Model Specification, Multicollinearity, and Model Selection

What Happens When You Leave Something Out, Put Too Much In, or Can't Tell Them Apart

Jake Anderson

March 21, 2026

Outline

Outline

The Scenario

You run a wage regression:

$$\text{wage}_i = \beta_1 + \beta_2 \, \text{educ}_i + e_i$$

The Scenario

You run a wage regression:

$$\text{wage}_i = \beta_1 + \beta_2 \, \text{educ}_i + e_i$$

You get $\hat{\beta}_2 = 2.3$: each extra year of education raises wages by \$2.30/hr.

The Scenario

You run a wage regression:

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + e_i$$

You get $\hat{\beta}_2 = 2.3$: each extra year of education raises wages by \$2.30/hr.

But **ability** also affects wages, and more able people tend to get more education.

Question: Is \$2.30 the true return to education, or is some of it actually the return to ability?

The True Model vs. What We Estimated

True model (what we should estimate):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\underbrace{\text{ability}_i}_{\text{omitted!}} + e_i$$

The True Model vs. What We Estimated

True model (what we should estimate):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\underbrace{\text{ability}_i}_{\text{omitted!}} + e_i$$

Estimated model (what we actually run):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \underbrace{(\beta_3\,\text{ability}_i + e_i)}_{v_i}$$

The True Model vs. What We Estimated

True model (what we should estimate):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\underbrace{\text{ability}_i}_{\text{omitted!}} + e_i$$

Estimated model (what we actually run):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \underbrace{(\beta_3\,\text{ability}_i + e_i)}_{v_i}$$

The omitted variable gets absorbed into the error term v_i.

The True Model vs. What We Estimated

True model (what we should estimate):

$$\text{wage}_i = \beta_1 + \beta_2 \, \text{educ}_i + \beta_3 \underbrace{\text{ability}_i}_{\text{omitted!}} + e_i$$

Estimated model (what we actually run):

$$\text{wage}_i = \beta_1 + \beta_2 \, \text{educ}_i + \underbrace{(\beta_3 \, \text{ability}_i + e_i)}_{v_i}$$

The omitted variable gets absorbed into the error term v_i.

If $\text{Cov}(\text{educ}, \text{ability}) \neq 0$, then $\text{Cov}(\text{educ}, v) \neq 0$.

The True Model vs. What We Estimated

True model (what we should estimate):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\underbrace{\text{ability}_i}_{\text{omitted!}} + e_i$$

Estimated model (what we actually run):

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \underbrace{(\beta_3\,\text{ability}_i + e_i)}_{v_i}$$

The omitted variable gets absorbed into the error term v_i.

If $\text{Cov}(\text{educ}, \text{ability}) \neq 0$, then $\text{Cov}(\text{educ}, v) \neq 0$.

$\Longrightarrow$ The exogeneity assumption $\text{Cov}(x, e) = 0$ is violated. OLS is **biased**.

The OVB Formula

How much bias? There is an exact formula. Let:

- b_2^* = the OLS estimate from the misspecified (short) regression
- β_2 = the true coefficient on education

The OVB Formula

How much bias? There is an exact formula. Let:

- b_2^* = the OLS estimate from the misspecified (short) regression
- β_2 = the true coefficient on education

Omitted variable bias:

$$\underbrace{E(b_2^*)}_{\text{what we get}} = \underbrace{\beta_2}_{\text{true effect}} + \underbrace{\beta_3}_{\substack{\text{effect of}\\ \text{ability on}\\ \text{wage}}} \times \underbrace{\frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}}_{\substack{\text{slope from regressing}\\ \text{ability on educ}}}$$

The OVB Formula

How much bias? There is an exact formula. Let:

- b_2^* = the OLS estimate from the misspecified (short) regression
- β_2 = the true coefficient on education

Omitted variable bias:

$$\underbrace{E(b_2^*)}_{\text{what we get}} = \underbrace{\beta_2}_{\text{true effect}} + \underbrace{\beta_3}_{\substack{\text{effect of}\\\text{ability on}\\\text{wage}}} \times \underbrace{\frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}}_{\substack{\text{slope from regressing}\\\text{ability on educ}}}$$

In shorthand: $\text{bias}(b_2^*) = \beta_3 \times \hat{\delta}_1$

The OVB Formula

How much bias? There is an exact formula. Let:

- b_2^* = the OLS estimate from the misspecified (short) regression
- β_2 = the true coefficient on education

Omitted variable bias:

$$\underbrace{E(b_2^*)}_{\text{what we get}} = \underbrace{\beta_2}_{\text{true effect}} + \underbrace{\beta_3}_{\text{effect of ability on wage}} \times \underbrace{\frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}}_{\text{slope from regressing ability on educ}}$$

In shorthand: $\text{bias}(b_2^*) = \beta_3 \times \hat{\delta}_1$

where $\hat{\delta}_1$ is the OLS slope from the **auxiliary regression** $\text{ability}_i = \delta_0 + \delta_1 \,\text{educ}_i + v_i$.

Signing the Bias

$$\text{bias}(b_2^*) = \underbrace{\beta_3}_{\text{sign?}} \times \underbrace{\hat{\delta}_1}_{\text{sign?}}$$

Signing the Bias

$$\text{bias}(b_2^*) = \underbrace{\beta_3}_{\text{sign?}} \times \underbrace{\hat{\delta}_1}_{\text{sign?}}$$

For the wage regression:

- $\beta_3 > 0$: higher ability $\implies$ higher wages
- $\hat{\delta}_1 > 0$: more able people get more education

Signing the Bias

$$\text{bias}(b_2^*) = \underbrace{\beta_3}_{\text{sign?}} \times \underbrace{\hat{\delta}_1}_{\text{sign?}}$$

For the wage regression:

- $\beta_3 > 0$: higher ability $\implies$ higher wages
- $\hat{\delta}_1 > 0$: more able people get more education

$\implies$ Bias is **positive**. The estimated return to education is **too large** because OLS attributes some of ability's effect to education.

Signing the Bias

$$\text{bias}(b_2^*) = \underbrace{\beta_3}_{\text{sign?}} \times \underbrace{\hat{\delta}_1}_{\text{sign?}}$$

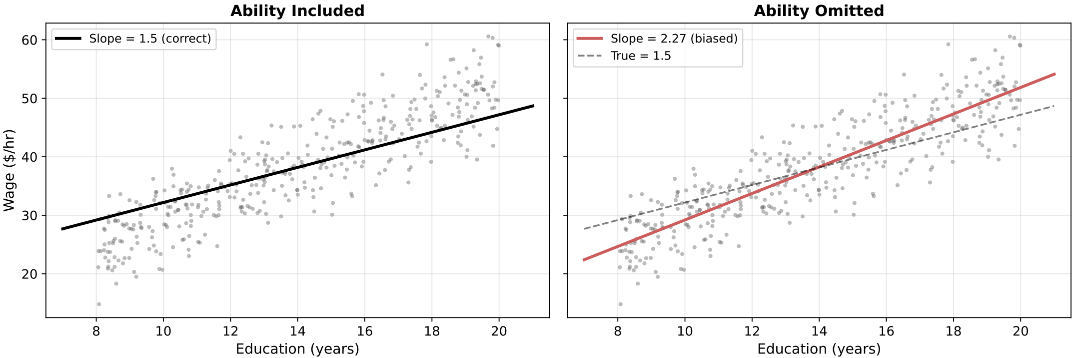
For the wage regression:

- $\beta_3 > 0$: higher ability $\implies$ higher wages
- $\hat{\delta}_1 > 0$: more able people get more education

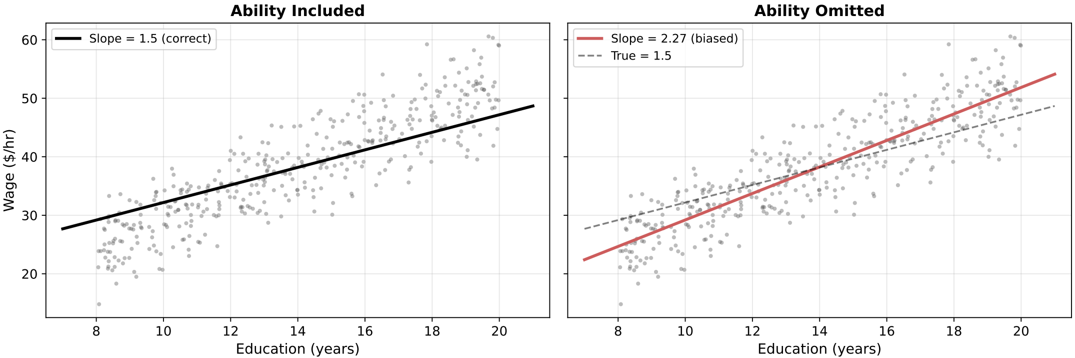
$\implies$ Bias is **positive**. The estimated return to education is **too large** because OLS attributes some of ability's effect to education.

	$\hat{\delta}_1 > 0$ (positive correlation)	$\hat{\delta}_1 < 0$ (negative correlation)
$\beta_3 > 0$ (OV helps y)	Upward bias	Downward bias
$\beta_3 < 0$ (OV hurts y)	Downward bias	Upward bias

Visualizing OVB



Visualizing OVB



The biased regression line is steeper: it picks up the indirect effect of ability through education.

OVB Is Not a Small-Sample Problem

OVB does **not** go away with more data. The inconsistency formula:

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}$$

OVB Is Not a Small-Sample Problem

OVB does **not** go away with more data. The inconsistency formula:

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}$$

As $N \to \infty$, the estimator converges to the **wrong value**.

OVB Is Not a Small-Sample Problem

OVB does **not** go away with more data. The inconsistency formula:

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}$$

As $N \to \infty$, the estimator converges to the **wrong value**.

$\implies$ This is not about imprecision (which N fixes). The model is wrong, and no amount of data corrects a wrong model.

OVB Is Not a Small-Sample Problem

OVB does **not** go away with more data. The inconsistency formula:

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{Cov(educ, ability)}}{\text{Var(educ)}}$$

As $N \to \infty$, the estimator converges to the **wrong value**.

$\implies$ This is not about imprecision (which N fixes). The model is wrong, and no amount of data corrects a wrong model.

Two conditions that eliminate OVB:

1. $\beta_3 = 0$ (the omitted variable doesn't affect y)
2. $\text{Cov}(x, z) = 0$ (the omitted variable is uncorrelated with the included regressor)

OVB Is Not a Small-Sample Problem

OVB does **not** go away with more data. The inconsistency formula:

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{Cov}(\text{educ}, \text{ability})}{\text{Var}(\text{educ})}$$

As $N \to \infty$, the estimator converges to the **wrong value**.

$\implies$ This is not about imprecision (which N fixes). The model is wrong, and no amount of data corrects a wrong model.

Two conditions that eliminate OVB:

1. $\beta_3 = 0$ (the omitted variable doesn't affect y)
2. $\text{Cov}(x, z) = 0$ (the omitted variable is uncorrelated with the included regressor)

$\implies$ An omitted variable is only a problem when **both** conditions fail.

OVB: Wage Regression Example

Using wage data ($N = 1{,}057$, Koop–Tobias):

	Without ability	With ability (SCORE)
Return to education	7.3%	5.9%

OVB: Wage Regression Example

Using wage data ($N = 1{,}057$, Koop–Tobias):

	Without ability	With ability (SCORE)
Return to education	7.3%	5.9%

Including an ability proxy reduces the education coefficient by 1.4 percentage points.

OVB: Wage Regression Example

Using wage data ($N = 1{,}057$, Koop–Tobias):

	Without ability	With ability (SCORE)
Return to education	7.3%	5.9%

Including an ability proxy reduces the education coefficient by 1.4 percentage points.

$\Longrightarrow$ Without the control, the estimated return to education was inflated by about 24% (relative to the controlled estimate).

OVB: Wage Regression Example

Using wage data ($N = 1{,}057$, Koop–Tobias):

	Without ability	With ability (SCORE)
Return to education	7.3%	5.9%

Including an ability proxy reduces the education coefficient by 1.4 percentage points.

$\implies$ Without the control, the estimated return to education was inflated by about 24% (relative to the controlled estimate).

Even this is only a partial fix: SCORE is a **proxy** for ability, not ability itself. Some bias likely remains.

Outline

The Opposite Mistake

What if, instead of leaving out a relevant variable, you include a variable that **doesn't belong**?

True model: $\text{wage}_i = \beta_1 + \beta_2 \, \text{educ}_i + \beta_3 \, \text{ability}_i + e_i$

The Opposite Mistake

What if, instead of leaving out a relevant variable, you include a variable that **doesn't belong**?

True model: $\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{ability}_i + e_i$

You estimate:

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{ability}_i + \beta_4 \underbrace{\text{shoe size}_i}_{\beta_4=0} + e_i$$

The Opposite Mistake

What if, instead of leaving out a relevant variable, you include a variable that **doesn't belong**?

True model: $\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{ability}_i + e_i$

You estimate:

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{ability}_i + \beta_4 \underbrace{\text{shoe size}_i}_{\beta_4=0} + e_i$$

Since $\beta_4 = 0$ in the true model, shoe size is **irrelevant**.

The Opposite Mistake

What if, instead of leaving out a relevant variable, you include a variable that **doesn't belong**?

True model: $\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{ability}_i + e_i$

You estimate:

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{ability}_i + \beta_4\,\underbrace{\text{shoe size}_i}_{\beta_4=0} + e_i$$

Since $\beta_4 = 0$ in the true model, shoe size is **irrelevant**.

What happens to the other estimates?

Irrelevant Variables: Unbiased but Inefficient

Good news: OLS estimates of β_1, β_2, β_3 remain **unbiased**.

Irrelevant Variables: Unbiased but Inefficient

Good news: OLS estimates of β_1, β_2, β_3 remain **unbiased**.

Bad news: Their standard errors **increase**.

Irrelevant Variables: Unbiased but Inefficient

Good news: OLS estimates of β_1, β_2, β_3 remain **unbiased**.

Bad news: Their standard errors **increase**.

Why? The variance formula for b_k includes a term $1/(1 - R_k^2)$ from the auxiliary regression of x_k on all other regressors. Adding shoe size to the regressor list increases R_k^2 (even slightly), which inflates $\text{Var}(b_k)$.

Irrelevant Variables: Unbiased but Inefficient

Good news: OLS estimates of β_1, β_2, β_3 remain **unbiased**.

Bad news: Their standard errors **increase**.

Why? The variance formula for b_k includes a term $1/(1 - R_k^2)$ from the auxiliary regression of x_k on all other regressors. Adding shoe size to the regressor list increases R_k^2 (even slightly), which inflates $\text{Var}(b_k)$.

$\implies$ Wider confidence intervals, larger p-values, less statistical power.

Irrelevant Variables: Unbiased but Inefficient

Good news: OLS estimates of $\beta_1, \beta_2, \beta_3$ remain **unbiased**.

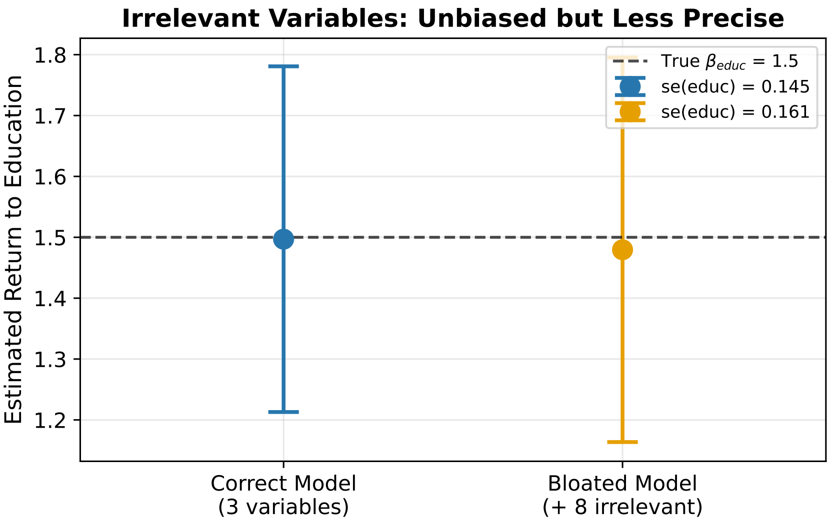
Bad news: Their standard errors **increase**.

Why? The variance formula for b_k includes a term $1/(1 - R_k^2)$ from the auxiliary regression of x_k on all other regressors. Adding shoe size to the regressor list increases R_k^2 (even slightly), which inflates $\text{Var}(b_k)$.

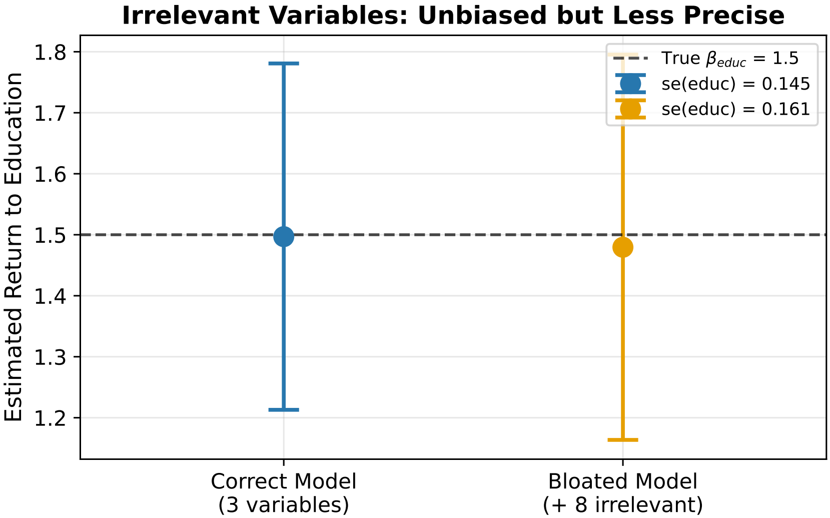
$\implies$ Wider confidence intervals, larger p-values, less statistical power.

You might fail to reject $H_0 : \beta_2 = 0$ not because education doesn't affect wages, but because you wasted degrees of freedom on shoe size.

Visualizing the Precision Loss



Visualizing the Precision Loss



OVB vs. Irrelevant Variables: The Tradeoff

	Omit relevant variable	Include irrelevant variable
Bias?	Yes (doesn't vanish)	No
Variance?	Lower	Higher
Consistency?	No	Yes

OVB vs. Irrelevant Variables: The Tradeoff

	Omit relevant variable	Include irrelevant variable
Bias?	Yes (doesn't vanish)	No
Variance?	Lower	Higher
Consistency?	No	Yes

$\Longrightarrow$ Given a choice between the two mistakes, including an irrelevant variable is the **lesser evil**. You pay with precision, not with bias.

OVB vs. Irrelevant Variables: The Tradeoff

	Omit relevant variable	**Include irrelevant variable**
Bias?	Yes (doesn't vanish)	No
Variance?	Lower	Higher
Consistency?	No	Yes

$\Longrightarrow$ Given a choice between the two mistakes, including an irrelevant variable is the **lesser evil**. You pay with precision, not with bias.

But this doesn't mean "throw everything in." More variables $=$ less precision $=$ wider intervals. The goal is a **well-specified** model.

Outline

Can We Test Whether Our Model Is Wrong?

The **RESET** (REgression Specification Error Test) checks for functional form misspecification and omitted variables.

Can We Test Whether Our Model Is Wrong?

The **RESET** (REgression Specification Error Test) checks for functional form misspecification and omitted variables.

Idea: If the model is correctly specified, then nonlinear functions of the fitted values $\hat{y}$ should not have additional explanatory power. Since $\hat{y}$ is a linear combination of the x's, $\hat{y}^2$ and $\hat{y}^3$ are polynomial functions of the regressors. They approximate many different nonlinear functional forms without specifying one.

Can We Test Whether Our Model Is Wrong?

The **RESET** (REgression Specification Error Test) checks for functional form misspecification and omitted variables.

Idea: If the model is correctly specified, then nonlinear functions of the fitted values $\hat{y}$ should not have additional explanatory power. Since $\hat{y}$ is a linear combination of the x's, $\hat{y}^2$ and $\hat{y}^3$ are polynomial functions of the regressors. They approximate many different nonlinear functional forms without specifying one.

Procedure:

1. Estimate your model: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$
2. Compute fitted values $\hat{y}$

Can We Test Whether Our Model Is Wrong?

The **RESET** (REgression Specification Error Test) checks for functional form misspecification and omitted variables.

Idea: If the model is correctly specified, then nonlinear functions of the fitted values $\hat{y}$ should not have additional explanatory power. Since $\hat{y}$ is a linear combination of the x's, $\hat{y}^2$ and $\hat{y}^3$ are polynomial functions of the regressors. They approximate many different nonlinear functional forms without specifying one.

Procedure:

1. Estimate your model: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$
2. Compute fitted values $\hat{y}$
3. Estimate the augmented model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e$$

Can We Test Whether Our Model Is Wrong?

The **RESET** (REgression Specification Error Test) checks for functional form misspecification and omitted variables.

Idea: If the model is correctly specified, then nonlinear functions of the fitted values $\hat{y}$ should not have additional explanatory power. Since $\hat{y}$ is a linear combination of the x's, $\hat{y}^2$ and $\hat{y}^3$ are polynomial functions of the regressors. They approximate many different nonlinear functional forms without specifying one.

Procedure:

1. Estimate your model: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$
2. Compute fitted values $\hat{y}$
3. Estimate the augmented model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e$$

4. Test $H_0 : \gamma_1 = \gamma_2 = 0$ using an F-test

Interpreting RESET

Reject H_0: The model is misspecified.

- Could be omitted variables, wrong functional form, or both
- RESET does not tell you *what* is wrong, only that *something* is

Interpreting RESET

Reject H_0**:** The model is misspecified.

- Could be omitted variables, wrong functional form, or both
- RESET does not tell you *what* is wrong, only that *something* is

Fail to reject H_0**:** Not conclusive evidence that the model is correct.

- RESET has limited power against certain alternatives
- A model can pass RESET and still be misspecified

Interpreting RESET

Reject H_0**:** The model is misspecified.

- Could be omitted variables, wrong functional form, or both
- RESET does not tell you *what* is wrong, only that *something* is

Fail to reject H_0**:** Not conclusive evidence that the model is correct.

- RESET has limited power against certain alternatives
- A model can pass RESET and still be misspecified

$\Longrightarrow$ RESET is a **general-purpose** diagnostic, not a test for a specific alternative.

RESET: Family Income Example

Applying RESET to the family income equation $\ln(\text{FAMINC}) = \beta_1 + \beta_2\,\text{HEDU} + \beta_3\,\text{WEDU} + e$:

RESET: Family Income Example

Applying RESET to the family income equation $\ln(\text{FAMINC}) = \beta_1 + \beta_2\,\text{HEDU} + \beta_3\,\text{WEDU} + e$:

Specification	**RESET *p*-value**	**Verdict**
Both HEDU and WEDU	0.75	Passes
WEDU omitted	0.04	**Rejects**
Both + children (KL6)	0.49	Passes
Both + irrelevant vars	0.43	Passes

RESET: Family Income Example

Applying RESET to the family income equation $\ln(\text{FAMINC}) = \beta_1 + \beta_2\,\text{HEDU} + \beta_3\,\text{WEDU} + e$:

Specification	**RESET p-value**	**Verdict**
Both HEDU and WEDU	0.75	Passes
WEDU omitted	0.04	**Rejects**
Both + children (KL6)	0.49	Passes
Both + irrelevant vars	0.43	Passes

RESET detects the omitted variable (WEDU) but does not flag irrelevant variables.

RESET: Family Income Example

Applying RESET to the family income equation $\ln(\text{FAMINC}) = \beta_1 + \beta_2\,\text{HEDU} + \beta_3\,\text{WEDU} + e$:

Specification	**RESET *p*-value**	**Verdict**
Both HEDU and WEDU	0.75	Passes
WEDU omitted	0.04	**Rejects**
Both + children (KL6)	0.49	Passes
Both + irrelevant vars	0.43	Passes

RESET detects the omitted variable (WEDU) but does not flag irrelevant variables.

$\Longrightarrow$ Useful as a diagnostic, but not a substitute for careful model building.

RESET: Family Income Example

Applying RESET to the family income equation $\ln(\text{FAMINC}) = \beta_1 + \beta_2\,\text{HEDU} + \beta_3\,\text{WEDU} + e$:

Specification	**RESET p-value**	**Verdict**
Both HEDU and WEDU	0.75	Passes
WEDU omitted	0.04	**Rejects**
Both + children (KL6)	0.49	Passes
Both + irrelevant vars	0.43	Passes

RESET detects the omitted variable (WEDU) but does not flag irrelevant variables.

$\implies$ Useful as a diagnostic, but not a substitute for careful model building.

We have checked whether the model form is correct. Next: what if the model is right but the data makes estimation difficult?

Outline

A Different Problem

New scenario: You include both education and experience in the wage regression:

$$\text{wage}_i = \beta_1 + \beta_2 \, \text{educ}_i + \beta_3 \, \text{exper}_i + e_i$$

A Different Problem

New scenario: You include both education and experience in the wage regression:

$$\text{wage}_i = \beta_1 + \beta_2 \,\text{educ}_i + \beta_3 \,\text{exper}_i + e_i$$

Both variables belong in the model. Neither is irrelevant.

A Different Problem

New scenario: You include both education and experience in the wage regression:

$$\text{wage}_i = \beta_1 + \beta_2 \,\text{educ}_i + \beta_3 \,\text{exper}_i + e_i$$

Both variables belong in the model. Neither is irrelevant.

But in your data, education and experience are **highly correlated**: people who stayed in school longer have fewer years of work experience.

A Different Problem

New scenario: You include both education and experience in the wage regression:

$$\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{exper}_i + e_i$$

Both variables belong in the model. Neither is irrelevant.

But in your data, education and experience are **highly correlated**: people who stayed in school longer have fewer years of work experience.

Question: What happens to your estimates when the regressors move together?

What Is Collinearity?

Collinearity (or multicollinearity): explanatory variables are highly correlated with each other.

What Is Collinearity?

Collinearity (or multicollinearity): explanatory variables are highly correlated with each other.

Exact collinearity ($r = \pm 1$): one regressor is a perfect linear function of another. OLS cannot compute unique estimates.

What Is Collinearity?

Collinearity (or multicollinearity): explanatory variables are highly correlated with each other.

Exact collinearity ($r = \pm 1$): one regressor is a perfect linear function of another. OLS cannot compute unique estimates.

Near collinearity ($|r|$ close to 1): OLS is defined but the estimates are very imprecise.

What Is Collinearity?

Collinearity (or multicollinearity): explanatory variables are highly correlated with each other.

Exact collinearity ($r = \pm 1$): one regressor is a perfect linear function of another. OLS cannot compute unique estimates.

Near collinearity ($|r|$ close to 1): OLS is defined but the estimates are very imprecise.

Not a violation of assumptions. The Gauss–Markov theorem still holds. OLS is still BLUE.

What Is Collinearity?

Collinearity (or multicollinearity): explanatory variables are highly correlated with each other.

Exact collinearity ($r = \pm 1$): one regressor is a perfect linear function of another. OLS cannot compute unique estimates.

Near collinearity ($|r|$ close to 1): OLS is defined but the estimates are very imprecise.

Not a violation of assumptions. The Gauss–Markov theorem still holds. OLS is still BLUE.

$\Longrightarrow$ The problem is not bias. The problem is that "Best Linear Unbiased" can still mean **very imprecise**.

The Variance Formula: Where Collinearity Hurts

For two regressors, the variance of b_2 is:

$$\mathrm{Var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2 \cdot \underbrace{(1 - r_{23}^2)}_{\text{collinearity factor}}}$$

The Variance Formula: Where Collinearity Hurts

For two regressors, the variance of b_2 is:

$$\mathrm{Var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2 \cdot \underbrace{(1 - r_{23}^2)}_{\text{collinearity factor}}}$$

- If $r_{23} = 0$: no collinearity, $(1 - r_{23}^2) = 1$, variance is at its minimum
- If $r_{23} = 0.9$: $(1 - r_{23}^2) = 0.19$, variance is **5.3 times larger**
- If $r_{23} = 0.99$: $(1 - r_{23}^2) = 0.02$, variance is **50 times larger**

The Variance Formula: Where Collinearity Hurts

For two regressors, the variance of b_2 is:

$$\text{Var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2 \cdot \underbrace{(1 - r_{23}^2)}_{\text{collinearity factor}}}$$

- If $r_{23} = 0$: no collinearity, $(1 - r_{23}^2) = 1$, variance is at its minimum
- If $r_{23} = 0.9$: $(1 - r_{23}^2) = 0.19$, variance is **5.3 times larger**
- If $r_{23} = 0.99$: $(1 - r_{23}^2) = 0.02$, variance is **50 times larger**

$\Longrightarrow$ High correlation between regressors inflates the variance of the coefficient estimates.

The Variance Inflation Factor (VIF)

With more than two regressors, the **auxiliary regression** replaces the simple correlation.

Regress x_2 on *all other* regressors: $x_{i2} = \delta_0 + \delta_1 x_{i3} + \cdots + \delta_{K-1} x_{iK} + v_i$

The Variance Inflation Factor (VIF)

With more than two regressors, the **auxiliary regression** replaces the simple correlation.

Regress x_2 on *all other* regressors: $x_{i2} = \delta_0 + \delta_1 x_{i3} + \cdots + \delta_{K-1} x_{iK} + v_i$

Let R_2^2 = the R^2 from this auxiliary regression. Then:

$$\text{Var}(b_2) = \frac{\sigma^2}{\sum (x_{i2} - \bar{x}_2)^2} \cdot \underbrace{\frac{1}{1 - R_2^2}}_{\text{VIF}}$$

The Variance Inflation Factor (VIF)

With more than two regressors, the **auxiliary regression** replaces the simple correlation.

Regress x_2 on *all other* regressors: $x_{i2} = \delta_0 + \delta_1 x_{i3} + \cdots + \delta_{K-1} x_{iK} + v_i$

Let R_2^2 = the R^2 from this auxiliary regression. Then:

$$\text{Var}(b_2) = \frac{\sigma^2}{\sum(x_{i2} - \bar{x}_2)^2} \cdot \underbrace{\frac{1}{1 - R_2^2}}_{\text{VIF}}$$

Variance Inflation Factor:

$$\text{VIF} = \frac{1}{1 - R_2^2}$$

The Variance Inflation Factor (VIF)

With more than two regressors, the **auxiliary regression** replaces the simple correlation.

Regress x_2 on *all other* regressors: $x_{i2} = \delta_0 + \delta_1 x_{i3} + \cdots + \delta_{K-1} x_{iK} + v_i$

Let R_2^2 = the R^2 from this auxiliary regression. Then:

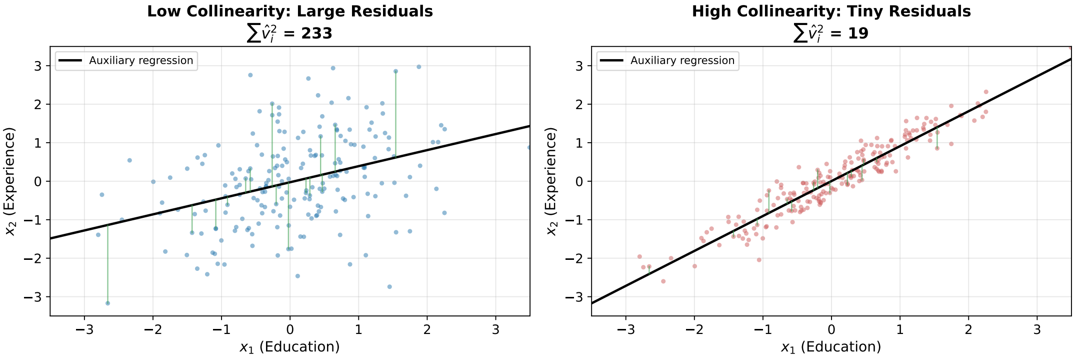
$$\text{Var}(b_2) = \frac{\sigma^2}{\sum (x_{i2} - \bar{x}_2)^2} \cdot \underbrace{\frac{1}{1 - R_2^2}}_{\text{VIF}}$$

Variance Inflation Factor:

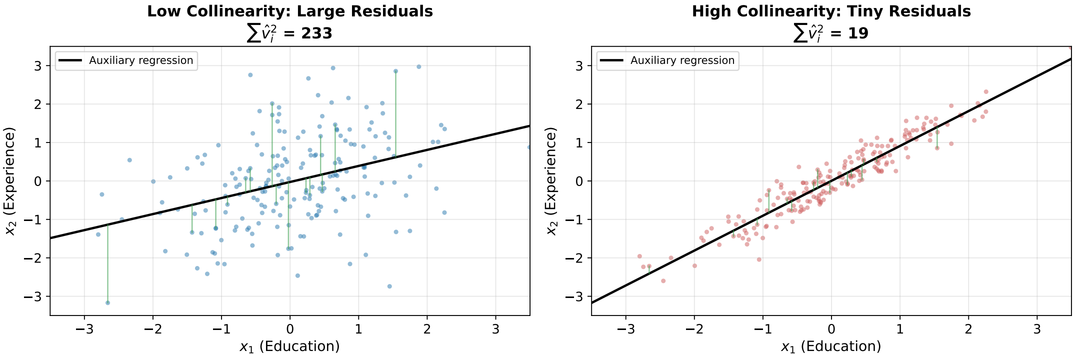
$$\text{VIF} = \frac{1}{1 - R_2^2}$$

- VIF = 1: no collinearity (baseline variance)
- VIF = 10: variance is 10 times the baseline ($R_2^2 = 0.9$)
- Rule of thumb: VIF > 10 signals problematic collinearity

Why Does This Happen? The Auxiliary Regression View



Why Does This Happen? The Auxiliary Regression View



OLS estimates β_2 using only the variation in x_2 that is **not explained by the other regressors**. When collinearity is high, very little unique variation remains.

Symptoms of Collinearity

How do you know collinearity is affecting your results?

Symptoms of Collinearity

How do you know collinearity is affecting your results?

1. **High R^2, insignificant t-statistics:** The model explains a lot, but you can't say which variables are doing the work

Symptoms of Collinearity

How do you know collinearity is affecting your results?

1. **High R^2, insignificant t-statistics:** The model explains a lot, but you can't say which variables are doing the work
2. **Large standard errors:** Confidence intervals for individual coefficients are very wide

Symptoms of Collinearity

How do you know collinearity is affecting your results?

1. **High R^2, insignificant t-statistics:** The model explains a lot, but you can't say which variables are doing the work
2. **Large standard errors:** Confidence intervals for individual coefficients are very wide
3. **Unstable coefficients:** Adding or removing one observation (or one variable) changes the estimates substantially

Symptoms of Collinearity

How do you know collinearity is affecting your results?

1. **High R^2, insignificant t-statistics:** The model explains a lot, but you can't say which variables are doing the work
2. **Large standard errors:** Confidence intervals for individual coefficients are very wide
3. **Unstable coefficients:** Adding or removing one observation (or one variable) changes the estimates substantially
4. **Joint significance but individual insignificance:** The F-test rejects $H_0 : \beta_2 = \beta_3 = 0$ even though neither t-test rejects individually

Symptoms of Collinearity

How do you know collinearity is affecting your results?

1. **High R^2, insignificant t-statistics:** The model explains a lot, but you can't say which variables are doing the work
2. **Large standard errors:** Confidence intervals for individual coefficients are very wide
3. **Unstable coefficients:** Adding or removing one observation (or one variable) changes the estimates substantially
4. **Joint significance but individual insignificance:** The F-test rejects $H_0 : \beta_2 = \beta_3 = 0$ even though neither t-test rejects individually

$\Longrightarrow$ If you see a high overall F but low individual t's, check for collinearity before dropping variables.

Rice Production Example

Philippine rice farmers ($N = 344$, 1994 data):

$$\ln(\text{PROD}) = \beta_1 + \beta_2 \ln(\text{AREA}) + \beta_3 \ln(\text{LABOR}) + \beta_4 \ln(\text{FERT}) + e$$

Rice Production Example

Philippine rice farmers ($N = 344$, 1994 data):

$$\ln(\text{PROD}) = \beta_1 + \beta_2 \ln(\text{AREA}) + \beta_3 \ln(\text{LABOR}) + \beta_4 \ln(\text{FERT}) + e$$

Variable	**Coeff**	***p*-value**	**VIF**	
ln(AREA)	0.339	0.198	9.2	insignificant
ln(LABOR)	0.175	0.465	17.9	insignificant
ln(FERT)	0.279	0.003	7.7	significant
Joint test: $H_0 : \beta_2 = \beta_3 = 0$			$p = 0.0021$	
Overall R^2			0.875	

Rice Production Example

Philippine rice farmers ($N = 344$, 1994 data):

$$\ln(\text{PROD}) = \beta_1 + \beta_2 \ln(\text{AREA}) + \beta_3 \ln(\text{LABOR}) + \beta_4 \ln(\text{FERT}) + e$$

Variable	**Coeff**	***p*-value**	**VIF**	
ln(AREA)	0.339	0.198	9.2	insignificant
ln(LABOR)	0.175	0.465	17.9	insignificant
ln(FERT)	0.279	0.003	7.7	significant
Joint test: $H_0 : \beta_2 = \beta_3 = 0$			$p = 0.0021$	
Overall R^2			0.875	

$\Longrightarrow$ Area and labor are individually insignificant, but jointly significant ($p = 0.002$). Dropping them would introduce OVB.

What Can You Do About Collinearity?

Option 1: Get more data

- More observations can help if the new data provide independent variation
- Combining 1993 and 1994 rice data: VIFs drop, all coefficients become significant

What Can You Do About Collinearity?

Option 1: Get more data

- More observations can help if the new data provide independent variation
- Combining 1993 and 1994 rice data: VIFs drop, all coefficients become significant

Option 2: Use nonsample information

- Impose restrictions from economic theory (e.g., constant returns to scale: $\beta_2 + \beta_3 + \beta_4 = 1$)
- This reduces the number of free parameters and narrows confidence intervals
- But if the restriction is wrong, you introduce bias

What Can You Do About Collinearity?

Option 1: Get more data

- More observations can help if the new data provide independent variation
- Combining 1993 and 1994 rice data: VIFs drop, all coefficients become significant

Option 2: Use nonsample information

- Impose restrictions from economic theory (e.g., constant returns to scale: $\beta_2 + \beta_3 + \beta_4 = 1$)
- This reduces the number of free parameters and narrows confidence intervals
- But if the restriction is wrong, you introduce bias

What you should not do:

- Do not drop a variable just because its t-statistic is insignificant
- If the variable belongs in the model theoretically, removing it causes OVB

Outline

Choosing Among Competing Models

You have several candidate models for y. How do you choose?

Choosing Among Competing Models

You have several candidate models for y. How do you choose?

R^2 **is not enough.** It always increases when you add variables, even irrelevant ones.

Choosing Among Competing Models

You have several candidate models for y. How do you choose?

R^2 **is not enough.** It always increases when you add variables, even irrelevant ones.

We need criteria that **penalize complexity**: reward a good fit, but punish you for using too many parameters.

Choosing Among Competing Models

You have several candidate models for y. How do you choose?

R^2 **is not enough.** It always increases when you add variables, even irrelevant ones.

We need criteria that **penalize complexity**: reward a good fit, but punish you for using too many parameters.

Three tools:

1. Adjusted R^2 ($\bar{R}^2$)
2. Akaike Information Criterion (AIC)
3. Schwarz / Bayesian Information Criterion (SC / BIC)

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

Unlike R^2, the adjusted R^2 does **not** automatically increase when a variable is added.

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

Unlike R^2, the adjusted R^2 does **not** automatically increase when a variable is added.

When does it increase? When the added variable's $|t| > 1$.

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

Unlike R^2, the adjusted R^2 does **not** automatically increase when a variable is added.

When does it increase? When the added variable's $|t| > 1$.

- $|t| > 1$: the variable reduces $SSE/(N-K)$ enough to compensate for the lost degree of freedom
- $|t| < 1$: the variable does not reduce $SSE/(N-K)$ enough; $\bar{R}^2$ falls

Adjusted R^2

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

Unlike R^2, the adjusted R^2 does **not** automatically increase when a variable is added.

When does it increase? When the added variable's $|t| > 1$.

- $|t| > 1$: the variable reduces $SSE/(N-K)$ enough to compensate for the lost degree of freedom
- $|t| < 1$: the variable does not reduce $SSE/(N-K)$ enough; $\bar{R}^2$ falls

Limitation: The threshold $|t| > 1$ corresponds to a significance level of about 32%, which is much weaker than the usual 5%. Adjusted R^2 is too generous with extra variables.

AIC and BIC

Akaike Information Criterion:

$$\text{AIC} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$$

AIC and BIC

Akaike Information Criterion:

$$\text{AIC} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$$

Schwarz / Bayesian Information Criterion:

$$\text{BIC} = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N}$$

AIC and BIC

Akaike Information Criterion:

$$\text{AIC} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$$

Schwarz / Bayesian Information Criterion:

$$\text{BIC} = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N}$$

Both have the form: $\underbrace{\ln(SSE/N)}_{\text{goodness of fit}} + \underbrace{\text{penalty}(K, N)}_{\text{complexity cost}}$

AIC and BIC

Akaike Information Criterion:

$$\text{AIC} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$$

Schwarz / Bayesian Information Criterion:

$$\text{BIC} = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N}$$

Both have the form: $\underbrace{\ln(SSE/N)}_{\text{goodness of fit}} + \underbrace{\text{penalty}(K, N)}_{\text{complexity cost}}$

- **Smaller is better** for both
- Adding a variable decreases *SSE* (improving the first term) but increases the penalty (worsening the second term)
- The preferred model balances fit against parsimony

AIC vs. BIC: Which Penalizes More?

Compare the penalty terms for one additional parameter:

- AIC penalty per parameter: $\frac{2}{N}$
- BIC penalty per parameter: $\frac{\ln(N)}{N}$

AIC vs. BIC: Which Penalizes More?

Compare the penalty terms for one additional parameter:

- AIC penalty per parameter: $\frac{2}{N}$
- BIC penalty per parameter: $\frac{\ln(N)}{N}$

BIC penalizes more heavily whenever $\ln(N) > 2$, i.e., $N > e^2 \approx 7.4$.

AIC vs. BIC: Which Penalizes More?

Compare the penalty terms for one additional parameter:

- AIC penalty per parameter: $\frac{2}{N}$
- BIC penalty per parameter: $\frac{\ln(N)}{N}$

BIC penalizes more heavily whenever $\ln(N) > 2$, i.e., $N > e^2 \approx 7.4$.

$\implies$ For any practical sample size ($N \geq 8$), BIC is **more conservative** than AIC.

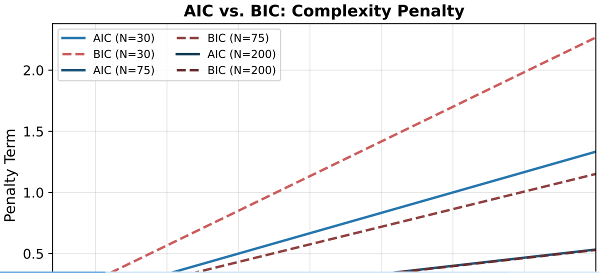
AIC vs. BIC: Which Penalizes More?

Compare the penalty terms for one additional parameter:

- AIC penalty per parameter: $\dfrac{2}{N}$
- BIC penalty per parameter: $\dfrac{\ln(N)}{N}$

BIC penalizes more heavily whenever $\ln(N) > 2$, i.e., $N > e^2 \approx 7.4$.

$\implies$ For any practical sample size ($N \geq 8$), BIC is **more conservative** than AIC.



Using Model Selection Criteria: House Price Example

Baton Rouge houses ($N = 800$ estimation, 100 held out):

Model	Variables	$\bar{R}^2$	AIC	BIC	RMSE
1	SQFT, AGE	0.463	–	–	0.327
2	SQFT, AGE, AGE^2	**0.476**	–	–	0.322
3	SQFT, AGE^2	0.473	–	–	0.326

Using Model Selection Criteria: House Price Example

Baton Rouge houses ($N = 800$ estimation, 100 held out):

Model	Variables	$\bar{R}^2$	AIC	BIC	RMSE
1	SQFT, AGE	0.463	–	–	0.327
2	SQFT, AGE, AGE^2	**0.476**	–	–	0.322
3	SQFT, AGE^2	0.473	–	–	0.326

Model 2 (with the AGE^2 term) is favored by $\bar{R}^2$, AIC, and BIC.

Using Model Selection Criteria: House Price Example

Baton Rouge houses ($N = 800$ estimation, 100 held out):

Model	Variables	$\bar{R}^2$	AIC	BIC	RMSE
1	SQFT, AGE	0.463	–	–	0.327
2	SQFT, AGE, AGE^2	**0.476**	–	–	0.322
3	SQFT, AGE^2	0.473	–	–	0.326

Model 2 (with the AGE^2 term) is favored by $\bar{R}^2$, AIC, and BIC.

Important restriction: You can only compare models with the **same dependent variable**. You cannot use AIC/BIC to compare a model for y against a model for $\ln(y)$.

Outline

What To Take Away

1. **Omit a relevant variable** $\implies$ biased and inconsistent estimates; bias direction = sign(β_3)$\times$ sign(Cov(x, z))

What To Take Away

1. **Omit a relevant variable** $\implies$ biased and inconsistent estimates; bias direction = $\text{sign}(\beta_3) \times \text{sign}(\text{Cov}(x, z))$
2. **Include an irrelevant variable** $\implies$ unbiased but larger standard errors (less precision)

What To Take Away

1. **Omit a relevant variable** $\implies$ biased and inconsistent estimates; bias direction = sign$(\beta_3)\times$ sign$(\text{Cov}(x, z))$
2. **Include an irrelevant variable** $\implies$ unbiased but larger standard errors (less precision)
3. **RESET** tests for misspecification by adding $\hat{y}^2, \hat{y}^3$; rejection signals a problem, but passing is not a guarantee

What To Take Away

1. **Omit a relevant variable** $\implies$ biased and inconsistent estimates; bias direction = sign(β_3)$\times$ sign(Cov(x, z))
2. **Include an irrelevant variable** $\implies$ unbiased but larger standard errors (less precision)
3. **RESET** tests for misspecification by adding $\hat{y}^2, \hat{y}^3$; rejection signals a problem, but passing is not a guarantee
4. **Multicollinearity** inflates variances but does not bias estimates; diagnose with VIF > 10; do not drop theoretically justified variables

What To Take Away

1. **Omit a relevant variable** $\implies$ biased and inconsistent estimates; bias direction = $\text{sign}(\beta_3) \times \text{sign}(\text{Cov}(x, z))$
2. **Include an irrelevant variable** $\implies$ unbiased but larger standard errors (less precision)
3. **RESET** tests for misspecification by adding $\hat{y}^2, \hat{y}^3$; rejection signals a problem, but passing is not a guarantee
4. **Multicollinearity** inflates variances but does not bias estimates; diagnose with VIF > 10; do not drop theoretically justified variables
5. **Model selection:** use adjusted R^2, AIC, or BIC (not R^2) to compare models; BIC penalizes complexity more heavily than AIC

Decision Checklist

When evaluating a regression model, ask:

1. Did I leave out a variable that affects y **and** correlates with x? $\implies$ OVB

Decision Checklist

When evaluating a regression model, ask:

1. Did I leave out a variable that affects y **and** correlates with x? $\implies$ OVB
2. Did I include variables that have no business being in the model? $\implies$ Inflated standard errors

Decision Checklist

When evaluating a regression model, ask:

1. Did I leave out a variable that affects y **and** correlates with x? $\Longrightarrow$ OVB
2. Did I include variables that have no business being in the model? $\Longrightarrow$ Inflated standard errors
3. Does RESET reject? $\Longrightarrow$ Something is wrong with the functional form or variable list

Decision Checklist

When evaluating a regression model, ask:

1. Did I leave out a variable that affects y **and** correlates with x? $\Longrightarrow$ OVB
2. Did I include variables that have no business being in the model? $\Longrightarrow$ Inflated standard errors
3. Does RESET reject? $\Longrightarrow$ Something is wrong with the functional form or variable list
4. Are any VIFs above 10? $\Longrightarrow$ Collinearity; do not confuse imprecision with irrelevance

Decision Checklist

When evaluating a regression model, ask:

1. Did I leave out a variable that affects y **and** correlates with x? $\Longrightarrow$ OVB
2. Did I include variables that have no business being in the model? $\Longrightarrow$ Inflated standard errors
3. Does RESET reject? $\Longrightarrow$ Something is wrong with the functional form or variable list
4. Are any VIFs above 10? $\Longrightarrow$ Collinearity; do not confuse imprecision with irrelevance
5. Am I comparing models with the same y? $\Longrightarrow$ Use AIC/BIC; lower is better

Thank you!

jakeanderson@g.ucla.edu