Regression Diagnostics and Troubleshooting

Jeffrey Arnold

May 3, 2016

Overview

1. Omitted Variable Bias
2. Measurement Error
3. Non-Normal Errors
4. Missing data

Omitted Variable Bias: Description

- The population is

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

- But we estimate a regression without X_2

$$y_i = \hat{\beta}_0 + \hat{\beta}_1^{(omit)} x_{1,i} + \varepsilon_i$$

Omitted Variable Bias: Problem

Coefficient Bias

$$E\left(\hat{\beta}_1^{(omit)}\right) = \beta_1 + \beta_2 \frac{\text{Cov}(X_2, X_1)}{\text{Var}(X_1)}$$

Bias Components

- β_2: Effect of omitted variable X_2 on Y
- $\frac{\text{Cov}(X_2, X_1)}{\text{Var}(X_1)}$: Association between X_2 and X_1

Omitted Variable Bias: Hueristic Diagnostic

- Heuristic: sensitivity of the coefficient to inclusion of controls
- If insensitive to inclusion of controls, OVB less plausible
- Note: sensitivity of **coefficient** not *p*-value.

"These controls do not change the coefficient estimates meaningfully, and the stability of the estimates from columns 4 through 7 suggests that controlling for the model and age of the car accounts for most of the relevant selection." (Lacetera et al. 2012)

Omitted Variable Bias: Diagnosing Statistic

- Suppose X and Z observed, and W unobserved in,

$$Y = \beta_0 + \beta_1 X + \gamma_2 Z + \beta_3 W + \varepsilon$$

- Statistic to assess importance of OVB

$$\delta = \frac{\text{Cov}(X, \beta_3 W)}{\text{Cov}(X, \beta_2 Z)} = \frac{\hat{\beta}_C}{\hat{\beta}_{NC} - \hat{\beta}_C}$$

- If Z representative of all controls, then large δ implies OVB implausible
- Example in Nunn and Wanthekon

Omitted Variable Bias: Reasoning about Bias

If know omitted variable, may be able to reason about its effect

$\text{Cov}(X_1, X_2)$	$\text{Cov}(X_2, Y) > 0$	$\text{Cov}(X_2, Y) = 0$	$\text{Cov}(X_2, Y) < 0$
> 0	$+$	0	-
0	0	0	0
< 0	-	0	$+$

Omitted Variable Bias: Solutions by Design

- OVB always a problem with methods relying on selection on observables
- Other methods (Matching, propensity scores) may be less model dependent, but still can have OVB
- Preference for methods relying on identification in other ways
  - experiments
  - instrumental tables
  - regression discontinuity
  - fixed effects/diff-in-diff

Measurement Error in X: Description

- We want to estimate

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- But we estimate

$$Y_i = \beta_0 + \beta_1 X_1^* + \beta_2 X_2 + \epsilon$$

- Where X_1^* is X_1 with measurement error

$$X_i^* = X_i + \delta$$

where $\mathrm{E}(delta) = 0$, and $\mathrm{Var}(\delta) = \sigma_\delta$.

Measurement Error in X: Problem

- Similar to OVB
- For variable with the measurement error
  - $\hat{\beta}_1$ biased towards zero (**attenuation bias**)
- For other variables:
  - $\hat{\beta}_2$ biased towards OVB bias.
  - When measurement error high, it's as if that variable is not controlled for

Measurement error in Y

- We want to estimate

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

- But we estimate

$$Y_i + \delta_i = \beta_0 + \beta_1 X_{1,i}^* + \beta_2 X_{2,i} + \varepsilon_i$$

- Not a problem. Regression with larger variance,

$$Y_i = \beta_0 + \beta_1 X_{1,i}^* + \beta_2 X_{2,i} + (\epsilon_i + \delta_i)$$

  where $\mathrm{E}(\epsilon_i + \delta_i) = 0$, and $\mathrm{Var}(\varepsilon_i + \delta_i) = \sigma_\varepsilon^2 + \sigma_\delta^2$.
- If δ_i has different variances, then heteroskedasticity

Measurement Error: Solutions

- If in treatment variable:
  - get better measure
- If in control variables:
  - include multiple measures. Multicollinearity less problematic than measurement error.
- Models for measurement error: Instrumental variables, structural equation models, Bayesian models, multiple imputation.

Non-Normal Errors

- Usually not-problematic
- Does not bias coefficients
- Only affects standard errors, only for small samples
- **But** may indicate
  - Model mis-specified
  - $\mathrm{E}(Y|X)$ is not a good summary
- Diagnose: QQ-plot of (Studentized) residuals

Missing Data

- Missing data in X
- **Listwise deletion:** Drop row with *any* missing values in Y or X
- Problem: If missingness correlated with X, coefficients biased
- Solution: Multiple imputation predicts missing values from non-missing data.
- Multiple imputation packages: **Amelia**, **mice**
- Imputation almost always better than listwise deletion
- What is does not solve: If Y truncated or censored, need more complicated models.