In-class Activity: Leverage and Influence February 22, 2107

Example: Consider the Crime data. The variables are

The crime.csv data set appears in Statistical Methods for Social Sciences, Third Edition by Alan Agresti and Barbara Finlay (Prentice Hall, 1997). The variables are **state id** (sid), **state name** (state), violent crimes per 100,000 people (**crime**), murders per 1,000,000 (**murder**), the percent of the population living in metropolitan areas (**pctmetro**), the percent of the population that is white (**pctwhite**), percent of population with a high school education or above (**pcths**), percent of population living under poverty line (**poverty**), and percent of population that are single parents (**single**). The goal of the analysis is to fit a multiple linear regression model by regressing the crime rate on the available predictors.

1. Make a pairwise scatter plot. Do you see any outliers? Can you identify them as either leverage or influential?
2. Fit a multiple linear regression model relating the number of games won to the tree covariates. Report the following:

* Adjusted R-square value
* MSE
* Can you identify any non-significant covariates?  
  Is this a wise decision to drop them from your resulting model?

1. Calculate the hii, the amount of leverage for each observation. Are there any leverage points?
2. Perform a residual analysis? Which residuals will be more appropriate to identify outliers?
3. Calculate the Cook’s distance. Can you identify influence points, if any?

Report the Cook’s distance corresponding to DC. How will you interpret this value?

1. Calculate DFBETAS.
2. Obtain the resulting model