AVERAGE TREATMENT EFFECT ESTIMATION: IV AND CONTROL FUNCTION APPROACHES

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. Homogeneous Treatment Effect and Treatment Effects a Function of Obervables
3. Heterogeneous Treatment Effects and LATE
4. Control Function Methods for Heterogeneous Treatment Effects
5. A Different Approach to Allowing Heterogeneous Treatment Effects

1. Introduction

- Now move the the case where unconfoundedness, or selection on observables, does not hold. Allow units to select into treatment based on unobservables that affect the response.
- Even if eligibility is randomly assigned, actual enrollment in programs may suffer from self selection. But randomized eligibility can often be used as an IV.
- Identification and estimation are very simple in the constant treatment effect case. Contemporary interest is often in the heterogeneous treatment effect case.

- Three general approaches: (i) Study simple IV estimators to see if they estimate anything of value under fairly weak assumptions; (ii) Add more assumptions and try to estimate a parameter such as τ_{ate}; (iii) Use "structural" economic models without parametric assumptions, and try to identify interesting policy or structural parameters.

2. Homogeneous Treatment Effect and Treatment Effects a Function of Obervables

- In the simplest case we assume $y_1 - y_0$ is constant. Then

$$\begin{aligned} y &= y_0 + w(y_1 - y_0) = y_0 + \tau w \\ &\equiv \mu_0 + \tau w + v_0 \end{aligned}$$

where $\mu_0 \equiv E(y_0)$ and $v_0 \equiv y_0 - \mu_0$. So, we have a simple regression model with a constant slope.

- Suppose we have a single instrumental variable, z. (At this point, z could be binary, continuous, or anything in between.) Recall the two requirments for an instrument:

$$\begin{aligned} Cov(z,w) &\neq 0 \\ Cov(z,v_0) &= Cov(z,y_0) = 0 \end{aligned}$$

- The first requirement (relevance) is fairly easily tested checked. The second (exogeneity) is maintained. Note that having z predetermined or randomly assigned does not guarantee its exogeneity. The value of z could affect y_0 through other channels.
- Conveniently, the nature of y is not restricted, and the $\boldsymbol{\tau}_{ate} = \boldsymbol{\tau}_{att}$ is consistently estimated by the IV estimator. [Remember, IV not usually unbiased.]
- Even if we restrict attention to consistency, it is **not** true that one should use a “slightly” endogenous instrument rather than OLS.

- Why? It is easy to show:

$$\text{plim } \hat{\tau}_{OLS} = \tau + \frac{\sigma_{v_0}}{\sigma_w} \cdot Corr(w, v_0)$$

and

$$\text{plim } \hat{\tau}_{IV} = \tau + \frac{\sigma_{v_0}}{\sigma_w} \cdot \frac{Corr(z, v_0)}{Corr(z, w)}$$

So, if $Corr(z, w)$ is small – that is, z is a “weak” instrument – then even a small correlation between z and v_0 can produce a larger asymptotic bias than OLS.

- In economics, very common to see IV estimates that are larger in magnitude than OLS estimates. Usually, other explanations are given other than bad IV. Measurement error and heterogeneous treatment effect (later) are among them.
- Weak instruments lead to large asymptotic standard errors, too:

$$Avar\sqrt{N}(\hat{\tau}_{IV} - \tau) = \frac{\sigma_{v_0}^2}{\sigma_w^2 \rho_{z,w}^2}.$$

When $\rho_{z,w}^2$ is small, the asymptotic variance can be very large. The formula for the OLS estimator omits $\rho_{z,w}^2$.

Adding Other Covariates and Instruments

- Suppose we need to add covariates before our instruments are appropriately exogenous:

$$y_g = \mu_g + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}_g + u_g, g = 0, 1$$

where $\mu_g = E(y_g)$ and $\boldsymbol{\mu}_{\mathbf{x}} = E(\mathbf{x})$.

- We have made a functional form assumption once we assume v_g has a zero mean, conditional on $\mathbf{x}$. We also assume that netting out $\mathbf{x}$ makes z exogenous. Easiest way to impose both:

$$E(u_g|\mathbf{x}, \mathbf{z}) = 0, g = 0, 1.$$

- In effect, we have made the standard exclusion restrictions plus a linear functional form. Write

$$
\begin{aligned}
y = y_0 + w(y_1 - y_0) &= \mu_0 + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}_0 + u_0 \\
&\quad + w(\mu_1 - \mu_0) + w(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + w(u_1 - u_0) \\
&= \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + u_0 + w(u_1 - u_0)
\end{aligned}
$$

where $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ and $\alpha_0 = \mu_0 - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta}_0$.

- The last term – the interaction of the treatment and the unobserved gain from treatment, $e \equiv u_1 - u_0$, causes difficulties.

- In the simplest case, there is no heterogeneity except in the intercept: $\boldsymbol{\delta} = \mathbf{0}, u_1 = u_0$. Then

$$y = \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + u_0,$$

and we can estimate the parameters by 2SLS using instruments $(1, \mathbf{x}, \mathbf{z})$.

- How might we exploit the binary nature of w?

1. (Not Recommended): Take the expected value conditional on all exogenous variables:

$$\begin{aligned} E(y|\mathbf{x},\mathbf{z}) &= \alpha_0 + \tau E(w|\mathbf{x},\mathbf{z}) + \mathbf{x}\boldsymbol{\beta}_0 + E(u_0|\mathbf{x},\mathbf{z}) \\ &= \alpha_0 + \tau P(w = 1|\mathbf{x},\mathbf{z}) + \mathbf{x}\boldsymbol{\beta}_0 \end{aligned}$$

because w is binary.

- Two-step procedure: (i) Estimate $P(w = 1|\mathbf{x}, \mathbf{z})$ by probit (or logit) and obtain the first-stage fitted probabilities, say $\hat{\Phi}_i = \Phi(\hat{\pi}_0 + \mathbf{x}_i\hat{\boldsymbol{\pi}}_1 + \mathbf{z}_i\hat{\boldsymbol{\pi}}_2)$. Should be able to reject $\boldsymbol{\pi}_2 = \mathbf{0}$ fairly strongly. (ii) Use the $\hat{\Phi}_i$ in place of w_i:

$$y_i \text{ on } 1, \hat{\Phi}_i, \mathbf{x}_i, i = 1, \ldots, N.$$

- Why not recommended? (a) Inconsistent generally if model (probit in this case) for $P(w = 1|\mathbf{x}, \mathbf{z})$ is wrong. Then $E(y|\mathbf{x}, \mathbf{z}) \neq \alpha_0 + \tau\Phi(\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2) + \mathbf{x}\boldsymbol{\beta}_0$; (b) Need to adjust standard errors for "generated regressor"; (c) No known efficiency properties; (d) Tempting to achieve identification off of the nonlinearity in the probit in the absense of instruments.

2. (Recommended): Use $\hat{\Phi}_i$ as an IV, not a regressor. That is, estimate

$$y_i = \alpha_0 + \tau w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + u_{i0},$$

by IV using instruments $(1, \hat{\Phi}_i, \mathbf{x}_i)$.

● Not the same as using $\hat{\Phi}_i$ as a regressor! The first stage when $\hat{\Phi}_i$ is used as an IV is

$$\hat{w}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{\Phi}_i + \mathbf{x}_i \hat{\boldsymbol{\gamma}}_2$$

and $\hat{\gamma}_0 \neq 0$, $\hat{\gamma}_1 \neq 1$, $\hat{\boldsymbol{\gamma}}_2 \neq \mathbf{0}$.

- Why recommended?

(a) Misspecification of probit model does not matter, provided w is partially correlated with $\Phi(\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2)$! So this method, like using $(1, \mathbf{x}_i, \mathbf{z}_i)$ as instruments, is robust to having the model for $P(w = 1|\mathbf{x}, \mathbf{z})$ wrong.

(b) No need to account for generated instruments.in standard errors – see Wooldridge (2002, Chapter 6).

(c) Estimator is efficient IV estimator if $Var(u_0|\mathbf{x}, \mathbf{z}) = Var(u_0)$ and probit model for w is correct.

- As a practical matter, there might not be much difference in the point estimates between (1) and (2). Certainly not if the coefficients in the regression w_i on $1, \hat{\Phi}_i, x_i$ are close to $0, 1, 0$. But using the fitted values as a regressor has no advantages.
- Easy to modify to allow treatment to interact with $\mathbf{x}$. After first stage probit, estimate

$$y_i = \alpha_0 + \tau w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + error_i$$

by IV, using instruments $[1, \hat{\Phi}_i, \mathbf{x}_i, \hat{\Phi}_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}})]$. Benefits same as without interaction.

- Can safely ignore estimation error in sample mean, $\bar{\mathbf{x}}$.

- Note that we have now allowed heterogeneous treatment effects but only as a function of **x**. In particular,

$$\hat{\tau}(\mathbf{x}) = \hat{\tau} + (\mathbf{x} - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}},$$

and we can insert different values for **x** to see how the ATE varies with observed characteristics.

3. Heterogeneous Treatment Effect and LATE

The Local Average Treatment Effect

- What does IV estimate, in general, if the gain from treatment, $y_1 - y_0$, is not constant? Imbens and Angrist (1994): Now treatment is counterfactual, too. Let z be the binary instrumental variable, which is often randomized eligibility, so that z is zero or one. Let w_0 be treatment status if $z = 0$ and let w_1 be treatment status if $z = 1$. (Some eligibles may not choose to participate; some non-eligibles may find other ways to "participate.")

- We only observe one of the counterfactual treatments:

$$w = (1 - z)w_0 + zw_1$$

Now write

$$y = (1 - w)y_0 + wy_1 = y_0 + w_0(y_1 - y_0) + z(w_1 - w_0)(y_1 - y_0)$$

Assumptions:

LATE.1: z is independent of (w_0, w_1, y_0, y_1)

LATE.2: $P(w = 1|z = 1) \neq P(w = 1|z = 0)$

LATE.3: $w_1 \geq w_0$ (no *defiers*, or monotonicity)

- This last condition means that if a unit would participates in the program when not eligible, they would participate if made eligible. (Vietnam draft lottery: a defier would be someone who would serve if not drafted but would not serve if drafted. LATE.3 rules this out.)
- Imbens and Angrist define the **local average treatment effect** (**LATE**) as

$$\tau_{late} = E(y_1 - y_0 | w_1 - w_0 = 1)$$

which is the average treatment effect for those induced into treatment by assignment. (That is, $w_0 = 0$ and $w_1 = 1$.)

- Imbens and Angrist call the subpopulation with $w_1 = 1$, $w_0 = 0$ the *compliers*. These individuals comply with assignment no matter what their eligibility. That is, if they are assigned to the control group, they would not participate, and if they are assigned to the treatment, they do participate.
- One criticism of τ_{late} is that it is a treatment effect for an unidentifiable segment of the population: we do not know, on the basis of observing (y_i, w_i, z_i) whether unit i is a complier.

- Consequently, LATE may not have “external validity,” that is, it may not apply to evaluations of other programs even if the population is the same.
- A related point is that two different binary instruments lead to different groups of compliers in the same population. For example, if y is log of earnings, w is college attendance, and z is living near a college, the compliers are those who would attend college if one is nearby, but not otherwise. But if z is whether a grant is offered, the compliers consist of those who would not attend college if they do not receive a grant, but would attend if they do.

- The key result of IA is that, under the three LATE assumptions, the usual IV estimator consistently estimates τ_{late}. To see this, use the independence assumption to write

$$E(y|z) = E(y_0) + E[w_0(y_1 - y_0)] + zE[(w_1 - w_0)(y_1 - y_0)]$$

so that

$$E(y|z = 1) - E(y|z = 0) = E[(w_1 - w_0)(y_1 - y_0)].$$

Now

$$\begin{aligned}
E[(w_1 - w_0)(y_1 - y_0)] &= 1 \bullet E(y_1 - y_0 | w_1 - w_0 = 1)P(w_1 - w_0 = 1) \\
+ 0 \bullet E(y_1 - y_0 | w_1 - w_0 &= 0)P(w_1 - w_0 = 0) \\
+ (-1) \bullet E(y_1 - y_0 | w_1 - w_0 &= -1)P(w_1 - w_0 = -1) \\
&= E(y_1 - y_0 | w_1 - w_0 = 1)P(w_1 - w_0 = 1) \\
- E(y_1 - y_0 | w_1 - w_0 &= -1)P(w_1 - w_0 = -1) \\
&= E(y_1 - y_0 | w_1 - w_0 = 1)P(w_1 - w_0 = 1)
\end{aligned}$$

because $P(w_1 - w_0 = -1) = 0$ by the no defiers assumption.

- Because $P(w_1 - w_0 = -1) = 0$, $w_1 - w_0$ is a binary variable, and

$$P(w_1 - w_0 = 1) = E(w_1 - w_0) = E(w_1) - E(w_0).$$

- But, again, z is independent of (w_0, w_1), and so

$$E(w|z) = (1 - z)E(w_0) + zE(w_1)$$

which implies

$$\begin{aligned} E(w_1) - E(w_0) &= E(w|z = 1) - E(w|z = 0) \\ &= P(w = 1|z = 1) - P(w = 1|z = 0) \neq 0 \end{aligned}$$

by LATE.2.

- We have shown that

$$E(y_1 - y_0|w_1 - w_0 = 1) = \frac{E(y|z = 1) - E(y|z = 0)}{P(w = 1|z = 1) - P(w = 1|z = 0)},$$

and this is the probability limit of the simple IV estimator when z and w are both binary. In fact, the IV estimate is just the Wald estimate,

$$\hat{\tau}_{late} = \frac{N_1^{-1} \sum_{i=1}^{N} z_i y_i - N_0^{-1} \sum_{i=1}^{N} (1 - z_i) y_i}{N_1^{-1} \sum_{i=1}^{N} z_i w_i - N_0^{-1} \sum_{i=1}^{N} (1 - w_i) y_i},$$

where $N_1 = \sum_{i=1}^{N} z_i$ and $N_0 = N - N_1$. (Sometimes called a “grouping estimator.”)

- The LATE interpretation of IV has had wide influence in the treatment effect literature. It is probably relied on too much to explain why IV estimate is larger than OLS. Sometimes one forgets the possibility that the IV is not exogenous.
- Literature is vast on allowing covariates, **x**, in the analysis and when **z** is not a binary scalar. Flavor of results carries through.

4. Control Function Methods for Heterogeneous Treatment Effects

- Suppose we think there are heterogeneous treatment effects but are not satisfied with estimating LATE (or possibly the necessary assumptions do not hold). The "old-fashioned" approach is to add more assumptions and apply control function methods. This leads to the so-called "switching regression" model, which is the same as a random coefficient model.

- Recall that we can write

$$y = \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + u_0 + w \cdot e$$

where $e = u_1 - u_0$. Note that the ATE is still the coefficient on w.

- We have to take seriously the underlying functional forms, $E(y_g|\mathbf{x}) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$, and we will further restrict the conditional distribution of y_g. In effect, this setup applies to y_g continuous. (Contrast the approaches based on unconfoundedness, particularly PS weighting and matching.)
- Even if we add the exogeneity assumptions

$$E(u_0|\mathbf{x},\mathbf{z}) = E(e|\mathbf{x},\mathbf{z}) = 0,$$

it is generally *not* true that

$$E(w \cdot e|\mathbf{x}, \mathbf{z}) = E(w \cdot e).$$

and this is the condition for earlier IV estimators to be consistent for τ if the interaction term $w \cdot e$ is in the error term – see Wooldridge (2003, *Economics Letters*).

- As it turns out, this condition can hold for continuous treatments, in which case only α_0 is estimated inconsistently. The IV estimator is consistent for $\tau = \tau_{ate}$.
- For common binary response models for w, this mean independence assumption is known not to hold.

- Instead, use a control function approach, that is, find $E(y|w,\mathbf{x},\mathbf{z})$:

$$\begin{aligned} E(y|w,\mathbf{x},\mathbf{z}) &= \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} \\ &\quad + E(u_0|w,\mathbf{x},\mathbf{z}) + wE(e|w,\mathbf{x},\mathbf{z}) \end{aligned}$$

- We can estimate the expectations if we model w. Probit is convenient, and we now have to take the model seriously.

$$w = 1[\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + c \geq 0]$$

and

$$(u_0, u_1, c)|\mathbf{x},\mathbf{z} \sim \text{Multivariate Normal}$$

- Can relax this a bit: $E(u_0|c,\mathbf{x},\mathbf{z})$ and $E(u_1|c,\mathbf{x},\mathbf{z})$ linear in c (and do not depend on $\mathbf{x},\mathbf{z}$) along with $D(c|\mathbf{x},\mathbf{z})$ standard normal.

- Expectations have the well-known “inverse Mills ratio” form:

$$\begin{aligned} E(y|w,\mathbf{x},\mathbf{z}) &= \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} \\ &\quad + \eta w[\phi(\mathbf{r}\boldsymbol{\pi})/\Phi(\mathbf{r}\boldsymbol{\pi})] \\ &\quad + \theta(1-w)\{\phi(\mathbf{r}\boldsymbol{\pi})/[1-\Phi(\mathbf{r}\boldsymbol{\pi})]\} \end{aligned}$$

where $\phi(\cdot)$ is the standard normal pdf and $\Phi(\cdot)$ is the standard normal cdf and $\mathbf{r} = (1,\mathbf{x},\mathbf{z})$.

- This is an estimating equation for τ, and $\boldsymbol{\delta}$, too, when we want $\tau(\mathbf{x}) = \tau + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta}$.

● Two-Step Method (Heckman, but for "switching regression," or "endogenous treatment," not sample selection):

(i) As before Run probit of w_i on $1, \mathbf{x}_i, \mathbf{z}_i$ and obtain $\hat{\phi}_i = \phi(\mathbf{r}_i\hat{\boldsymbol{\pi}})$ and $\hat{\Phi}_i = \Phi(\mathbf{r}_i\hat{\boldsymbol{\pi}})$.

(2) Run the OLS regression

$$y_i \text{ on } 1, w_i, \mathbf{x}_i, w_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}), w_i \cdot (\hat{\phi}_i/\hat{\Phi}_i),\ (1 - w_i) \cdot [\hat{\phi}_i/(1 - \hat{\Phi}_i)]$$

● There is a "generated regressor" problem. Need to adjust standard errors and inference for first-stage estimation. Can use "heckit" in Stata applied to treated and nontreated.

- If, say, we estimate the coefficients $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0')'$ and $(\hat{\alpha}_1, \hat{\boldsymbol{\beta}}_1')'$ using heckit on the $w_i = 0$ and $w_i = 1$ subgroups, important to form

$$\hat{\tau}(\mathbf{x}) = (\hat{\alpha}_1 - \hat{\alpha}_0) + \mathbf{x}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)$$

and, as a special case,

$$\hat{\tau} = \hat{\tau}_{ate} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0).$$

It makes no sense to just look at $\hat{\alpha}_1 - \hat{\alpha}_0$ unless $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0$ has been imposed or $\bar{\mathbf{x}}$ is, coincidentally, close to zero.

- The ATEs are never obtained by looking at $\hat{E}(y_i|w_i = 1, \mathbf{x}_i, \mathbf{z}_i) - \hat{E}(y_i|w_i = 0, \mathbf{x}_i, \mathbf{z}_i)$ and averaging these differences across i. We already know how τ and $\tau(\mathbf{x})$ depend on the parameters. They do not depend on $\mathbf{z}$!
- Again, we use the formual for $E(y|w, \mathbf{x}, \mathbf{z})$ as an estimating equation
- Whether separate estimations are carried out or the pooled regression is used, can use the bootstrap for inference.

- The Stata command “treatreg” (typically using MLE, but sometimes using the CF approach) does not apply to the full switching regression case. In fact, it seems to apply only to the simple model with homogeneous effect,

$$y = \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + u_0,$$

in which case more robust IV methods are available. The CF regression in this case is

$$y_i \text{ on } 1, w_i, \mathbf{x}_i, w_i \cdot (\hat{\phi}_i/\hat{\Phi}_i) + (1 - w_i) \cdot [\hat{\phi}_i/(1 - \hat{\Phi}_i)]$$

- That is, treatreg imposes the restriction $\eta = \theta$.

- As usual, need make sure **z** has at least one element that predicts treatment. (Use the first-stage probit.)
- In the pooled estimation in the general case, can use a joint test of the two terms $w_i \cdot (\hat{\phi}_i/\hat{\Phi}_i)$, $(1 - w_i) \cdot [\hat{\phi}_i/(1 - \hat{\Phi}_i)]$ to test the null that w is exogenous. Under the null, do not need to account for generated regressors, so use a heteroskedasticity-robust Wald test.

- Can estimate τ_{att}, too, but must be careful. First find $\tau_{att}(\mathbf{x},\mathbf{z})$ and then average out $(\mathbf{x},\mathbf{z})$. (Heckman, Tobias, Vytlacil).
- In general,

$$\begin{aligned}\tau_{att}(\mathbf{x},\mathbf{z}) &= E(y_1 - y_0|w = 1,\mathbf{x},\mathbf{z}) \\ &= E(y_1|w = 1,\mathbf{x},\mathbf{z}) - E(y_0|w = 1,\mathbf{x},\mathbf{z})\end{aligned}$$

- In the current context, $E(y_1|w = 1,\mathbf{x},\mathbf{z})$ is obtained directly from the Heckit using $w_i = 1$:

$$\hat{E}(y_1|w = 1,\mathbf{x},\mathbf{z}) = \hat{\alpha}_1 + \mathbf{x}\hat{\boldsymbol{\beta}}_1 + \hat{\gamma}_1\lambda(\mathbf{r}\hat{\boldsymbol{\pi}})$$

where $\hat{\gamma}_1$ is the consistent estimator of $Corr(u_1,e) \cdot \sigma_1$ and $\mathbf{r} = (1,\mathbf{x},\mathbf{z})$.

- For $E(y_0|w = 1, \mathbf{x}, \mathbf{z})$, we first run Heckit using $w_i = 0$. But we do not want $E(y_0|w = 0, \mathbf{x}, \mathbf{z})$. Instead,

$$\hat{E}(y_0|w = 1, \mathbf{x}, \mathbf{z}) = \hat{\alpha}_0 + \mathbf{x}\hat{\boldsymbol{\beta}}_0 + \hat{\gamma}_0\lambda(\mathbf{r}\hat{\boldsymbol{\pi}})$$

It follows that

$$\hat{\tau}_{att}(\mathbf{x}, \mathbf{z}) = (\hat{\alpha}_1 - \hat{\alpha}_0) + \mathbf{x}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\lambda(\mathbf{r}\hat{\boldsymbol{\pi}})$$

Then $\hat{\tau}_{att}$ can be obtained by averaging $\hat{\tau}_{att}(\mathbf{x}_i, \mathbf{z}_i)$ across the $w_i = 1$ observations:

$$\hat{\tau}_{att} = N_1^{-1} \sum_{i=1}^{N} w_i[(\hat{\alpha}_1 - \hat{\alpha}_0) + \mathbf{x}_i(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\lambda(\mathbf{r}_i\hat{\boldsymbol{\pi}})].$$

- If $\gamma_1 = \gamma_0$, $\tau_{att}(\mathbf{x},\mathbf{z}) = \tau_{ate}(\mathbf{x})$, which of course implies. τ_{att} and τ_{ate} still generally differ in this case because $\tau_{att} = E[\tau_{ate}(\mathbf{x})|w = 1]$ while $\tau_{ate} = E[\tau_{ate}(\mathbf{x})]$.
- If we have no need for the full switching regression – in particular, if $u_1 = u_0$ – then the IMR does not appear in calculation of τ_{att}.

- If y has discreteness, we can try to exploit its nature. For example, suppose y_g are binary, and we are willing to assume, for $g = 0, 1$,

$$y_g = 1[\alpha_g + \mathbf{x}\boldsymbol{\beta}_g + u_g > 0]$$

$$u_g|\mathbf{x} \sim Normal(0, 1)$$

- We know that the ATE as a function of **x** is

$$\tau(\mathbf{x}) = \Phi(\alpha_1 + \mathbf{x}\boldsymbol{\beta}_1) - \Phi(\alpha_0 + \mathbf{x}\boldsymbol{\beta}_0),$$

which is easy to estimate given estimates of the parameters.

- If we again assume

$$w = 1[\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + c \geq 0]$$

and

$$(u_0, u_1, c)|\mathbf{x}, \mathbf{z} \sim \text{Multivariate Normal},$$

then estimation using standard software is straightforward: apply the Heckman selection approach to the probit model for the control $(w_i = 0)$ and treated $(w_i = 1)$ groups separately.

- In Stata, “heckprob.” This gives $\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0, \hat{\alpha}_1, \hat{\boldsymbol{\beta}}_1$ and then we have

$$\hat{\tau}(\mathbf{x}) = \Phi(\hat{\alpha}_1 + \mathbf{x}\hat{\boldsymbol{\beta}}_1) - \Phi(\hat{\alpha}_0 + \mathbf{x}\hat{\boldsymbol{\beta}}_0),$$

which we can study at various values of **x**, or average across subpopulations. The estimate of

$$\tau_{ate} = E_{\mathbf{x}}[\Phi(\alpha_1 + \mathbf{x}\boldsymbol{\beta}_1) - \Phi(\alpha_0 + \mathbf{x}\boldsymbol{\beta}_0)]$$

is

$$\hat{\tau}_{ate} = N^{-1}\sum_{i=1}^{N}[\Phi(\hat{\alpha}_1 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_1) - \Phi(\hat{\alpha}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_0)].$$

- Estimation of τ_{att} is more complicated.

5. A Different Approach to Allowing Heterogeneous Treatment Effects

- Based on Wooldridge (2007, *Advances in Econometrics*). Go back to the linear model:

$$y = \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + u_0 + w \cdot e$$

- Rather than find $E(u_0|w,\mathbf{x},\mathbf{z})$ and $E(e|w,\mathbf{x},\mathbf{z})$, now write $w \cdot e = E(w \cdot e|\mathbf{x},\mathbf{z}) + a,$ where $E(a|\mathbf{x},\mathbf{z}) = 0$ by definition. Then

$$y = \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + E(w \cdot e|\mathbf{x},\mathbf{z}) + u_0 + a,$$

where now the composite error, $c = u_0 + a$, has zero mean conditional on $(\mathbf{x},\mathbf{z})$.

- Turns out $E(w \cdot e|\mathbf{x},\mathbf{z})$ is remarkably simple:

$$E(w \cdot e|\mathbf{x},\mathbf{z}) = \rho\phi(\mathbf{r}\boldsymbol{\pi}) = \rho\phi(\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2)$$

- The estimating equation is

$$\begin{aligned} y &= \alpha_0 + \tau w + \mathbf{x}\boldsymbol{\beta}_0 + w \cdot (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + \rho\phi(\mathbf{r}\boldsymbol{\pi}) + c \\ E(c|\mathbf{x},\mathbf{z}) &= 0. \end{aligned}$$

- But w is still endogenous in this equation, so we still need an instrument. (This is why the method differs from standard control function approach.)
- The added regressor, $\phi(\mathbf{r}\boldsymbol{\pi}),$ is exogenous, and acts as its own instrument. The natural instrument for w is $\Phi(\mathbf{r}\boldsymbol{\pi})$.

- Two-step procedure:

(i) As before, probit of w on $1, \mathbf{x}, \mathbf{z}$. Obtain $\hat{\phi}_i = \phi(\mathbf{r}_i\hat{\boldsymbol{\pi}})$ and $\hat{\Phi}_i = \Phi(\mathbf{r}_i\hat{\boldsymbol{\pi}})$

(ii) Estimate

$$y_i = \alpha_0 + \tau w_i + \mathbf{x}_i\boldsymbol{\beta}_0 + w_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + \rho\hat{\phi}_i + error_i$$

by IV, using instruments $[1, \hat{\Phi}_i, \mathbf{x}_i, \hat{\Phi}_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}), \hat{\phi}_i]$.

- Generated regressor problem with $\hat{\phi}_i$, so use delta method or bootstrap. (Do not need to worry about generated instruments.)

- A test of whether $w \cdot (u_1 - u_0)$ is needed is just the usual (heteroskedasticity-robust) t statistic on $\hat{\phi}_i$, after IV estimation. Not a test of endogeneity of w because w can still be correlated with u_0.
- Drawback to this approach: Does not work with binary z without covariates: $\Phi(\pi_0 + \pi_2 z)$ and $\phi(\pi_0 + \pi_2 z)$ are perfectly collinear when z is binary.
- Generally, might be sensitive to weak instruments. Need z to vary sufficiently. Control function approach does work with binary z, at least in some cases. But in CF case, test of interaction not robust to nonnormality (endogeneity and functional form tied together).

- Can learn other things by studying the estimating equation. Assume we have no covariates, so drop **x** and write

$$\begin{aligned} y &= \alpha_0 + \tau w + \rho\phi(\pi_0 + \pi_2 z) + c \\ E(c|z) &= 0. \end{aligned}$$

Via simulations, Angrist (1990, NBER Technical Working Paper) studied the properties of the usual IV estimator in the equation without $\rho\phi(\pi_0 + \pi_2 z)$. In other words, the implicit error term is $\rho\phi(\pi_0 + \pi_2 z) + c$. In general, the IV, z [which is what Angrist used, not $\Phi(\pi_0 + \pi_2 z)$], is correlated with $\phi(\pi_0 + \pi_2 z)$. But not always. In fact, in Angrist's simulations, $\phi(\pi_0 + \pi_2 z) = \phi(\pi_2 z)$ and z has a symmetric distribution about zero.

- Just like z and z^2 are uncorrelated when z is symmetrically distributed about 0,

$$Cov[z, \phi(\pi_2 z)] = 0$$

[because $\phi(\bullet)$ is symmetric about zero]. We already know $Cov(z, c) = 0$, and so z is uncorrelated with $\rho\phi(\pi_0 + \pi_2 z) + c$.

- Likely explains why Angrist's simulation results for the simple IV estimator look so promising for estimating τ.