Econ 103: Introduction to Econometrics

Lecture 18 — Model Specification, Multicollinearity & Model Selection

Ryan Longmuir

UCLA

Summer Session C, 2026

Reading: Hill, Griffiths & Lim (5th ed.), §4.3, 6.3; Stock & Watson (4th ed.), §6.1, 7.5, 9.2.

Where we are

We can now estimate, interpret, and test multiple regressions. But every result has assumed we already *have* the right model. Choosing it is the hard part, and it turns on one central tension:

- Omit a relevant variable ⇒ bias (Lecture 13's problem, now with a formula);
- Include an irrelevant one ⇒ inflated variance (imprecision).

Today we navigate that trade-off and pick up tools to diagnose and select:

- the **omitted-variable bias** formula and its direction,
- irrelevant variables, control variables, and causal vs. prediction,
- adjusted R^2, **AIC/BIC**, **RESET**, **Jarque–Bera**, VIF.

Today's plan

1. Omitted-variable bias, quantified
2. Irrelevant variables & control variables
3. Causal vs. prediction
4. Diagnostic & selection tools

Part 1

Omitted-variable bias, quantified

The bias formula

Suppose the true model is $y = \beta_1 + \beta_2 x + \beta_3 z + e$, but we omit z and run $y = \beta_1 + \beta_2 x + v$. Then (HGL §6.3.2)

$$\text{bias}(b_2) = \mathbb{E}(b_2) - \beta_2 = \beta_3 \frac{\widehat{\text{Cov}}(x, z)}{\widehat{\text{Var}}(x)}$$

- The ratio $\widehat{\text{Cov}}(x, z)/\widehat{\text{Var}}(x)$ is the slope of regressing z on x — how the omitted variable tracks the included one.
- Bias vanishes only if $\beta_3 = 0$ (z irrelevant) *or* $\text{Cov}(x, z) = 0$ (z uncorrelated with x) — the two conditions for OVB from Lecture 13.

Signing the bias — and seeing it

The direction of the bias is the product of two signs:

$$\text{sign}\big(\text{bias}\big) = \text{sign}(\beta_3) \times \text{sign}\big(\text{Cov}(x, z)\big).$$

Example (Family income (HGL, `edu_inc`))

$\ln(\textit{FAMINC}) = \beta_1 + \beta_2\,\textit{HEDU} + \beta_3\,\textit{WEDU} + e$. *Both education effects are positive; husband's and wife's education are positively correlated. Omitting WEDU:*

$$\textit{HEDU coefficient: } 0.044 \longrightarrow 0.061.$$

Wife's education effect gets misattributed to the husband — an upward bias, exactly as $\beta_3 > 0$ *and* $\text{Cov}(\textit{HEDU}, \textit{WEDU}) > 0$ *predict.*

Knowing the likely signs lets you reason about bias *even when you lack data on z* — a routine move in applied work.

Part 2

Irrelevant variables & control variables

The opposite error: irrelevant variables

If omitting matters, why not include *everything*? Because an **irrelevant** variable (true coefficient 0) that is correlated with your regressors inflates their variances.

Example (Family income, continued)

Add two artificial regressors correlated with HEDU/WEDU but with no effect on income. Their coefficients are insignificant (good) — but the standard errors on HEDU and WEDU rise, and precision falls.

The bias–variance trade-off

	bias	variance
omit a *relevant* variable	biased	lower
include an *irrelevant* variable	unbiased	inflated

Control variables and proxies

You can't include a confounder you can't *measure* (ability, in a wage equation). The fix: a **control variable** or **proxy** that stands in for it.

Example (A proxy for ability (HGL, Koop–Tobias))

ln(*WAGE*) *on EDUC, EXPER — omitting ability biases the return to education up. Add SCORE (an aptitude-test proxy):*

$$\textit{EDUC return: } 7.3\% \longrightarrow 5.9\%.$$

The proxy soaks up ability, shrinking the overstated education effect.

For a proxy to work it must satisfy conditional mean independence: once you control for the proxy, the included regressor is "as if" randomly assigned with respect to the omitted factor. The proxy's *own* coefficient is not causal — it's just there to clean up the coefficient you care about.

Part 3

Causal vs. prediction

Two purposes, two rulebooks

Which variables belong depends entirely on *why* you built the model.

Causal inference

Goal: an unbiased *effect*. OVB is the enemy — include every confounder / control. A low R^2 is fine; what matters is that "other things" are truly held constant.

Prediction

Goal: accurate $\hat{y}$. Want regressors highly correlated with y and a high R^2. There is no "held constant," so OVB doesn't apply — a good predictor needn't be causal.

Why it matters

The selection tools below (adjusted R^2, AIC/BIC, hold-out RMSE) chase *predictive* fit. They are useful for prediction, but a high-scoring predictive model can still be badly biased for a causal question. Never let a fit criterion overrule economic theory about confounders.

Part 4

Diagnostic & selection tools

Fit criteria that penalize size

Plain R^2 never falls when you add a regressor — useless for choosing how many to include. Three penalized alternatives (smaller SSE good, more variables bad):

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(N-K)}{\text{SST}/(N-1)}, \quad \text{AIC} = \ln\frac{\text{SSE}}{N} + \frac{2K}{N}, \quad \text{BIC} = \ln\frac{\text{SSE}}{N} + \frac{K\ln N}{N}.$$

- **Adjusted R^2**: rises only if a new variable's $|t| > 1$ — a *weak* penalty (loses the "% explained" meaning).
- **AIC / BIC**: pick the model that minimizes the criterion. BIC's penalty $K \ln N/N$ is harsher than AIC's $2K/N$ (for $N \geq 8$), so BIC favors smaller models.
- Valid only across models with the *same* dependent variable (not y vs. $\ln y$).

RESET: is the functional form wrong?

The **RESET** test (Regression Specification Error Test) hunts for omitted variables or wrong functional form. After fitting and getting $\hat{y}$, add powers of the *fitted values*:

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

then test $H_0 : \gamma_1 = \gamma_2 = 0$ (an F-test).

- $\hat{y}^2, \hat{y}^3$ are polynomials in the x's, so if a curve or interaction is missing, they improve the fit and γ's turn nonzero.
- Reject $\Rightarrow$ the model is misspecified — go look for a missing term or transformation.
- Asymmetric warning: failing to reject does *not* certify the model; RESET just didn't catch anything.

Jarque–Bera: are the errors normal?

Exact small-sample t and F inference leans on SR6/MR6 (normal errors). Check it with the **Jarque–Bera** test on the residuals, built from skewness S and kurtosis K:

$$\mathrm{JB} = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \sim \chi^2_{(2)} \quad \text{under normality.}$$

- Normal $\Rightarrow S = 0$, $K = 3$, so JB near 0. Reject if $\mathrm{JB} > \chi^2_{(0.95,2)} = 5.99$.
- Food data: $S = -0.10$, $K = 2.99$, $\mathrm{JB} = 0.06$ ($p = 0.97$) — don't reject; normality is plausible.

If errors aren't normal, all is not lost: in *large* samples the CLT makes t/F inference approximately valid anyway (Lecture 9).

And collinearity: the variance inflation factor

From Lecture 14, near-collinear regressors blow up standard errors. Quantify it per regressor with the **variance inflation factor**:

$$\mathrm{Var}(b_2 \mid \mathbf{X}) = \frac{\sigma^2}{\sum (x_{i2} - \bar{x}_2)^2} \cdot \underbrace{\frac{1}{1 - R^2_{2\bullet}}}_{\text{VIF}},$$

where $R^2_{2\bullet}$ is the R^2 from regressing x_2 on *all the other* regressors.

- No collinearity ($R^2_{2\bullet} = 0$): VIF $= 1$.
- $R^2_{2\bullet} = 0.9$: VIF $= 10$ — the variance is ten times larger.
- Rule of thumb: VIF > 10 flags a worrying degree of collinearity.

A word of caution on model selection

These tools *inform*; they do not *decide*.

- Start from economic theory and the model's purpose — not from whatever maximizes a statistic.
- Use several signals together: signs/magnitudes, t and F tests, RESET, residual plots, robustness across specifications, AIC/BIC, a hold-out sample for prediction.
- Don't data-mine. Running dozens of models and reporting only the "significant" one invalidates the inference. Disclose your search.

The honest standard

A good specification is defensible on *theory* and *robust* across reasonable alternatives — not merely the one with the prettiest R^2.

Recap

The core trade-off

- omit relevant: $\text{bias}(b_2) = \beta_3 \dfrac{\text{Cov}(x,z)}{\text{Var}(x)}$ (family income $0.044 \to 0.061$)
- include irrelevant: inflated variance
- proxy/control for unobservables (ability $\to$ SCORE)

Purpose decides

- causal: kill OVB; prediction: maximize fit

Diagnostics & selection

- $\bar{R}^2$, AIC, BIC (penalize size; BIC harsher)
- RESET: add $\hat{y}^2, \hat{y}^3 \to$ form/omission
- Jarque–Bera: residual normality (food JB $= 0.06$)
- VIF $= 1/(1 - R^2_{2\bullet})$ for collinearity
- theory first; don't data-mine

Next time (Lecture 19): indicator (dummy) variables

Many regressors are *categorical* — gender, region, treatment status. We encode them as 0/1 dummies: intercept shifts, slope dummies, and the linear probability model when y itself is binary.

Questions?