# Econ 361: Advanced Econometrics

"Endogeneity"

# Partial Derivative and *Ceteris Paribus*

$$\frac{dE[Y|X]}{dX_j} = \beta_j + \sum_{s \neq j} \underbrace{\frac{\partial E[Y|X]}{\partial X_s}}_{\beta_s} \frac{\partial X_s}{\partial X_j}$$

Here, $(Y, X)$ refers to the **population** random variables.

$\beta_j$ might be thought as the average impact of $X_j$ on $Y$ after controlling ("fixing") the values of the other $X$'s. By controlling the other $X's$, $\frac{\partial X_s}{\partial X_j} = 0$. this idea of controlling the values of the other $X$'s is analogous to the assumption of *ceteris paribus* often maintained in economic models

# "Violations" of the Linearity Condition

The Linearity Condition, $E[Y|X] = X\beta$, may be "violated" for the regression model *available* to the researcher. There are two main manners by which the "violation" can occur:

1. $E[Y|X]$ is not linear in the parameters
2. $E[Y|X]$ is not linear in the *desired* parameters

Only the first is a real violation. The second just implies that the OLS and GLS models are estimators of parameters *different* from those that are desired.

**Aside:** "Linear in the parameters" ? Consider
$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 (X)^2 + \beta_3 (X)^3$ and $E[Y|X] = \beta_0 + \beta_1 ln(X)$

# $E[Y|X]$ not linear in parameters

In general, there is no reason to believe that $E[Y|X]$ should be linear in the parameters. So this violation may be prevalent.

One comforting result is that if the joint distribution of $\{Y, X\}$ is multivariate Normal, then the conditional distribution $Y|X$ is multivariate Normal, as well, with a conditional mean that is linear in the parameters. See Goldberger Chapter 18.2 for details.

Violations of this form suggest one of two "solutions"

- Run OLS / GLS but interpret the estimates as estimates of the coefficients of the $BLP_{MSE}(Y|X)$
- Use a different, *non-linear* estimation method

We will discuss a few examples of non-linear estimation methods at the end of this course.

# BLP as a Linear Approximation of $E[Y|X]$

We will revisit this perspective later.

Sometimes innocuous, other times not (e.g. linear probability model)

# $E[Y|X]$ not linear in **desired** parameters

The second "violation" is a problem in that neither OLS nor GLS provide estimates of the parameters of economic interest. In the empirical economics literature, this is the most commonly considered "violation" of the Linearity Condition. The problems associated with this "violation" are usually lumped under the title of "endogeneity problem."

Three of the most popular forms of this "endogeneity" are

1. Measurement Error
2. Omitted Variables
3. Simultaneity

# Measurement Error

Measurement error refers to the situation where you do not observe the conditioning variables $X$ but, rather, the conditioning variables with noise $\tilde{X}$. Without loss of generality,

$$\tilde{X} \quad = \quad X + \eta$$

where $\eta$ is the matrix of "noise" contaminating your data on $X$.

# Measurement Error

As you only observe $\tilde{X}$ (and not $X$), you cannot filter out the noise. This implies that a regression model cannot be built around $E[Y|X]$. At best, one can be built around $E[Y|\tilde{X}]$. But even if $E[Y|X] = X\beta$, $E[Y|\tilde{X}] \neq \tilde{X}\beta$ is possible. So applying OLS to $\{Y, \tilde{X}\}$ may not (and generally will not) yield estimates of the desired parameters $\beta$.

Moreover, even if $E[Y|\tilde{X}]$ is linear in $\tilde{X}$, the coefficients for $\tilde{X}$ may not be the same as the coefficients for $X$ in $E[Y|X]$: $E[Y|X] = X\beta$, $E[Y|\tilde{X}] = \tilde{X}\gamma$, and $\gamma \neq \beta$. So applying OLS in this situation provides estimates of $\gamma$ which may not coincide with $\beta$.

# Measurement Error: "White Noise" Example (I of II)

$$\tilde{X} \quad = \quad X + \eta$$

where $\eta$ is statistically independent of $(X, Y)$ and $E[\eta] = 0$ and Var$[\eta] = \sigma_\eta^2 > 0$

Using OLS to regress $Y$ on $\tilde{X}$ yields estimates of the coefficients for the BLP$_{MSE}(Y|\tilde{X}) = \tilde{\alpha} + \tilde{\beta}X$ where

$$\begin{aligned} \tilde{\alpha} &= E[Y] - \tilde{\beta}E[\tilde{X} = E[Y] - \tilde{\beta}E[X] \\ \tilde{\beta} &= \frac{\text{Cov}(\tilde{X}, Y)}{\text{Var}(\tilde{X})} = \frac{\text{Cov}(X, Y)}{\text{Var}(X) + \sigma_\eta^2} \end{aligned}$$

# Measurement Error: "White Noise" Example (II of II)

Note that BLP$_{MSE}(Y|X) = \alpha^* + \beta^*$ where

$$\alpha^* = E[Y] - \beta^* E[X] \qquad \beta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

And $||\tilde{\beta}|| < ||\beta^*||$ as $\sigma_\eta^2 > 0$

As using OLS to regress $Y$ on $\tilde{X}$ results in an estimate of $\tilde{\beta}$, the same OLS regression can be said to result in an estimate of $\beta^*$ that is downward biased (in magnitude). This downward bias in magnitude due to (white noise) measurement error in the $X$ variable is referred to as "**attenuation bias**"

# Omitted Variables

Suppose that the Linearity Condition of economic interest is $E[Y|X_1, X_2] = X_1\beta_1 + X_2\beta_2$. However, you only observe $X_1$ (not $X_2$). Then, a regression model can be built around $E[Y|X_1]$ but not $E[Y|X_1, X_2]$. Using the Law of Iterated Expectations

$$\begin{aligned} E[Y|X_1] &= E_{X_2|X_1}\left[\, E[Y|X_1, X_2]\,\right] = E_{X_2|X_1}[X_1\beta_1 + X_2\beta_2] \\ &= X_1\beta_1 + E[X_2|X_1]\beta_2 \end{aligned}$$

If $E[X_2|X_1]$ is not linear in $X_1$ then neither is $E[Y|X_1]$.

# Omitted Variables

If $E[X_2|X_1]$ is linear in $X_1$, say $E[X_2|X_1] = X_1\gamma_1$ then
$E[Y|X_1] = X_1(\beta_1 + \gamma_1\beta_2)$ and OLS yields estimates not of $\beta_1$ but $\beta_1 + \gamma\beta_2$.

There are two special cases under which $\beta_1 = \beta_1 + \gamma_1\beta_2$:

1. $\beta_2 = 0$ (conditioning on $X_2$ does not matter)
2. $\gamma_1 = 0$ ("$X_2$ is uncorrelated with $X_1$")

   Note that if $X_1$ includes a constant, the constant will be distorted even if $\gamma_1 = 0$. We also need $\gamma_0 = 0$

See Goldberger Chapter 17.5 for details.

# Simultaneity

Related to Omitted Variables.

We will discuss more fully through “Simultaneous Equations Model”

Most famous example of simultaneity in economics? Market equilibrium price and quantity in the standard “supply and demand” model of a competitive market

# Endogenous and Exogenous Variables

Variables in a regression model are categorized not only as $\{$dependent, independent/explanatory$\}$ but also as $\{$endogeneous, exogenous$\}$. Endogeneous/exogenous refers to whether the value of the variable is determined within the regression model or outside the regression model.

- **Endogenous**: value determined within the model
- **Exogenous**: value determined outside the model.

The value of an exogenous variable is **fixed** within the regression model. Changes in the value of other variables (or parameters) within the model do not alter the value of the exogenous variable.

# Endogenous and Exogenous Variables

For "well behaved" regression models, the only endogenous variable should be the dependent variable ("$Y$"). All of the explanatory (conditioning) variables should be exogenous. However, the problems outlined above each imply that one (or more) of the explanatory variables behave *like* an endogenous variable.

Consider the Omitted Variables problem. When the observed $X_1$ is correlated with the unobserved $X_2$, then $X_1$ is forced to play "double duty" in the regression model; $X_1$ must reflect not only how $X_1$ is related to $Y$ but also how $X_2$ is related to $Y$. It is this double duty that leads to OLS estimating a confounded set of parameters, rather than the desired $\beta_1$. Therefore, $X_1$ in the shortened regression model behaves like an endogenous variable whose effective value is determined partly by the value of $X_2$.

Hence, such "violations" of the Linearity Condition are often dubbed "endogeneity problems"

# Instruments and Fixed Effects

The two primary ways by which empirical economists have addressed "endogeneity problems" in recent years

1. Fixed Effects

   includes differencing (e.g. "diff-in-diff")

2. Instruments

   includes randomization and regression discontinuity methods

Both of which work only if you have the "right" kind of data