Regression Diagnostics and Troubleshooting II

Jeffrey Arnold

May 5, 2016

Topics for Today

1. Heteroskedasticity and correlated errors
2. Bootstrapping
3. Outliers

- What are they?
- What problems do they cause? Biased $\hat{\beta}$? Biased or large se?

Heteroskedasticity and Correlated errors

Classical Linear Regression Model

$$y = \boldsymbol{X}\beta + \varepsilon$$

where

$$\begin{aligned} \mathsf{E}(\varepsilon_i) &= 0 \\ \mathsf{Var}(\varepsilon_i) = \sigma^2 \, \mathsf{Cov}(\varepsilon_i, \varepsilon_j) &= 0 \quad \text{for all } i \neq j. \end{aligned}$$

Examples of Heteroskedasticity and Correlated errors

1. Clustering: variables in same geographic area or people in group are Correlated
2. Autocorrelation: observations close in time are correlated
3.

General Methods to Deal with It

1. Weighted Least Squared
2. Adjusted Variance-Covariance Matrices ("Robust" standard errors)
3. Time-series or spatial weighting methods

Weighted Least Squares

- Minimize the 3we

$$\arg\min_{b} \sum_{i} w_i (y_i - \boldsymbol{x}_i' \beta)^2$$

- If right weighting: unbiased **and** efficient
- Different estimates than OLS, $\hat{\beta}_{WLS} \neq \hat{\beta}_{OLS}$
- Most often used:
  - different populations
  - known measurement errors
- If don't know weights: use adjusted standard standard errors

"Robust" Standard errors

- Run OLS, but use different variance covariance matrix
  1. OLS $\hat{\beta}$
  2. Different Var($\hat{\beta}$)
- Different methods for different correlations
- Heteroskedasticity

Bootstrapping

- Flexible way of generating standard errors by resampling data
  1. Parametric: resample from a model
  2. Non-parametric: resample data itself
- Why? Can calculate standard errors for unusual functions or complicated DGPs
- See examples in Assignment 03

Outliers

Three concepts:

1. Leverage: unusual points in X
2. Outliers: large errors ε_i
3. Influential points:
   - points with a large effect on $\hat{\beta}$
   - influential = leverage * outlier

What to do about unusual data?

- Can have large effects on $\hat{\beta}$ and se
- Whether these are "wrong" depends on the DGP of those points
- Do we drop them? Generally **not**
  1. Why is it unusual? Is it bad data?
  2. Reformulate the model
  3. Learn from what the model is not capturing
  4. Robust and resistant methods - least absolute deviations