

Assignment 12

Question 1

The manager of a URL commercial address is interested in predicting the number of megabytes downloaded, `megasd`, by clients according to the number of minutes they are connected, `mconnected`. The manager randomly selects (megabyte, minute) pairs, records the data, and stores the pairs (`megasd`, `mconnected`) in the file `URLADDRESS`. The goal here is to explore the relationship between `megasd` and `mconnected`.

- (a) What assumptions need to be satisfied in order to use the linear model for inferential purposes?
- (b) Are there any outlying observations?
- (c) Are there any influential observations? Compute and graph Cook's distances, `DFFITs`, and `DFBETAS` to answer this question. Create a bubble plot of studentized residuals versus leverage values with plotted points proportional to Cook's distance using the function `influencePlot()` from the `car` package. Does the bubble plot confirm your answer with respect to influential observations?
- (d) Estimate the mean value of megabytes downloaded by clients spending 5, 10, and 15 minutes on line. Construct the corresponding 90% confidence intervals.
- (e) Predict the megabytes downloaded by a client spending 30 minutes on line. Construct the corresponding 90% prediction interval.

Question 2

The data frame `KINDER` contains the height in inches and weight in pounds of 20 children from a kindergarten class. Use all 20 observations and construct a regression model where the results are stored in the object `mod` by regressing weight on height.

- (a) Create a scatterplot of weight versus height to verify a possible linear relationship between the two variables.
- (b) Compute and display the hat values for `mod` in a graph. Use the graph to identify the two largest hat values. Superimpose a horizontal line at $2p/n$. Remove the values that exceed $2p/n$ and regress weight on height, storing the results in an object named `modk`.
- (c) Remove case 19 from the original data frame `KINDER` and regress weight on height, storing the results in `modk19`. Is the child with the largest hat value an influential observation if one considers the 19 observations without case 19 from the original data frame? Compute and consider Cook's D_i , `DFFITsi`, and `DFBETASk(i)`, in reaching a conclusion. Specifically, produce a graph showing h_{ii} , the differences in $\hat{\beta}_{1(i)} - \hat{\beta}_1$, `DFBETASk(i)`, studentized residuals, `DFFITsi`, Cook's D_i , and a bubble-plot of studentized residuals versus leverage values with plotted points proportional to Cook's distance along with the corresponding values that flag observations for further scrutiny assuming $\alpha = 0.10$. (Hint: Use the functions `fortify()` from the `ggplot2` package and `lm.influence()`.)
- (d) Remove case 20 from the data frame `KINDER` and regress weight on height, storing the results in `modk20`. Is the child with the largest hat value an influential observation if one considers the 19 observations without case 20 from the original data frame? Compute and consider Cook's D_i , `DFFITsi`, and `DFBETASk(i)` in reaching a conclusion. Specifically, produce a graph showing h_{ii} , the differences in $\hat{\beta}_{1(i)} - \hat{\beta}_1$, `DFBETASk(i)`, studentized residuals, `DFFITsi`, Cook's D_i , and a bubble-plot of studentized residuals versus leverage values with plotted points proportional to Cook's distance along with the corresponding values that flag observations for further scrutiny assuming $\alpha = 0.10$.

- (e) Create a scatterplot showing all 20 children. Use a solid circle to identify case 19 and a solid triangle to identify case 20. Superimpose the lines for models `mod` (`lty = 1`), `modk` (`lty = 2`), `mod19` (`lty = 3`), and `mod20` (`lty = 4`).