University of Illinois Department of Economics
Spring 2017 Econ 574 Roger Koenker

Lecture 21
"Treating the Treated: Varieties of Causal Inference"

The last decade has seen a vigorous revival of interest in models of causal inference in econometrics and statistics more generally. In this lecture, I would like offer a few comments on this literature that try to draw out some themes from earlier aspects of the course. In this survey we will meet again several old friends: instrumental variables, quantile regression, and perhaps most fundamentally – identification.

1. A Prototype: The elementary supply and demand model

   Heckman (2000), in a valuable retrospective on causal analysis in econometrics, characterizes economists as "people of the model". We rely on models as coherent constructs within which we can evaluate the consequences of various "thought experiments." Econometrics seeks to quantify the qualitative conclusions of the model.

   The canonical example is evitably the elementary supply and demand model,

$$\begin{aligned} Q^D &= Q^D(P^D, Z^D) \\ Q^S &= Q^S(P^S, Z^S) \end{aligned} \tag{1}$$

   where Z^D and Z^S denote other variables that shift demand and supply, respectively. From a statistical standpoint you can think of the two equations as representing the conditional expectations of Q^D and Q^S. If Z^D and Z^S contain the same variables and we impose the equilibrium conditions

$$Q^D = Q^S \text{ and } P^D = P^S \tag{2}$$

   then we have no way of recovering the *structural* relationships (1). If, however, we have independent variation in Z^D that doesn't appear in Z^S, say a variable Z_i^D, then even in the equilibrium setting of the model we *can* recover some information about marginal structural effects. To see this, write,

$$\frac{\partial Q^S}{\partial Z_i^D} = \frac{\partial Q^S}{\partial P^S} \; \frac{\partial P^S}{\partial Z_i^D}$$

   and using (2), and provided $\partial P^D/\partial Z_i^D = \partial P/\partial X_i^D \neq 0$, we have

$$\frac{\partial Q^S}{\partial P^S} = \frac{\partial Q/\partial Z_i^D}{\partial P/\partial Z_i^D} \tag{3}$$

   Similarly, if we have Z_j^S appearing independently in the supply equation we obtain marginal structure effects on demand:

$$\frac{\partial Q^D}{\partial P^D} = \frac{\partial Q/\partial Z_j^S}{\partial P/\partial Z_j^S} \tag{4}$$

Note that we can view the foregoing as motivating the IV estimator of these effects. In the linear setting the natural estimator of the ratio on the *rhs* of (3) and (4) would be simply the ratio of the OLS estimators of these derivatives from the "reduced form" model,

$$\begin{aligned} P &= P^R(Z^D, Z^S) \\ Q &= Q^R(Z^D, Z^S) \end{aligned} \tag{5}$$

Obviously, if models are nonlinear, then we need to be more careful about local interpretations of these derivatives and their corresponding estimates.

2. IV for Errors in Variables.

It is valuable to return to the misty early history of instrumental variables to explore the original motivation of the technique. The first paper seems to be Wald (1940), which dealt with errors in variables models. This estimator has received considerable recent attention in the literature, so it is particularly useful to consider it.

The simplest version of the Wald estimator is the bivariate errors in variables model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i \\ \tilde{x}_i &= x_i + v_i \end{aligned}$$

The model hypothesizes an unobservable covariate x_i for which we have a "proxy" observable variable $\tilde{x}_i$. Least-squares estimation of $\beta = (\beta_0, \beta_1)$ is unsatisfactory. Applying least squares to the model,

$$y_i = \beta + \beta_1 \tilde{x}_i + e_i$$

where

$$e_i = u_i - \beta_1 v_i$$

clearly leads to inconsistent estimates of the parameters since the condition $e_i \perp\!\!\!\perp x_i$ fails. Thus, the condition expectation of y_i given x_i is different than $E(y_i|\tilde{x}_i)$. Wald's suggestion was to define the instrumental variable,

$$z_i = \text{ sgn } (\tilde{x}_i - \hat{\mu})$$

and using $\hat{\mu} = \text{median}_i\{\tilde{x}_i\}$ compute,

$$\hat{\beta}_{IV} = \frac{\sum y_i z_i}{\sum \tilde{x}_i z_i}$$

It is helpful to view this geometrically. We split the sample into two halves, denoting the right and left halves by R, L and then we can write,

$$\hat{\beta}_{IV} = \frac{\bar{y}_R - \bar{y}_L}{\bar{x}_R - \bar{x}_L}$$

Having averaged over both $\tilde{x}$ and y we get consistency. Of course, if $\tilde{x}_i = x_i$, so there is no errors in variables problem we sacrifice efficiency relative to OLS.

Various other similar proposals were made to improve the efficiency problem. Bartlett (1949) suggested dividing the $\tilde{x}$ axis into 3 parts and ignoring the middle piece in computing the Wald estimator. Durbin (1954) suggested using the ranks of the $\tilde{x}_i$'s as the IV, so

$$\hat{\beta}_{IV} = \frac{\sum i y_{(i)}}{\sum i \tilde{x}_{(i)}}$$

where $\tilde{x}_{(i)}$ are the order statistics of the $\tilde{x}_{(i)}$'s and $y_{(i)}$ are ordered similarly.

Wald's approach was criticized by Neyman and Scott (1951) and others. Fuller(1987) argues that the crucial condition for the validity of the Wald estimator – that the variable z_i indicating the classification of observations into the two groups is independent of the u_i's – is rather implausible. To the extent that the errors of observation lead to misclassification across the median x boundary we will still have bias. A nice paper by Pakes (1982) shows that in the strictly Gaussian case there is no improvement in bias from the Wald procedure.

An interesting exercise is to compute asymptotic relative efficiencies for these alternatives for various distributions of (x_i, y_i).

An interesting connection, at least to me, to quantile regression can be made. Hogg (1975) proposed a method of quantile regression analogous to Wald's IV estimator. The technique was graphical, but can be interpreted formally in light of Wald. Let $\psi_\tau(u) = \tau - I(u < 0)$ and consider $\hat{\beta}_{IV}(\tau)$ "solving",

$$\sum z_i \psi(y_i - \beta_0 - \beta_1 \tilde{x}_i) = 0$$

where $z_i = \text{ sgn } (\tilde{x}_i - \text{ med } (\tilde{x}_i))$, so we have the pair of equations

$$\begin{aligned} \sum \psi(u_i) &= 0 \\ \sum z_i \psi(u_i) &= 0 \end{aligned}$$

and everything is determined by the counts of residuals in the four quadrants
$\tilde{x}_i \lesseqgtr \text{ med}\{\tilde{x}_i\}$ and $\hat{u}_i \lesseqgtr 0$.

Let $P = \#\{\hat{u}_i > 0\}, N = \#\{\hat{u}_i < 0\}$ and $P^+ = \#\{\hat{u}_i > 0, \tilde{x}_i > \text{ med } \{\tilde{x}_i\}\}$, etc.

Rewriting our pair of equations we have

$$\tau P = (1 - \tau) N \tag{1}$$

$$\tau(P^+ - P^-) = (1 - \tau)(N^+ - N^-) \tag{2}$$

If we now add these two equations together, and then subtract them, canceling the resulting twos on both sides we obtain the pair of equations,

$$\tau P^+ = (1 - \tau) N^+$$

and

$$\tau P^- = (1 - \tau) N^-$$

> and we have shown that Hogg's proposal is just the Wald IV version of quantile regression.
>
> Clearly, this can be extended to higher dimensions, but there are some real questions about how the z_i should be defined in the multivariate case.

Treatment Effects

Undoubtedly the most basic statistical question is: do two populations differ or are they really the same? This requires some evaluation of how to measure the difference between the two populations. The classical setting for these questions is the two sample treatment control problem. We have a treatment population and a control population, and suppose for the moment that subjects are randomly assigned to the two populations. And suppose that we have a continuously measured response variable. A very general notion of the treatment effect has been proposed by Lehmann(1974):

> "Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment responses is that of the random variable $X + \Delta(X)$ where X is distributed according to F."

Doksum (1974) defines $\Delta(x)$ as the "horizontal distance" between F and G at x, *i.e.*

$$F(x) = G(x + \Delta(x)).$$

Then $\Delta(x)$ is uniquely defined as

$$\Delta(x) = G^{-1}(F(x)) - x.$$

This is exactly what is plotted in the conventional statistical QQ plot. Changing variables so $\tau = F(x)$ we have the quantile treatment effect (QTE):

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$

An example is illustrated in Figure 1 where we illustrate a standard normal control distribution for F and a shifted G which is $\mathcal{N}(1, 4)$. The dotted lines illustrate the mean and median treatment effect, which are identical in this case, but also clearly shows that the QTE at other quantiles can be very different from the mean treatment effect.

In very simple special cases we have the location shift model in which,

$$\delta(\tau) = \delta_0,$$

a constant, or the scale shift model,

$$\delta(\tau) = \delta_1 F^{-1}(\tau)$$

or the location-scale shift model,

$$\delta(\tau) = \delta_0 + \delta_1 F^{-1}(\tau)$$

These cases are illustrated in Figure 2, where we plot distribution functions, densities, and the QTE's for examples of all three models.

The Lehmann QTE is naturally estimable by

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau)$$

where $\hat{G}_n$ and $\hat{F}_m$ denote the empirical distribution functions of the treatment and control observations, Consider the quantile regression model

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i$$

where D_i denotes the treatment indicator, and $Y_i = h(T_i)$, *e.g.* $Y_i = \log T_i$, which can be estimated by solving,

$$\min \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \delta D_i)$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$.

An important aspect of this simple version of the treatment control setting is that we cannot identify anything about the joint distribution of the pair of random variables that would describe the response of subjects under both the treatment and control. Since we never observe subjects in both states it is hard to see how we can be expected to learn anything about this joint distribution. We see the two marginals but the so-called copula function,

$$\varphi(u, v) = H(F^{-1}(u), G^{-1}(v))$$

is *not* identified by the marginal distributions of the control, F, and treatment, G, distributions. The nature of the dependence is shrouded in mystery. The Lehmann QTE characterizes the difference in the marginal distributions, but it cannot reveal anything about the joint distribution, H.

Effects with Imperfect Compliance

In most social/economic programs, even in experimental settings, we have imperfect compliance. For example, individuals are randomized into the treatment group, but this only *allows* them to be treated, it cannot *compel* them to be treated. The classical treatment control model considers

$$Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$$

where D_i is the treatment indicator. Elaborating, we may have a latent variable model for D_i, so

$$\begin{aligned} D_i^* &= \mu_D(Z_i) + U_i \\ D_i &= I(D_i^* > 0) \end{aligned}$$

This is basically the Roy (1951) model of occupational choice. In the simplest setting we would like to estimate the average treatment effect

$$\Delta^{ATE}(x) = E(\Delta = Y_1 - Y_0|x)$$

but this is difficult to identify so more typically we try to estimate the ATE on the treated,

$$\Delta^{TT}(x) = E(\Delta|x, \;\; D = 1)$$

To explore this Rubin, Heckman and others consider the propensity score,

$$P(z) = P(D = 1|Z = z).$$

This is just the probability of being treated given characteristics z.

Now define, following Heckman and Vytlacil (1999),

$$\begin{aligned}&\Delta^{LATE}(x, P(z), P(z + \Delta z)) \\ &\qquad = \frac{E(Y|X = x, P(Z) = P(z)) - E(Y|X = x, P(Z) = P(z + \Delta z)}{P(z) - P(z + \Delta z)}\end{aligned}$$

and taking limits let,

$$\begin{aligned}\Delta^{LIV}(x, P(z)) &= \lim_{\Delta z \to 0} \Delta^{LATE}(x, P(z), P(z + \Delta z)) \\ &= \frac{\partial E(Y|X = x, P(Z) = P(z))}{\partial P(z)}.\end{aligned}$$

This provides, when it exists, a local IV notion; integrating with respect to various weighting functions over z we obtain a variety of notions that appear in the literature.

Average Treatment Effects

To expand on these ideas a bit and clarify the various notions of "treatment effect" in the recent literature we will briefly survey some recent work on estimation of average treatment effects. The discussion here is based closely on Wooldridge (2002, Chapter 18).

The basic framework will be the "potential outcomes" scheme of Rubin (1974). We have an outcome, or response, variable, y, and a binary treatment variable, w. Observations are identified as either treatment $w = 1$, or control $w = 0$ according to whether they receive the treatment and we observe

$$y = wy_1 + (1 - w)y_0.$$

The Rubin scheme is *counterfactual* in the sense that it posits the existence of a response in both treatment and control states even though we can only observe one of these responses.

In most of the literature the focus is on estimating the average treatment effect (ATE)

$$\beta = E(y_1 - y_0)$$

or the average treatment effect on the treated

$$\gamma = E(y_1 - y_0|w = 1)$$

Clearly, we need further assumptions since we lack data on the pair (y_1, y_0). To anticipate somewhat more general settings we can also consider estimating conditional versions of these effects

$$\begin{aligned}\beta(x) &= E(y_1 - y_0|x) \\ \gamma(x) &= E(y_1 - y_0|x, w = 1)\end{aligned}$$

by disaggregating in x over the entire population.

Suppose, to begin that the treatment is randomly assigned so

$$w \perp\!\!\!\perp (y_1, y_0) \tag{1}$$

then, it is easy to see that $\beta = \gamma$, and we can estimate these effects by the difference in sample means, since

$$\begin{aligned} E(y|w=1) &= E(y_1|w=1) = Ey_1 \\ E(y|w=0) &= E(y_0|w=0) = Ey_0 \end{aligned}$$

However, effective randomization is difficult in many evaluation settings. we have self-selection into treatment, or partial compliance with treatment and these create problems for the simple estimator based on unconditional means.

A somewhat weaker condition under which we can estimate γ, but not β, is that

$$E(y_0|w) = Ey_0 \tag{2}$$

So the control response is "mean independent" of the treatment indicator, this certainly doesn't seem very plausible in most self selection settings, but may have some plausibility in other settings. Note that since

$$y = y_0 + w(y_1 - y_0)$$

we have under the condition (1)

$$\begin{aligned} E(y|w=1) - E(y|w=0) &= E(y_0|w=1) - E(y_0|w=0) + E(y_1 - y_0|w=1) \\ &= \gamma \end{aligned}$$

While condition (1) isn't very plausible, it may be more plausible to assume that some conditional version of it holds,

$$w \perp\!\!\!\perp (y_0, y_1) \quad \text{conditional on } x \tag{3}$$

that is, for same vector of observables, x, we have conditional mean independence. This is sometimes called "selection on observables" since we can view it as a device for assigning treatment based on the variables x. If knowing x, allows us to predict w and the assumption 3 then implies there is no further role played by the response. This would certainly be the case if there were a a deterministic rule to assign w based on x, but also if there were a random assignment as long as the random component conditional on x, was independent of the response. We can weaken (3) to its mean independent form

$$E(y_i|x, w) = E(y_i|w) \qquad i = 0, 1 \tag{4}$$

We can then estimate the average treatment effects β and γ by estimating,

$$r_i(s) = E(y|x, w = i) \qquad i = 0, 1$$

these two functions can, in principle, be consistently estimated nonparametrically. Then we may simply average over the relevant population to obtain,

$$\beta = n^{-1}\sum_{i=1}^{n}(\hat{r}_1(x_i) - \hat{r}_0(x_i))$$

and

$$\gamma = (\sum w_i)^{-1}\sum_{i=1}^{n} w_i(\hat{r}_1(x_i) - \hat{r}_0(x_i))$$

This presupposes that it is really practical to estimate $\hat{r}_i(\cdot)\ \ i = 0, 1$ nonparametrically.

The prior strategy breaks down in the case that there is no overlap in the (x, w) distribution. To see this in an extreme version, suppose that there is one binary x and that *in the population* $x = 1$ implies $w = 1$. Then we cannot estimate $E(y|x = 1, w = 0)$ for example, and this makes impossible to implement the foregoing strategy since it relied on the fact that

$$E(y|x, w = 1) - E(y|x, w = 0) = E(y_1|x) - E(y_0|x) = \beta(x)$$

One can also take a parametric approach to the estimation of ATE. Assume condition 3 holds and that

$$y_i = \mu_i + v_i \qquad i = 0, 1$$

with $Ev_i = 0$, this gives the switching regression or mixture model,

$$y = \mu_0 + (\mu_1 - \mu_0)w + v_0 + w(v_1 - v_0)$$

Now assume that $E(v_1 - v_0)|x = 0$, then

$$E(y|w, x) = \mu_0 + \beta w + g_0(x)$$

where $g_0(x) = E(v_0|x)$.

In this formulation, $g_0(x)$ is sometimes referred to as a "control function". In parametric models we can view this as adding covariates that "control for" the effect of the endogeneity in w.

If we were to view $g_0(x)$ as expandable as a linear function in parameters, say $Z\delta$, then we would have

$$\hat{\beta} = (w'Q_{\tilde{Z}}w)^{-1}w'Q_{\tilde{Z}}y$$

where $\tilde{Z} = [1\vdots Z]$, and $Q_Z = I - Z(Z'Z)^{-1}Z'$. Another way to think about this is that this is 2SLS in which we have used the orthogonal complement of the space spanned by Z as instrumental variables. Similar caveats are in order for situations in which we have perfect predictability of w based on x.

Again we can weaken our assumptions somewhat and consider models of the form,

$$E(y|x, w) = \mu + \beta w + g_0(x) + \omega(g_1(x) - g_0(x))$$

so we have new interaction terms, again in the linear case we have a simple estimation strategy. Note that in this case we should center the interaction effect as estimating the effect at the mean. A special case of this model is the regression discontinuity design: suppose $w = f(x)$ for example

$$w = I(x_1 > \xi_0)$$

If the rule is really non-stochastic like this, so ξ_0 is a known, fixed constant, then the conditioning of our assumptions must apply and then the only problem is finding a sufficiently flexible form for estimating the functions g_0 and g_1.

Binary Treatment Models and Randomization

The simplest experimental treatment model is the following

$$y_i = \alpha + \beta D_i + u_i$$

where D_i is 1 if the subject is "treated", and 0 if the subject is a control. In this model the least squares estimator of β is,

$$\hat{\beta} = \bar{y}_1 - \bar{y}_0$$

and

$$\hat{\alpha} = \bar{y}_0$$

Why? If, as is common, the response y_i is really a *change* in something after a treatment is completed, then we have instead

$$\Delta y_i = \alpha + \beta D_i + u_i$$

and

$$\hat{\beta} = \overline{\Delta y_1} - \overline{\Delta y_0}$$

This is the beloved diff-in-diff model. It has many elaborations; a good overview can be found in Angrist and Pischke's *Mostly Harmless*. It need not be the case that we have only two periods and a binary treatment. If we consider the standard panel data setup with an indicator variable that flips from 0 to 1 at some known date, then when we consider the usual Frisch Waugh treatment of the model we see that effectively the other effects in the model are based on with estimates that are differences in means before and after the flip. Individual specific effects, geographic effects are similar. As implied by the epigraph in the relevant section of *Mostly Harmless* however, one should be aware that all of these intercept shifting effects leave the typical assumption that the slope effect corresponding to the treatment variable intact, so there is an (often implicit) parallelism condition that may be implausible. At some point it is worthwhile to consider interaction effects to explore this.

The focus on mean effects and linearity thus far, although representative of the literature, may leave the faulty impression that nonlinearities are either impossible to deal with, or unimportant. It is worth remembering that even in the standard wage equation model that we have conveniently estimated in log-linear form, the transition back to dollars and cents requires a perilous change of variable that violates the niceties of computing expectations of linear functions.

When we step away from conditional mean models these issues come to the forefront, as can be seen in the literature that seeks to decompose changes in the wage distribution into components due to changes in characteristics of the labor force and changes in the remuneration of these characteristics, as for example in DiNardo, Fortin and Lemieux (1996) and Machado and Mata (2006). There is more recent discussion of quantile treatment effects in the diff-in-diff context in Athey and Imbens (2006) and Stewart (2012).

A Case Study A classical example is the Lanarkshire milk experiment described by Student (1931). In an effort to improve nutrition for elementary school children in a relatively poor region of Scotland an experiment was undertaken to provide milk in schools. The intention was to *randomly select* between 200-400 kids in each of 67 schools, of which half would get milk each day; the other half would not. Evaluation of the effectiveness of the "treatment" was exactly the diff-in-diff strategy which would be done as a t-test. The response, y, was change in weight.

What went wrong? Teachers decided who got the milk and presumably gave the milk to the poorer, smaller "more deserving" kids. We can check this by noting with randomization the treated and control kids would have the same initial weight but they didn't; treated kids were lighter by approximately 3 months growth, and shorter by 4 months growth in height. Since the initial weighing occurred in February and the final weighing in June, and children were weighed with their clothes on, the real weight response is confounded with the change in the weight of the clothes. Again, if the randomization were done properly this would not be a problem, a source of additional variability of course, but not of bias. As it was, it is a serious bias consideration. *Could this be corrected?* Not really after the fact. Student suggests using a smaller trial with only twins, in a future experiment.

The Wald Estimator

In many instances of the treatment-control experiment, there is randomization in what has been called "intention to treat," but often there cannot be any way to force people to accept the treatment. So we have to distinguish *compliance* from *intent to treat.* In the simplest setting this gives rise to a simple form of the IV estimator. Suppose x_i is actual treatment/control as before and z_i is the intent to treat variable, then in our simple original setup we can use the Wald Estimator. The simplest way to obtain the Wald estimator is to consider the model

$$y_i = \alpha + \beta x_i + u_i$$

Suppose $Ez_i u_i = 0$ so we have the moment equations, recalling that z_i is binary,

$$E(y_i|z_i = 1) = \alpha + \beta E(x_i|z_i = 1)$$
$$E(y_i|z_i = 0) = \alpha + \beta E(x_i|z_i = 0)$$

now subtract one from the other to obtain.

$$\beta = \frac{E(y|z_i = 1) - E(y_i|z_i = 0)}{E(x_i|z_i = 1) - E(x_i|z_i = 0)}$$

so a natural estimator would replace these population quantities by their sample analogues. This is the Ur-iv estimator. Angrist calls it the mother of all IV estimators. In some heuristic sense we "see" the relationship between y and x "through the looking glass" as reflected by the IV z_i. When x_i is binary, say D_i to use our prior notation, then $E(D_i|z_i = j) = \Pr(D_i = 1|z_i = j) \equiv \pi_j$ for $j = 0, 1$, so the denominator is the difference in these probabilities. Note that, focusing on the denominator, we might expect that in many situations that the term $E(x_i|z_i = 0)$ would be zero, since subject who aren't "intended to be treated" may find it difficult to *be* treated. On the other hand, $E(x_i|z_i = 1)$ is generally likely to be somewhat less than one, since some of those randomized into the treatment may decide that they don't want to be treated. In the extreme case that the proposed IV z_i doesn't impact the mean of mean of the x_i's, then we have a classical failure of the IV strategy and division by zero.

Returning to the pure randomization model for a moment, there is often, even in well randomized experiments, a temptation to include other covariates in the model, e.g.

$$y_i = \alpha + x_i'\beta + \delta D_i + u_i$$

so D_i is an randomized treatment indicator and x_i denotes a vector of other variables. Now, the randomization implies that

$$x_i \perp D_i$$

and this assumption can be checked. (This is usually done by computing conditional means of the x's with respect to D.) What is the advantage of including the additional covariates? We know that given their orthogonality with D that they shouldn't change our $\hat{\delta}$, so why bother?

The usual answer to this question, exemplified by Gertler (2004) is that including x_i's "improves the power of the estimates". Gertler is analyzing the effect of PROGRESSA the conditional cash transfer program in Mexico. In many respects this program is like the Lanarkshire milk experiment except that cash is distributed directly to households according to a randomized scheme. But children's heights and weights are still the principle measures of program effect. What does "improves the power of the estimates" mean? Presumably, it means "reduces their standard errors". Since $D \perp x$ this has nothing to do with $X'X$, but only with $\hat{\sigma}^2$. Clearly if x's are effective in "explaining" y, then their inclusion will reduce $\hat{\sigma}^2$ and thereby reduce the standard error of $\hat{\delta}$. What's not to like about this?

The case against including covariates is laid out nicely in Freedman (2009). He argues that the presumption that the linear specification is a good approximation can be dangerous. Freedman adopts what he calls the Neymann (1923) model. It seems to be a precursor of what is now usually called the Rubin "potential outcomes" model. We have a response variable y and several treatment levels, individual subjects are assigned, in the simplest case, to one and only one of the treatment options. Each individual has a potential outcome associated with each of the treatments, but we only observe one of these, for the treatment that is actually assigned. We would like to estimate the "average" treatment effect for each of the treatments, or alternatively the differential treatment effects, treatment level i's average response minus, say the average response under the control treatment. This is essentially a random coefficient model in which each subject has an individualized response to each of the treatments. The structure is quite distinct from the usual regression model where we tend to automatically assume that treatment effects are constant across subjects and additive. In Freedman's context inclusion of other covariates is potentially dangerous. Depending upon whether we have additivity and balanced design there are possible biases introduced by inclusion of covariates. Generally, with treatment randomization these biases can be show to be asymptotically negligible, but nevertheless they may be significant in particular finite sample settings, and Freedman recommends that the simpler model-free approach to estimating treatment effects be considered as a "more robust" alternative.

Visual Instrumental Variables

As a final installment in this rather loosely organized lecture, I'd like to try to describe a technique for visualizing the IV estimator in a scatter plot. What follows is my attempt to formalize somewhat the discussion in Section 4.1.3 of Angrist and Pischke (2009).

Consider the following simple model

$$y = \alpha + z\beta + u.$$

Suppose that z should be considered endogenous and for simplicity assume it is scalar. Supppose too that we have another variable, say f, that we would like to act like an instrumental variable. In `R` terminology f is a "factor," i.e., it takes discrete values $f \in \{1, \dots, J\}$. From f we can create a matrix F of indicator (dummy) variables $F_{ij} = 1$ if $f_i = j$, and $F_{ij} = 0$ otherwise. For example, f might be an occupational indicator, or in Angrist's context it might some grouped version of individual i's draft lottery number.

We have the following "reduced form" estimated equations

$$\begin{aligned} \hat{y} &= F\hat{\gamma} \\ \hat{z} &= F\hat{\delta} \end{aligned}$$

so we have two J-vectors $\hat{\gamma}$ and $\hat{\delta}$. Note that these estimates are simply the group means for y and z respectively determined by the F groups. That is, $\hat{\gamma}_j$ is just the mean of the n_j observations y_i that have $f_i = j$ for $j = 1, \cdots J$.

We now plot the J points $\bar{y} = \hat{\gamma}$ vs $\bar{z} = \hat{\delta}$ and overplot some sort of least squares line obtained from the regression,

$$(*) \qquad \hat{\gamma}_i = a + b\hat{\delta}_i + v_i$$

The question is what sort of least squares line would deliver an slope estimate equivalent to the 2SLS estimator? Suppose we consider OLS as a naive first thought:

$$\|\bar{y} - a - \bar{z}b\|^2 = \|(F^\top F)^{-1}F^\top y - a - (F^\top F)^{-1}F^\top zb\|^2$$

This is rather a mess, but if we modify it slightly, to do the GLS version,

$$\|\bar{y} - a - \bar{z}b\|^2_{(F^\top F)} = \|(F^\top F)^{-1}F^\top y - a - (F^\top F)^{-1}F^\top zb\|^2_{(F^\top F)} = \|y - a - zb\|^2_{P_F}$$

which is indeed the 2SLS estimator. This argument can be generalized somewhat to replace the intercept in our simple model with a vector of coefficients associated with some exogonous covariates and then apply the always useful Frisch-Waugh result to reduce the situation back to our simple case.

Obviously, the case that the IVs are just a set of discrete covariates is somewhat special, but it is a useful case to illustrate somewhat more geometrically how IV estimation works.

By binning continuous covariates one can construct approximations to more general cases as well.

Propensity Score Methods

Rosenbaum and Rubin (1983) introduced an alternative strategy for identifying and estimating the ATE. They suggested focusing on the propensity score,

$$p(x) = P(w = 1|x)$$

i.e., the conditional probability of treatment given covariates x.

Prop (Rosenbaum and Rubin) Suppose 4 holds and that $0 < p(x) < 1$ for all x, then

$$\beta = E((w - p(x))y)/(p(x)(1 - p(x)))$$

Proof Write,

$$\begin{aligned}(w - p(x))y &= (w - p(x)((1 - w)y_0 + wy_1) \\ &= wy_1 - p(x)(1 - w)y_0 - p(x)wy_1\end{aligned}$$

Now, let $m_j(x) = E(y_j|x)$, $j = 0, 1$ and computing expectations of the above conditional on (w, x) gives the right hand side as,

$$wm_1(x) - p(x)(1 - w)m_0(x) - p(x)wm_1(x)$$

Now, taking expectations with respect to x yields,

$$\begin{aligned}p(x)m_1(x) &- p(x)(1 - p(x))m_0(x) - p^2(x)m_1(x) \\ &= p(x)(1 - p(x))(m_1(x) - m_0(x))\end{aligned}$$

The assertion of the proposition is

$$\beta = m_1(x) - m_0(x) = \mu_1 - \mu_0$$

but this now follows by iterated expectations.

In practice we need of course an estimate of $p(x)$ and Rosenbaum and Rubin suggest logit specifications, but there are many possibilities here, eventually we have estimators of β of the form,

$$\hat{\beta} = n^{-1} \sum (w_i - \hat{p}(x_i))y_i/[\hat{p}(x_i)(1 - \hat{p}(x_i))]$$

As further evidence that there is nothing new under the sun, this is just a version of the Horvitz and Thompson (1952) estimator for handing mean estimation with nonrandom sampling. There is considerable work on an appropriate choice of estimation strategy for $p(x)$, and no general agreement.

Digression on the Horvitz-Thompson estimator In sample surveys a common problem is estimating a total from a sample on a finite population. Suppose we denote Y_i as the response of the ith observation and D_i as the indicator of whether the ith observation was sampled. Typically, we have an explicit sampling plan so we know $p_i = P(D_i = 1)$, so the HT estimator of the total is simply

$$T_n = \sum_{i=1}^{n} D_i Y_i / p_i.$$

Taking expectations conditionally shows that this gives an unbiased estimate of the total. In fact it can be shown to be the UMVUE. To see the connection with $\hat{\beta}$ above suppose we want to estimate averages now, and we think of D_i as a treatment rather than a sampling decision, then the same argument shows that $ED_iY_i/p_i = EY_1i$ and also that $E(1 - D_i)Y_i/(1 - p_i) = EY_0i$, and combining these facts we have our expression for $\hat{\beta}$.

There is a nice cautionary tale told by Basu (1971) about the HT estimator: A circus impresario has to ship his 50 elephants and needs a rough estimate of their total weight. He speaks with his trainer who suggests weighing Sambo the elephant who is roughly middle size and then multiplying Sambo's weight by 50. Sambo ways 5,000 kg, so the total would be 250,000 kg. But the circus statistician intervenes and says "No, we need a sampling plan." So he proposes chosing Sambo with probability 99/100 and assigning the rest of the mass uniformly to the rest of the elephants. So the impresario agrees and they randomly draw a $U[0, 1]$ which turns out to be less than .99 and so they weigh Sambo. The impresario is about to multiply by 50 when the statistician says "No, stop, the Horvitz Thompson estimator is known to be UMVUE and it tells us to divide Sambo's weight by $p_i = 99/100$, so we want to multiply by 100/99 not 50, so our estimate of the total becomes 5,050 kg, instead of 250,000 kg. Moral: Unbiasedness isn't always such a great property.

Commentary: Note that if, by chance, we would have chosen Jumbo the biggest elephant whose p_i was only 1/5000, we would have multiplied his weight by 5000, since Jumbo weighed 7000 kg our estimation of the total would have been 35,000,000kg which would have compensated for the drastic underestimate obtained with Sambo. But of course we are only doing this procedure once, so the fact that repeatedly doing it gives us something unbiased is not much comfort. Each time we would be getting something quite stupid for an answer. Wooldridge suggests treating $p(x)$ as a control function in a linear regression.

References

Angrist J. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*, Princeton U. Press.

Athey, S. and G. Imbens, (2006) Identification and Inference in Nonlinear Difference-in-Differences Models, 74, 431-497.

Bartlett (1949), The fitting of straight lines if both variables are subject to error, *Biometrics*, 207-12.

Basu, D, (1971) AN ESSAY ON THE LOGICAL FOUNDATIONS OF SURVEY SAMPLING, *Foundations of statistical inference*, eds. Godambe, V.P. and Sprott, D.A. Holt, Reinhardt and Winston.

DiNardo, J., N. Fortin, T. Lemieux, (1996) Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach, *Econometrica*, 64, 1001-1044.

Doksum, K. (1974) Empirical probability plots and statistical inference for nonlinear models in the two sample case, *Annals of Statistics*, **2**, 267-77.

Durbin, J. (1954), “Errors in Variables,” *Intl Stat Review*, 23-32.

Freedman, D. (2009). On Regression Adjustment in Experiments with Several Treatments, *Annals of Applied Stat.* 2, 176-196.

Fuller, W. (1987), *Measurement Error Models*, Wiley.

Gertler, M. (2004). Do Conditional Cash Transfers Improve Child Health? *AER* 94, 336-341.

Heckman, J. (2000), “Causal parameters and policy analysis in economics, ” *QJE*, 45-97.

Heckman, J. and E.J. Vytlacil (1999), “Local IV,” preprint.

Hogg, R. (1975), “Estimates of percentile regression lines using salary data,” *JASA*,

Lehmann, E. (1974), *Nonparametric Statistics*, Holden Day.

Machado, J. A. F. and Mata, J. (2005), Counterfactual decomposition of changes in wage distributions using quantile regression. *J. Appl. Econ.*, 20, 445465.

Neyman, J. and E. Scott, (1951) On certain methods of estimating the linear structural relation, *Annals of Math Stat*, 22, 351-361.

Pakes, A. (1982), On the Asymptotic Bias of Wald-Type Estimates of a Straight Line when both Variables are Subject to Error, *International Economic Review*, 23, 491-497.

Stewart, M.B. (2012) Quantile estimates of counterfactual distribution shifts and the impact of minimum wage increases on the wage distribution, JRSS(A),

Student (1931), The Lanarkshire Milk Experiment, *Biometrika*, 23, 398-406.

Wald, A. (1940), “The fitting of straight lines if both variables are subject to error,” *Annals*,” 284-300.

Figure 1: Lehmann Quantile Treatment Effect

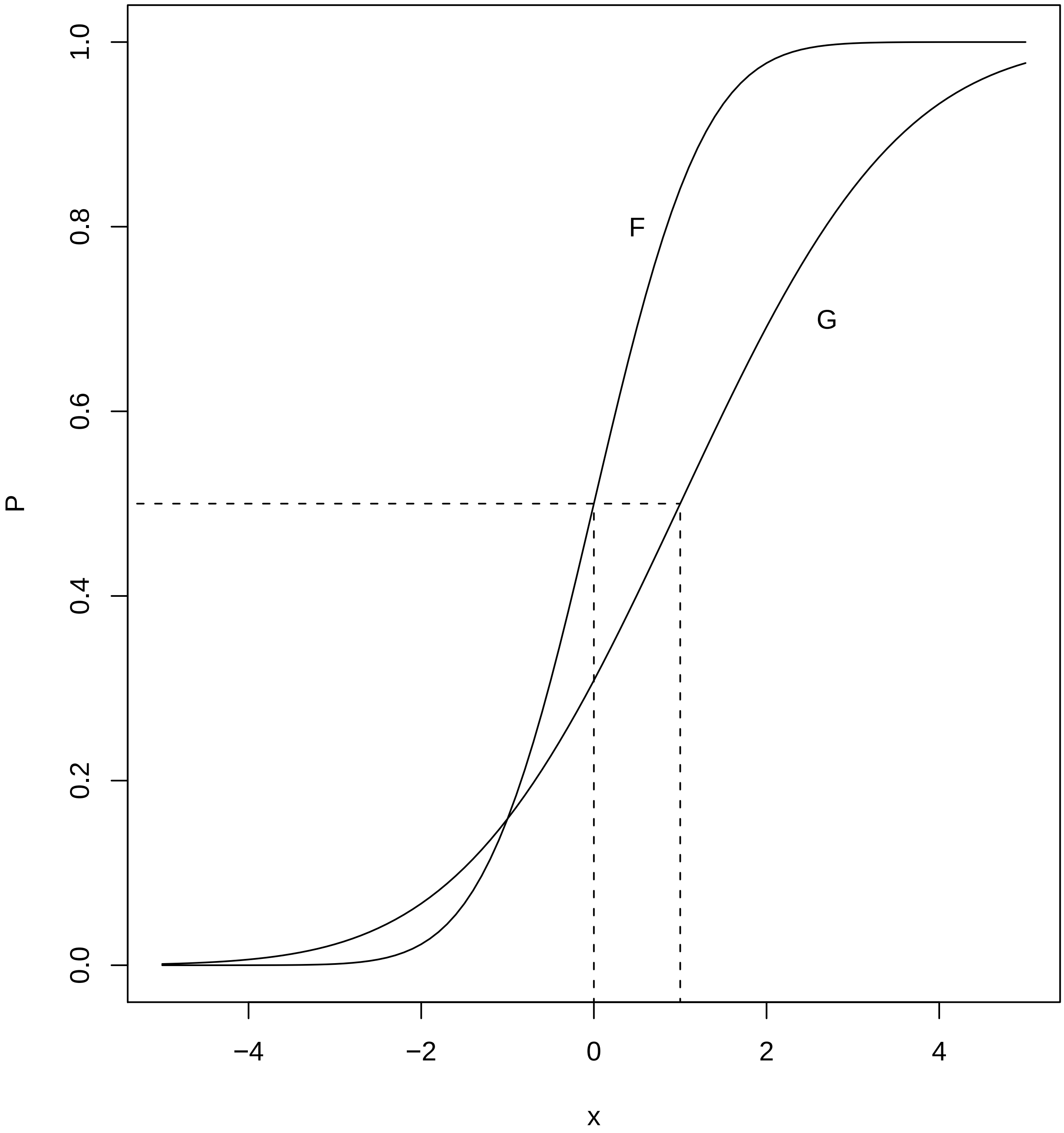
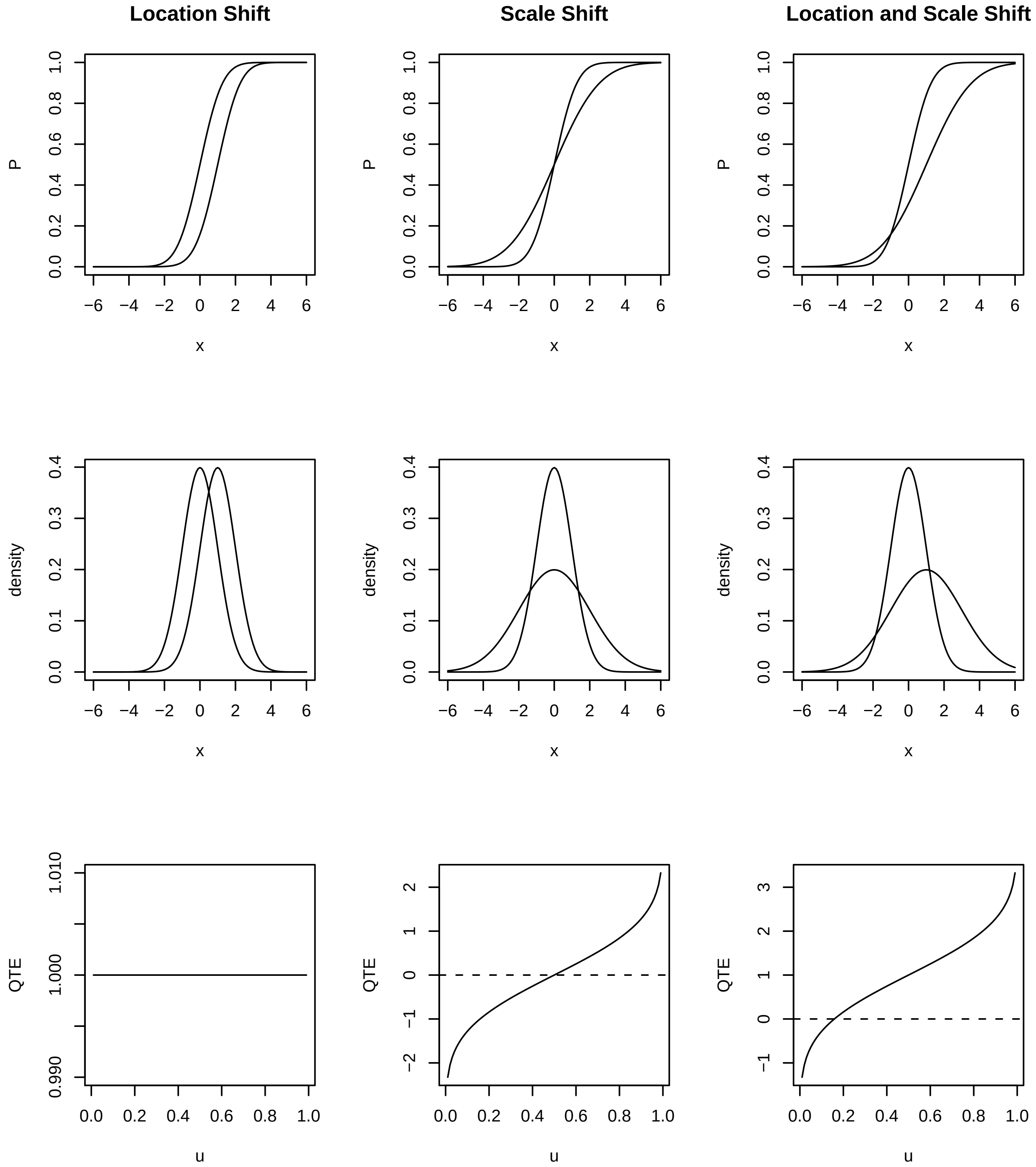


Figure 2: Lehmann Quantile Treatment Effect: Examples


Location Shift
Scale Shift
Location and Scale Shift
P
density
QTE
x
u