STAT 1221/2220
Homework 10-11

Name: Stephen Skasko

Problems 1a, 1c, 1e, 1f, 1g, 1m, and 1o require the use of R. Be sure to install the MASS package before beginning this assignment. Upload the knitted R file with the completed problems.

1. Recall the dataset from Homeworks 7, 8, and 9 that analyzed the strength of 44 one cubic meter concrete mixtures, measured in megapascals. The best model for the data is the one containing the amounts of cement (in kg) (X_1), water (in liters) (X_2), and superplasticizer (in kg) (X_3). Read the dataset **concrete.csv** into R to answer the following questions.

   (a) Fit the model that uses all three predictors to describe strength. Print the model summary.

   (b) What is the cutoff for which an observation would be considered high leverage? $\frac{2(3+1)}{44} = 0.18$

   (c) Calculate the leverage for each observation using R and save the values to the data frame. Then create and print the leverage plot with a horizontal line at the cutoff you calculated in part (b).

   (d) Click on the data frame in R to open it up. Sort the data in decreasing order according to leverage by clicking on the leverage column that you added in part (b). The observation number is listed in the first column. Report the observation numbers for mixtures that are considered high leverage as well as their leverage values.

   #43 with leverage point 0.1866

   #7 with leverage point 0.1849

   (e) Calculate and save the following values to the data frame: (1) the predicted values, (2) the pure residuals (i.e. difference between observed and expected), (3) the studentized residuals, and (4) the jackknife residuals. Then create a residual plot of the **jackknife** residuals against the predicted values using R.

   (f) Create a normal probability plot of the jackknife residuals using R.

   (g) Run the Shapiro-Wilk test in R on the jackknife residuals and print the test results.

   (h) Are linearity, homoscedasticity, and normality of the residuals satisfied? Justify your answer using the results from parts (e), (f), and (g) above.

   Linearity: Aside from a few outliers, the other points of the residual are evenly scattered above and below 0 with no curves patterns.

   Homoscedasticity: Spread of residual points is about the same at all predicted strength points, consistantly ranging between -2 and 2.

   Normality: Most of the data follows a straintline but at the top and bottom of the data points it shows a deviation, but not enough to violate normality and the Shapiro test shows $p = 0.54 > 0.05$

   (i) Report the critical values that would be used to determine if an observation is an outlier using a 5% level of significance. Include the degrees of freedom.

   $t = \pm 2.02$ on full dataset at 40 df qt(0.025, 40)

(j) Click on the data frame in R to open it up. Sort the data in decreasing order according to the jackknife residuals by clicking on the column name. Report the observation numbers for mixtures that are considered outliers as well as their jackknife residuals. (Remember to look at both the positive and negative values.)

#42 at -2.32

#18 at 2.52

#22 at 2.68

(k) Open the data frame in R and look at the values for observation 22. Why does it make sense that the jackknife residual is so much larger in magnitude than the studentized residual?

It makes sense because if the observation has small leverage at 0.048 then $S_{Y|X(i)}$ will be greater than $S_{Y|X}$ since the jackknife residual removes an observation that fit well.

(l) What is the cutoff for which an observation would be considered an influential point? $4/44 = 0.091$

(m) Calculate the Cook's distance value for each observation using R and save the values to the data frame. Create and print a plot of the Cook's distance values with a horizontal line at the cutoff you calculated in part (l).

(n) Click on the data frame in R to open it up. Sort the data in decreasing order according to Cook's distance. Report the observation numbers for mixtures that are considered influential points as well as their Cook's distance values.

#13, 0.15

#18, 0.23

#48, 0.27

(o) Remove the observations from the dataset that were identified as influential points in part (n). Then fit the model with cement, water, and superplasticizer on this dataset. Print the model summary.

(p) Calculate the *DFBETAS* value for each coefficient using the outputs from parts (a) and (o).

Intercept: $\left|\frac{79.25 - 86.037}{21.13}\right| = 0.32$ **Cement:** $\left|\frac{0.0389 - 0.63875}{0.0246}\right| = 0.061$

Water: $\left|\frac{-0.334 - (-0.366)}{0.0997}\right| = 0.020$ **Superplasticizer:** $\left|\frac{1.6624 - 1.571}{0.550}\right| = 0.166$

(q) Based on the values from part (p), which model should be used: the one with or without the influential points? Briefly justify your answer.

We should use the model with the influential points since no coefficient changed more than one standard error.