Ch07. Endogeneity and Instrumental Variables

Ping Yu

HKU Business School
The University of Hong Kong

Endogeneity

Endogeneity

- In the linear regression
$$y_i = \mathbf{x}_i'\beta + u_i, \tag{1}$$
if $E[\mathbf{x}_i u_i] \neq \mathbf{0}$, there is endogeneity.
- In this case, the LSE will be asymptotically biased.
- Here, β in (1) is the structural parameter rather than the linear projection coefficient of y on $span(\mathbf{x})$ since from Chapter 2 we can always find a β such that $E[\mathbf{x}_i u_i] = \mathbf{0}$.
  - Example: in the return to schooling, if $\mathbf{x}$ is the education level and u is the ability, then given the ith individual's education level $\mathbf{x}_i$ and ability u_i, we can determine his/her wage level from equation (1). That is, equation (1) describes an economic reality for each individual rather than only a statistical relationship.
- The analysis of data with endogenous regressors is arguably the main contribution of econometrics to statistical science.

Five Sources of Endogeneity

- Simultaneous causality.
  - Example: Do higher hotel prices decrease occupancy rates? Do Cigarette taxes reduce smoking? Does putting criminals in jail reduce crime? Does declining fertility explain increasing female labor supply?
  - Solution: using instrumental variables (IVs), and designing and implementing a randomized controlled trial (RCT) in which the reverse causality channel is nullified
- Omitted variables.
  - Example: in the model on returns to schooling, ability is an important variable that is correlated to years of education, but is not observable so is included in the error term.
  - Solution: using IVs, using panel data and using RCTs.

History of RCT

Michael Kremer (1964-, Chicago), Esther Duflo (1972-, MIT)
and Abhijit Banerjee (1961-, MIT), NP2019

- They won the Nobel Prize in 2019 because they successfully applies RCT to improve our ability to fight global poverty.

Continue...

- Errors in variables. This term refers to the phenomenon that an otherwise exogenous regressor becomes endogenous when measured with error.
  - Example: in the returns-to-schooling model, the records for years of education are fraught with errors owing to lack of recall, typographical mistakes, or other reasons.
  - Solution: using IVs (e.g., exogenous determinants of the error ridden explanatory variables, or multiple indicators of the same outcome, i.e., repeated measurements).
- Sample selection.
  - Example: in the analysis of returns to schooling, only wages for employed workers are available, but we want to know the effect of education for the general population.
  - Solution: Heckman's control function approach.
- Functional form misspecification. $E[y|\mathbf{x}]$ may not be linear in $\mathbf{x}$.
  - Solution: nonparametric methods.

Simultaneous Causality

- Philip Green Wright (1928) considered to estimate the elasticity of butter demand, which is critical in the policy decision on the tariff of butter.
- Define $p_i = \ln P_i$ and $q_i = \ln Q_i$, and the demand equation is

$$q_i = \alpha_0 + \alpha_1 p_i + u_i, \tag{2}$$

where u_i represents other factors besides price that affect demand, such as income and consumer taste. But the supply equation is in the same form as (2):

$$q_i = \beta_0 + \beta_1 p_i + v_i, \tag{3}$$

where v_i represents the factors that affect supply, such as weather conditions, factor prices, and union status.
- So p_i and q_i are determined "within" the model, and they are endogenous. Rigorously, note that

$$\begin{aligned} p_i &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, \\ q_i &= \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1}, \end{aligned}$$

by solving two simultaneous equations (2) and (3).

- Suppose $Cov(u_i, v_i) = 0$, then

$$Cov(p_i, u_i) = -\frac{Var(u_i)}{\alpha_1 - \beta_1}, Cov(p_i, v_i) = \frac{Var(v_i)}{\alpha_1 - \beta_1},$$

  which are not zero. If $\alpha_1 < 0$ and $\beta_1 > 0$, then $Cov(p_i, u_i) > 0$ and $Cov(p_i, v_i) < 0$, which is intuitively right (why?).

- If regress q_i on p_i, then the slope estimator converges to

$$\frac{Cov(p_i, q_i)}{Var(p_i)} = \alpha_1 + \frac{Cov(p_i, u_i)}{Var(p_i)} = \beta_1 + \frac{Cov(p_i, v_i)}{Var(p_i)}$$
$$\stackrel{?}{=} \frac{\alpha_1 Var(v_i) + \beta_1 Var(u_i)}{Var(v_i) + Var(u_i)} \in (\alpha_1, \beta_1).$$

- So the LSE is neither α_1 nor β_1, but a weighted average of them. Such a bias is called the simultaneous equations bias. The LSE cannot consistently estimate α_1 or β_1 because both curves are shifted by other factors besides price, and we cannot tell from data whether the change in price and quantity is due to a demand shift or a supply shift.

- If $u_i = 0$, that is, the demand curve stays still, then the equilibrium prices and quantities will trace out the demand curve and the LSE is consistent to α_1.[1] The following figure illustrates the discussion above intuitively.

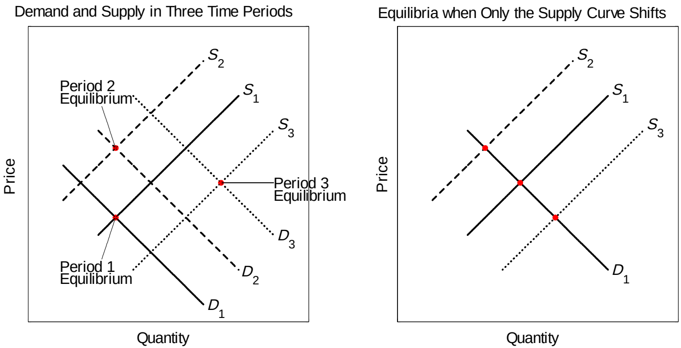


Figure: Endogeneity and Identification of Instrument Variables

[1] When $Var(u_i) = 0$, then $\frac{Cov(p_i, q_i)}{Var(p_i)} = \alpha_1$. Note that the supply curve is still not identifiable because essentially, only one point on the supply curve is observed.

- From above, we can see that p_i has one part which is correlated with u_i $\left(-\frac{u_i}{\alpha_1-\beta_1}\right)$ and one part is not $\left(\frac{v_i}{\alpha_1-\beta_1}\right)$. If we can isolate the second part, then we can focus on those variations in p_i that are uncorrelated with u_i and disregard the variations in p_i that bias the LSE.
- Take one supply shifter z_i, e.g., weather, which can be considered to be uncorrelated with the demand shifter u_i such as consumer's tastes, then

$$Cov(z_i, u_i) = 0, \text{ and } Cov(z_i, p_i) \neq 0.$$

So

$$Cov(z_i, q_i) = \alpha_1 \cdot Cov(z_i, p_i),$$

and

$$\alpha_1 = \frac{Cov(z_i, q_i)}{Cov(z_i, p_i)}.$$

- A natural estimator is

$$\widehat{\alpha}_1 = \frac{\widehat{Cov}(z_i, q_i)}{\widehat{Cov}(z_i, p_i)},$$

which is the IV estimator implicitly defined in Appendix B of Philip (1928).

History of the Instrumental Variable Regression

Philip G. Wright (1861-1934), Lombard College

Sewall G. Wright (1889-1988), Chicago

- Sewall Wright is also famous for path analysis (a key step for causal effects evaluation), which will be touched at the end of this chapter.

- Another method to estimate α_1 as suggested above is to run regression

$$q_i = \alpha_0 + \alpha_1 \widehat{p}_i + \widetilde{u}_i,$$

where $\widehat{p}_i$ is the predicted value from the following regression:

$$p_i = \gamma_0 + \gamma_1 z_i + \eta_i,$$

and $\widetilde{u}_i = \alpha_1 (p_i - \widehat{p}_i) + u_i$.

- It is easy to show that $Cov(\widehat{p}_i, \widetilde{u}_i) = 0$, so the estimation is consistent.
- Such a procedure is called two-stage least squares (2SLS) for an obvious reason.
- In this case, the IV estimator and the 2SLS estimator are numerically equivalent.

- Mundlak (1961) considered the production function estimation, where the error term includes factors that are observable to the economic agent under study but unobservable to the econometrician, and endogeneity arises when regressors are decisions made by the agent on the basis of such factors.
- Suppose that a farmer is producing a product with a Cobb-Douglas technology:

$$Q_i = A_i \cdot (L_i)^{\phi_1} \cdot \exp(v_i), 0 < \phi_1 < 1, \tag{4}$$

  where Q_i is the output on the ith farm, L_i is a variable input (labor), A_i represents an input that is fixed over time (soil quality), and v_i represents a stochastic input (rainfall), which is not under the farmer's control.
- We shall assume that the farmer knows the product price p and input price w, which do not depend on his decisions, and that he knows A_i but econometricians do not.
- The factor input decision is made before knowing v_i, and so L_i is chosen to maximize expected profits. The factor demand equation is

$$L_i = \left(\frac{w}{p}\right)^{\frac{1}{\phi_1 - 1}} (A_i B \phi_1)^{\frac{1}{1-\phi_1}}, \tag{5}$$

  so a better farm induces more labors on it.

- We assume that (A_i, ν_i) is i.i.d. over farms, and A_i is independent of ν_i for each i, so $B = E[\exp(\nu_i)]$ is the same for all i, and the level of output the farm expects when it chooses L_i is $A_i \cdot (L_i)^{\phi_1} \cdot B$.
- Take logarithm on both sides of (4), we have a log-linear production function:

$$\log Q_i = \log A_i + \phi_1 \cdot \log(L_i) + \nu_i.$$

$\log A_i$ is an omitted variable. Equivalently, each farm has a different intercept.

- The LSE of ϕ_1 will converge to

$$\frac{Cov(\log Q_i, \log(L_i))}{Var(\log(L_i))} = \phi_1 + \frac{Cov(\log A_i, \log(L_i))}{Var(\log(L_i))},$$

which is not ϕ_1 since there is correlation between $\log A_i$ and $\log(L_i)$ as shown in (5).

- The following figure shows the effect of $\log A_i$ on ϕ_1 by drawing $E[\log Q | \log L, \log A]$ for two farms. In the figure, the OLS regression line passes through points AB with slope $\frac{\log Q_1 - \log Q_2}{\log L_1 - \log L_2}$, but the true ϕ_1 is $\frac{D-C}{\log L_1 - \log L_2}$. Their difference is $\frac{A-D}{\log L_1 - \log L_2} = \frac{\log A_1 - \log A_2}{\log L_1 - \log L_2}$, which is the bias introduced by the endogeneity of $\log A_i$.

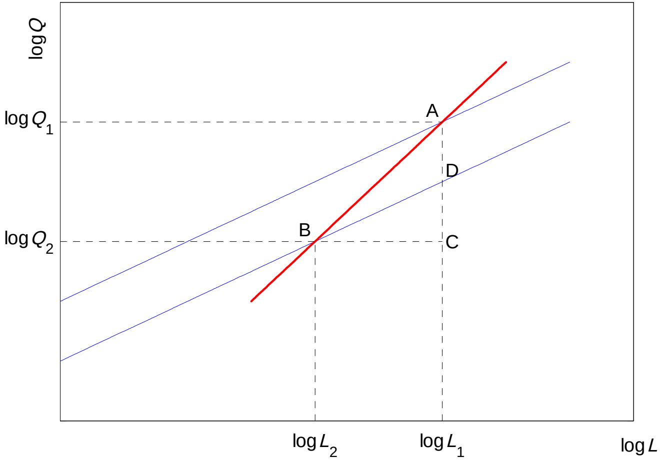


Figure: Effect of Soil Quality on Labor Input

- Rigorously, let $u_i = \log(A_i) - E[\log(A_i)]$, and $\phi_0 = E[\log(A_i)]$, then $E[u_i] = 0$ and $A_i = \exp(\phi_0 + u_i)$.
- (4) and (5) can be written as

$$\log Q_i = \phi_0 + \phi_1 \cdot \log(L_i) + \nu_i + u_i, \tag{6}$$

$$\log L_i = \beta_0 + \frac{1}{1-\phi_1} u_i, \tag{7}$$

where $\beta_0 = \frac{1}{1-\phi_1}\left(\phi_0 + \log(B\phi_1) - \log\left(\frac{w}{p}\right)\right)$ is a constant for all farms.
- It is obvious that $\log L_i$ is correlated with $(\nu_i + u_i)$. Thus, the LSE of ϕ_1 in the estimation of log-linear production function confounds the contribution to output of u_i with the contribution of labor. Actually,

$$\widehat{\phi}_{1,OLS} \xrightarrow{p} 1,$$

because substituting (7) into (6), we get

$$\log Q_i = \phi_0 - (1-\phi_1)\beta_0 + \mathbf{1} \cdot \log(L_i) + \nu_i.$$

- The lesson from this example is that a variable chosen by the agent taking into account some error component unobservable to the econometrician can induce endogeneity.

History of the Hidden Information

Yair Mundlak (1927-2015), Chicago

Errors in Variables

- Measurement errors are embodied in regression analysis from the beginning. Galton (1889) analyzed the relationship between the height of sons and the height of fathers. More specifically,
$$S_i^* = \alpha + \beta F_i^* + u_i, \tag{8}$$
where S_i^* and F_i^* are the heights of sons and fathers, respectively.
- Even if S_i^* should perfectly match F_i^* (that is, $\alpha_0 = 0$, $\beta_0 = 1$ and $u_i = 0$, or $S_i^* = F_i^*$), the OLS estimator would be smaller than 1 if there are environmental factors or measurement errors that affect S_i^* and F_i^*.
- Suppose the observables are $S_i = S_i^* + s_i$, and $F_i = F_i^* + f_i$, where s_i and f_i are the mean-zero environmental factors; then our regression becomes
$$S_i = \alpha + \beta (F_i - f_i) + s_i = \alpha + \beta F_i + s_i - \beta f_i.$$
- The OLS estimator of β will converge to
$$\frac{Cov(F_i, S_i)}{Var(F_i)} = \frac{Var(F_i^*)}{Var(F_i^*) + Var(f_i)} < 1$$
where $\frac{Var(F_i^*)}{Var(F_i^*)+Var(f_i)} \equiv \rho$ is called the reliability coefficient. In Galton's analysis, this coefficient is about $2/3$. He termed this phenomenon as "regression towards mediocrity". [intuition here]

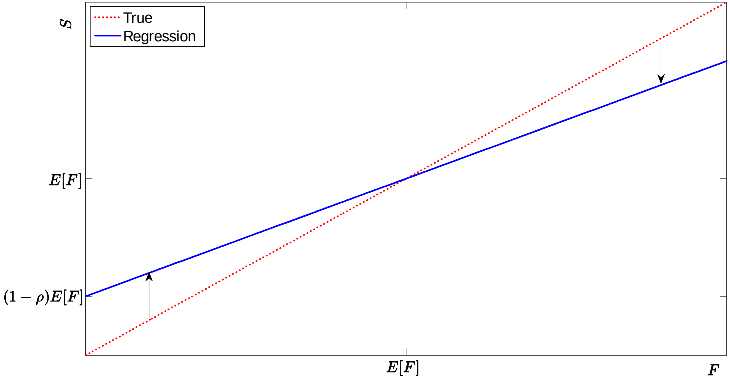
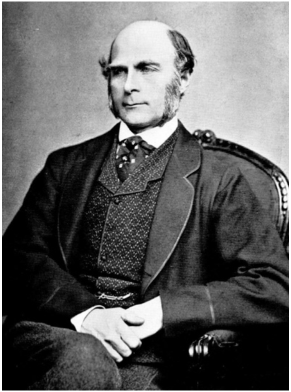


Figure: Relationship Between the Height of Sons and Fathers

History of "Regression"



Sir Francis Galton (1822-1911), English[2]

[2]Galton was Charles Darwin (1809-1882)'s half-cousin, sharing the common grandparent. He was also the advisor of Karl Pearson, (South West) African explorer, and inventor of fingerprinting.

Instrumental Variables

- $y_i = \mathbf{x}_i'\beta + u_i$ is called the structural equation or primary equation. In matrix notation, it can be written as
$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}. \tag{9}$$
- Any solution to the problem of endogeneity requires additional information which we call instrumental variables (or simply instruments).
- The $l \times 1$ random vector $\mathbf{z}_i$ is an instrument for (1) if $E[\mathbf{z}_i u_i] = \mathbf{0}$. This condition cannot be tested in practice since u_i cannot be observed.
- In a typical set-up, some regressors in $\mathbf{x}_i$ will be uncorrelated with u_i (for example, at least the intercept). Thus we make the partition
$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}, \tag{10}$$
where $E[\mathbf{x}_{1i} u_i] = \mathbf{0}$ yet $E[\mathbf{x}_{2i} u_i] \neq \mathbf{0}$. We call $\mathbf{x}_{1i}$ exogenous and $\mathbf{x}_{2i}$ endogenous.

continue...

- By the above definition, $\mathbf{x}_{1i}$ is an instrumental variable, so should be included in $\mathbf{z}_i$, giving the partition

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix} , \tag{11}$$

  where $\mathbf{x}_{1i} = \mathbf{z}_{1i}$ are the included exogenous variables, and $\mathbf{z}_{2i}$ are the excluded exogenous variables.
- In other words, $\mathbf{z}_{2i}$ are variables which could be included in the equation for y_i (in the sense that they are uncorrelated with u_i) yet can be excluded, as they would have true zero coefficients in the equation which means that certain directions of causation are ruled out a priori.
- The model is just-identified if $l = k$ (i.e., if $l_2 = k_2$) and over-identified if $l > k$ (i.e., if $l_2 > k_2$). We have noted that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

Reduced Form

Reduced Form

- The reduced form relationship between the variables or "regressors" $\mathbf{x}_i$ and the instruments $\mathbf{z}_i$ is found by linear projection. Let

$$\boldsymbol{\Gamma} = E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1} E\left[\mathbf{z}_i\mathbf{x}_i'\right]$$

be the $l \times k$ matrix of coefficients from a projection of $\mathbf{x}_i$ on $\mathbf{z}_i$.

- Define

$$\mathbf{v}_i = \mathbf{x}_i - \boldsymbol{\Gamma}'\mathbf{z}_i$$

as the projection error. Note that $\mathbf{v}_i$ must be correlated with u_i. (why?)

- The reduced form linear relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is the instrumental equation

$$\mathbf{x}_i = \boldsymbol{\Gamma}'\mathbf{z}_i + \mathbf{v}_i. \tag{12}$$

In matrix notation,

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}, \tag{13}$$

where $\mathbf{V}$ is a $n \times k$ matrix.

- By construction, $E\left[\mathbf{z}_i\mathbf{v}_i'\right] = \mathbf{0}$, so (12) is a projection and can be estimated by OLS:

$$\mathbf{X} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}} + \widehat{\mathbf{V}}, \widehat{\boldsymbol{\Gamma}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{X}\right).$$

- Substituting (13) into (9), we find

$$\mathbf{y} = (\mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V})\,\beta + \mathbf{u} = \mathbf{Z}\lambda + \mathbf{e} \tag{14}$$

  where $\lambda = \boldsymbol{\Gamma}\beta$ and $\mathbf{e} = \mathbf{u} + \mathbf{V}\beta$.
- Observe that

$$E\left[\mathbf{z}e\right] = E\left[\mathbf{z}\mathbf{v}'\right]\beta + E\left[\mathbf{z}u\right] = \mathbf{0}. \tag{15}$$

  Thus (14) is a projection equation and may be estimated by OLS. This is

$$\mathbf{y} = \mathbf{Z}\widehat{\lambda} + \widehat{\mathbf{e}}, \widehat{\lambda} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right).$$

- The equation (14) is the reduced form for $\mathbf{y}$. (13) and (14) together are the reduced form equations for the system

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\lambda + \mathbf{e}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}. \end{aligned}$$

- The system of equations

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{u}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{V}, \end{aligned}$$

  are called triangular (or recursive) simultaneous equations because the second part of equations do not depend on $\mathbf{y}$.

An Economic Example of Triangular Simultaenous Equations

- Let Y denote individual lifetime earnings and X denote level of education, where we use the capital letters such as X to denote random variables and the corresponding lower case letters such as x denote the potential values they may take.
- The value of X is chosen first, as a function of expected but not of realized Y. The value of Y is determined next, as a function of X, as well as of other observable and unobservable variables.
- In the simple version of such a model,

$$X = \arg\max_{x} \{E[m_1(x,U)|Z,V] - c(x,Z)\},$$

  where $Y = m_1(X,U)$, U is productivity (or ability), $c(X,Z)$ is the cost of education, Z determines the cost of a unit of education, and V is an imperfect signal of U (so correlated with U).
- The solution X is a function, m_2, of Z and V, so we have a recursive model,

$$\begin{aligned} Y &= m_1(X,U), \\ X &= m_2(Z,V). \end{aligned} \tag{16}$$

Identification

Identification

- The structural parameter β relates to $(\lambda, \boldsymbol{\Gamma})$ by $\lambda = \boldsymbol{\Gamma}\beta$.
- This relation can be derived directly by using the orthogonal condition $E\left[\mathbf{z}_i\left(y_i - \mathbf{x}_i'\beta\right)\right] = \mathbf{0}$ which is equivalent to

$$E\left[\mathbf{z}_i y_i\right] = E\left[\mathbf{z}_i \mathbf{x}_i'\right]\beta. \tag{17}$$

  Multiplying each side by an invertible matrix $E\left[\mathbf{z}_i \mathbf{z}_i'\right]^{-1}$, we have $\lambda = \boldsymbol{\Gamma}\beta$.
- The parameter is identified, meaning that it can be uniquely recovered from the reduced form, if the rank condition

$$\operatorname{rank}(\boldsymbol{\Gamma}) = k \tag{18}$$

  holds. Intuitively, this condition requires that $\mathbf{z}$ can perturb $\mathbf{x}$ in all directions.
- If $\operatorname{rank}(E\left[\mathbf{z}_i \mathbf{z}_i'\right]) = l$ (this is trivial), and $\operatorname{rank}(E\left[\mathbf{z}_i \mathbf{x}_i'\right]) = k$ (this is crucial), this condition is satisfied.
- Assume that (18) holds. If $l = k$, then $\beta = \boldsymbol{\Gamma}^{-1}\lambda$. If $l > k$, then for any $\mathbf{A} > 0$, $\beta = \left(\boldsymbol{\Gamma}'\mathbf{A}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\mathbf{A}\lambda$.
- If (18) is not satisfied, then β cannot be uniquely recovered from $(\lambda, \boldsymbol{\Gamma})$.
- Note that a necessary (although not sufficient) condition for (18) is the order condition $l \geq k$.

continue...

- Since $\mathbf{Z}$ and $\mathbf{X}$ have the common variables $\mathbf{X}_1$, we can rewrite some of the expressions.
- Using (10) and (11) to make the matrix partitions $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ and $\mathbf{X} = [\mathbf{Z}_1, \mathbf{X}_2]$, we can partition $\mathbf{\Gamma}$ as

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{pmatrix} = \underset{\quad k_1 \qquad k_2}{\begin{pmatrix} \mathbf{I} & \mathbf{\Gamma}_{12} \\ \mathbf{0} & \mathbf{\Gamma}_{22} \end{pmatrix}} \begin{matrix} k_1 \\ l_2 \end{matrix} .$$

- (13) can be rewritten as

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{Z}_1 \\ \mathbf{X}_2 &= \mathbf{Z}_1 \mathbf{\Gamma}_{12} + \mathbf{Z}_2 \mathbf{\Gamma}_{22} + \mathbf{V}_2 . \end{aligned}$$

- β is identified if $\text{rank}(\mathbf{\Gamma}) = k$, which is true if and only if $\text{rank}(\mathbf{\Gamma}_{22}) = k_2$ (by the upper-diagonal structure of $\mathbf{\Gamma}$). Thus the key to identification of the model rests on the $l_2 \times k_2$ matrix $\mathbf{\Gamma}_{22}$.

(*) What Variable Is Qualified to Be An IV?

- It is often suggested to select an instrumental variable that is

$$\text{(i) uncorrelated with } u\text{; (ii) correlated with endogenous variables.}^3 \qquad (19)$$

- (i) is the instrument exogeneity condition, which says that the instruments can correlate with the dependent variable only indirectly through the endogenous variable.
- (ii) intends to repeat the instrument relevance condition which says that $\mathbf{X}_1$ and the predicted value of $\mathbf{X}_2$ from the regression of $\mathbf{X}_2$ on $\mathbf{Z}_2$ and $\mathbf{X}_1$ are not perfectly multicollinear; in other words, there must be "enough" extra variation in $\widehat{\mathbf{x}}_2$ that can not be explained by $\mathbf{x}_1$. Such a condition is required in the second stage regression.
- Sometimes (19) is misleading.
- Check the following example with only one endogenous variable:

$$\begin{aligned} y &= x_1\beta_1 + x_2\beta_2 + u, \\ E[x_1 u] &= 0,\ E[x_2 u] \neq 0,\ Cov(x_1, x_2) \neq 0. \end{aligned}$$

[3] Of course, we also require the instrument to be excluded from the outcome equation. But mathematically, if z should be included in the outcome equation but is omitted, then $E[zu] = 0$ cannot hold. Maybe this is the most important case of $E[zu] \neq 0$ in practice. See below for the difference in DAG representations of violation of exclusion and exogeneity.

- One may suggest the following instrument for x_2, say, $z = x_1 + \varepsilon$, where ε is some computer-generated random variable independent of the system.[4]
- Now, $E[zu] = 0$ and $Cov(z, x_2) = Cov(x_1, x_2) \neq 0$. It seems that z is a valid instrument, but intuition tells us that it is NOT, since it includes the same useful information as x_1.
- What is missing? We know the right conditions for a random variable to be a valid instrument are

$$\begin{aligned} E[zu] &= 0, \\ x_2 &= x_1\gamma_1 + z\gamma_2 + v \text{ with } \gamma_2 \neq 0. \end{aligned} \tag{20}$$

  In this example, $x_2 = x_1\gamma_1 + z\gamma_2 + v = x_1(\gamma_1 + \gamma_2) + (\varepsilon\gamma_2 + v)$, γ_2 is not identified![5]
- The arguments above indicate that (19) is not sufficient, is it necessary? The answer is still NO!
- For this simple example, can we find some z such that

$$\gamma_2 \neq 0 \text{ but } Cov(z, x_2) = 0?$$

---

[4]WLOG, assume $E[x_1] = E[\varepsilon] = 0$ so that $E[zu] = Cov(z, u)$.

[5]Actually, from the formula of linear projection, γ_2 can be identified as 0.

- Observe that $Cov(z,x_2) = Cov(z,x_1\gamma_1 + z\gamma_2 + v) = Cov(z,x_1)\gamma_1 + Var(z)\gamma_2$, so if $\frac{Cov(z,x_1)}{Var(z)} = -\frac{\gamma_2}{\gamma_1}$, this could happen.
- That is, although z is not correlated with x_2, z is correlated with x_1, and x_1 is correlated with x_2. In mathematical language, $Cov(z,x_1) \neq 0$, $\gamma_1 \neq 0$.
- In such a case, z is related to x_2 only indirectly through x_1. If we assume $Cov(z,x_1) = 0$, or $\gamma_1 = 0$, then the assumption $Cov(z,x_2) \neq 0$ is the right condition for z to be a valid instrument.
- So the right condition should be that z is **partially** correlated with x_2 after netting out the effect of x_1.
- In general, a **necessary** condition for a set of qualified instruments is that at least one (need not be the same one) instrument appears in each of the first-stage regression.
  - When $k = l$, each instrument must appear in at least one endogenous regression (why?).

How to Select Instruments?

- Generally speaking, good instruments are not selected based on mathematics, but based on economic theory.
- In the return to schooling example, the usual practice in the literature is to seek instruments which proxy, or are correlated with, costs of schooling.
  - Angrist and Krueger (1991) propose using quarter of birth as an IV for education in the analysis of returns to schooling because of a mechanical interaction between compulsory school attendance laws and age at school entry.[6]
  - Butcher and Case (1994) use the sex of siblings, in particular whether a girl has any sisters, as an IV to estimate the schooling return to women because the gender of siblings may affect the cost of investing in a child's human capital through the existence of borrowing constraints if there are exogenous gender differences in the return to human capital.[7]

[6] Children born earlier in the year enter school at an older age (e.g., for many states, children turning six by January 1 can enter the primary school on September 1) and are therefore allowed to drop out (on their 16th or 17th birthday) after having completed less schooling than children born later in the year. The exclusion condition may fail because children born in the first quarter are a few months older than other children, and at vey young ages a difference of a few months might be an advantage in performance in school. This indicates that the estimator based on this IV may underestimate the return to schooling (exercise).

[7] Surprisingly, they find that girls who have any sisters, conditional on the number of siblings, have lower school attainment than do girls with no sisters; on the other hand, the school attainment of boys is found to be unrelated to gender composition. This may be because parents prefer a "gender mix".

- - Card (1995) uses college proximity as an instrument to identify the returns to schooling, noting that living close to a college during childhood may induce some children to go to college but is unlikely to directly affect the wages earned in their adulthood.
- In development economics,
  - Acemoglu, Johnson and Robinson (2001) use the mortality rates (of soldiers, bishops, and sailors) as an IV to estimate the effect of property rights and institutions on economic development.
- In political economics,
  - Levitt (1997) uses the timing of mayoral and gubernatorial elections as an IV to identify the causal effect of police on crime by arguing that after controlling some economic variables such as state unemployment rates and spending on public welfare or education this IV does not affect the crime rate but will affect the number of police officers.
- Deaton (2010): Exogeneity is different from externality (not set or caused by the variables in the model). The former is not guaranteed by the latter.
  - The instruments above are external, but exogenous? [see the draft lottery example below]

History of Institutions on Growth

Daron Acemoglu (1967-), MIT, NP2024

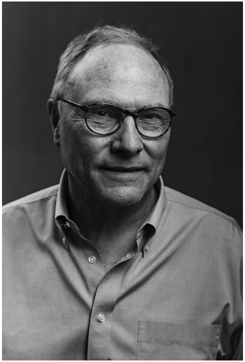
Simon Johnson (1963-), MIT, NP2024

James A. Robinson (1960-), Chicago, NP2024

- Acemoglu, Johnson and Robinson (2001) is their most cited paper.

Three Other Nobel Laureates

Joshua D. Angrist (1960-), MIT, NP2021



David Card (1956-), Berkeley, NP2021[8]

Angus Deaton (1945-), Princeton, NP2015

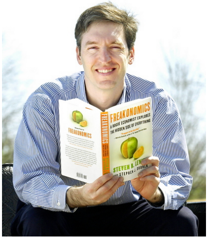
[8] Both Card and Angrist were supervised by Orley Ashenfelter (1942-) at Princeton. Card has many good students, e.g., David S. Lee, Justin R. McCrary, Thomas Lemieux, Michael B. Greenstone, Kenneth Chay, and Kristin Butcher among others; he is also the second advisor of Angrist.

Authors Except Nobel Laureates

Alan B. Krueger (1960-2019), Princeton

Kristin F. Butcher (?-), Fed. of Chicago

Anne Case (1958-), Princeton



Steven D. Levitt (1967-), Chicago

Estimation: Two-Stage Least Squares

IV Estimator

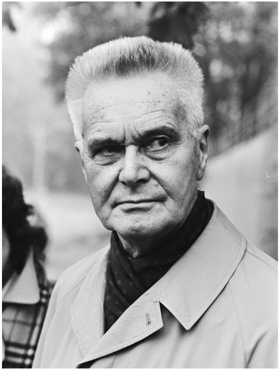
- If $l = k$, then the moment condition is $E\left[\mathbf{z}_i\left(y_i - \mathbf{x}_i'\beta\right)\right] = \mathbf{0}$, and the corresponding IV estimator is a MoM estimator:

$$\widehat{\beta}_{IV} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right).$$

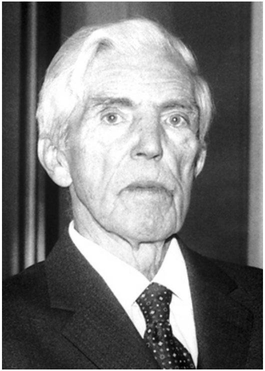
- Another interpretation stems from the fact that since $\beta = \mathbf{\Gamma}^{-1}\lambda$, we can construct the Indirect Least Squares (ILS) estimator:

$$\widehat{\beta} = \widehat{\mathbf{\Gamma}}^{-1}\widehat{\lambda} = \left(\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}\right) = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right).$$

History of the ILS



Jan Tinbergen (1903-1994), Dutch, NP1969



Trygve Haavelmo (1911-1999), Oslo, NP1989

2SLS Estimator as An IV Estimator

- When $l > k$, the two-stage least squares (2SLS) estimator can be used.
- Given any k instruments out of $\mathbf{z}$ or its linear combinations can be used to identify β, the 2SLS chooses those that are most highly (linearly) correlated with $\mathbf{x}$.
- It is the sample analog of the following implication of $E[\mathbf{z}u] = \mathbf{0}$:

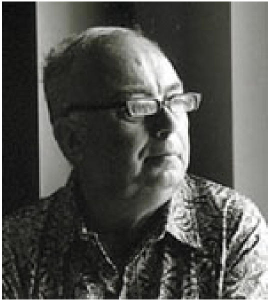
$$\mathbf{0} = E\left[E^*\left[\mathbf{x}|\mathbf{z}\right]u\right] = E\left[\mathbf{\Gamma}'\mathbf{z}u\right] = E\left[\mathbf{\Gamma}'\mathbf{z}(y - \mathbf{x}'\beta)\right], \tag{21}$$

  where $E^*\left[\mathbf{x}|\mathbf{z}\right]$ is the linear projection of $\mathbf{x}$ on $\mathbf{z}$.
- Replacing population expectations with sample averages in (21) yields

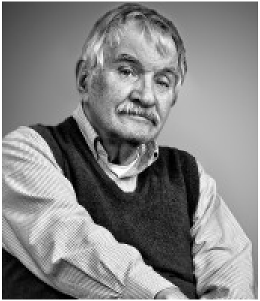
$$\widehat{\beta}_{2SLS} = \left(\widehat{\mathbf{X}}'\mathbf{X}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{y},$$

  where $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\mathbf{\Gamma}} \equiv \mathbf{PX}$ with $\widehat{\mathbf{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$ and $\mathbf{P} = \mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. In other words, the 2SLS estimator is an IV estimator with the IVs being $\widehat{\mathbf{x}}_i$.
- When $l = k$, the 2SLS estimator and the IV estimator are numerically equivalent (why?).

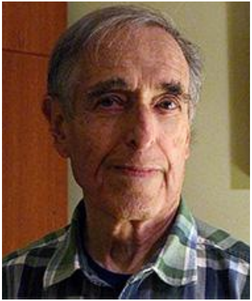
History of the 2SLS



Henri Theil (1924-2000)
Chicago and Florida



Robert Basmann (1926-2024)
TAMU and Bringhamton



Lester Telser (1931-2022)
Chicago

Theil (1953)'s Formulation of 2SLS

- The source of the name "two-stage" is from Theil (1953)'s formulation of 2SLS.
- From (15),

$$\mathbf{0} = E\left[E^*[\mathbf{x}|\mathbf{z}](u + \mathbf{v}'\beta)\right] = E\left[\left(\mathbf{\Gamma}'\mathbf{z}\right)\left(y - \mathbf{z}'\mathbf{\Gamma}\beta\right)\right],$$

  i.e., β is the least squares regression coefficients of the regression of y on fitted values of $\mathbf{\Gamma}'\mathbf{z}$, so this method is often called the fitted-value method.
- The sample analogue is the following two-step procedure:
  1. First, regress $\mathbf{X}$ on $\mathbf{Z}$ to get $\widehat{\mathbf{X}}$.
  2. Second, regress $\mathbf{y}$ on $\widehat{\mathbf{X}}$ to get

$$\widehat{\beta}_{2SLS} = \left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{y} = \left(\mathbf{X}'\mathbf{P}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}\mathbf{y}\right). \tag{22}$$

Basmann (1957)'s version of 2SLS

- Basmann (1957)'s version of 2SLS is motivated by observing that $E[\mathbf{z}u] = \mathbf{0}$ implies

$$0 = E^*[u|\mathbf{z}] = E^*[y|\mathbf{z}] - E^*[\mathbf{x}|\mathbf{z}]'\beta,$$

so

$$\widehat{\beta}_{2SLS} = \left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\widehat{\mathbf{y}}.$$

- Equivalently, $\widehat{\beta}_{2SLS} = \arg\min_{\beta}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{P}_{\mathbf{Z}}(\mathbf{y} - \mathbf{X}\beta)$, which is a GLS estimator.
- Intuitively, $\mathbf{P}_{\mathbf{Z}}(\mathbf{y} - \mathbf{X}\beta)$ should converge in probability to zero because $E[\mathbf{z}u] = \mathbf{0}$, so we try to find some β value such that the length of $\mathbf{P}_{\mathbf{Z}}(\mathbf{y} - \mathbf{X}\beta)$ is as close to zero as possible.

Telser (1964)'s version of 2SLS

- Telser (1964)'s control function formulation:

$$\begin{pmatrix} \widehat{\beta}_{2SLS} \\ \widehat{\rho}_{2SLS} \end{pmatrix} = \left(\widehat{\mathbf{W}}'\widehat{\mathbf{W}}\right)^{-1}\widehat{\mathbf{W}}'\mathbf{y},$$

  where $\widehat{\mathbf{W}} = [\mathbf{X}, \widehat{\mathbf{V}}]$.
- This construction exploits another implication of $E[\mathbf{z}u] = \mathbf{0}$:

$$E^*[u|\mathbf{x},\mathbf{z}] = E^*[u|\Gamma'\mathbf{z}+\mathbf{v},\mathbf{z}] = E^*[u|\mathbf{v},\mathbf{z}] = E^*[u|\mathbf{v}] \equiv \mathbf{v}'\rho$$

  for some coefficient vector ρ, where the third equality follows from the orthogonality of both error terms u and $\mathbf{v}$ with $\mathbf{z}$ (why? Exercise).
- So

$$E^*[y|\mathbf{x},\mathbf{z}] = E^*[\mathbf{x}'\beta + u|\mathbf{x},\mathbf{z}] = \mathbf{x}'\beta + E^*[u|\mathbf{x},\mathbf{z}] = \mathbf{x}'\beta + \mathbf{v}'\rho.$$

- Thus, this particular linear combination of the first-stage errors $\mathbf{v}$ is a function that controls for the endogeneity of the regressors $\mathbf{x}$; one can think of $\mathbf{v}$ as proxying for the factors in u that are correlated with $\mathbf{x}$.
- From the FWL theorem, $\widehat{\beta}_{2SLS}$ is the effect of the net variation in $\mathbf{x}$ on y after excluding the variation in $\widehat{\mathbf{v}}$, while the net variation in $\mathbf{x}$ comes from $\mathbf{z}$ because $\mathbf{x} = \widehat{\Gamma}'\mathbf{z} + \widehat{\mathbf{v}}$.

- Recall that $\mathbf{Z} = [\mathbf{X}_1, \mathbf{Z}_2]$ and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, so

$$\widehat{\mathbf{X}} = [\mathbf{P}\mathbf{X}_1, \mathbf{P}\mathbf{X}_2] = [\mathbf{X}_1, \mathbf{P}\mathbf{X}_2] = \left[\mathbf{X}_1, \widehat{\mathbf{X}}_2\right],$$

  since $\mathbf{X}_1$ lies in the span of $\mathbf{Z}$.
- Thus in the second stage, we regress $\mathbf{y}$ on $\mathbf{X}_1$ and $\widehat{\mathbf{X}}_2$. So only the endogenous variables $\mathbf{X}_2$ are replaced by their fitted values:

$$\widehat{\mathbf{X}}_2 = \mathbf{Z}_1\widehat{\mathbf{\Gamma}}_{12} + \mathbf{Z}_2\widehat{\mathbf{\Gamma}}_{22}.$$

- Note that as a linear combination of $\mathbf{z}$, $\widehat{\mathbf{x}}_2$ is not correlated with u and it is often interpreted as the part of $\mathbf{x}_2$ that is uncorrelated with u.

- In the control function formulation of 2SLS, only $\widehat{\mathbf{v}}_2 = \mathbf{x}_2 - \widehat{\mathbf{x}}_2$ should be added to the regression since $\widehat{\mathbf{v}}_1 = \mathbf{x}_1 - \widehat{\mathbf{x}}_1 = \mathbf{x}_1 - \mathbf{x}_1 = \mathbf{0}$.
- $\mathbf{x}_2 = \mathbf{\Gamma}'_{12}\mathbf{z}_1 + \mathbf{\Gamma}'_{22}\mathbf{z}_2 + \mathbf{v}_2$ implies $\mathbf{v}_2 = \mathbf{x}_2 - \mathbf{\Gamma}'_{12}\mathbf{z}_1 - \mathbf{\Gamma}'_{22}\mathbf{z}_2$, so the rank condition that $\text{rank}(\mathbf{\Gamma}_{22}) = k_2$ guarantees that there is separate variation in $\mathbf{v}_2$ from $\mathbf{x} = \left(\mathbf{z}'_1, \mathbf{x}'_2\right)'$ in the regression of y on $\mathbf{x}$ and $\mathbf{v}_2$.

The Wald (1940) Estimator - A Special IV Estimator

- The Wald estimator is a special IV estimator when the single instrument z is binary.
- Suppose we have the model

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u, \, Cov(x,u) \neq 0, \\ x &= \gamma_0 + \gamma_1 z + v. \end{aligned}$$

- The identification conditions are

$$Cov(z,x) \neq 0, Cov(z,u) = 0. \text{(why?)} \tag{23}$$

- It can be shown that the IV estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n} (z_i - \overline{z})(x_i - \overline{x})}.$$

History of the Wald Estimator

Abraham Wald (1902-1950), Columbia

continue...

- If z is binary that takes the value 1 for n_1 of the n observations and 0 for the remaining n_0 observations, then it can be shown that $\widehat{\beta}_1$ is equivalent to

$$\widehat{\beta}_{Wald} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0} \xrightarrow{p} \frac{E\left[y|z=1\right] - E\left[y|z=0\right]}{E\left[x|z=1\right] - E\left[x|z=0\right]},$$

  where $\overline{y}_1$ is mean of y across the n_1 observations with $z=1$, $\overline{y}_0$ is the mean of y across the n_0 observations with $z=0$, and analogously for x.

- Note that the numerator and denominator of $\text{plim}\left(\widehat{\beta}_{Wald}\right)$ are exactly the slope coefficients in the reduced form equations:

$$\begin{aligned} y &= \lambda_0 + \lambda_1 z + e, \\ x &= \gamma_0 + \gamma_1 z + v, \end{aligned}$$

  so the form of $\widehat{\beta}_{Wald}$ is a direct application of ILS.

- A simple interpretation of this estimator is to take the effect of z on y and divide by the effect of z on x.
- The following figure provides some intuition for the identification scheme of the Wald estimator in the linear demand/supply system - the shift in p by z devided by the shift in q by z is indeed a reasonable slope estimator of the demand curve.

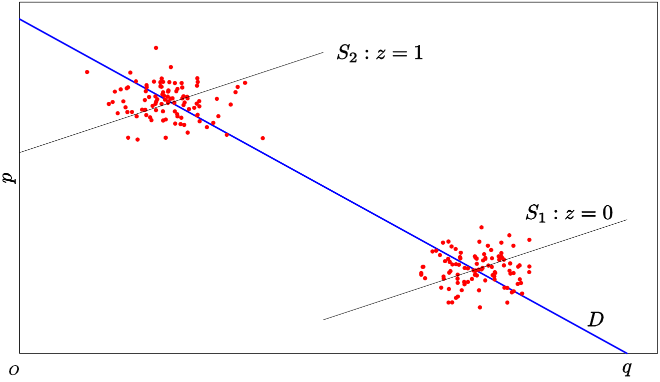


Figure: Intuition for the Wald Estimator in the Linear Demand/Supply System

Some Popular Examples of the Wald Estimator

- In Card (1995), y is the log weekly wage, x is years of schooling S, and z is a dummy which equals 1 if born in the neighborhood of an university and 0 otherwise.
- In studying the returns to schooling in China, Giles et al. (2003) used a dummy indicator of living through the Cultural Revolution or not as z.
- Angrist and Evans (1998) use the dummy of whether the sexes of the first two children are the same, which indicates the parental preferences for a mixed sibling-sex composition, (and also a twin second birth) as the instrument to study the effect of a third child on employment, hours worked and labor income.
- Angrist (1990) uses the Vietnam era draft lottery as an instrument for veteran status to identify the effects of mandatory military conscription on subsequent civilian mortality and earnings (via college deferment).[9]

[9] See Heckman (1997) for a critique on the validity of this instrument. Suppose $z \neq x$ is because $x = 0$ although $z = 1$, i.e., draft evaders ($x = 1$ while $z = 0$, the volunteers, seem fine with exclusion although they may anticipate high earnings gains from military service). If this is for medical reasons, or more generally reasons that make these candidates ineligible to serve, then the exclusion assumption seems plausible. If, on the other hand these are individuals fit but unwilling to serve, they may have had to take actions to stay out of the military that could have affected their subsequent civilian labor market careers. Such actions may include extending their educational career, or temporarily leaving the country. Note that these issues are not addressed by the random assignment of the instrument.

Examples of the Wald Estimator

Joshua D. Angrist (1960-), MIT, NP2021

William N. Evans (?-), Notre Dame

James J. Heckman (1944-), Chicago, NP2000

Angus Deaton (1945-), Princeton, NP2015

Interpretation of the IV Estimator

The IV Estimation as a Projection

- For simplicity,
  - assume $k = k_2 = 1$ and $l = l_2 = 1$;
  - discuss the population version of the IV estimator instead of the sample version and denote $\text{plim}\left(\widehat{\beta}_{IV}\right)$ as β_{IV}.
- In this simple case, $x\beta_{IV}$ is the projection of y onto $span(x)$ along $span^{\perp}(z)$; this can be easily seen from $x\beta_{IV} = xE[zx]^{-1}E[zy] \equiv \mathbf{P}_{x\perp z}(y)$
- Since $z \perp u$, this is also the projection of y onto $span(x)$ along u if $\dim\left(span^{\perp}(z)\right) = 1$ as in the following figure.
- In the following figure, $\mathbf{P}_{x\perp z}(y)$ is very different from the orthogonal projection of y onto $span(x)$ - $\mathbf{P}_x(y) \equiv xE[x^2]^{-1}E[xy]$, because z is different from x (otherwise, $E[zu] \neq 0$ since $E[xu] > 0$ in the figure).
- On the other hand, z cannot be orthogonal to x in the figure (which corresponds to the rank condition); otherwise, $\mathbf{P}_{x\perp z}(y)$ is not well defined.
- So z must stay between x and $x^{\perp}$, just as shown in the figure.

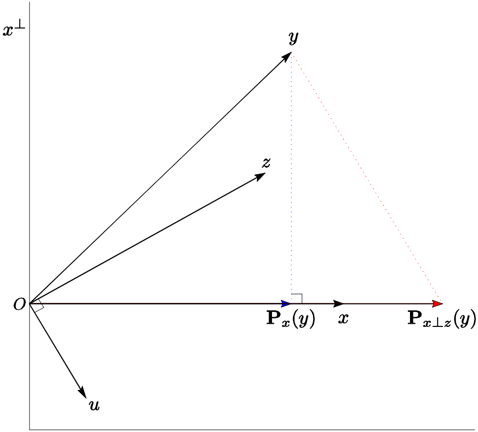


Figure: Projection Interpretation of the IV Estimator

(*) What is the IV Estimator Estimating?

- In the linear model, the IV estimator is estimating β, the *constant* effect of $\mathbf{x}$ on y.
- In a generally nonseparable model (e.g., the equations (16), or $\mathbf{x}$ and y are both binary),

$$\begin{aligned} y &= m(\mathbf{x}, \mathbf{u}), \\ \mathbf{x} &= h(\mathbf{z}, \mathbf{v}), \end{aligned}$$

  the effect of $\mathbf{x}$ on y is heterogenous.[10]
- What is the IV estimator estimating? The local average treatment effect (LATE).
  - Imbens and Angrist (1994) show that the IV estimator is estimating the average treatment effect for those individuals whose $\mathbf{x}$ status is affected by $\mathbf{z}$.
  - This implies that the interpretation of the IV estimator depends on the choice of instruments.

[10] From the discussions below, you will see that the "heterogeneity" here means that the effect of $\mathbf{x}$ on y depends on $\mathbf{v}$.

Joshua D. Angrist (1960-), MIT, NP2021

Guido W. Imbens (1963-), Stanford, NP2021[11]

[11] Imbens is a Dutch econometrician who won the Nobel prize in 2021. Famous Dutch econometricians include Tinbergen, who won the Nobel prize in 1969, Tjalling C. Koopmans (1910-1985), who won the Nobel prize in 1975, Theil, who invented $\bar{R}^2$, 2SLS, k-class estimators, and the multinomial logit model, Herman K. van Dijk, Frank Kleibergen, and Paul Bekker.

Three Traditions of Treatment Effects Evaluation

- Statistics: Potential Outcomes (PO) Approach.
- Economics: Simultaneous Equation Model.[12]
- Computer Science: Structural Causal Models (SCMs) based on path diagrams or graphs especially directed acyclic graphs (DAGs).
  - I will use the LATE to show the differences in the three languages.

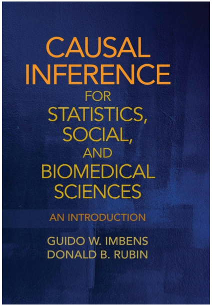
Among econometricians:

- Fresh water: Heckman and his co-authors; emphasizes "causes of effects"; more structural (combining PO and Simultaenous Equation).
- Salt water: Imbens, Card, Angrist, Abandie, $\cdots$; emphasizes "effects of causes"; more reduced-formed (mainly PO).

[12] In sociology, this is termed as Structural Equation Model (SEM) which considers only the linear case. Note also that different from SEMs, simultaneous equations are nondirectional, so are not causal relationships. This is why econometricians have to borrow potential outcome notations from statistics to represent causality.

History of the Potential Outcome Approach

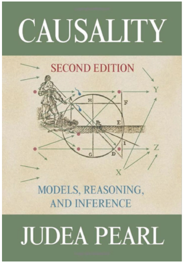
- The potential outcome framework was proposed in Neyman (1923) and Fisher (1925) in experimental studies and was extended to observational studies by Rubin (1974).
- Holland (1986) called this framework as the Rubin Causal Model (RCM).
- For an introduction of RCM, see Rubin (2005, 2008), and for more details, see Imbens and Rubin (2015):



- For the fresh water tradition, see Heckman and Vytlacil (2007a, b) in *Handbook of Econometrics*.

History of Path Analysis

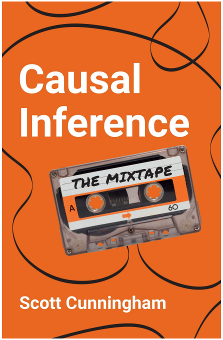
Judea Pearl (1936-), UCLA, Turing2011



Pearl (2009)

- Pearl, J. and D. Mackenzie, 2018, *The Book of Why: The New Science of Cause and Effect*, New York: Basic Books.
- Pearl, J, M. Glymour, and N.P. Jewell, 2016, *Causal Inference in Statistics: A Primer*, West Sussex, England : Wiley.
- Hernán M.A. and J. M. Robins, 2020, *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.
- Morgan, S.L. and C. Winship, 2015, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd edition, New York: Cambridge. [Social Science]
- Peters, J., D. Janzing, and B. Schölkopf, 2017, *Elements of Causal Inference: Foundations and Learning Algorithms*, Cambridge, MA: MIT Press. [Machine Learning]

Econometric Literature on Path Analysis



Cunningham (2021)

- Heckman, J.J., and R. Pinto, 2015, Causal Analysis after Haavelmo, *Econometric Theory*, 31, 115-151.
- Imbens, G.W., 2020, Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics, *Journal of Economic Literature*, 58, 1129-1179.

Notations

- Following the literature, we use Y for y, D (sometimes, T or W) for the treatment x (like the endogenous variable), X for the confounders (like the included IVs), Z for the instruments, and the corresponding lower case letters such as y, d, x and z for the values they may potentially take.
  - In the PO framework, the difference between D and X is that D is manipulable while X is some noncausal attributes. Which variable is D and which is X depends on your purpose, e.g., *gender* or *race* is D or X?[13]
- Treatment means *Ceteris Paribus*, i.e., with all other factors fixed, D changes from 0 to 1 (or generally, increases by one unit).

[13] Instead of changing gender or race, econometricians study the causal effects of interventions like hiding the gender of the job candidate at the time of interview, e.g., Goldin and Rouse (2000) study the effect of blind audition (behind curtains) for orchestras on the hiring of female musicians, or manipulation of the perception of race by changing names from Caucasian-sounding (like Emily and Greg) to African American-sounding ones (like Lakisha and Jamal), e.g., Bertrand and Mullainathan (2004).

How to express the treatment or intervention of D?

- Define Y_d (or $Y(d)$) as potential outcomes as D is assigned d; Y_1 and Y_0 are potential outcomes for the treated group ($D=1$) and control group ($D=0$).
  - $Y(d)$ are sometimes called counterfactuals. A value is counterfactual if it cannot be observed, that is, if it is entirely hypothetical. In this sense, the term "counterfactual" here is not very appropriate since which of $Y(0)$ and $Y(1)$ is counterfactual is not predetermined.
  - The observed outcome $Y=DY_1+(1-D)Y_0$, so only one of Y_1 and Y_0 can be observed. In other words, causal inference is basically a missing data problem, which is referred to as "the fundamental problem of causal inference" in Holland (1986).
  - The average treatment effect (ATE) or average causal effect (ACE) [see more discussions below] is

$$E[Y_1-Y_0]=E[Y_1]-E[Y_0].$$

Continued

- Remove the defining equation of D (i.e., the equation with D on the left side), and change the the value of D in all other equations from 0 to 1.
  - Assume $Y = f_Y(D, X, U_Y)$; then the ATE equals the following burdensome expression
  $$E[f_Y(1, X, U_Y)] - E[f_Y(0, X, U_Y)].$$
  - This is why econometricians borrow the potential outcome notation and write
  $$Y_1 = \mu_1(X, U_1) \text{ and } Y_0 = \mu_0(X, U_0), \tag{24}$$
  where $\mu_d(\cdot,\cdot) = f_Y(d,\cdot,\cdot)$.[14]
- Remove all edges directed into D, and change the value of D from 0 to 1.
  - The ATE is equal to
  $$E[Y|do(D=1)] - E[Y|do(D=0)].$$

[14] The distribution of (X, U_Y) may depend on $D = d$, so we explicitly write out this dependence for U_Y, and implicitly assume $P(X_1 = X_0) = 1$; often, X includes some pretreatment variables, or some characteristics which are not affected by the treatment, like age, sex, etc.

Three Tenets of RCM

- The estimands of causal effects are comparison of potential outcomes on one common set of units, not the treatment potential outcomes for one set of units and the control potential outcomes for a different set.
  - Take $\{X_i, Y_i(0), Y_i(1)\}$ as the "science", and the estimands can be $E[Y_i(1) - Y_i(0)]$ above, $med(Y_i(1) - Y_i(0))$, $\{med(Y_i(1)) - med(Y_i(0))\}|X_i = male$, or $E[\log Y_i(1) - \log Y_i(0)]$, etc.

Units	Covariates X	Potential outcomes: Treatment Y(1)	Potential outcomes: Control Y(0)	Unit-level Causal effects	Summary Causal effects
1	X_1	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ v. $Y_1(0)$	Comparison of $Y_i(1)$ v. $Y_i(0)$ for a common set of units
⋮	⋮	⋮	⋮	⋮	
i	X_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ v. $Y_i(0)$	
⋮	⋮	⋮	⋮	⋮	
N	X_N	$Y_N(1)$	$Y_N(0)$	$Y_N(1)$ v. $Y_N(0)$	

Figure 1. "Science"—The Causal Estimand.

  - In the notations above, we implicitly assume stable unit treatment value assumption (SUTVA) which comprises (i) no interference between units, i.e., $Y_i(\mathbf{D}) = Y_i(\mathbf{D}')$ as long as $D_i = D_i'$, where $\mathbf{D} = (D_1, \cdots, D_n)'$, (ii) no hidden versions of treatments, i.e., $Y_i(\mathbf{D}, \mathbf{V}) = Y_i(\mathbf{D}', \mathbf{V}')$ as long as $D_i = D_i'$, where $\mathbf{V} = (V_1, \cdots, V_n)'$ is the versions of treatments for these n units.

Continued

- We need to posit an assignment mechanism, a model for how units were assigned the treatments they received, i.e., $P(D|X, Y(0), Y(1))$.
  - For inference of treatment effects, this assignment mechanism is enough, and a model for the underlying data, $P(X, Y(0), Y(1))$ is not required. Of course, if $P(X, Y(0), Y(1))$ is assumed, we can do more, e.g., derive the distribution of $P(Y_{\text{mis}}|X, Y_{\text{obs}}, D)$ and make Bayesian prediction on the distribution of causal effects.
  - For inference of treatment effects, $\{X_i, Y_i(0), Y_i(1)\}_{i=1}^n$ are treated as fixed, and only $\{D_i\}_{i=1}^n$ are random.
  - Some popular assignment mechanisms include (i) completely randomized experiments with n units among which n_1 treated:

$$P(\mathbf{D}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} 1/C_{n_1}^n, & \text{if } \sum_{i=1}^n D_i = n_1, \\ 0, & \text{otherwise.} \end{cases}$$

  (ii) unconfounded assignment mechanism: $P(D|X, Y(0), Y(1)) = P(D|X)$. (iii) ignorable assignment mechanism: $P(D|X, Y(0), Y(1)) = P(D|X, Y_{\text{obs}})$.[15]
- We need to be explicit about assumptions because human beings are very bad at dealing with uncertainty (which is why there are many paradoxes).

[15] In LATE, D depends on both $Y(0)$ and $Y(1)$ given X, but in a special way. Heckman terms the bias of OLS resulting from $Cov(D, Y(0)|X) \neq 0$ as selection bias and $Cov(D, Y(1) - Y(0)|X) \neq 0$ as essential heterogeneity.

Comparison of Weaknesses of the Three Traditions

- Treat counterfactuals as abstract mathematical objects that are managed by algebraic machinery but not derived from a model (i.e., model-free): view causal inference as a missing-data problem (which is misleading?); hard to explain and test the assumptions, e.g., the unconfoundedness assumption[16] is expressed as $D \perp\!\!\!\perp (Y_1, Y_0)\,|X$, which means that for any d, y_d, y_d' and x,

$$P(D = d|Y_1 = y_1, Y_0 = y_0, X = x) = P\left(D = d|Y_1 = y_1', Y_0 = y_0', X = x\right),$$

  where $\perp\!\!\!\perp$ is read as "is independent of", and $|$ is read as "conditional on".
  - The unconfounedness assumptions are usually made because they justify the use of available statistical methods, not because they are truly believed.
  - In the simultaneous equation tradition, assume $D = \mu_D(X, U_D)$; then unconfoundedness means $U_D \perp\!\!\!\perp (U_1, U_0)\,|X$, which is easier to understand.
- In complicated models, it is hard to identify the causal effects as in SCMs through backdoor and frontdoor criteria in the following slides (or more rigorously, the *do*-calculus).
- Some information is not easy to be embodied in a graph, e.g., linearity, mean independence (rather than conditional independence), simultaneity, shape restrictions like monotonicity and concavity, etc. [I will provide an example at the end of this chapter]

[16] Other names for unconfoundedness include exogeneity, selection-on-observables, ignorability, or simply conditional independence.

Comparison of Strengths of the Three Traditions

- PO and Simultaneous Equation: (i) Weakness of DAG is the strength of PO. (ii) Connect easily to traditional approaches to economic models, such as supply and demand settings where potential outcome functions are the natural primitives. (iii) Many of currently popular identification strategies focus on models with relatively few (sets of) variables, where identification questions have been worked out once and for all. (iv) Account well for treatment effect heterogeneity and incorporate such heterogeneity in estimation and design of optimal policy functions. (v) Connect well with questions of study design, estimation of causal effects, and inference for such effects.
- DAG: (i) Pedagogical: formulating the critical assumptions in a form that captures the way some researchers think of causal relationships, and being a powerful way of illustrating the key assumptions underlying causal models. (ii) Mathematical: the *do*-calculus developed by Pearl can be used to answer causal identification questions in a novel way, particularly for questions in complex models with a large number of variables.
  - The DAG is assumed, but how to create it and is it an accurate description on how this world works (e.g., why is an arrow absent instead of present?)?
- Why is DAG lack of adoption in economics? (i) The merits of PO. (ii) Although DAG is potentially powerful, it lacks substantive empirical examples.

Path Diagram for Unconfoundedness

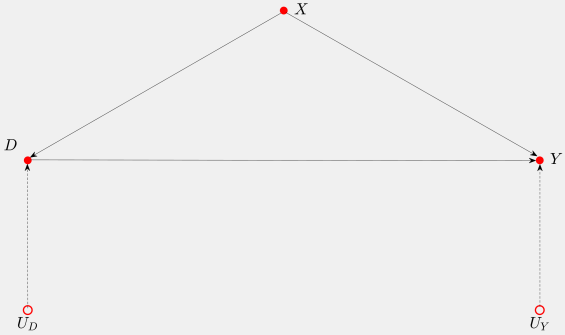


Figure: Causal Diagram for Unconfoundedness

- Solid dot: observable; circle: unobservable.
- U_D and U_Y are independent, and the presence of them does not affect any conclusions, so often omitted.

Assumptions Implied by the Path Diagram

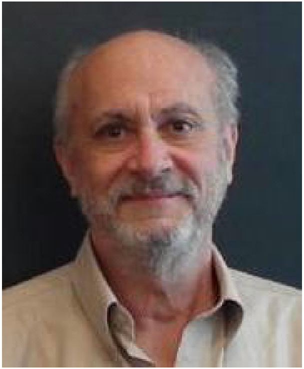
- The DAGs are usually imposed the conditional independence assumptions implied by the so-called d-separation, where d stands for "directional".
- d-separation: A path p is blocked by a set of nodes Z if and only if (i) p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (i.e., B is conditioned on), or (ii) p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z. If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

  - These assumptions are equivalent to the rule of recursive product decomposition, which simplifies the expression of the joint distribution of the variables in the model.
- In the DAG of last slide, the d-separation implies that $U_D \perp\!\!\!\perp U_Y$ and $(U_D, U_Y) \perp\!\!\!\perp X$, which imply $U_D \perp\!\!\!\perp U_Y | X$, i.e., the graph implies some stronger relationships than the conditional independence.

  - No differentiation of Y_1 vs. Y_0 or U_1 vs. U_0, but only Y and U_Y; anyway, the messages intended to deliver are the same.
- $D \perp\!\!\!\perp (Y_1, Y_0) | X$, in combination with the auxiliary assumption that $0 < p(X) := E[D|X] < 1$ (i.e., probabilistic assignment), is referred to as strong ignorability (i.e., probabilistic unconfounded), and D is named "conditionally ignorable given X" in Rosenbaum and Rubin (1983), where $p(X)$ is called the propensity score.

History of Ignorability

Paul R. Rosenbaum (1953-), UPenn



Donald B. Rubin (1943-), Harvard[17]

[17] Rubin has many famous students including Rosenbaum.

Five Weapons in Treatment Effects Evaluation

1. Randomized Controlled Trial (RCT): is the most scientifically rigorous method of hypothesis testing (so-called A/B testing) available, and is regarded as the **gold standard** trial for evaluating the effectiveness of interventions.
2. Backdoor and Frontdoor Criteria: X satisfying the backdoor criterion is exactly the X in the unconfoundedness assumption.
   - Popular estimators under unconfoundedness include, *inter alia*, the matching, subclassification, propensity score weighting and double robust estimators.
3. Instrumental Variables.
4. Difference in Differences (DID): a special panel data solution.
5. Regression Discontinuity Designs (RDDs): a special natural experiment or quasi-experiment.

Recent New Problems:

- Interactions, Spillovers and Peer Effects. [SUTVA fails]
- Big Data: searching for needles in a haystack.

Recent New Tools:

- Machine Learning.
- Synthetic Control.

David Card (1956-), Berkeley, NP2021

LATE: Assumptions

- Suppose for simplicity that Z is a binary assignment. The treatment status $D \neq Z$ is due to noncompliance.
- Assumptions:

1. **Instrument Exclusion**: $P(Y_{d1} = Y_{d0}|X) = 1$ for $d = 1, 0$, where Y_{dz} is the potential outcome of Y when $Z = z$ and $D = d$.
2. **Random Assignment**: $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1) \perp\!\!\!\perp Z|X$, where D_z is the potential treatment status when $Z = z$.
3. Monotonicity (or Uniformity): $P(D_1 \geq D_0|X) = 1$ (or $P(D_1 \leq D_0|X) = 1$). [see the table in the next slide]

- These three assumptions strengthens the instrument exclusion, instrument exogeneity and instrument relevance conditions in linear models.
  - Strictly speaking, the instrument relevance condition if monotonicity is imposed should be $P(D_1 > D_0|X) > 0$.
- $D = Z \cdot D_1 + (1 - Z) \cdot D_0 = D_0 + (D_1 - D_0)Z$, and similarly, $Y = Y_0 + (Y_1 - Y_0)D$.

LATE: Identification

		D_0	
		0	1
D_1	0	$Y_0 - Y_0 = 0$ Never-taker	$Y_0 - Y_1 = -(Y_1 - Y_0)$ Defier
	1	$Y_1 - Y_0$ Complier	$Y_1 - Y_1 = 0$ Always-taker

Table: Causal Effect of Z on Y, $Y_{D_1} - Y_{D_0}$ Classified by D_0 and D_1

- Suppress the dependence on X for simplicity.
- What is the IV estimator (in this case, the Wald estimator, $\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}$) estimating?
- The denominator is

$$E[D|Z=1] - E[D|Z=0] = E[D_1|Z=1] - E[D_0|Z=0]$$
$$\overset{A2}{=} E[D_1 - D_0] = \sum_{\Delta=-1,0,1} \Delta \cdot P(D_1 - D_0 = \Delta) \overset{A3}{=} P(D_1 - D_0 = 1),$$

which is the probability of compliers.
- Note that we can figure out the probability of compliers, but cannot tell whether a specific individual is a complier or not since we can only observe either D_1 or D_0.
- If excluding defiers, then from the table, Z have effects on Y only for compliers.

- The numerator is the intention-to-treat (ITT) effect:

$$
\begin{aligned}
& E[Y|Z=1]-E[Y|Z=0] \\
&= E[Y_0+(Y_1-Y_0)D_1|Z=1]-E[Y_0+(Y_1-Y_0)D_0|Z=0] \\
&\overset{A2}{=} E[Y_0+(Y_1-Y_0)D_1]-E[Y_0+(Y_1-Y_0)D_0] \\
&= E[(Y_1-Y_0)\cdot(D_1-D_0)] \\
&= \sum_{\Delta=-1,0,1} E[(Y_1-Y_0)\cdot(D_1-D_0)\,|D_1-D_0=\Delta]\,P(D_1-D_0=\Delta) \\
&\overset{A3}{=} E[Y_1-Y_0|D_1-D_0=1]\,P(D_1-D_0=1)\,.
\end{aligned}
$$

- In summary, the Wald estimator converges to

$$
\frac{E[Y_1-Y_0|D_1-D_0=1]\,P(D_1-D_0=1)}{P(D_1-D_0=1)}=E[Y_1-Y_0|\text{Compliers}]\,,
$$

the treatment effect for the compliers, which is called the LATE.
- "local" in LATE is relative to the "global" treatment effect of ATE.

Three Assumptions of LATE in Equation Form

1. $Y_1 = \mu_1(X, U_1)$ and $Y_0 = \mu_0(X, U_0)$.
2. $(U_1, U_0, U_D) \perp\!\!\!\perp Z | X$.
3. The participation decision

$$D = 1(U_D \le p(X, Z)),^{18} \tag{25}$$

where $U_D | X, Z \sim U(0,1)$ and the propensity score $p(X, Z)$ satisfies $p(X, 1) \ge p(X, 0)$ a.s. P_X.
- Vytlacil (2002) shows that (25) is equivalent to the monotonicity assumption.

- The three groups of individuals are re-expressed (after suppressing X) as

$$\begin{aligned} \text{Complier} &= \{p(0) < U_D \le p(1)\}, \\ \text{Never-taker} &= \{U_D > p(1)\}, \\ \text{Always-taker} &= \{U_D \le p(0)\}. \end{aligned}$$

[18] Recall that the most general specification of D should be $D = \mu_D(X, Z, U_D)$.

Marginal Treatment Effect (MTE)

- $MTE(u_D) = E[Y_1 - Y_0 | U_D = u_D]$ is the treatment effect for individuals who would be indifferent between treatment or not if they were exogenously assigned a value of Z, say z, such that $p(z) = u_D$.
- The MTE can unify all kinds of treatment effects. For example,

$$LATE = \frac{1}{p(1) - p(0)} \int_{p(0)}^{p(1)} MTE(u_D)\, du_D$$

and

$$ATE = \int_0^1 MTE(u_D)\, du_D.$$

- The average treatment effect on the treated (ATT) can also be expressed in the MTE:

$$ATT = E[Y_1 - Y_0 | D = 1] = \int_0^1 MTE(u_D)\, \omega(u_D)\, du_D,$$

where

$$\omega(u_D) = \frac{1 - F_{p(Z)}(u_D)}{\int_0^1 \left(1 - F_{p(Z)}(t)\right) dt}.$$

- Only if $p(Z) \geq U_D$, $D = 1$, so $\omega(u_D) = P(p(Z) \geq u_D) / P(D = 1)$, which overweights those individuals with low values of u_D that make them more likely to participate in the program.

MTE: Identification by Local IV

- From the expression of LATE as a function of MTE, we can see

$$MTE\left(u_D\right) = \lim_{u_D' \uparrow u_D} \frac{1}{u_D - u_D'} \int_{u_D'}^{u_D} MTE\left(u\right) du \equiv \lim_{u_D' \uparrow u_D} LATE\left(u_D', u_D\right).$$

- Alternatively, since

$$\begin{aligned} E\left[Y|p\left(Z\right)=p\right] &= E\left[Y_1|p\left(Z\right)=p, D=1\right] P\left(D=1|P\left(Z\right)=p\right) \\ &\quad + E\left[Y_0|p\left(Z\right)=p, D=0\right] P\left(D=0|P\left(Z\right)=p\right) \\ &= \int_0^p E\left[Y_1|U_D=u\right] du + \int_p^1 E\left[Y_0|U_D=u\right] du, \end{aligned}$$

where the second equality is because

$$\begin{aligned} E\left[Y_1|p\left(Z\right)=p, D=1\right] &\overset{A3}{=} E\left[Y_1|p\left(Z\right)=p, U_D \le p\left(Z\right)\right] \\ &\overset{A2}{=} E\left[Y_1|U_D \le p\right] = \tfrac{1}{p}\int_0^p E\left[Y_1|U_D=u\right] du, \end{aligned}$$

and similarly for $E\left[Y_0|p\left(Z\right)=p, D=0\right]$, we have

$$MTE\left(u_D\right) = \left.\frac{\partial E\left[Y|p\left(Z\right)=p\right]}{\partial p}\right|_{p=u_D}.$$

- This implies that $MTE\left(u_D\right)$ can be identified only for $u_D \in \text{supp}\left(p\left(Z\right)\right)$, the support of $p\left(Z\right)$.

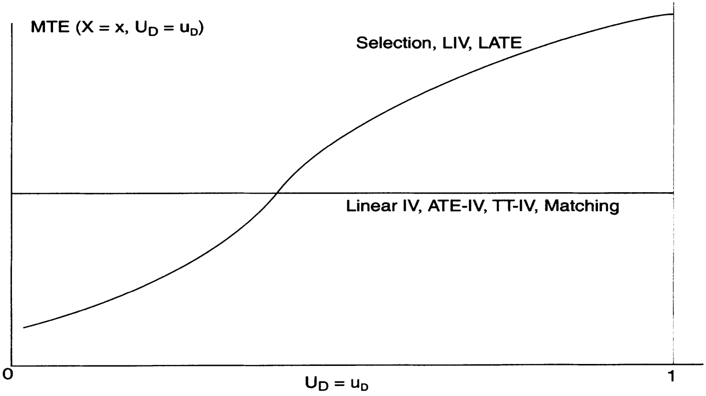


Figure: Comparison of $MTE\left(u_D\right)$ Under Unconfoundedness and Essential Heterogeneity

- $E\left[Y|p\left(Z\right)=p\right]$ is a straight line for linear IV.
- $MTE\left(u_D\right)$ is increasing because a smaller U_D need only a smaller treatment effect to induce participation.

History of the MTE

James J. Heckman (1944-), Chicago, NP2000[19]

Edward J. Vytlacil (1971-), Yale

[19] Heckman has many famous students including Vytlacil.

Path Diagram for LATE

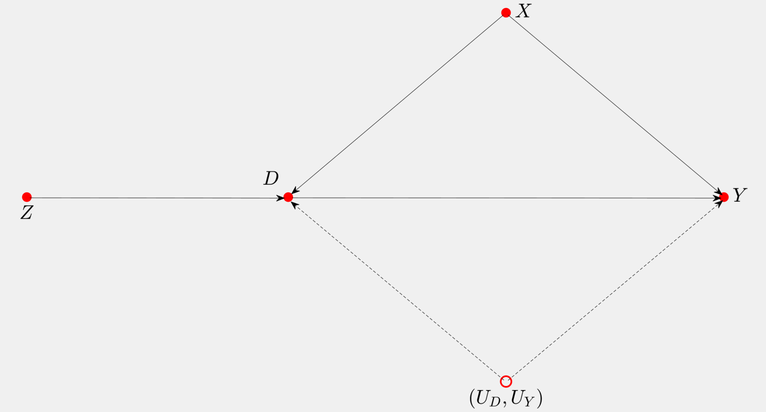


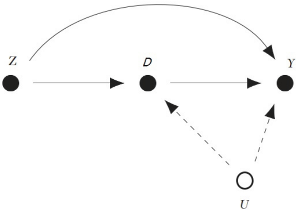
Figure: Causal Diagram for an RCT with Noncompliance

Assumptions Implied by the Path Diagram

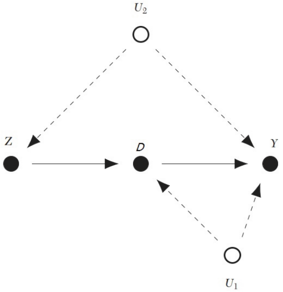
- **Instrument Exclusion**: no edge from node Z to node Y.
- **Random Assignment**: the d-separation implies $(U_D, U_Y) \perp\!\!\!\perp Z$, and even $X \perp\!\!\!\perp (Z, U_D, U_Y)$, so $(U_D, U_Y) \perp\!\!\!\perp Z | X$ hold.
  - (U_D, U_Y) are unobserved confounders for D and Y, which implies U_D and U_Y may be correlated.
  - $D \perp\!\!\!\perp Y | X$ does not hold, so popular estimators under unconfoundedness cannot apply.
- **Monotonicity**: although we can claim Z affects D, the path diagram cannot express Z affects D in the way of $D = 1(U_D \leq p(X, Z))$.

- In the linear case, suppose the ATE of Z on D is a, and that of D on Y is b, then the ITT of Z on Y is ab. In other words, b can be identified by the ITT of Z on Y divided by the ATE of Z on D, which is exactly the Wald estimator when Z is binary.

Violation of Exclusion and Exogeneity Assumptions



Violation of Exclusion



Violation of Exogeneity