New York University
Wilf Family Department of Politics

Quantitative Research in Political Science I

Professor Patrick Egan

Omitted Variables Bias (OVB)

BIVARIATE REGRESSION

linear model	$y = \beta_0 + \beta_1 x + u$
least-squares estimator	$\widehat{\beta}_1 \equiv \frac{cov(x,y)}{var(x)}$
expected value of estimator	$E\left(\widehat{\beta}_1\right) = \beta_1 + \frac{1}{SST_x} E\left[\sum (x_i - \overline{x}) u_i\right]$
conditional independence assumption	$E(u\|x) = 0$
but what if true DGP[1] is:	$y = \beta_0 + \beta_1 x + \mathbf{z}'\gamma + v$
then the OVB formula is:	$\widehat{\beta}_1 = \beta_1 + \gamma' \delta_{zx}$

The scalar $\gamma'\delta_{zx}$ is the product of:

- δ_{zx}, the vector of coefficients from the separate regressions of each of the variables in $\mathbf{z}$ on x, and
- γ, a vector whose elements are the coefficients from the regressions of y on each of the elements of $\mathbf{z}$.

[1] DGP = "data generating process": the (at least partially unobserved) social process that generates y.

MULTIPLE REGRESSION

linear model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{X}$ is an $N \times K$ matrix
least-squares estimator	$\widehat{\boldsymbol{\beta}} \equiv \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$
expected value of estimator	$E\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} + E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\right]$
conditional independence assumption	$E\left(\mathbf{u}\|\mathbf{X}\right) = 0$
but what if true DGP is:	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{v}$, where $\mathbf{Z}$ is an $N \times J$ matrix
then the OVB formula is:	$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\right)\boldsymbol{\gamma}$

Here, $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\right)\boldsymbol{\gamma}$ is a $K \times 1$ "correction vector," the product of:

- $\boldsymbol{\gamma}$, a $J \times 1$ vector whose elements are the coefficients from the regressions of y on each of the J elements of $\mathbf{z}$, and
- $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\right)$, the $K \times J$ matrix whose columns are the coefficients from the J separate regressions of each of the variables in $\mathbf{Z}$ on the K variables in $\mathbf{X}$. We might write this as

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\right) = \begin{bmatrix} \delta_{11} & \delta_{12} & ... & \delta_{1J} \\ \delta_{21} & \ddots & & \\ \vdots & & & \\ \delta_{K1} & & & \delta_{KJ} \end{bmatrix},$$

- Thus we can write

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\right)\boldsymbol{\gamma} \\ \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_K \end{bmatrix} &= \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \delta_{11} & \delta_{12} & ... & \delta_{1J} \\ \delta_{21} & \ddots & & \\ \vdots & & & \\ \delta_{K1} & & & \delta_{KJ} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_J \end{bmatrix} \end{aligned}$$

So what does this tell us about the OVB formula for some generic element of $\widehat{\boldsymbol{\beta}}$, $\widehat{\beta}_k$?

$$\widehat{\beta}_k = \beta_k + \delta_{k1}\gamma_1 + \delta_{k2}\gamma_2 + ...\delta_{kJ}\gamma_J$$

- Thus $\widehat{\beta}_k$ is biased by the sum of J products consisting of:
  - δ_{kj}, the coefficient on x_k in the regression of z_j on $\mathbf{x}$, multiplied by
  - γ_j, the coefficient on z_j in the regression of y on $\mathbf{z}$.