2i}, estimate the first stage as:

$$d_i = \gamma_1 z_{1i} + \gamma_2 z_{2i} + x_i\gamma_3 + u_i$$

Then use $\hat{d}_i$ as the regressor in the second stage instead of d_i. In matrices, the formula remains the same: $\hat{\beta}_{2SLS} = (D'P_Z D)^{-1}(D'P_Z Y)$. Now Z contains more columns than D, but that doesn't affect the conformability of P_Z (which is an $N \times N$ matrix) with D. Under Gauss-Markov type assumptions, 2SLS efficiently combines all of the instruments to estimate β.

6.5 Applications

We now consider two important applications of instrumental variables. These applications are particularly helpful when studying IV in the context of heterogeneous treatment effects and the "weak instruments" issue (both of which we will cover).

6.5.1 Medical Trials

For a variety of reasons, medical trials are a fantastic example of an application of instrumental variables – I would argue the best, in fact. First of all, they are socially important (perhaps the most important application of IV to date). Furthermore, they are very clean in terms of experimental design, so they make a great teaching example for conveying the intuition behind what the IV estimator is doing. My personal recommendation would be to use this example whenever possible to guide you in understanding how IV operates.

The model for a medical trial is the same simple regression model that we are accustomed to: $y_i = \beta_0 + \beta_1 d_i + \varepsilon_i$. In this case, y_i represents a medical outcome, which could either be

a continuous variable such as blood pressure or cholesterol level or a discrete variable such as whether or not you survive (e.g., 1 if you survive, 0 if you do not). The variable d_i is generally a dummy variable that is 1 if you receive the treatment and 0 if you do not. It could alternatively be continuous (for example, it could be the dosage in milligrams of the drug that you receive), but in this example we will assume it is binary (you either take the pill or you do not take the pill). The error term ε_i represents all other factors that affect the health outcome. Note that the regression model corresponds to the potential outcomes model with constant treatment effects ($y = dy_1 + (1-d)y_0,\ y_0 = \beta_0 + \varepsilon,\ y_1 = y_0 + \beta_1$).

At this point I will switch to a specific example in order to make the discussion clearer. Let y_i be blood pressure, and let d_i represent a pill that is designed to treat high blood pressure, so $d_i = 1$ if individual i takes the pill and $d_i = 0$ if individual i does not take the pill. Our goal is to estimate the effect that the pill has on lowering blood pressure – our hope is that β_1 is large and negative. One way to estimate the effect is to start selling the drug to the general population and then collect some data and run a regression of blood pressure on whether or not you take the pill. However, this estimate will clearly suffer from a selection issue – people who take the pill are the ones who have high blood pressure to begin with! We will likely get a positive estimate of β_1 from this procedure, even if the true β_1 is large and negative. This may be true even after we condition on observable covariates using one of the selection on observables designs we discussed earlier. Therefore, in order to accurately estimate β_1, we design a medical trial in which we randomly assign some patients to the treatment group and assign other patients to the control group. The patients assigned to the treatment group are then given the pill and told to take it, while the patients assigned to the control group are given a placebo (or nothing at all).

Back in the old days (perhaps even older than me), people estimated the effect of the drug by simply subtracting the mean of y_i for the control group from the mean of y_i for the treatment group (in other words, regressing y_i on a variable that is 1 if you are in the treatment group and 0 if you are in the control group). This is what is known as an "intention to treat" analysis, because you are taking the difference between the group that

you intend to treat and the group that you do not intend to treat. But there was the problem of "non-compliance" – some people in the treatment group would fail to take the pill and others in the control group would obtain the pill from another source, even though they were not supposed to. This non-compliance can cause a bias in the estimate of β_1, and it was not immediately clear how to fix this bias until it became obvious that what we were looking at was actually a simple IV problem.

In this case, the instrument z_i is the intention to treat, i.e. $z_i = 1$ if you are assigned to treatment group (we intend to treat you), and $z_i = 0$ if you are assigned to the control group (we do not intend to treat you). It is easy to see that z_i satisfies the two properties of a good instrument. First of all, z_i is uncorrelated with ε_i by construction, because whether you are assigned to the treatment group or the control group is randomly determined, so $\text{Cov}(z_i, \varepsilon_i) = 0$. Second, z_i is correlated with d_i, because you are going to be more likely to take the pill if you are in the treatment group, so $\text{Cov}(z_i, d_i) \neq 0$. Therefore, z_i is a valid instrument for d_i, and the IV estimator gives us a consistent estimate of β_1, the effect of taking the pill on blood pressure.

How does this fix the non-compliance problem that we discussed before? To facilitate understanding, assume that the non-compliance problem only exists for the people in the treatment group. That is to say, assume that nobody in the control group takes the pill, but also assume that only half the people in the treatment group take the pill (i.e., half of the treatment group fails to comply and does not take the pill, while the other half takes the pill, as they were supposed to). What will the IV estimate look like?

The first stage will regress d_i on z_i, i.e. regress whether you took the pill on whether you were in the treatment group. So the first stage is:

$$d_i = \gamma_1 z_i + u_i$$

Since zero people in the control group took the pill while half the people in the treatment group took the pill, it should be intuitively clear that our estimate for γ_1 will be 0.5 (being

in the treatment group raises your probability of taking the pill by 50 percentage points, so $\hat{\gamma}_1 = 0.5$).

Now recall that the IV estimate is the reduced form rescaled by the first stage. In this case, the reduced form is a regression of y_i (your blood pressure) on z_i (whether you were assigned to the treatment or control group). So the reduced form is:

$$y_i = \pi_1 z_i + v_i$$

Therefore, our IV estimate is $\hat{\beta}_{1IV} = \hat{\pi}_1/\hat{\gamma}_1 = \hat{\pi}_1/0.5$. How is this fixing the non-complier problem? Well, we know that the reduced form estimates the causal effect of the instrument on y_i, so in our case the reduced form is estimating the effect that being assigned to the treatment group has on blood pressure. If there were a perfect correlation between being assigned to the treatment group and taking the pill (i.e. everyone in the treatment group took the pill, and nobody in the control group took the pill), then the reduced form estimate would be the effect of taking the pill on blood pressure. In that case the first stage would give us $\hat{\gamma}_1 = 1$, and the IV would be $\hat{\beta}_{1IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \hat{\pi}_1$. In other words, the IV would be the same as the reduced form (which is what we would expect, since both are supposed to be estimating the same thing in this case, i.e., the effect of the pill on blood pressure).

In our example, however, there is not a perfect correlation between being assigned to the treatment group and taking the pill, which is why our first stage estimate is $\hat{\gamma}_1 = 0.5$, not $\hat{\gamma}_1 = 1$. So in our case, the reduced form is estimating the effect on your blood pressure of increasing the probability that you take the pill by 50 percentage points. This means that the reduced form is not going to be estimating the full effect of taking the pill. Instead, it's estimating half of the effect of taking the pill. If it helps, imagine that there are 10 people in the treatment group, 5 of whom take the pill and 5 of whom do not, and 10 people in the control group, 0 of whom take the pill. The (expected) mean blood pressure for the treatment group will be $\frac{5 \cdot \beta_0 + 5 \cdot (\beta_0 + \beta_1)}{10} = \beta_0 + \frac{\beta_1}{2}$, while the (expected) mean blood pressure for the control group will just be β_0. So the reduced form coefficient, $\hat{\pi}_1$, will be the difference

of means between the treatment and control groups, or $\frac{\beta_1}{2}$. This is, of course, half the effect of taking the pill.

Therefore, the (plim of the) IV estimate will be $\beta_{1IV} = \frac{\pi_1}{\gamma_1} = \frac{0.5\beta_1}{0.5} = \beta_1$, which is exactly what we want. We can see that the IV estimate gives us a consistent estimate precisely because it is rescaling the reduced form by the first stage. In our example what this means in practice is that we are rescaling the reduced form to account for the fact that being in the treatment group only increases your probability of taking the pill by 50 percentage points, not by a full 100 percentage points. So the reduced form only represents half the effect of taking the pill, and it must be rescaled by (divided by) 0.5 in order to estimate the full effect of taking the pill.

More generally, what this example demonstrates is that IV functions by taking the estimated causal effect of z_i on y_i (the reduced form) and rescaling it by the estimated causal effect of z_i on d_i (the first stage).

Before we move on, I should note how IV is different than simply taking the mean of y_i for the people in the treatment group who took the pill and subtracting the mean of y_i for the people in the control group who did not take the pill (which, in our example, is the entire control group). The estimator I just described, which I will refer to as the naïve estimator, is affected by the same selection issues as a simple OLS regression of y_i on d_i. Specifically, it may be the case that the people in the treatment group who choose not to take the pill do so because their blood pressure was not very high to begin with. Thus the group of people that actually took the pill are the ones that all had high blood pressure to begin with, and we will tend to estimate that the pill does not have much of an effect (because its downward effect is being counteracted by the fact that the people who select to take it all had high blood pressure to begin with).

The IV estimator does not suffer from this selection problem because it does not release the people in the treatment group who choose not to take the pill. To understand this, imagine for the moment that there are two types of people in our sample: high blood

pressure types and low blood pressure types. Assume that they occur with equal frequency, so that when we randomly assign our sample to the treatment and control groups, half of the treatment group is high blood pressure, half of the treatment group is low blood pressure, half of the control group is high blood pressure, and half of the control group is low blood pressure. The half of the treatment group that takes the pill all have high blood pressure, so when we apply the naïve estimator and compare their average blood pressure to the average blood pressure of the control group, we underestimate the effect of the pill because we are comparing a group of high blood pressure people (who took the pill) to a group that is a 50/50 mix of high blood pressure and low blood pressure people (who did not take the pill). In contrast, what IV does is compare the mean of the treatment group (which is half high blood pressure people and half low blood pressure people) to the mean of the control group (which is half high blood pressure people and half low blood pressure people) in the reduced form. It then rescales this difference in means by the first stage to account for the fact that not all of the treated group took the pill. So unlike the naïve estimator, which deceptively compares a high blood pressure group to a half-high/half-low blood pressure group, IV compares two comparable groups, and that is why it gives us a consistent estimate of the effect of the pill.

6.5.2 Quarter of Birth

The quarter of birth application is perhaps the most-studied example of IV in the economics literature. This example is taken from Angrist and Krueger (1991). I will discuss the basic framework and idea; for more details see the article itself. The purpose is to demonstrate a nice application of IV/2SLS (which may help you think about what a good instrument looks like) and to familiarize you with the canonical example used in the "weak instruments" literature.

The question addressed with this instrument is a familiar one: what is the return to an additional year of schooling? One way to answer this question is to run a standard regression, $y_i = \beta_0 + \beta_1 d_i + x_i\beta_2 + \varepsilon_i$, where y_i is log wages, d_i is years of school, and x_i is a vector of covariates. However, as we know, this regression is likely to give us a biased estimate of

β_1 for a variety of reasons, including selection bias and measurement error. The problem is that $\text{Cov}(d_i, \varepsilon_i) \neq 0$; one way to address this problem is to find an instrument z_i that is correlated with d_i (schooling) but uncorrelated with ε_i.

Angrist and Krueger suggest quarter of birth as the instrument. Why use this as an instrument for schooling? The idea is that states have mandatory schooling laws stipulating that students must stay in school until a given age (say age 16, for simplicity). However, the key thing is that these laws dictate the *age* at which a student may leave school, not how many *years of schooling* a student must get. Therefore, if a student starts school at age 6, she will be legally required to receive 10 years of schooling. However, if she starts school at age 5, she will be legally required to receive 11 years of schooling. Thus, variations in the age at which a student starts school will result in variations in the amount of schooling that student is legally required to receive. While this will not make a difference for most people (because most people do not drop out of high school as soon as they are no longer required to be there), it will make a difference for some people, so there should be a nonzero correlation (albeit a modest one) between the age one starts school and how many years of schooling one receives.

How does this all pertain to quarter of birth? The quarter in which a student is born can have a large effect on what age the student starts school because the academic calendar begins in September regardless of quarter of birth. Many states require children to start school in the calendar year in which they turn 6. So, for example, a child born in December (fourth quarter) might start school at age 5.7, and thus be required by law to receive a minimum of 10.3 years of schooling (16 minus 5.7). However, a child born in January (first quarter) might start school at age 6.7, and thus be required by law to receive a minimum of only 9.3 years of schooling (16 minus 6.7). Quarter of birth is thereby correlated with legally required years of schooling, and thus quarter of birth is also correlated with actual years of schooling. Quarter of birth therefore satisfies the first property of a good instrument, $\text{Cov}(d_i, \varepsilon_i) \neq 0$.

Does quarter of birth satisfy the second property of a good instrument, i.e. $\text{Cov}(z_i, \varepsilon_i) =$

0? Potentially, yes (though it turns out not). It seems plausible that the quarter in which one is born might not causally affect one's future wages, except through its effect on schooling (though there could be some strange weather effect on young babies). There is also no obvious reason to think that quarter of birth should be spuriously correlated with anything that affects future wages, particularly if we think that the time of conception is determined in a random manner. It is therefore plausible that quarter of birth and future wages are uncorrelated (except through changes in schooling).

How would we implement the quarter of birth instrument in practice? We would probably use three instruments: a dummy for the first quarter (z_1), a dummy for the second quarter (z_2), and a dummy for the third quarter (z_3) (we exclude the fourth quarter to avoid the dummy variable trap, i.e. to avoid perfect colinearity with the constant term). So the first stage would be to regress schooling on quarter of birth (assuming there are no additional covariates that we are including):

$$d_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + u_i$$

Then take the predicted $\hat{d}_i$ from the first stage and use them in the second stage to run the regression:

$$y_i = \beta_0 + \beta_1 \hat{d}_i + u_i$$

The value of $\hat{\beta}_1$ from this regression is our estimate of the effect of schooling on wages. If the two IV assumptions are true ($\text{Cov}(d_i, \varepsilon_i) \neq 0$ and $\text{Cov}(z_i, \varepsilon_i) = 0$), then this will be a consistent estimate of the effect of schooling on wages.

There are a couple of things to note in this application. First, Angrist and Krueger implement IV in a couple of different ways. They begin with a Wald estimator which compares only two groups, people born in the first quarter and people born in the second through fourth quarters. The Wald estimator divides the difference in mean earnings for the two groups by the difference in mean schooling. Given our previous discussion of IV, it

should be clear that this procedure is equivalent to doing IV with a binary instrument and no covariates. With this estimator, Angrist and Krueger estimate the return to schooling to be around 0.10 (i.e., one additional year of schooling raises wages by 10 percent) in the 1980 Census. This is higher than the OLS estimate from the same sample, which they find to be around 0.07. However, Angrist and Krueger also implement a 2SLS procedure in which they use dozens of instruments. They produce these instruments by interacting quarter of birth with year of birth (since the effect of quarter of birth on schooling might vary across years). In their 2SLS regressions, they frequently find coefficients closer to the OLS estimate of 0.07 than to the Wald estimate of 0.10. Unbeknown to them, the culprit behind this pattern is the "weak instruments" problem, which we will discuss in a subsequent section.

Second, while the quarter of birth instrument is much better than most instruments you will come across (at least in terms of satisfying the exclusion restriction), it is still not impervious to criticism. For example, many babies are conceived shortly after people get married. Some couples are likely to wait until the summer to get married, while other couples are more likely to get married quickly or when it is most convenient. Therefore, couples of the first type would be more likely to have children in the first or second quarter, whereas couples of the latter type would be equally likely to have children in any quarter. If couples of the first type are different in some important way (e.g., perhaps they have higher income on average) than couples of the second type, then that could introduce a correlation between quarter of birth and future wages. Any nonzero correlation between z_i and ε_i would be particularly problematic in this case because the first stage is relatively weak (again, we will discuss this issue in a subsequent section).

6.6 Heterogenous Treatment Effects: AIR (1996) and LATE

All of the discussion above concentrates on IV in the context of homogeneous treatment effects. This was the focus of IV estimation for the first 50 years, but it doesn't fit in with our discussion of heterogeneous treatment effects at the beginning of the course. Recall the distinction between ATE – the average treatment effect for a randomly drawn individual in

our sample – and TOT – the average treatment effect for a randomly drawn treated individual in our sample. With IV, these distinctions become more interesting. We have sidestepped this discussion so far by assuming homogeneous treatment effects, so ATE is equal to TOT, and both are equal to the average treatment effect for any other sub-population one might think of. If we allow for heterogeneous treatment effects, however, what is it that IV actually estimates? ATE? TOT? The answer, presented in Angrist, Imbens, and Rubin's seminal 1996 paper (henceforth AIR 1996), is "neither."

6.6.1 Intuition

I focus on explaining intuitively what IV estimates in the context of heterogeneous treatment effects. I also have notes available presenting the mathematical proof, which I am happy to share if there is interest. However, I believe that understanding the terms and concepts is more important than seeing the proof. What IV generally estimates is the "local average treatment effect," or LATE. LATE is the average treatment effect of d_i on y_i for the units for whom changing the instrument (changing z_i) changes their treatment status (changes d_i). This is somewhat abstract, but it should become clearer in the context of our three examples, the medical trial, the draft lottery, and the quarter of birth instrument.

What does it mean to say that IV estimates the average treatment effect of d_i on y_i for the units for whom changing z_i changes d_i? In practice, this is best illustrated in the medical trial example. In this example, there are four potential types of people. Note that not all of these types need exist in practice; in fact, we will explicitly rule out one type by assumption when we do the proof. The first type are people who always take the pill, regardless of whether they are assigned to the treatment group or the control group.[20] In the language of AIR 1996, we call these people "always-takers." The second type are people who never take the pill, regardless of whether they are assigned to the treatment group or the control group. We call these people "never-takers." The third type are people that take the pill if and only if

[20] You might wonder how the control group could get the pill. Think about terms like "black market" or "prescription abuse."

they are assigned to the treatment group. We call these people "LATE-compliers." Finally, the fourth type are people who take the pill if and only if they are in the control group. We call this perverse group the "LATE-defiers," and we rule them out by assumption.

The people "for whom changing z_i changes their value of d_i" are the people who take the pill if and only if they are in the treatment group, i.e. the LATE-compliers (recall that assignment to treatment versus control group is the instrument in this example). The always-takers are unaffected by the instrument, because they take the treatment regardless of whether they are in the treatment or control group. Likewise, the never-takers are also unaffected by the instrument, because they eschew the treatment regardless of whether they are in the treatment or control group. The defiers are ruled out by assumption. Therefore, the IV estimator estimates the effect of the pill on blood pressure for the people who take the pill if they are in the treatment group but do not take it if they are in the control group. If the effect is homogeneous, then this distinction is irrelevant, but if the effect varies across individuals, then this distinction can become important.

Suppose that there are two types of people: people who respond to the pill and people who do not respond to the pill. This is not a far-fetched assumption – most medical trials find that the treatment is successful in treating some cases, but unsuccessful in treating other cases. So β_1 is negative for people who respond to the pill (remember that we think the pill should lower blood pressure), and β_1 is zero for people who do not respond to the pill. Further suppose that people who respond to the pill know that they will respond to it (don't ask me how), so they always take it, regardless of whether they are in the treatment or the control group. However, the people for whom the treatment has no effect take the pill only if they are in treatment group (when they are given the pill for free), and not if they are in the control group. We know that IV estimates the effect of the treatment on the LATE-compliers, i.e. the people that take it if and only if they are in the treatment group. Therefore, in this case, IV estimates the effect of the treatment on the people for whom the treatment has no effect, because they are the only ones for whom the instrument changes whether or not they take the pill. So IV will estimate $\beta_1 = 0$ in this example, despite the

fact that the average treatment effect is negative.[21]

Does this mean that IV is inconsistent? Not really – it is simply providing a consistent estimate of the local average treatment effect (the average effect for the people for whom changing the instrument changed d_i), not the average treatment effect for the entire population or sample. As long as you interpret IV correctly, then it is not inconsistent. Of course, it may not estimate what you want to estimate (which might be ATE or TOT), but that's the way the cookie crumbles. So the lesson here is that IV is consistent, but that you have to be careful in thinking about exactly what it is estimating. Importantly, IV estimates the average treatment effect for individuals that "comply" with the instrument. Since different instruments will have different sets of compliers, it follows that different instruments can plim to different values, even if all the instruments under consideration meet the two criteria for valid instruments. This result basically invalidates overidentification tests as a valid scientific testing procedure and has implications for instrumenting for multiple endogenous variables simultaneously.

Why does IV estimate LATE in our example? As I have reiterated many times, the IV estimator is the reduced form divided by the first stage. So if the IV estimate is 0 in the example I discussed above, that means that the reduced form must be 0. In the medical trial example, the reduced form is the mean blood pressure for the treatment group minus the mean blood pressure for the control group. Since the always-takers take the pill when they are in the treatment group and when they are in the control group, their mean blood pressure will not be any different when they are in the treatment group than it is when they are in the control group. So those people will never contribute anything to moving the reduced form away from zero. The people who can potentially move the reduced form away from zero are the people who take the treatment when they're in the treatment group but do not take it when they are in the control group. But we assumed that those were the people for whom the pill had no effect, so of course their mean blood pressure in the treatment

[21] Of course, we could alternatively construct a scenario in which the individuals with no treatment effect are the never-takers and the individuals with a negative treatment effect are the LATE-compliers. In that scenario, IV would produce a negative estimate of β_1, but the magnitude would be larger than ATE.

group is not any different than their mean blood pressure in the control group. Thus we get a reduced form of 0 in our example.

If the pill did have an effect for these people, then the reduced form would be capturing that effect, and we would get a nonzero coefficient estimate. That coefficient would represent the total effect of the pill averaged over all of the individuals in the treatment group. In fact, however, only the LATE-compliers were affected. The IV thus rescales the reduced form by the first stage because the first stage estimates, in our example, the fraction of the sample that are LATE-compliers (i.e., the fraction that changed their value of d_i in response to being assigned to the treatment group).

To reiterate, the always-takers and the never-takers do not, in expectation, contribute anything to moving the reduced form away from zero, because for them the treatment indicator is always the same in the treatment group and the control group (and the random assignment procedure balances them, on average, across treatment and control). Thus their mean blood pressure is no different in the treatment group than it is in the control group. Therefore, the only group of people who can move the reduced form away from zero is the group of LATE-compliers, because for them the treatment level actually varies depending on whether they are in the treatment group or in the control group. So if the treatment has an effect for them, then their mean blood pressure will be different in the treatment group than it is in the control group. But by definition, the LATE-compliers are the people for whom changing z_i changes d_i. Thus IV estimates the average treatment effect for the people for whom changing z_i changes d_i, because those are the people who drive the reduced form, and IV is just the reduced form rescaled by the first stage.

Now consider how LATE applies in the Vietnam draft lottery example. Compliers are those who are induced to serve due to becoming draft eligible. It is perhaps easiest to consider the groups that are *not* compliers. First there are volunteers. These are "always takers" in the sense that they will serve regardless of whether they are chosen to be draft eligible. Then there are men who are draft eligible but are not called up because manpower requirements are met before they are needed. These are "never takers" in the sense that they do not serve

regardless of draft eligibility. Finally consider those who enlist in the National Guard, flee the country, or go to college to avoid service. These are also never takers, but they may represent a violation of the exclusion restriction (assuming that participating in the National Guard, fleeing the country, or going to college affects earnings). LATE compliers are those who don't volunteer, who do get called up, and who don't avoid service when being called up. Apparently this group represented only about 10 to 16 percent of young males during the Vietnam era (i.e., the size of Angrist's first stage estimates).

Finally consider how LATE applies to the quarter of birth example. Recall that in the quarter of birth example, the instrument works because some people stay in school only as long as they are legally required to, and then they drop out as soon as they reach age 16. These are the people for whom the instrument z_i (quarter of birth) has an effect on d_i (years of school). If it helps, you could literally imagine a 15.5 year old potential dropout who was born in the third quarter thinking to himself, "If only I had been born in the first quarter, then I would be able to drop out of school right now, because I'd already be 16. But instead I have to stay in school until the third quarter and receive 11 years of schooling instead of 10.5 years of schooling!" These people are the equivalent of the LATE-compliers (they don't have to actually think in this manner though!). In contrast, however, for the vast majority of people the instrument (quarter of birth) has no effect on how long they stay in school, because they plan to stay in school long past the age at which they can legally dropout. They are the equivalent of the always-takers.[22]

Since IV estimates the causal effect of d_i (schooling) on y_i (wages) for the people for whom the instrument z_i (quarter of birth) changes their value of d_i, the IV estimate gives us the average effect of schooling on wages for people who drop out as soon as they are no longer legally required to stay in school. So the quarter of birth instrument is really estimating the average effect of an additional year of schooling on wages for high school dropouts. Is there any reason to believe that this is the same effect of schooling that the "average" person would have? Probably not. On the one hand, it may overestimate the "average" effect of schooling

[22]The never-takers would be the ones that disregard the law entirely and drop out of school long before they are legally allowed to.

if we believe that wages are a concave function of schooling, so that the return to schooling falls as you get more schooling.[23] On the other hand, it may underestimate the "average" effect of schooling if we believe that high school dropouts don't apply themselves in school anyway, so they don't get much out of being in school. Either way, the point is that the IV regression is estimating the average effect of schooling on wages for high school dropouts rather than for the entire population. It consistently estimates this effect, but this effect is probably different than the population average effect of schooling on wages. Thus we need to be careful about how we interpret the result. Finally, note that the reason IV estimates the average effect of schooling on wages for high school dropouts is not because our sample only consists of high school dropouts. The sample is taken from the entire population, but IV only estimates the average effect of d_i on y_i for the LATE-compliers (i.e., the high school dropouts), not the average effect for the entire population. However, if our policy interest pertains to students at risk of dropping out, the average effect for LATE-compliers may be very informative.

6.6.2 Discussion

We have seen in this section that IV estimates the "local average treatment effect," or LATE. This is the average treatment effect for units that are induced by the instrument to change their treatment status. The clear application of this finding is that it allows us to think more precisely about which group of individuals our treatment effect estimate applies to. There are, however, other important implications.

Most importantly, the LATE result implies that, in the presence of treatment effect heterogeneity, different instruments should produce different estimates, even in arbitrarily large samples. The choice of instrument defines the group of LATE-compliers; different instruments therefore estimate the average treatment effect for different groups of LATE-

[23] This phenomenon has been referred to as "discount rate bias" because a simple human capital model implies that an individual should stay in school until her return to schooling equals her discount rate. Students that drop out early do so because they have higher discount rates, and their marginal return to schooling is higher. However, the term "discount rate bias" is somewhat deceptive in the sense that it's not really an issue of bias but rather an issue of heterogeneous treatment effects and external validity.

compliers. There is no reason why these averages need be equal for different groups.

The fact that different instruments can produce different treatment effect estimates (even absent sampling error) calls into question the general utility of overidentification tests. These tests compare coefficient estimates produced by different instruments – the idea is that if the instruments are all valid, all the estimates should be equal (up to sampling error). If some instruments are invalid, however, the estimates produced by different instruments may differ. In the context of heterogeneous treatment effects, however, we know that different instruments can produce different coefficient estimates even if all of the instruments are internally valid. Thus it is impossible to ever "reject" the validity of the instruments, making the overidentification tests scientifically questionable. The same critique holds for the Hausman test, which compares the IV estimate to the OLS estimate. With heterogeneous treatment effects, there is no reason that OLS (which, under ideal conditions, will estimate ATE or TOT) need equal IV (which estimates LATE).

Heterogeneous treatment effects also complicate matters when you have multiple endogenous variables that you want to instrument for. Consider, for example, a simple case in which you wish to simultaneously estimate the effect of education (d_1) and experience (d_2) on earnings (y). The model might look like:

$$y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \varepsilon_i$$

Both "treatments" are subject to selection issues and are endogenously determined. Instrumenting for education and controlling for experience as a covariate will not give consistent estimates of the effect of education on earnings – it is inappropriate to control for a variable that is affected by the treatment (in general, getting more education will mean getting less job experience). The correct way to estimate the causal effect of education on earnings is to instrument for education and include as covariates only predetermined variables.

If, however, you want to estimate a structural model that contains both education and experience, i.e., you want to know the effect of education when holding experience constant

(even though we may not be able to imagine such a scenario in real life), you might find two instruments, one for education (call it z_1) and one for experience (call it z_2).[24] You can then identify β_1 and β_2 by running 2SLS, using both z_1 and z_2 as instruments. Intuitively, 2SLS is using z_2 to estimate β_2, and then using this estimate of β_2 to adjust for the fact that z_1 affects both d_1 and d_2 when estimating β_1 (i.e., the effect of education on earnings holding experience constant).

With homogenous treatment effects, this strategy is valid. With heterogenous treatment effects, however, we know that different instruments generally estimate different local average treatment effects. Assuming that z_1 estimates the same treatment effect for d_2 that z_2 estimates is therefore unjustified.[25] In principle, the effect of manipulating education while holding experience constant could be positive for all individuals, yet the 2SLS procedure could generate a negative estimate of β_1 (even ignoring sampling error).

6.7 Weak Instruments

Weak instruments – that is to say, instruments that are only weakly correlated with the treatment of interest – pose a special set of problems. First, and most importantly, a weak first stage implies that any bias in the reduced form will be amplified in the IV estimate. This is true regardless of the number of instruments one uses. When using many weak instruments, however, a finite sample issue arises and 2SLS becomes biased towards the OLS estimate (conventional standard errors are also inaccurate). Though these issues have been known to some degree for several decades, they were brought to the attention of applied researchers by Bound, Jaeger, and Baker (1995) (henceforth BJB 1995). I focus on the first issue – that any bias in the reduced form may be amplified in the IV estimate – here because I believe it is the more important of the two.

[24] Of course, the education instrument will invariably affect experience. In principle, however, the experience instrument need not affect education.

[25] This is equivalent to assuming that z_1 and z_2 should both produce identical estimates of β_2.

6.7.1 Omitted Variables Bias

Consider a case with a single endogenous variable, d_i, one or more instruments, z_i, and no covariates.[26] We are interested in the causal relationship between d_i and y_i, summarized as

$$y_i = \alpha + \beta d_i + \varepsilon_i$$

We have an instrument z_i that we use to predict d_i

$$d_i = z_i\gamma + u_i$$

Consider the consistency of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$. For OLS,

$$\text{plim } \hat{\beta}_{OLS} = \frac{\text{Cov}(d_i, y_i)}{\text{Var}(d_i)} = \frac{\text{Cov}(d_i, \beta d_i + \varepsilon_i)}{\text{Var}(d_i)} = \beta + \frac{\sigma_{d\varepsilon}}{\sigma_{dd}}$$

The plim for 2SLS relies on the fact that $\hat{d}_i$ plims to $z_i\gamma$,

$$\text{plim } \hat{\beta}_{2SLS} = \frac{\text{Cov}(\hat{d}_i, y_i)}{\text{Var}(\hat{d}_i)} = \frac{\text{Cov}(\hat{d}_i, \beta d_i + \varepsilon_i)}{\text{Var}(\hat{d}_i)} = \frac{\text{Cov}(z_i\gamma, \beta(z_i\gamma + u_i) + \varepsilon_i)}{\text{Var}(\hat{d}_i)} = \beta + \frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{\hat{d}\hat{d}}}$$

If there is zero covariance between d and ε then OLS will consistently estimate β. If there is zero covariance between z and ε then 2SLS will consistently estimate β (note that 2SLS is never unbiased because it is a ratio of two random variables). What happens when these covariances are nonzero, however? Under what conditions will one estimator be more or less inconsistent than the other?

The ratio of the inconsistency in the IV estimator to the inconsistency in the OLS estimator is:

$$\frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}} \cdot \frac{\sigma_{dd}}{\sigma_{\hat{d}\hat{d}}} = \frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}} \cdot \frac{1}{R^2_{FS}}$$

[26] As per BJB 1995, the core results remained unchanged by the addition of covariates.

R^2_{FS} is the R^2 of the first stage; the equality holds because $R^2_{FS} = SSR/SST = \sigma_{\hat{d}\hat{d}}/\sigma_{dd}$. If we had covariates in the model, the R^2_{FS} term would be the partial R^2 from the first stage, i.e., the R^2 from running d_i on z_i after the covariates have been partialled out from both.[27]

From the result above, we see that the relative inconsistency of IV vis a vis OLS depends on two quantities. First, it depends on the covariance of $\hat{d}_i$ and ε_i relative to the covariance of d_i and ε_i. If the covariance of the error term and $\hat{d}_i$ increases (relative to the covariance of the error term and d_i), then the inconsistency of IV increases – this is quite intuitive. More interestingly, the relative inconsistency of IV also depends on the inverse of the R^2 (or partial R^2, if you have covariates) of the first stage. Thus, if the first stage is weak (i.e., low R^2), any violation of the exclusion restriction will be amplified, and IV can become very inconsistent. A first stage (partial) R^2 of 0.1, for example, will inflate the ratio $\frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}}$ by a factor of 10. Except that things aren't quite that simple.

The complication is that $\sigma_{\hat{d}\varepsilon}$ is itself affected by the strength of the first stage. If the first stage is weak, then by definition the variance of $\hat{d}$ will be relatively low, and so the covariance $\sigma_{\hat{d}\varepsilon}$ will tend to be low as well. For tractability, and because it covers the preponderance of meaningful cases, suppose that z_i contains only one instrument. In that case:

$$\frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}} \cdot \frac{1}{R^2_{FS}} = \frac{\text{Cov}(\gamma z, \varepsilon)}{\text{Cov}(d, \varepsilon)} \cdot \frac{\text{Var}(d)}{\text{Var}(\gamma z)} = \frac{\sigma_{z\varepsilon}/\sigma_z\sigma_\varepsilon}{\sigma_{d\varepsilon}/\sigma_d\sigma_\varepsilon} \cdot \frac{\sigma_d}{\sigma_{\gamma z}} = \frac{\rho_{z\varepsilon}}{\rho_{d\varepsilon}} \cdot \frac{1}{R_{FS}}$$

The last expression is more useful in the sense that it is expressed in terms that do not depend on the units of measurement for any of the variables in question. The second term, $\frac{1}{R_{FS}}$, confirms that a weak first stage does exacerbate the relative inconsistency of IV vis a vis OLS, but the degree of bias is not as strong as originally implied. With a first stage (partial) R^2 of 0.1, for example, IV will be less inconsistent than OLS as long as the correlation between the instrument, z_i, and the error term, ε_i, is approximately three times less than the correlation between d_i and ε_i. With a first stage (partial) R^2 of 0.01, however, the correlation between z_i and ε_i needs to be ten times less than the correlation between d_i

[27] With covariates in the model, the $\sigma_{\hat{d}\varepsilon}$ and $\sigma_{d\varepsilon}$ terms are also calculated after the covariates have been partialled out from d_i and z_i.

and ε_i in order for IV to be preferable to OLS.

So, if the first stage is relatively weak, then you should think carefully about whether your exclusion restriction ($\text{Cov}(z, \varepsilon) = 0$) holds. Even a modest correlation between the instrument and the structural error term can make IV highly inconsistent if the first stage (partial) R^2 is low. This is true regardless of whether you have one instrument or many instruments.

BJB 1995 analyze the potential for omitted variables bias in Angrist and Krueger (1991) using the just-identified case. Quarter of birth is parameterized as a single indicator variable that equals zero if an individual is born in the first quarter and unity if an individual is born in the second through fourth quarters. With this parameterization, Angrist and Krueger report a first stage coefficient of 0.1 – people born in the first quarter have 0.1 years less education than those born in the second through fourth quarters. This is a fairly small effect, but the coefficient is highly significant since the sample numbers in the hundreds of thousands.

BJB note that the difference in mean log per capita family income for young children born in the second through fourth quarters versus those born in the first quarter is 0.024 – families of children born in the first quarter have per capita income that is about 2.4% lower than families of children born in the second through fourth quarters. Using an intergenerational correlation coefficient of 0.4 (the standard in the literature at that time – now it is estimated to be even higher), BJB infer that omitted factors might lead to a difference in mean log income of 0.01 between individuals born in the second through fourth quarters and individuals born in the first quarter. Though this differential is quite small, it is important to remember that the first stage is also very small, with a coefficient of 0.1. Thus the bias in the IV estimate will be 10 times the bias in the reduced form estimate – a reduced form bias of 0.01 translates to an IV bias of 0.10. Interestingly, this is very close to the return to education that Angrist and Krueger estimate using the quarter of birth instrument. I am not claiming that their estimate is necessarily wrong, but the relatively weak first stage does mean that the quarter of birth design is not quite as clean as it first appears.

7 Additional References

Arceneaux, Kevin, Alan Gerber, and Donald Green. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis*, 2006, 14, 3762.

Angrist, Joshua, Guido Imbens, and Donald Rubin. "Identication of Causal E?ects Using Instrumental Variables." *Journal of the American Statistical Association*, 1996, 91, 444455.

Angrist, Joshua and Alan Krueger. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 1991, 106, 9791014.

Card, David. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 1990, 43, 245257.

Dehejia, Rajeev and Sadek Wahba. "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, 94, 1999, 10531062.

Freedman, David. "Statistical Models and Shoe Leather." *Sociological Methodology*, 1991, 21, 291313.

Geiser, Saul and Maria Veronica Santelices. "Validity Of High-School Grades In Predicting Student Success Beyond The Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes." *Center for Studies in Higher Education Research and Occasional Paper Series CSHE.6.07*, 2007.

Heckman, James. "Dummy Endogenous Variables in a Simultaneous Equations System." *Econometrica*, 1978, 46, 931-59.

Kellogg, Ryan and Hendrik Wolff. "Daylight Time and Energy: Evidence from an Australian Experiment." *Journal of Environmental Economics and Management*, 2008, 56, 207220.

Rubin, Donald. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 1974, 66, 688-701.

Shadish, William, M. H. Clark, and Peter Steiner. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments." *Journal of the American Statistical Association*, 2008, 103, 13341356.