CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 22
Influence in Regression

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Influence During Regression

- Outliers are data with an extreme value of the response variable (Y)
- Leverage points are data with an extreme value of the predictor variable (X)
- Some combination of extreme Y (outlier) and extreme X (leverage) makes a data point influential
- An influential data point: removing the data point substantially changes the regression results
  – How do we define "substantial"?

Influence: Cook's Distance

- Delete the i[th] data point, then look at the difference in predicted y-values

Predicted j[th] response with i[th] data point removed

Cook's Distance: $$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{p s_e^2}$$

More convenient form:

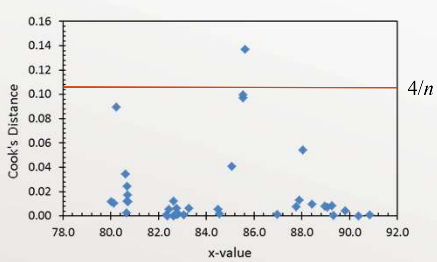
$$D_i = \frac{{e_i}^2}{p s_e^2}\frac{h_{ii}}{(1-h_{ii})^2} = \frac{{isr_i}^2}{p}\frac{h_{ii}}{(1-h_{ii})}$$

Measuring Influence

- The Cook's Distance is a measure of influence, but it is not a statistical test
  – Outliers are not necessarily influential, and influential points are not necessarily outliers
  – We don't remove or adjust highly influential points
  – Our goal is to identify influential points
  – We worry about fragility: our conclusions depend only on 1 or 2 data points

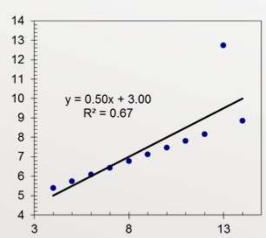
R.D. Cook, "Detection of influential observation in linear regression", *Technometrics*, **19**(1), 15–18 (1977).
R.D. Cook, S. Weisberg, "Characterizations of an empirical influence function for detecting influential cases in regression", *Technometrics*, **22**(4), 495–508 (1980).

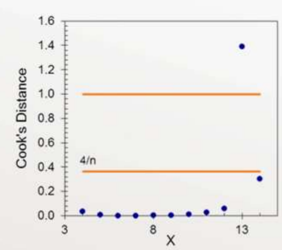
Cook's Distance



The Cook's Distance is considered significant if it is bigger than about $4/n$ (alternately, use the 50th percentile of the $F_{p;n-p}$ distribution)

Anscombe Revisited





More Influence Measures

- For each β_k of interest, find its estimate with and without the i^{th} data point

$$DFBETA_{k,i} = \frac{b_k - b_{k(i)}}{SE(b_{k(i)})}$$

Considered significant if DFBETA is bigger than about $2/\sqrt{n}$

- A measure similar to the Cook's Distance is

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{\varepsilon(i)}\sqrt{h_{ii}}} = esr_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

- Also, Mahalanobis Distance (we won't discuss)

Review of Influence Measures

Metric	Equation	Small Sample Criterion	Large Sample Criterion
Cook's Distance	$D_i = \frac{isr_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$	1	$4/n$, F.INV(0.5)
Difference in Beta	$DFBETA_{k,i} = \frac{b_k - b_{k(i)}}{SE(b_{k(i)})}$	1	$\sqrt{4/n}$
Difference in Fit	$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{\varepsilon(i)}\sqrt{h_{ii}}} = esr_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$	1	$\sqrt{4p/n}$

"small" means about n = 20 or less

Multiple Regression

- When regressing on two or more predictor variables, it is best to let a statistical software package calculate quantities like $isr_i, esr_i, h_{ii}, D_i, DFBETA_{k,i}$, etc.
- Additionally, with multiple regression we have to worry about correlations between predictor variables
  - We'll cover this later

Experimental Design

- An important goal of Design of Experiments (DoE) is to equalize the leverage of every point during multiple regression
  - We want to make $h_{ii} = \frac{p}{n}$ for every i
  - More on DoE later

Conclusions

- When regressing, outliers and high leverage data points are important to consider, but it is influential data that matters most
- When regressing, calculate for every data point: $isr_i, esr_i, h_{ii}, D_i, DFFITS_i$, etc.
- Use the Williams graph and graphs of the Cook's Distance to get a feel for influence
- Consider deleting or altering outliers only if they are influential
- If your results are fragile, consider collecting more data to reduce the influence of the few data points that make your results fragile

Lecture 22: What have we learned?

- Define influence
- Name several metrics of influence
- Explain what is meant by a "fragile" regression
- How does a measure of influence affect the way we approach outliers?