“What’s Wrong With My OLS Model?”

- Omitted variable bias
- Multicollinearity
- Heteroskedasticity
- Outliers, “unusual data,” influential observations
- Non-normality of error term
- Measurement error
- Endogeneity

Omitted Variable Bias

- **<u>Standard assumptions for the multiple regression model (cont.)</u>**
- **Assumption MLR.4 (Zero conditional mean)**

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

The value of the explanatory variables must contain no information about the mean of the unobserved factors

  - In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error

- **Example: Average test scores**

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

If avginc was not included in the regression, it would end up in the error term; it would then be hard to defend that expend is uncorrelated with the error

Omitted Variable Bias

- **Omitting relevant variables: the simple case**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

True model (contains x_1 and x_2)

$$y = \alpha_0 + \alpha_1 x_1 + w$$

Specified model (x_2 is omitted)

- **Omitted variable bias**

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

$$= \boxed{(\beta_0 + \beta_2\delta_0)} + \boxed{(\beta_1 + \beta_2\delta_1)}x_1 + \boxed{(\beta_2 v + u)}$$

If y is only regressed on x_1 this will be the estimated intercept

If y is only regressed on x_1, this will be the estimated slope on x_1

error term

- **Conclusion:** **All estimated coefficients will be biased**

Omitted Variable Bias

- **Example: Omitting ability in a wage equation**

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

Will both be positive

$$abil = \delta_0 + \delta_1 educ + v$$

$$wage = (\beta_0 + \beta_2\delta_0) + \boxed{(\beta_1 + \beta_2\delta_1)} educ + (\beta_2 v + u)$$

The return to education β_1 will be <u>overestimated</u> because $\beta_2\delta_1 > 0$ It will look as if people with many years of education earn very high wages, but this is partly due to the fact that people with more education are also more able on average.

- **When is there *NO* omitted variable bias?**
  - If the omitted variable is **irrelevant** ($\beta = 0$)
  - If the omitted variable is **uncorrelated with the included x**

Omitted Variable Bias

- **Omitted variable bias: more general cases**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

True model (contains x_1, x_2 and x_3)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w$$

Estimated model (x_3 is omitted)

  - No general statements possible about direction of bias
  - Analysis as in simple case if one regressor uncorrelated with others

- **Example: Omitting ability in a wage equation**

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

If exper is approximately uncorrelated with educ and abil, then the direction
of the omitted variable bias can be analyzed as in the simple two variable case.

Multicollinearity

- Recall that if a perfect linear combination exists among the IVs, OLS estimates are not identified (no unique estimates).

- We also know that as the correlation increases among the IVs, the standard errors increase, and therefore, the precision of our estimates decreases.
  - This is multicollinearity (or often simply “collinearity”): strong correlations among our IVs.
  - Multicollinearity is an informational problem; efficiency problem.

Multicollinearity

- Sometimes it's easy to point the finger at multicollinearity , but oftentimes other data-specific issues may induce the imprecision that we often attribute to multicollinearity :
  - "Weak data"; small n
  - Lack of variation in IVs
  - Bad fit, and therefore, high error variance.

- Bottom line: We need to detect whether and the degree to which the existence of multicollinearity affects the standard errors (efficiency).

- First steps in diagnosis:
  - Look at correlations among the IVs.
  - Are there especially large correlations?

Primary Diagnostic: Variance Inflation Factor

- Recall that the sampling variance (standard error squared) of a slope coefficient can be written as:

$$Var(\hat{\beta}_k) = \frac{1}{1 - R_k^2} \times \frac{\hat{\sigma}^2}{(n-1)Var(x_k)}$$

- The *variance inflation factor* (VIF) is: $\frac{1}{1 - R_k^2}$
- Tells you the extent to which multicollinearity is influencing *sampling variance* of a coefficient for a particular x.
- We can use **sqrt(VIF)** to examine directly how multicollinearity affects *standard errors*.
- Start worrying when sqrt(VIF) > 2; indicates multicollinearity is inflating SE by a factor of 2.

Treatment? No, Management

- Best practice: Don't think of "treating" multicollinearity. Instead, *manage* it.

- Think about underlying factors that may be inducing multicollinearity for which a different model specification may be more appropriate.

Treatment? No, Management

- Steps in management:
  1. Think about your data; how “informative” is it? Sample size, variation in X’s.
  2. Look at the correlations among X’s
  3. Examine VIFs and sqrt(VIF)

4. Model (Re)specification

- How is the overall fit? Is error variance a problem? Are we missing some important variables that might explain a substantial portion of the variance in Y?

- Alternatively, are there extraneous variables in the model that are inducing multicollinearity but that could be excluded?
  - Respecify the model excluding "problem" variables.
  - Now, how do the SEs for the variables of interest look?

5. Latent Variable Underlying Correlated X's?

- Are the correlated X's actually indicators of an underlying latent concept?

- Measurement model; factor analysis.

6. More Complex Structural Model?

- Is OLS the right modeling strategy?

- The issue may be more serious than multicollinearity: could be endogeneity; reciprocal causation.

7. Joint Hypothesis Test

- If there's no way out of including two or three highly correlated variables, you can use a joint hypothesis test (F-test) to see if at least the correlated variables are *jointly* statistically significant (if not statistically significant on their own).

- Example: Party and ideology in congressional research; or in voting behavior research.

- Recall the null hypothesis in joint hypothesis test: $\beta_1=\beta_2=0$

Heteroscedasticity

- Recall the Gauss-Markov Theorem: Under the assumptions of *linearity*, *constant variance*, and *error independence*, the OLS estimator has minimum variance compared to all linear unbiased estimators.
- Note that a violation of the constant variance assumption is an *efficiency* problem. Ultimately, these violations will affect the sampling variance of our estimators, and thus our standard errors which affect our statistical inferences.
- But, even in the face of heteroscadasticity, OLS remains unbiased.

Detecting Heteroscedasticity

- Again, we’ll look at this visually.
- Error variance can increase as a function of:
  - Y-hat (so plot residuals against y-hats)
  - X’s (so plot residuals against the X’s)
- We can also plot “studentized residuals” against each.

Formal Tests for Detecting Heteroscedasticity

- There are quite a few of these, but we'll focus on the Breusch-Pagan test (aka, Breusch-Pagan-Godfrey test or Breusch-Pagan – Cook-Weisberg test).

- Goal: Develop a test for whether the error variance changes as a function of our independent variables.
  - One can do a global test and/or covariate-specific tests.

Formal Tests for Detecting Heteroscedasticity

- Steps in executing the BPG test:
  1. Estimate your specified model and save the residuals ($\hat{u}_i$).
  2. Calculate the *ML* error variance σ^2 = SSR/n.
  3. Generate the following variable: $g_i = \hat{u}_i^2 / \sigma^2$
  4. Then, regress g_i on your independent variables in the original model.
  5. Your test statistic will be ***SSE/ 2*** from the regression in step 4.
  6. The test statistic follows a chi-squared distribution with df=k, where k is the number of independent variables.

Formal Tests for Detecting Heteroscedasticity

- The null hypothesis for the test is that there is homoscedasticity, so if you reject H_0, there’s trouble.

- The BPG test is a global test. One can also execute covariate- specific tests, e.g., for one covariate or a subset of covariates.

Correcting for Heteroscedasticity

- If there's significant heterosc., then our standard errors will be biased. So we need to think of corrections.
- Most common correction: *robust standard errors*, or "heteroscedastic-consistent standard errors," or "White standard errors," named after Halbert White (1980, *Econometrica*).
- Think of derivation of var(beta-hat)….

Correcting for Heteroscedasticity

- White showed that if errors are heteroscedastic (yet still independent), a consistent var-cov matrix of the $\hat{\beta}$'s can be estimated as:

  $\text{var}(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$

- where Σ is a diagonal matrix with $\hat{u}_i^2$ on the main diagonal.
- Note this method corrects for heterosc. of an unknown form; other methods (WLS/GLS) correct for a specific form of heterosc.
- In R….

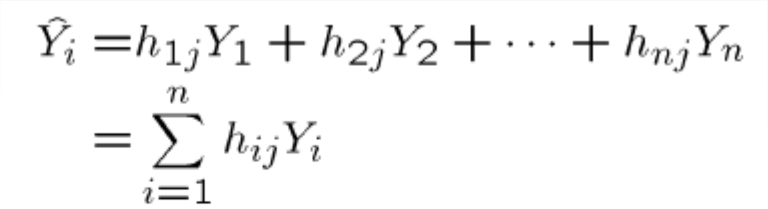
Thinking About Outliers, Leverage, Influence

- In regression, “an **outlier** is an observation whose dependent variable value is conditionally unusual given the value of the independent variable.” (Fox)
  - “off the beaten path”
  - Unusually large residuals
- Depending on where the outlier is on the X distribution, it will exert some degree of **leverage** (potentially altering the slope, compared to if that observation were deleted).
  - **Leverage** relates to how distant an observation is from the mean of X.
- Leverage and outlyingness (discrepancy) combine to ***influence*** the regression estimates (partial slopes).
  - *Influence* = f (*outlyingness*, *leverage*)

Measuring Leverage: Hat Values

- Hat values give you an idea of how far a particular observation deviates from the central part of the distribution for each X.

- Hat values capture the weight of y's contribution to y-hat.



$$\begin{aligned}\hat{Y}_i &= h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{nj}Y_n \\ &= \sum_{i=1}^{n} h_{ij}Y_i\end{aligned}$$

Measuring Leverage: Hat Values

- In simple regression, hat values represent distance from the mean of the independent variable.

- Multiple regression: distance from centroid, or the “point of means” of the X’s.

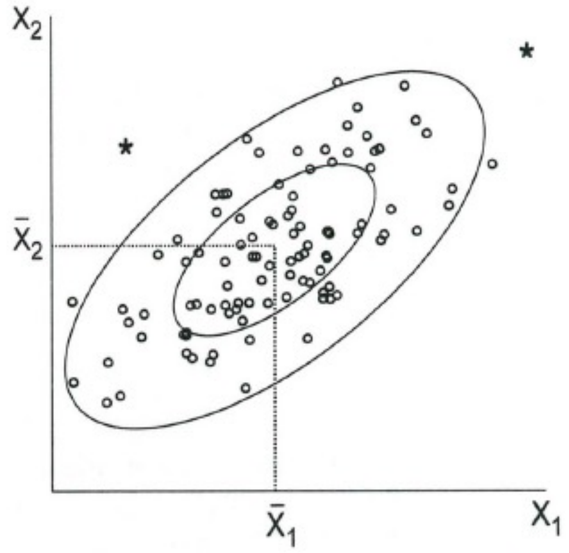


Figure 11.3 from Fox (1997)

Detecting Outliers

- Recall that in detecting outliers, we’re interested in the *conditional* unusualness of a particular Y. That is, conditional on the X’s
  - This relates to Y-hats and residuals, which are conditional on the X’s.

- Outliers will have large residuals; that is, our regression does not predict these observations well, thus, the distance between Y and Yhat will be large.

Measuring Leverage: Hat Values

$$\begin{aligned}\hat{\mathrm{y}} &= X\mathrm{b} \\ &= X(X'X)^{-1}X'\mathrm{y} \\ &= \mathrm{Hy}\end{aligned}$$

$$\underset{(n\times n)}{\mathrm{H}} = X(X'X)^{-1}X'$$

- Average hat value, h-bar=(k+1)/n (where k=number of independent variables).
- Guidelines for what constitutes severe leverage:
  - When h_i > 2 x h-bar
  - For smaller samples, use a higher threshold: when h_i > 3 x h-bar
- Use graphs to examine.

Detecting Outliers: Studentized Residuals

- We use *studentized residuals*: $u_i^* = \hat{u}_i / (\sigma_{(-i)} * \text{sqrt}(1 - h_i))$
  - To compute $\sigma_{(-1)}$, delete the given observation and calculate σ using n-1 observations. This follows the t-distribution.

- Guidelines. Since these follow a t-distribution, which converges to the normal in large samples, we can connect the idea of "large residuals" to the properties of the normal. So about 95% of the residuals are going to be within about 2 standard deviations from zero.

- Thus, large outliers will have: $|u_i^*| > 2$.

Measuring Influence I: DFBETAS

- DFBETAS assess the influence of deleting an observation on the coefficients.
- Directly measures influence: How much would my coefficients change if I were to delete observation *i*?

$$D_{ij}^{*} = \frac{b_j - b_{j(-1)}}{SE_{(-i)}(b_j)}$$

- For each coefficient, you get *n* DFBETAS, for a total of *n***k*.
- Graphically: For each coefficient, produce a scatter of the D's (on the y-axis) against observation index.
- Guidelines: Take notice if | D | > 2/sqrt(n)

Measuring Influence II: Cook's D

- Intuition: Assess the distance between b_j using the full sample and b_j after deleting an observation.
- Unlike DFBETAS, Cook's D is a *global* measure of influence. Thus, we calculate one of these per observation.
- Calculate a statistic analogous to an F-statistic testing the hypothesis: $\beta_j = \beta_{j(-1)}$

$$D_i = \frac{u_i'^2}{k+1} * \frac{h_i}{1-h_i}$$

- First term is a measure of discrepancy, or outlyingness, since it uses standardized residuals. The second term is a measure of leverage, since it uses hat values.
- Guidelines: Take notice if $D_i > 4 / (n-k-1)$

What to do with Outliers and Influential Obs?

- Think about *why* an observation might be unusual.
- Outliers: Can motivate further thinking about omitted variable(s) that might explain that outlier.
- Influential or outlying obs can motivate respecification.
- Robustness checks.
- In short, *know your data*. Stare at these graphical plots and try to understand what’s going on, any patterns, etc.

Non-Normality of Errors

- Again, unbiasedness property of OLS is robust when errors aren’t normal.

- But when normality breaks down, we lose the minimum variance property.

- Particularly bothersome are heavy-tailed error distributions – efficiency loss.

- We can also look for skewed error distributions, which might suggest the need to transform the data.

- Graphical diagnosis using OLS or studentized residuals.

Normality Plots

- Graphs

Jarque-Bera Test for Normality of Errors

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right)$$

n = sample size

s = skewness

k = kurtosis

Use OLS residuals

JB statistic follows chi-square distribution with 2 degrees of freedom.

Null hypothesis: Errors are normally distributed.

Measurement Error

- **<u>Properties of OLS under measurement error</u>**
- **Measurement error in the dependent variable**

$$y = y^* + \boxed{e_0}$$

Mismeasured value = True value + Measurement error

$$\boxed{y^*} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Population regression

$$\Rightarrow \boxed{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \boxed{(u + e_0)}$$

Estimated regression

- **Consequences of measurement error in the dependent variable**
  - Estimates will be less precise (higher standard errors) because the error variance is higher
  - Otherwise, OLS will be unbiased and consistent (as long as the measurement error is unrelated to the values of the explanatory variables)

Measurement Error

- **Measurement error in an independent variable**

$$x_1 = x_1^* + \boxed{e_1}$$

Mismeasured value = True value + Measurement error

$$y = \beta_0 + \beta_1 \boxed{x_1^*} + \cdots + \beta_k x_k + u$$

Population regression

$$\Rightarrow \quad y = \beta_0 + \beta_1 \boxed{x_1} + \cdots + \beta_k x_k + \boxed{(u - \beta_1 e_1)}$$

Estimated regression

<u>Classical errors-in-variables assumption</u>: $\boxed{Cov(x_1^*, e_1) = 0}$

Error unrelated to true value

$$\Rightarrow \quad Cov(x_1, e_1) = Cov(x_1^*, e_1) + Cov(e_1, e_1) = \sigma_{e_1}^2$$

$$\Rightarrow \quad Cov(\boxed{x_1}, \boxed{u - \beta_1 e_1}) = -\beta_1 Cov(x_1, e_1) = \beta_1 \sigma_{e_1}^2$$

<u>The mismeasured variable x$_1$ is correlated with the error term!</u>

Measurement Error

- **Consequences of measurement error in an explanatory variable**
  - Under the classical errors-in-variables assumption, OLS is biased and inconsistent because the mismeasured variable is endogenous (corr. with error term)
  - One can show that the inconsistency is of the following form:

$$plim\ \hat{\beta}_1 = \beta_1 \left(\frac{\sigma^2_{r_1^*}}{\sigma^2_{r_1^*} + \sigma^2_{e_1}} \right)$$

This factor (which involves the error variance of a regression of the true value of x_1 on the other explanatory variables) will always be between zero and one

  - The effect of the mismeasured variable suffers from ***attenuation bias***, i.e. the magnitude of the effect will be attenuated towards zero
  - In addition, the effects of the other explanatory variables will be biased