Assignment 4 - 0679576 Graham Eckel

Q3

```
library(car)
library(MASS)
set.seed(2019-11-13)
dir = "C:\\Users\\graha\\Google Drive\\1 Math Undergrad\\1 UoGuelph\\2_Fall_19\\Applied Regression Analysis\\Assignment 4\\"
file1 = "3240_F19_solar.csv"
dfSolar = read.table(file=paste(dir,file1, sep=""), header=TRUE, sep=',')

Flux = dfSolar$flux
Insolation = dfSolar$insolation
East = dfSolar$east
South = dfSolar$south
North = dfSolar$north

mlrFlux = lm(Flux~Insolation+East+South+North)
summary(mlrFlux) # 0.1911
```

a)

```
vif(mlrFlux)
Insolation       East      South      North
  2.532924   1.338372   1.674447   2.860636
```

The variance inflation factor is 1/(1-R^2) and explains the amount of variance that is inflated because of a linear dependence with other predictors. So, in the cause of Insolation, the variance is a little over 2.5x larger than it would be if there was no correlation between it and the remaining predictors. Given the relatively high inflation of variance in Insolation and North and that the remaining predictors have a VIF > 1, there exists some multicollinearity.

b)

max(hatvalues(mlrFlux)) = 0.4207657 = observation with the greatest leverage
mean(hatvalues(mlrFlux)) = 0.2 = average hat values

Since 0.4207657 > 2(0.2) = 0.4, which is our rough guideline of high leverage, the observation can be said to have high leverage.

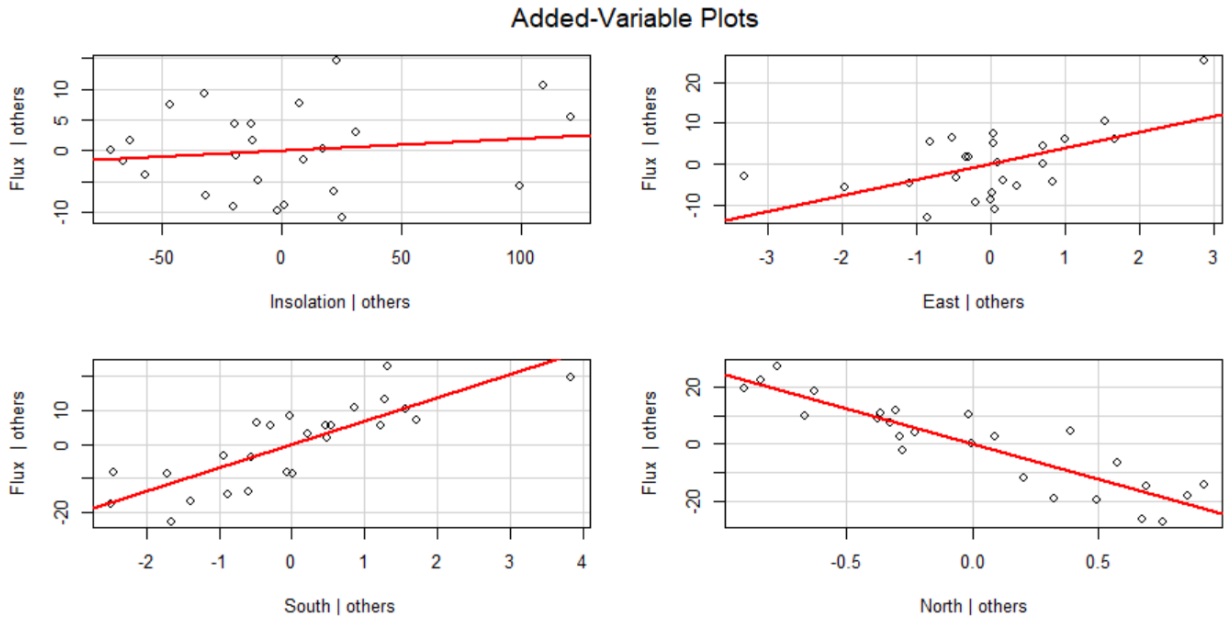
c)
max(abs(rstudent(mlrFlux))) = 2.568805 = internally studentized residual with highest magnitude

As a rough guideline, if the observation with the highest externally studentized residual has a magnitude greater than 2, then we have a bit of an outlier, if it is greater than 3, then we can consider a more troubling outlier. Since we have 2.568805 < 3, we have a bit of an outlier.

d)
max(cooks.distance(mlrFlux)) = 0.4791133

As a rough guideline, a very influential observation has a Cook's distance greater than one. This observation is around half that so it can't be said that it is very influential, but at almost 0.5, it's starting to become influential.

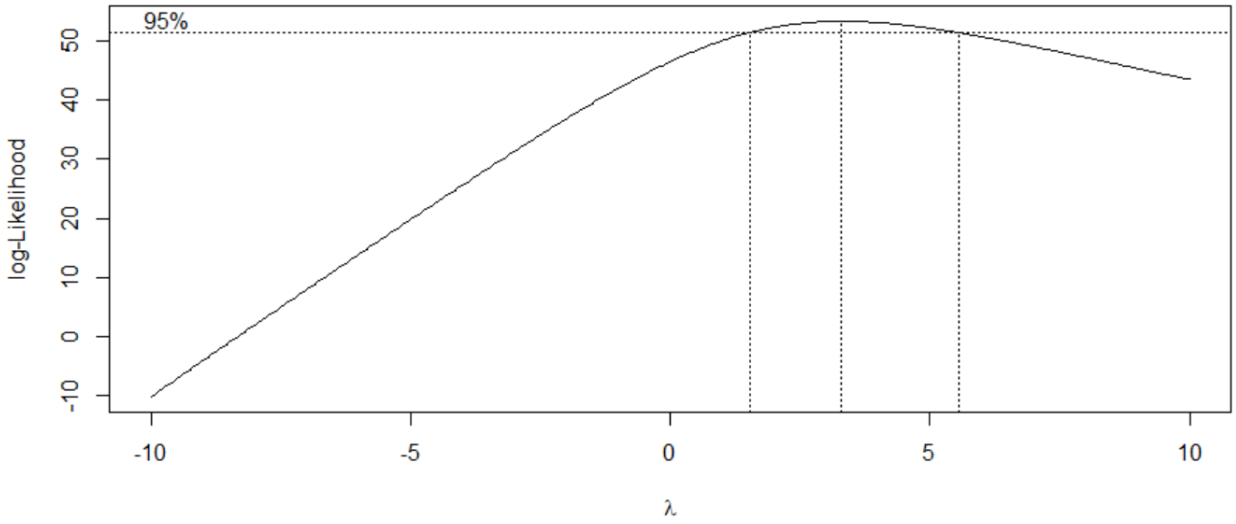
e)
avPlots(mlrFlux)



The partial regression plots illustrate the relationship between a predictor variable and response variable after holding the remaining predictor variables constant. This helps us identify each predictor variables contribution to the model. For example, North has a larger contribution than Isolation when added last to the model. It is also useful in illustrating outliers and points with high leverage in each predictor variable. For example, East has a point that is extreme in x-space and has high leverage. We also note that each partial regression plot has the same slope.

We can also see that generally there aren't problems with curvature or variance in each of the partial regressions. The observations cluster fairly evenly around the regression line instead of curving or blowing up.

f)
test = boxcox(mlrFlux, lambda = seq(-10,10,0.05))
test$x[which.max(test$y)]



The approximate confidence interval for lambda has a lower bound of 1.51 and an upper bound of 6.51.

With a lambda of 3.3, this doesn't immediately imply a simple or convenient transformation on y. It is outside the commonly chosen bounds of -2 and 2. If we were to do a transformation, we could use a (y^3.3 -1)/3.3, or we could consider cubic transformation. y^3.

However, since the Boxcox value is outside the common bounds and relatively close to a Boxcox value of 3, and I would probably look a little closer at the data and see if it's in a shape the could be unwound by a lamba of 3. Considering our exploration of multicolinearity in Q3a-d, we might see a cubic distribution, possibly something with increasing variance at the extremes.