# Endogenous Regressors

Jake Anderson

May 16, 2026

# The OLS Assumptions Under Random Sampling

| Assumption | Description |
|---|---|
| RS1 | The model is linear: $y_i = \beta_1 + \beta_2 x_i + e_i$ |
| RS2 | The data $(y_i, x_i)$ are i.i.d. |
| **RS3** | **Exogeneity:** $E(e_i \mid x_i) = 0$ |
| RS4 | Homoskedasticity: $\text{Var}(e_i \mid x_i) = \sigma^2$ |
| RS5 | Rank condition: $x_i$ takes at least two values |
| RS6 | Normality: $e_i \sim N(0, \sigma^2)$ |

We have relaxed RS4 (heteroskedasticity), RS6 (non-normality), and RS2 (time series).

$\implies$ Now we tackle **RS3**: what happens when $E(e_i \mid x_i) \neq 0$?

Three sources of endogeneity:

1. Omitted variable bias
2. Measurement error
3. Simultaneity

# Motivation: Test Scores

What predicts test scores? Two plausible factors:

- $X_1$ = prior grade (foundation from last term)
- $X_2$ = study time in current course

The true model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Three researchers run different regressions:

1. "The past is irrelevant!" $\implies$ $Y = \beta_0 + \beta_1 X_1 + e_1$
2. "Effort is all that counts!" $\implies$ $Y = \beta_0 + \beta_2 X_2 + e_2$
3. "Both matter!" $\implies$ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

**Question:** Will $\hat{\beta}_1$ be the same in Model 1 and Model 3?

# A Counterexample: The Room Game

Imagine a game:

- You walk down a hallway with rooms on each side
- Each room you enter gives you 1 point
- Rooms are independent of each other
- Each player randomly decides which rooms to enter (coin flip)

Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where $\beta_1 = \beta_2 = 1$.

Now run the short regression: $Y = \beta_0 + \beta_1 x_1 + e$

$\implies \hat{\beta}_1 \approx 1$ (no bias!)

Why? Because $x_1 \perp x_2$ (room choices are independent). The omitted variable $x_2$ does not bias $\hat{\beta}_1$ when it is **uncorrelated** with $x_1$.

$\implies$ OVB requires both (1) omitted variable affects $Y$, and (2) omitted variable is correlated with included $X$.

# The OVB Formula

**Short regression** (omits $X_2$):

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$$

**Long regression** (includes $X_2$):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

**Auxiliary regression** (how are $X_1$ and $X_2$ related?):

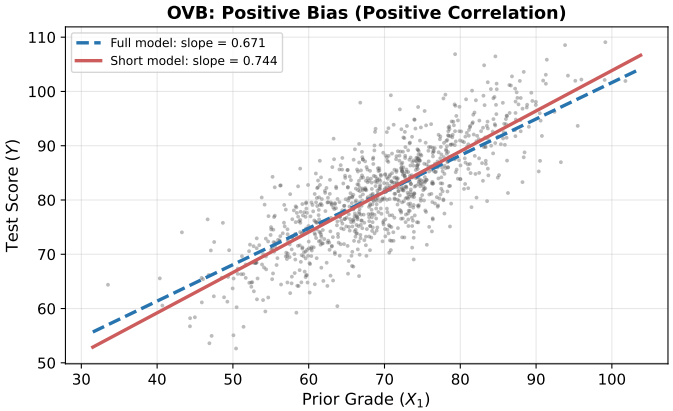
$$X_{2i} = \delta_0 + \delta_1 X_{1i} + v_i$$

**The OVB formula:**

$$\hat{\beta}_1^{\text{short}} = \hat{\beta}_1^{\text{long}} + \underbrace{\hat{\beta}_2}_{\substack{\text{effect of } X_2 \\ \text{on } Y}} \times \underbrace{\hat{\delta}_1}_{\substack{\text{relationship} \\ \text{of } X_2 \text{ to } X_1}}$$

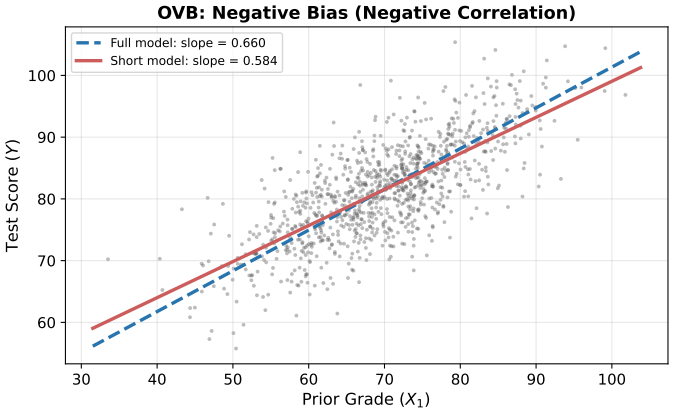
$\implies$ The bias is $\hat{\beta}_2 \times \hat{\delta}_1$. If either is zero, there is no bias.

# Visualizing OVB: Positive Bias




When $X_1$ and $X_2$ are **positively** correlated and $\beta_2 > 0$: the short model overestimates $\beta_1$ because it gives $X_1$ credit for the effect of $X_2$.

# Visualizing OVB: Negative Bias




When $X_1$ and $X_2$ are **negatively** correlated and $\beta_2 > 0$: the short model underestimates $\beta_1$ because the omitted variable works against $X_1$.

# Direction of Bias: The Sign Table

$$\hat{\beta}_1^{\text{short}} = \hat{\beta}_1^{\text{long}} + \hat{\beta}_2 \times \hat{\delta}_1$$

| | $\text{Corr}(X_{\text{inc}}, X_{\text{omit}}) > 0$ | $\text{Corr}(X_{\text{inc}}, X_{\text{omit}}) < 0$ |
|---|---|---|
| $\beta_{\text{omit}} > 0$ | **Positive bias (overestimate)** | **Negative bias (underestimate)** |
| $\beta_{\text{omit}} < 0$ | **Negative bias (underestimate)** | **Positive bias (overestimate)** |

The direction of OVB depends on two signs: (1) the effect of the omitted variable on $Y$, and (2) the correlation between included and omitted regressors.

# OVB Simulation: Test Scores

True DGP: $Y = 30 + 0.65X_1 + 0.40X_2 + \varepsilon$, with $\text{Corr}(X_1, X_2) \approx 0.4$

| | **Model 1** ($X_1$ only) | **Model 2** ($X_2$ only) | **Model 3** (Full) |
|---|---|---|---|
| $\hat{\beta}_1$ (Prior Grade) | $\approx 0.73$ | — | $\approx 0.65$ |
| $\hat{\beta}_2$ (Study Time) | — | $\approx 0.55$ | $\approx 0.40$ |

- Model 1 **overestimates** $\beta_1$: it attributes some of study time's effect to prior grade
- Model 2 **overestimates** $\beta_2$: it attributes some of prior grade's effect to study time
- Model 3 recovers estimates close to the true values

$\Longrightarrow$ Both $\beta > 0$ and $\text{Corr} > 0$, so the bias is positive (overestimation) in both short models.

# Motivation: Study Time and Office Hours

Suppose test scores depend on **true study time** ($x_i^*$):

$$y_i = \beta_1 + \beta_2 x_i^* + v_i$$

But we cannot observe $x_i^*$ directly. Instead, we use a proxy: **office hours attendance** ($x_i$).

The proxy measures true study time with error:

$$x_i = x_i^* + u_i$$

where $u_i$ is measurement error with $E(u_i) = 0$ and $\text{Var}(u_i) = \sigma_u^2$.

Some students study a lot but never come to office hours. Others show up frequently but don't study much otherwise.

$\implies$ Office hours attendance is a **noisy** version of the true variable we care about.

# Why Measurement Error Causes Endogeneity

Substitute $x_i^* = x_i - u_i$ into the true model:

$$y_i = \beta_1 + \beta_2(x_i - u_i) + v_i = \beta_1 + \beta_2 x_i + \underbrace{(v_i - \beta_2 u_i)}_{e_i}$$

Now check whether the regressor $x_i$ is correlated with the composite error $e_i$:

$$\begin{aligned}
\text{Cov}(x_i, e_i) &= \text{Cov}(x_i^* + u_i,\ v_i - \beta_2 u_i) \\
&= \underbrace{\text{Cov}(x_i^*, v_i)}_{=0} - \beta_2 \underbrace{\text{Cov}(x_i^*, u_i)}_{=0} + \underbrace{\text{Cov}(u_i, v_i)}_{=0} - \beta_2 \underbrace{\text{Cov}(u_i, u_i)}_{=\sigma_u^2} \\
&= -\beta_2 \sigma_u^2 \neq 0
\end{aligned}$$

$\implies$ If $\beta_2 > 0$, there is a **negative** correlation between $x_i$ and $e_i$. OLS underestimates $\beta_2$.

# Attenuation Bias Formula

As $N \to \infty$, the OLS estimator converges to:

$$b_2 \xrightarrow{p} \beta_2 \cdot \underbrace{\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_u}}_{\lambda}$$

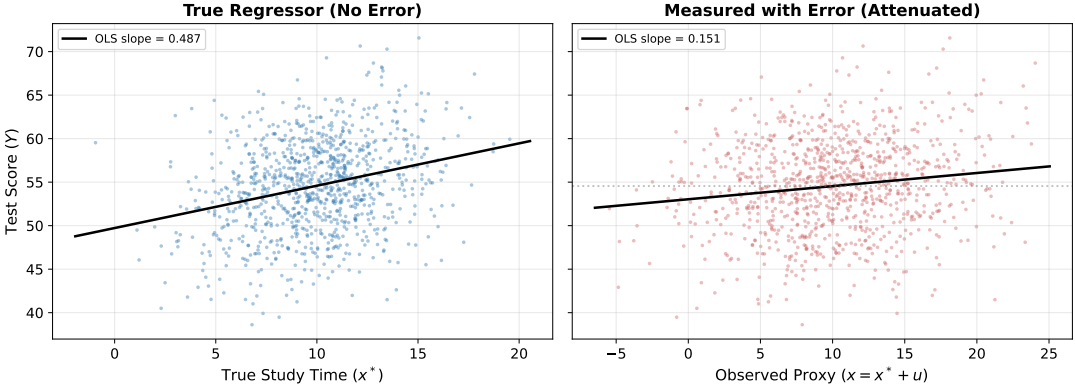
where $\lambda$ is the **reliability ratio**, always between 0 and 1.

Two extreme cases:

- If $\sigma^2_u = 0$ (no error): $\lambda = 1$ and $b_2 \to \beta_2$ (no bias)
- If $\sigma^2_u \to \infty$ (pure noise): $\lambda \to 0$ and $b_2 \to 0$

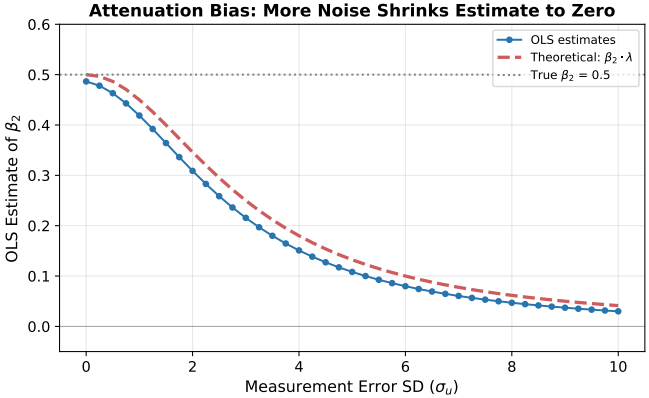
$\implies$ Measurement error **always shrinks the coefficient toward zero**. This is called **attenuation bias**. More data does not help.

# Visualizing Measurement Error



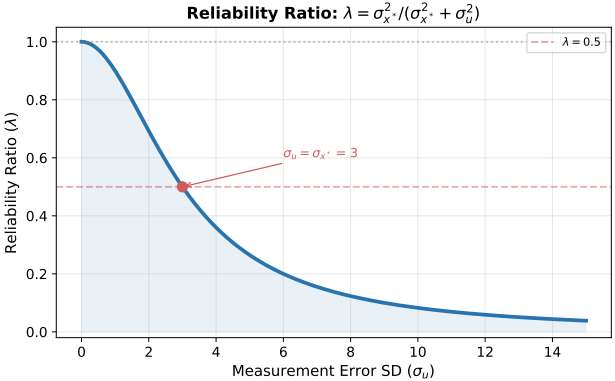

Left: using true $x^*$, OLS recovers the correct slope. Right: using observed $x$, the scatter is wider and the slope is attenuated.

# Simulation: Attenuation Bias in Action



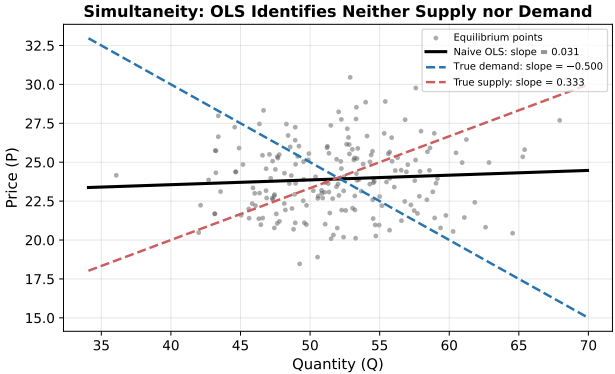

As measurement error ($\sigma_u$) increases, the OLS estimate shrinks toward zero. The theoretical curve $\beta_2 \cdot \lambda$ matches the simulated estimates.

# The Reliability Ratio




When $\sigma_u = \sigma_{x^*}$, the reliability ratio drops to 0.5: OLS captures only half of the true effect. As measurement error grows, the regressor becomes uninformative.

# Simultaneity: A Brief Introduction




When price and quantity are determined simultaneously in equilibrium, we only observe equilibrium points. Naive OLS traces out **neither** the supply curve nor the demand curve.

$\Longrightarrow$ Covered in detail in the **Simultaneous Equations** chapter. Solution: instrumental variables.

# Lagged Dependent Variable Models

A common dynamic model:

$$y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t + e_t$$

$y_{t-1}$ is a **random regressor**. Whether OLS works depends on the errors:

**Case 1: i.i.d. errors** ($e_t$ uncorrelated across time)

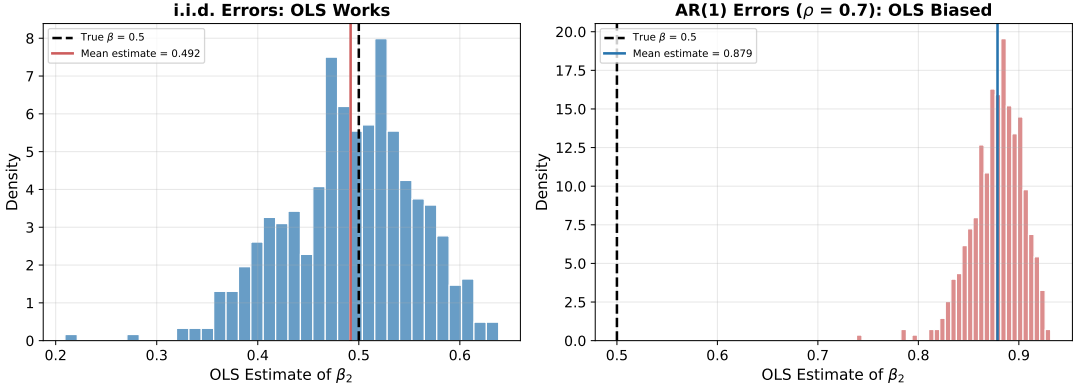
$y_{t-1}$ was determined before $e_t$ is realized $\implies$ $\text{Cov}(y_{t-1}, e_t) = 0$ $\implies$ OLS is consistent.

**Case 2: AR(1) errors** ($e_t = \rho\, e_{t-1} + v_t$)

1. $y_{t-1}$ depends on $e_{t-1}$ (from the equation at time $t-1$)
2. $e_t$ depends on $e_{t-1}$ (from the AR(1) structure)
3. Therefore $\text{Cov}(y_{t-1}, e_t) \neq 0$ when $\rho \neq 0$

$\implies$ Serial correlation in errors makes $y_{t-1}$ endogenous.

# Simulation: Lagged DV with Serial Correlation




Left: with i.i.d. errors, OLS is centered on the true $\beta$. Right: with AR(1) errors ($\rho = 0.7$), OLS is **biased upward**. OLS attributes error persistence to $\beta_2$.

# Testing for Serial Correlation

**Problem:** The standard Durbin-Watson test is invalid when $y_{t-1}$ is a regressor.

**Solution:** Use the **Breusch-Godfrey** test instead.

Procedure:

1. Estimate the model and obtain residuals $\hat{e}_t$
2. Regress $\hat{e}_t$ on all original regressors *plus* $\hat{e}_{t-1}$ (and possibly more lags)
3. Test whether the coefficients on lagged residuals are jointly zero ($\chi^2$ or $F$ test)

If the test rejects:

- The lagged dependent variable is endogenous
- OLS is inconsistent
- $\Longrightarrow$ Consider instrumental variables or other methods

# Three Sources of Endogeneity

| Source | Problem | Bias Direction |
|---|---|---|
| Omitted variables | Correlated omitted factor | Depends on signs |
| Measurement error | Noisy proxy for true $X$ | Toward zero (attenuation) |
| Simultaneity | $X$ and $Y$ jointly determined | Ambiguous |
| Lagged DV + AR(1) | $y_{t-1}$ correlated with $e_t$ | Upward (when $\rho > 0$) |

In every case, $\text{Cov}(x_i, e_i) \neq 0$, violating RS3 (exogeneity).

$\Longrightarrow$ More data does not fix any of these problems. The bias persists even as $N \to \infty$.

**What can we do?**

- **Omitted variables**: include the variable, or use instrumental variables / fixed effects
- **Measurement error**: find a better measure, or use IV
- **Simultaneity**: use simultaneous equations methods (2SLS)
- **Lagged DV**: test with Breusch-Godfrey, use IV if errors are autocorrelated

Thank you!
jakeanderson@g.ucla.edu