Name: ___SOLUTION_____

CHE384 Data to Decisions
Fall, 2018, Chris A. Mack

Exam #2 – In Class Portion (25%)

open book, open notes, calculators and laptops allowed

1) For the first model that you generated during the take-home portion, calculate the variance inflation factors for each regressor variable. What can you conclude? Do the same for the second (reduced) model. What changes?

First Model:

```
Variance Inflation Factors (>4=start worrying; >10=do something?):
Material1  Material2 Material3 Condition
 1.762293  1.554770  2.346033  2.114003
Is the mean VIF much bigger than 1?
mean VIF =  1.944275
```

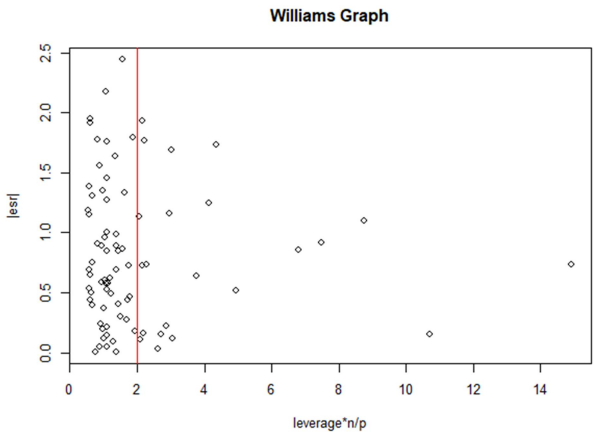
None of the individual VIFs are greater than 4, so none are a major problem. The mean VIF is moderately large, so multicoliniearity definitely exists.

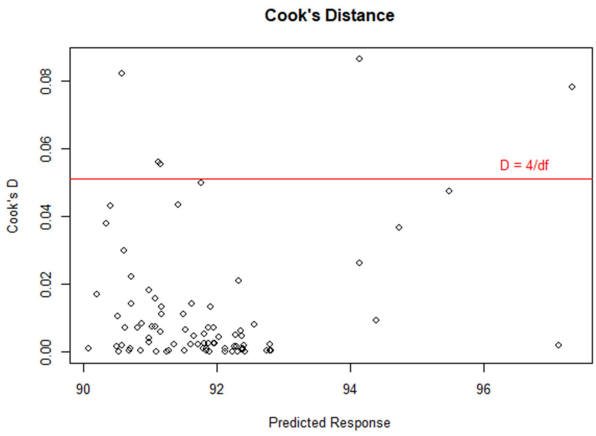
Second Model:

```
Variance Inflation Factors (>4=start worrying; >10=do something?):
Material1 Material2 Condition
 1.687848  1.536171  1.130783
Is the mean VIF much bigger than 1?
mean VIF =  1.451601
```

None of the individual VIFs are greater than 4, and are smaller than the first model. The mean VIF is also smaller. So, multicollinearity is reduced in the subset model compared to the first model. In particular, the VIF for the Condition variable is significantly reduced when Material 3 is removed from the model, the result of strong correlation between Material 3 and Condition.

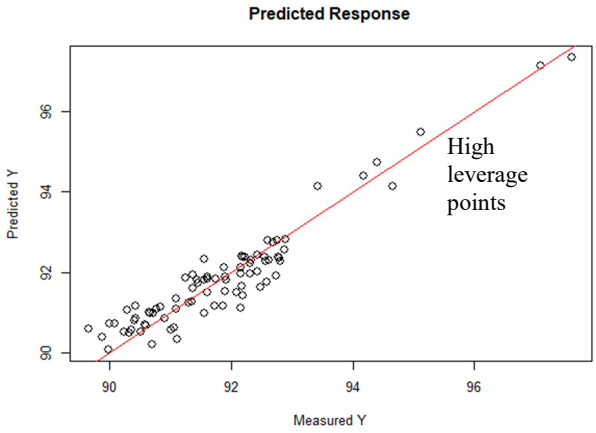
2) Consider influential data points in your model fit. Do you have any data that you might consider highly influential? If so, why are they influential? What might you do about that?



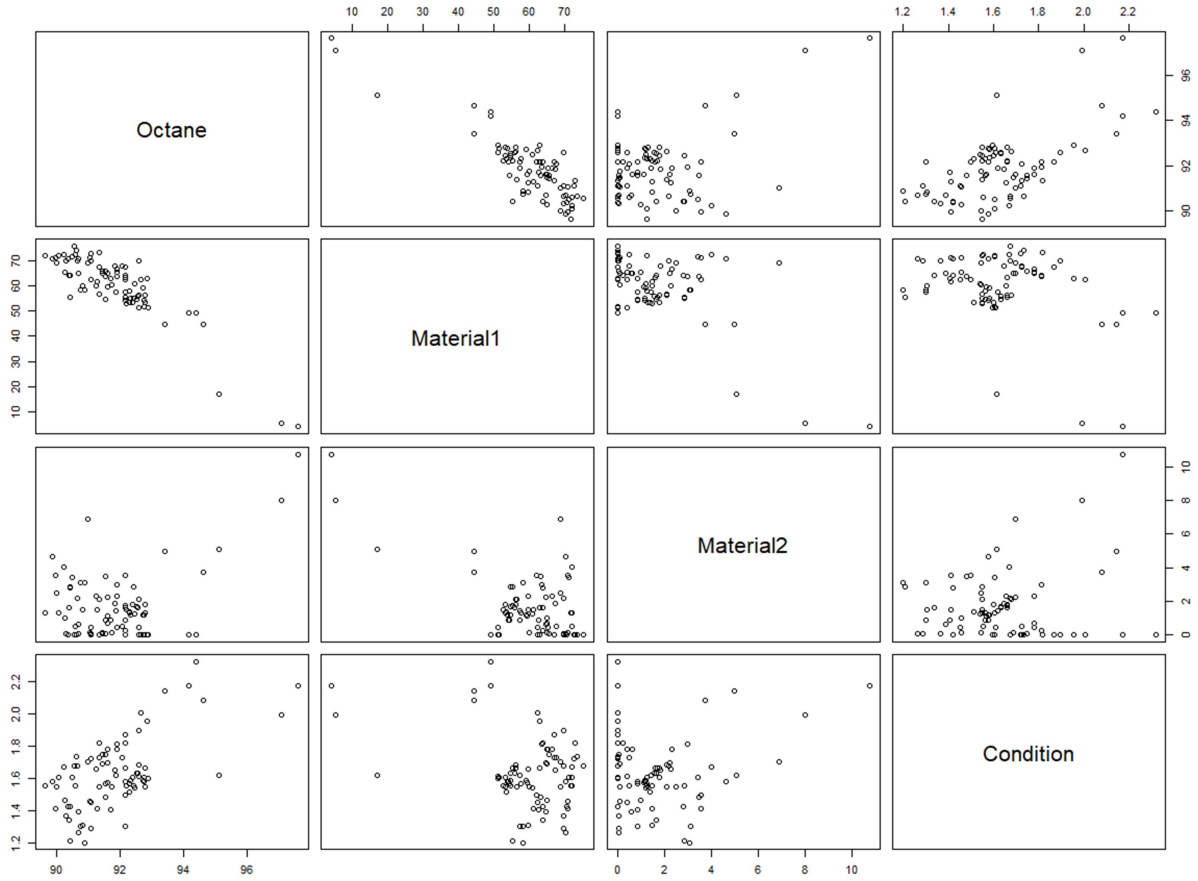


The William graph and Cook's Distance plots show several high leverage points and a few moderately high influence points (shown here for the subset model – similar results are obtained for the full model).

The high leverage points come from the high octane responses. Two of the three highest Cook's distance values are also at high octane responses. The way to reduce their leverage is to increase the number of data collected at high octane values.



3) Consider the sampled material 1 and material 2 parameter space. Are there regions of that space that are not adequately sampled? What are they? What might sampling in those regions do for your regression?



Looking at the scatter plots of each input variable, we see that there are very few data points that have low amounts of Material 1 (< 50) and very few data points with high amounts of Material 2 (> 4). These are also the parts of the sample space that produce higher octane responses (the two highest octane values use the two most extreme values for Materials 1 and 2). Thus, sampling these parts of the sample space could be valuable in building a better model.