Group 2 Questions

library(readr)  
sleep = read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv")

## Rows: 253 Columns: 27

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): LarkOwl, DepressionStatus, AnxietyStatus, Stress, AlcoholUse  
## dbl (22): Gender, ClassYear, NumEarlyClass, EarlyClass, GPA, ClassesMissed, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1. Start by building a model that predicts GPA based on the number of classes a student has missed, ClassesMissed

mod = lm(GPA~ClassesMissed, data = sleep)

1. Find and display the top 10 standardized residuals (be sure to look at absolute values when sorting). Of these, which data points may be considered influential?

Any values greater than 2 / less than -2 are considered to be influential. In our case, the top 9 data points have high standardized residuals.

head(sort(abs(rstandard(mod)), decreasing = TRUE), 10)

## 55 67 203 126 89 224 34 72   
## 3.198291 3.198291 3.140456 2.273283 2.250347 2.077591 2.037753 2.021602   
## 169 50   
## 2.021602 1.998586

1. Find and display the top 10 studentized residuals (be sure to look at absolute values when sorting). Of these, which data points may be considered influential?

Values we consider to be high: Any values greater than 2 / less than -2 In our case, all of the top 10 data points have high standardized residuals.

head(sort(abs(rstudent(mod)), decreasing = TRUE), 10)

## 55 67 203 126 89 224 34 72   
## 3.259012 3.259012 3.197646 2.292473 2.268865 2.091510 2.050724 2.034199   
## 169 50   
## 2.034199 2.010663

1. Based on only the information from the previous two questions, do we have any indications about data points that may exert significant influence on the model?

For question 4, I think the answer should be we have no evidence that the data points may have significant influence because standardized residuals and studentized residuals only tell us how far away each point deviates from the mean, they don’t tell us how influential each point is.

1. Briefly explain the difference between influence and leverage.

Influence is the effect of a single data point on the regression line, and represents how well the data point matches the of the rest of the points. Leverage represents how much influence a data point has on the regression line. It explains how far away the data point is from the other points (explains how much of an outlier it is)

For question 5, a better explanation of leverage is the “potential” to exert influence.

1. Briefly explain the difference between standardized and studentized residuals.

Standardized: Residuals of the data are standardized If a rstandard value is greater than 2 (or less than -2) it is a mild outlier. If the value is greater than 3 (or less than -3), it is a strong case of being an outlier. Studentized: An outlier is removed from the data, then standardized.

For question 6, the studentized residual is calculated by removing the data point whose studentized value you want to calculate, not removing an outlier.

1. Find and display the top 10 points with highest leverage. Do any points have significantly large leverage?

None of these data points appear to have high leverage.

2/12 # Average leverage (2/n)

## [1] 0.1666667

2 \* 2/12 # Points we might need to worry about

## [1] 0.3333333

3 \* 2/12 # Points we might need to worry about

## [1] 0.5

head(sort(hatvalues(mod), decreasing=TRUE), 10)

## 20 26 85 163 65 204 148   
## 0.12357263 0.12357263 0.06578311 0.06578311 0.05649286 0.05649286 0.04795851   
## 3 27 66   
## 0.04018004 0.02689076 0.02689076

1. Calculate the Cook’s Distance for the points in our data set. How many of these points may be considered influential outliers?

From class, we said to study any case with Di > 0.5; worry if Di > 1.0. In our data set, none of the cook’s distances exceed 0.5, so we have 0 influential outliers in our dataset.

head(sort(cooks.distance(mod), decreasing=TRUE),10)

## 65 26 55 67 163 245 34   
## 0.09610117 0.02662156 0.02314754 0.02314754 0.02079628 0.02033074 0.01966539   
## 203 89 126   
## 0.01965083 0.01476538 0.01341317