Assignment 7 - Diagnostics

your name goes here

Due: Wednesday, March 21, 2018

Summary

Primarily from the topics in Chapter 10 of your text, this homework assignment gives you practice thinking about influential observations (specific cases in your data set) and also about the multicollinearity among variables.

Assignment

1. A student asked: "Why is it necessary to perform diagnostic tests when R^2 is large?" Comment.
2. A student suggested: "If extremely influential outlying cases are detected in a data set, simply discard these cases from the data set." Comment (note that the end of the notes discusses what to do with outliers).
3. Describe several informal methods that can be helpful in identifying multicollinearity among the explanatory variables in a multiple regression model.
4. Consider again the College data set (which you might remember had an unusual data point...) In this exercise, we will investigate the regression diagnostics. Again, predict number of applications (Apps); use Accept, Top10perc, Private, Outstate, and PhD as explanatory variables.

```
require(mosaic); require(ISLR)
require(dplyr)
data(College)
col.lm <- lm(Apps~ Accept + Top10perc + Private + Outstate + PhD,
             data=College)
```

   (a) Obtain the studentizded deleted residuals, and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .01$: that is, the level of significance should be $\alpha_B =$ (0.01/ number of points). State the decision rule and conclusion.

   (b) Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

   (c) The researcher wishes to estimate the number of 1995 applications for a private university (private "yes" is coded as 2) who Accepts 1000 applications, has 75% of their faculty with PhDs, has 5% of the students from the top 10% of their high school class, and has an out of state tuition of \$15,000. Use (10.29) $[h_{new,new} = X^t_{new}(X^tX)^{-1}X_{new}]$ to determine whether this estimate will involve a hidden extrapolation.

   (d) Cases 462 and 484 appear to be outlying X observations, and cases 251, 460, and 484 appear to be outlying Y observations. Obtain the DFFITS, DFBETAS, and Cook's distance values for each case to assess its influence. What do you conclude?

(e) Calculate Cook's distance D_i for each case and prepare an index plot. Are any additional cases influential according to this measure?

(f) Remove the most influential case from the data set and re-run the linear model. Does the significance of any of the variables change? Comment on how you would proceed given what you've discovered in your diagnostic analysis.

(g) What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables? (Use the `pairs` function or `ggpairs` in the `GGally` package on the matrix containing the variables in the model.)

(h) Obtain the five variance inflation factors. Do they indicate that a serious multicollinearity problem exists here?