Confounding Variables and Endogeneity

POLS 602

Dr. Mike Burnham
Texas A&M Political Science

Endogeneity

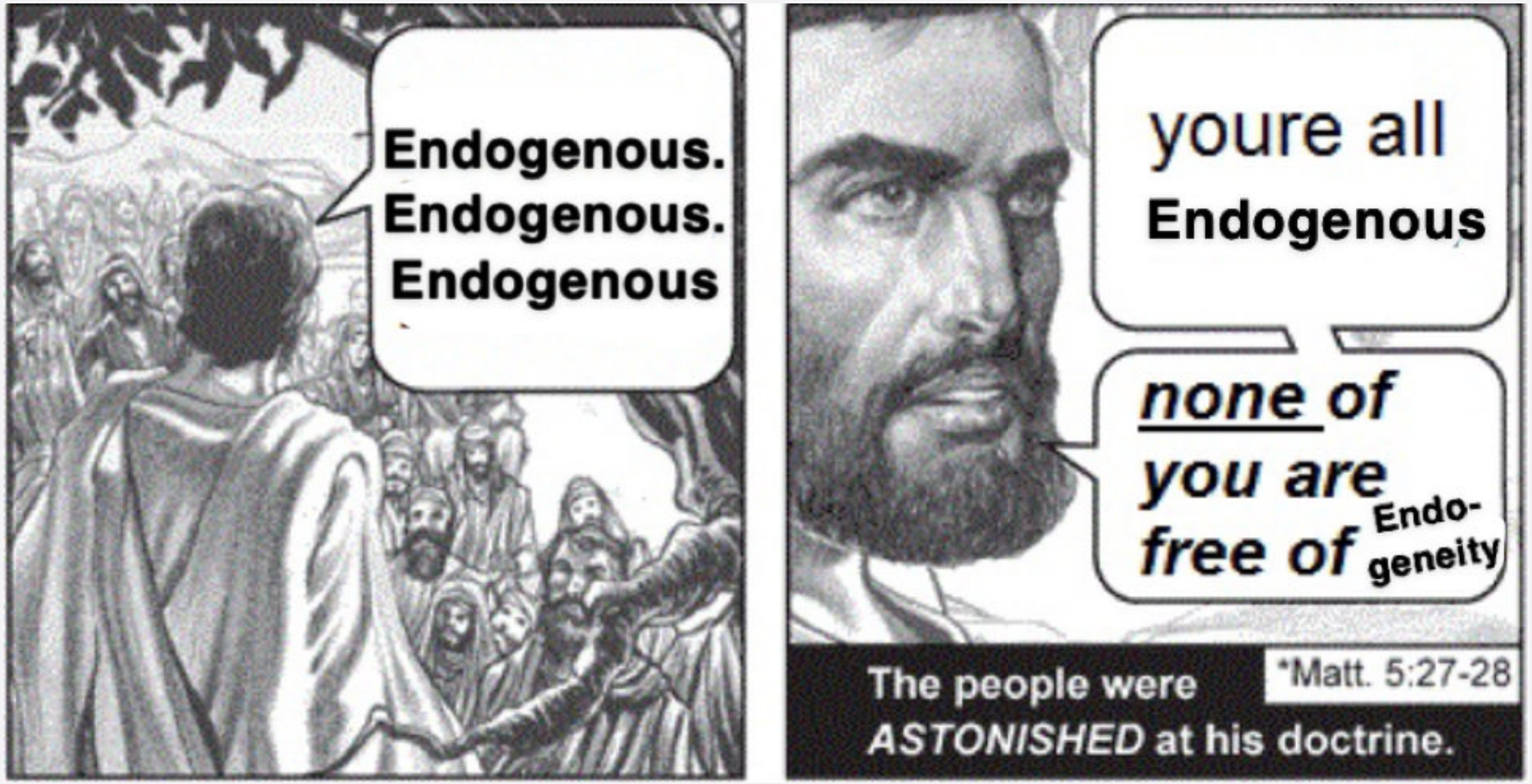
Endogeneity

When explanatory variables are correlated with the error term

Endogeneity

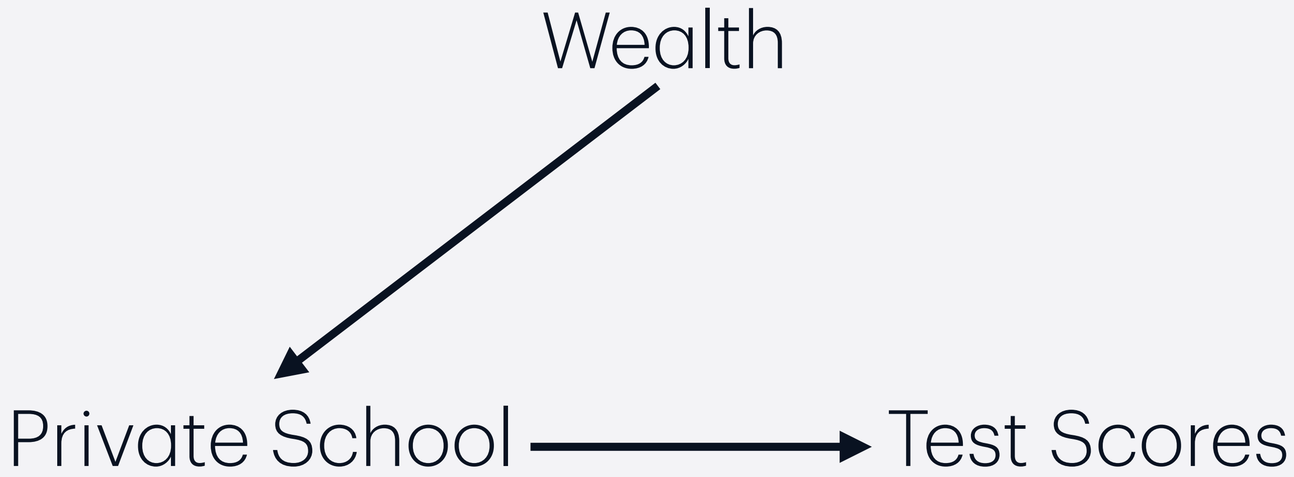
- Means “determined within the model”
- Often associated with a feedback loop between X and Y
- Can be caused by other issues such as omitted variables and measurement error
- When endogeneity is present, your coefficient estimates are **biased** and **inconsistent**
- **Exogeneity** means something is determined outside of the model

Endogeneity is everywhere in observational data

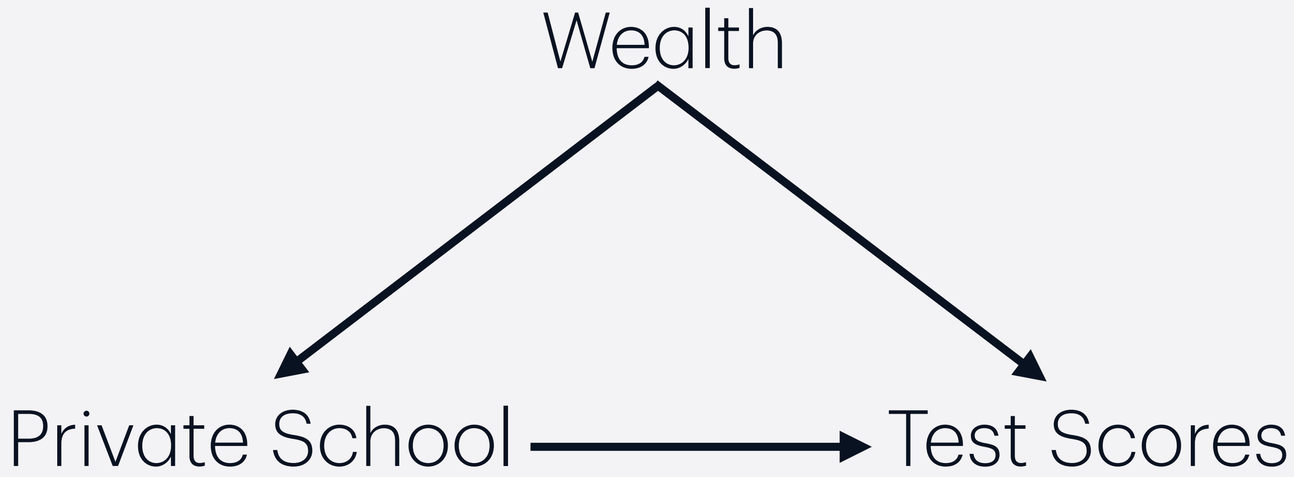


Example: Confounders

Example: Confounders



Example: Confounders



Example: Confounders

1. Suppose the true model is: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ where y_i is a student's test score, x_i is private school, and z_i is wealth.

2. Suppose z_i is omitted from the model: $y_i = \beta_0 + \beta_1 x_i + u_i$

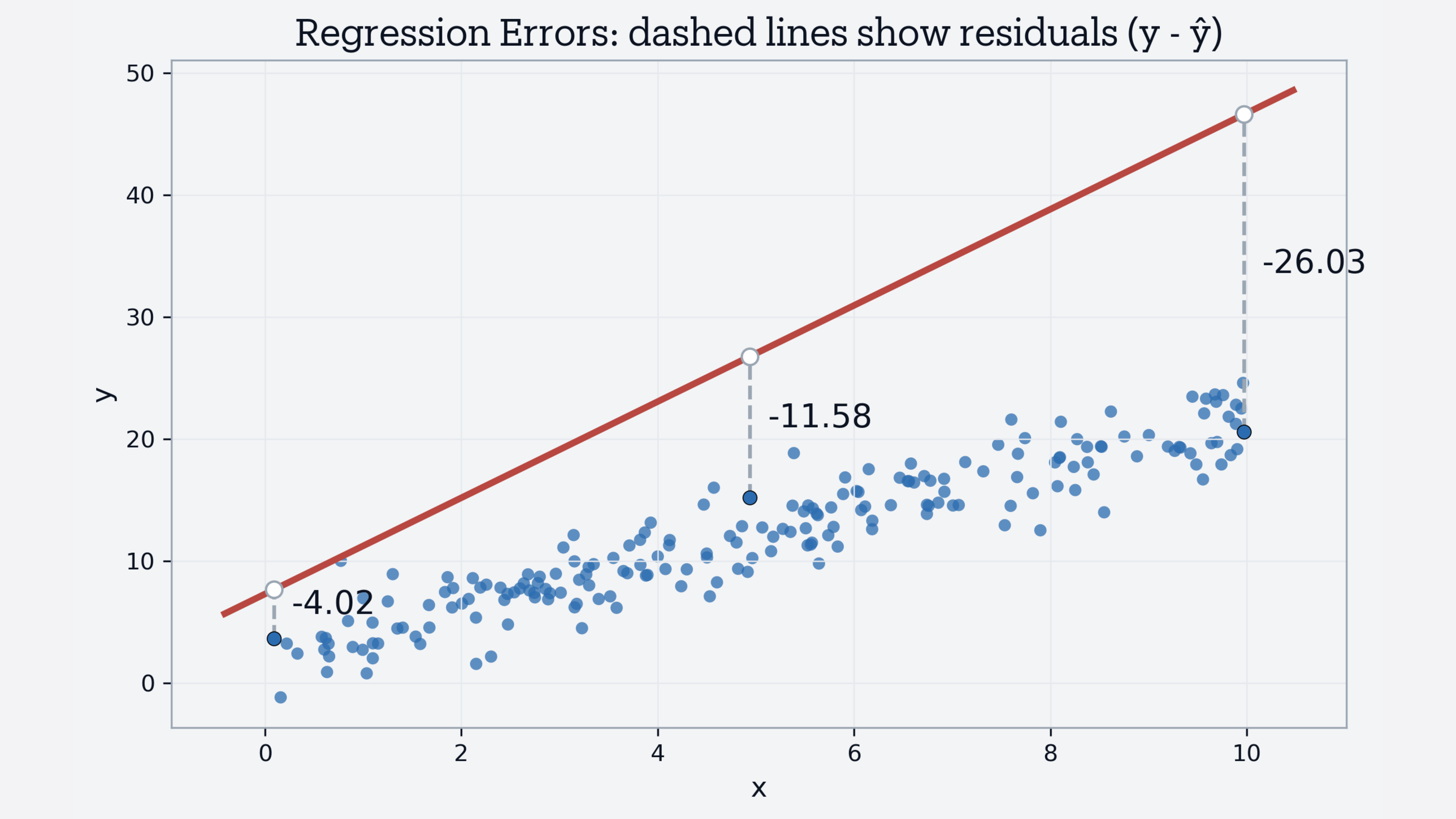
3. Thus: $u_i = \beta_1 z_i + \epsilon_i$

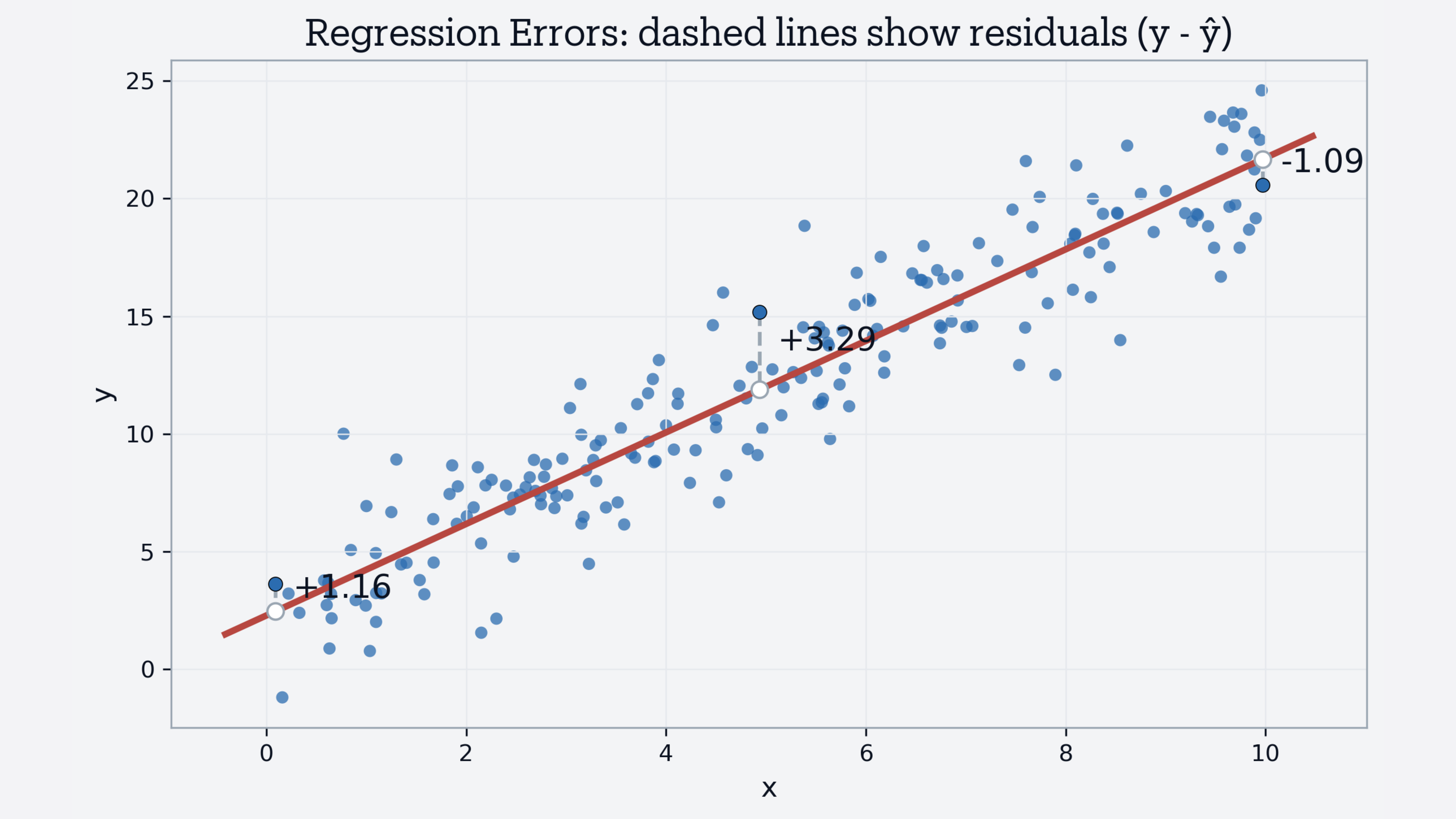
4. If $\boldsymbol{COR}(X, Z) \neq 0$, then $\boldsymbol{COR}(X, u) \neq 0$

5. $\hat{\beta}_1$ is a biased estimate of β_1

Example: Confounders

- Note: Endogeneity is correlation with the *true* error term (ϵ), not the estimated error term, or residuals ($\hat{\epsilon}$).
- When using OLS, $\boldsymbol{COR}(X, \hat{\epsilon}) \neq 0$ by definition

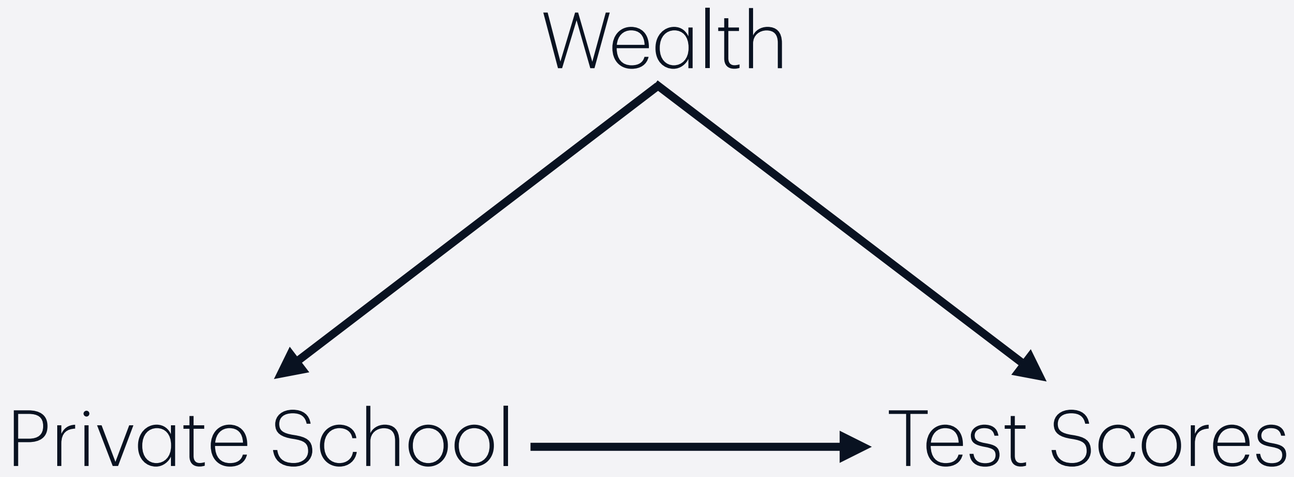

Regression Errors: dashed lines show residuals (y - ŷ)
50
40
30
20
10
0
y
0
2
4
6
8
10
x
-4.02
-11.58
-26.03


Regression Errors: dashed lines show residuals (y - ŷ)
25
20
15
10
5
0
y
0
2
4
6
8
10
x
+1.16
+3.29
-1.09

Endogeneity is why statistics is hard

- There are no fool-proof statistical tests for endogeneity
- Endogeneity is primarily a question of your theoretical assumptions, and identification strategy

Example: Confounders



Solution: Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Solution: Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$