CHE384 Data to Decisions SOLUTION
Chris Mack, Fall 2018

Exam 2 – Take home portion (75%)

Notes:

- Bring your printed exam (this document filled in with answers to the following questions) to class, to be turned in at the end of class Tuesday. Try to make this document stand-alone, so that it makes sense without a spreadsheet or R code (e.g., insert graphs and modeling results as appropriate, explain what you did to get your results).
- Please name the file with supporting calculations using this format: Exam2_yourname.xlsx or Exam2_yourname.R, and email the files by the start of class on Tuesday to: chris@lithoguru.com.

The Octane data found in Data_Sets_2.xlsx shows how three different materials in the feed stock and a composite variable describing processing conditions affect the octane rating of refined gasoline. Since higher octane is worth a lot of money to a refinery, we wish to build a multiple regression model to predict resulting octane depending on feed stock composition and processing conditions.

1. Generate an OLS model with all main effects included. Perform standard regression diagnostics on this model. What can you conclude?

2. Next, generate a subset model with the least significant main effect excluded. Compare these two models using all of the model comparison tools we have learned. What can you conclude?

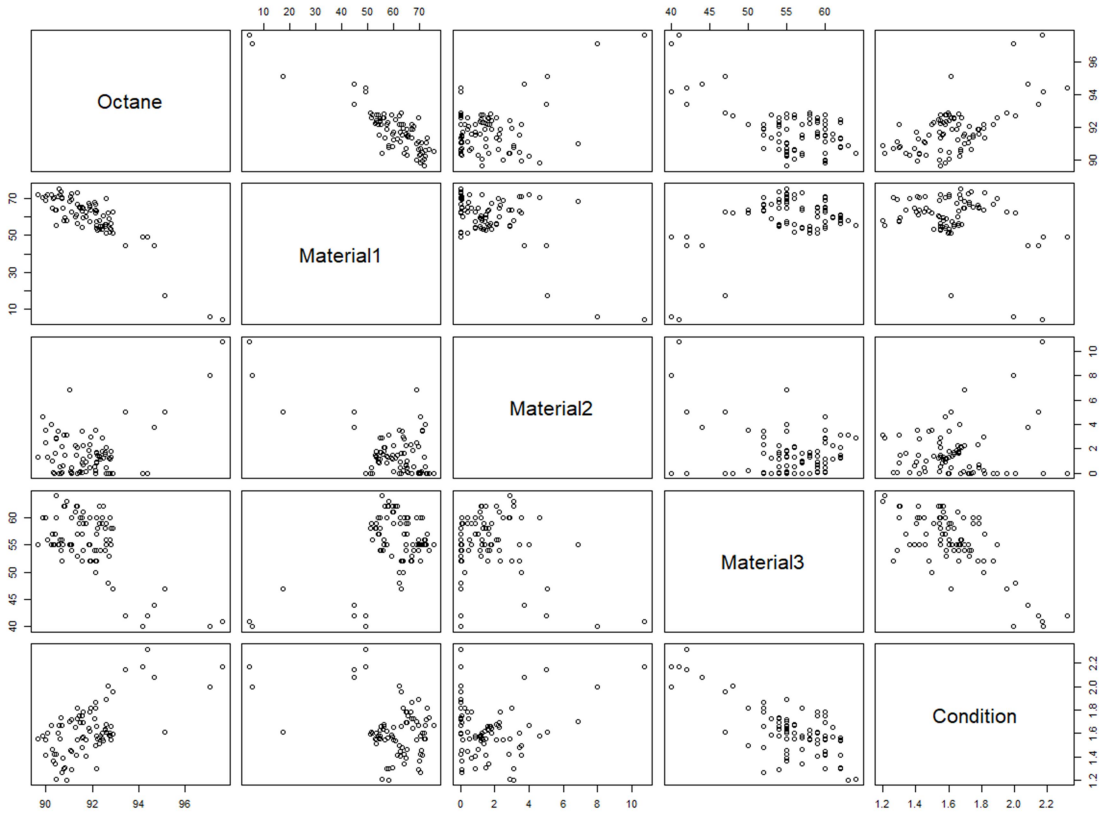
3. If your goal was to produce gasoline at an octane rating of 95, pick one set of operating conditions that would do so. Make sure that this operating condition set is within the scope of the model (that is, within the ranges for each variable used to build the model).

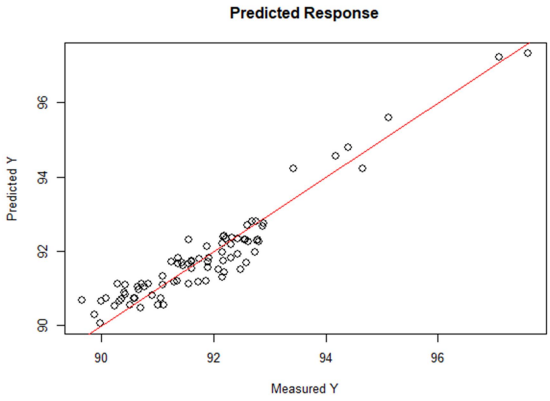
1. Generate an OLS model with all main effects included. Perform standard regression diagnostics on this model. What can you conclude?

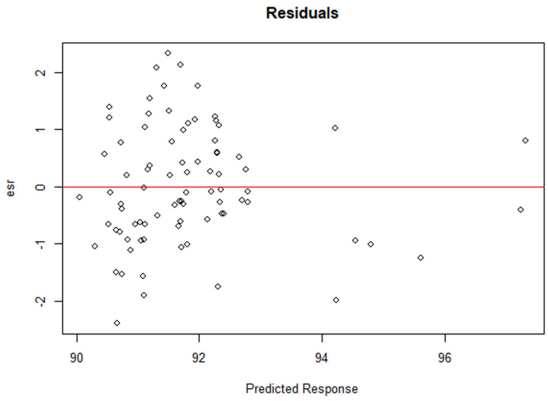
```
Model:
lm(formula = Octane ~ Material1 + Material2 + Material3 + Condition, data =
octanedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00612 -0.28588 -0.04679  0.32159  0.98069

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 95.853150   1.224877  78.255  < 2e-16
Material1   -0.092821   0.005235 -17.729  < 2e-16
Material2   -0.126798   0.032157  -3.943 0.000176
Material3   -0.025381   0.013971  -1.817 0.073160
Condition    1.967603   0.324573   6.062 4.65e-08

Residual standard error: 0.4415 on 77 degrees of freedom
Multiple R-squared:  0.9056,   Adjusted R-squared:  0.9007
F-statistic: 184.7 on 4 and 77 DF,  p-value: < 2.2e-16

                  2.5 %      97.5 %
(Intercept) 93.41410886 98.29219160
Material1   -0.10324577 -0.08239537
Material2   -0.19083204 -0.06276466
Material3   -0.05320054  0.00243918
Condition    1.32129610  2.61390984

AIC = 105.4492     BIC = 119.8896
PRESS = 17.09105      Predictive R^2 = 0.8925193

------------------------
Testing for Multicollinearity:
Correlation Matrix:
              Octane  Material1  Material2  Material3  Condition
Octane     1.0000000 -0.8701592  0.3923593 -0.6384901  0.6287297
Material1 -0.8701592  1.0000000 -0.5894513  0.4487330 -0.3369172
Material2  0.3923593 -0.5894513  1.0000000 -0.2983958  0.1611956
Material3 -0.6384901  0.4487330 -0.2983958  1.0000000 -0.7217482
Condition  0.6287297 -0.3369172  0.1611956 -0.7217482  1.0000000

Eigenvalues:
[1] 3.08903039 1.05946720 0.53554867 0.26620139 0.04975235
Condition Number (is it large, > 100 or so?):
[1] 62.08814

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
------------------------
Testing for Outliers:
Grubbs' Critical Value (alpha = 0.01) = 3.709667
```

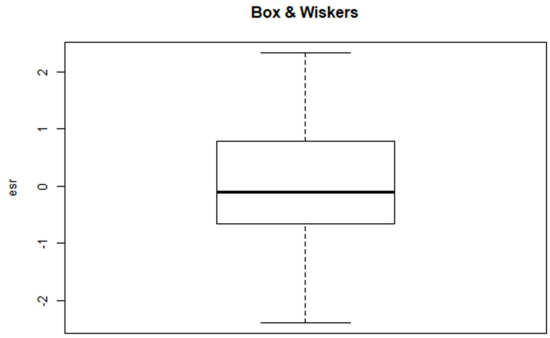
```
Grubbs' test for one outlier using outlierTest from the car package:

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
61 -2.387102           0.019469           NA
------------------------
Testing Externally Studentized Residuals (esr) for Normality:
(Small p-values means we can reject the assumption of esr normality)
Skewness Z =  0.4448 ,   p-value =  0.656452
Kurtosis Z =  -0.7148 ,   p-value =  0.4747441

        Shapiro-Wilk normality test

data:  esr
W = 0.98994, p-value = 0.7789

------------------------
Testing Externally Studentized Residuals (esr) for Influence:
Cook's Distance cut-off (4/df) =  0.05194805
Maximum Cook's Distance =  0.09961329
  which occurs at index =  73
Maximum DFFITS =  0.7189754
  which occurs at index =  73
Intercept Maximum DFBETA =  0.5103312
  which occurs at index =  82
Material1 Maximum DFBETA =  0.002342116
  which occurs at index =  77
Material2 Maximum DFBETA =  0.01919062
  which occurs at index =  44
Material3 Maximum DFBETA =  0.006803743
  which occurs at index =  82
Condition Maximum DFBETA =  0.1359357
  which occurs at index =  82
------------------------
Testing for Homoscedasticity:
(Data sorted by y-hat and split in half)
(Small p-value indicates heteroscedasticity)

Breusch-Pagan test from bptest, lmtest package:

        studentized Breusch-Pagan test

data:  model
BP = 2.8947, df = 4, p-value = 0.5756

Barlett test from bartlett.test, stats package:

        Bartlett test of homogeneity of variances

data:  esr_sorted and as.factor(group)
Bartlett's K-squared = 1.6749, df = 1, p-value = 0.1956

Brown-Forsythe test from levene.test, lawstat package:

        modified robust Brown-Forsythe Levene-type test based on the
absolute deviations from the median
```

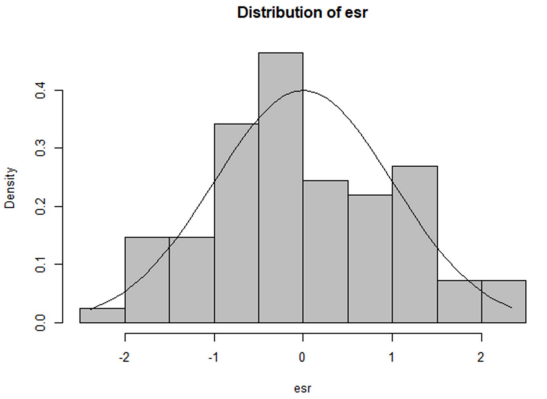
```
data:  esr_sorted
Test Statistic = 1.3143, p-value = 0.255
```

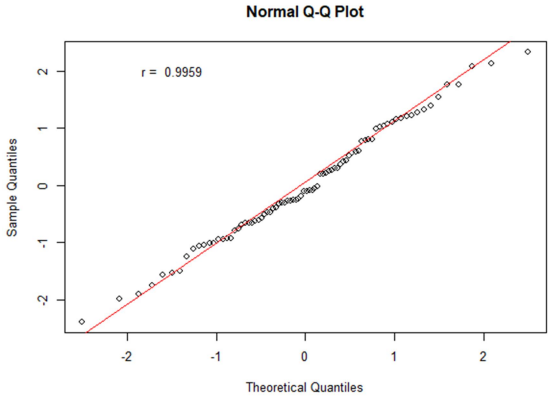


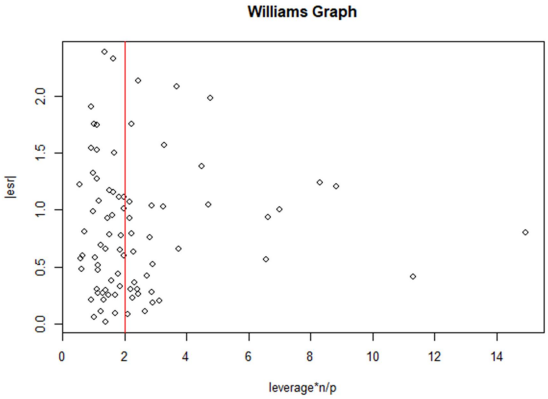

Predicted Response
Predicted Y
Measured Y
90
92
94
96


Residuals
esr
Predicted Response
-2
-1
0
1
2
90
92
94
96

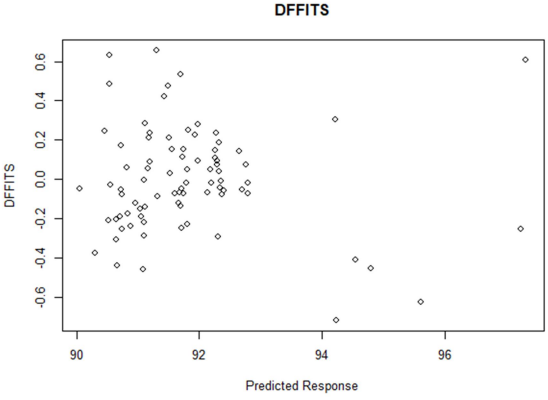

Box & Wiskers
esr
-2
-1
0
1
2


Distribution of esr
Density
esr
0.0
0.1
0.2
0.3
0.4
-2
-1
0
1
2


Normal Q-Q Plot
r = 0.9959
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2


Williams Graph
|esr|
leverage*n/p
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8
10
12
14


Cook's Distance
Cook's D
D = 4/df
Predicted Response
0.00
0.02
0.04
0.06
0.08
0.10
90
92
94
96


DFFITS
DFFITS
Predicted Response
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
90
92
94
96

From the graphs, the model appears appropriate with no obvious model error. The Grubbs test did not reveal an outlier at the 0.01 significance level. The skewness, kurtosis, and Shipiro-Wilk tests did not allow us to reject the assumption of normally distributed residuals ($\alpha = 0.05$). The Breusch-Pagan, Bartlett, and Brown-Forsythe tests could not reject the assumption of homoscedasticity at $\alpha = 0.05$. There were, however, several highly influential data points (two with leverage above 10) occurring at high octane values. No remediation is warranted, however. There is some multicollinearity in the model, but the condition number and variance inflation factors are not overly concerning.

2. Next, generate a subset model with the least significant main effect excluded. Compare these two models using all of the model comparison tools we have learned. What can you conclude?

The Material 3 model coefficient had a p-value of 0.073 and so could reasonably be considered statistically equivalent to zero. This term can be removed to create a subset model.

```
Model:
lm(formula = Octane ~ Material1 + Material2 + Condition, data = octanedata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93875 -0.27262 -0.05275  0.33081  1.04178

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.898362   0.593786 158.135  < 2e-16
Material1   -0.094775   0.005199 -18.231  < 2e-16
Material2   -0.120409   0.032432  -3.713 0.000383
Condition    2.369727   0.240857   9.839 2.58e-15

Residual standard error: 0.4479 on 78 degrees of freedom
Multiple R-squared:  0.9016,   Adjusted R-squared:  0.8978
F-statistic: 238.2 on 3 and 78 DF,  p-value: < 2.2e-16

                 2.5 %      97.5 %
(Intercept) 92.7162254 95.08049928
Material1   -0.1051253 -0.08442556
Material2   -0.1849769 -0.05584077
Condition    1.8902163  2.84923699

AIC = 106.8906     BIC = 118.9242
PRESS = 17.19044      Predictive R^2 = 0.8918943
------------------------
```

For this subset model, all coefficients are statistically significant. The conclusions about outliers, normality, homoscedasticity, and influence are unchanged from the full model.

	Full Model	Subset Model
Adjusted R2	0.9007	0.8978
AIC	105.45	106.89
BIC	119.89	118.92

Comparing these two models using the Adjusted R^2, AIC, and BIC gives mixed results. The Adjusted R^2 and AIC show that the full model is better, but the BIC gives the nod to the subset model.

An ANOVA table was generated comparing the two models using a partial F-test (null hypothesis: the extra parameters in the full model have coefficients=0). A p-value of 0.073 means that we cannot reject the null hypothesis at an $\alpha = 0.05$ significance level (this was also

obvious from the studentized t-test of the model coefficient for Material 3, which has the same p-value since that is the only term that is different between the models).

```
Analysis of Variance Table

Model 1: Octane ~ Material1 + Material2 + Material3 + Condition
Model 2: Octane ~ Material1 + Material2 + Condition
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     77 15.006
2     78 15.649 -1  -0.64318 3.3003 0.07316
```

3. If your goal was to produce gasoline at an octane rating of 95, pick one set of operating conditions that would do so. Make sure that this operating condition set is within the scope of the model (that is, within the ranges for each variable used to build the model).

Using the subset model,

```
Octane = 93.898 -0.094775*Material1 -0.120409*Material2 + 2.369727*Condition
```

The range of each parameter in the dataset is:

Material 1: 4.23 – 75.54
Material 2: 0.00 – 10.76
Condition: 1.19975 – 2.31909

Form the model we see that adding Material 1 and Material 2 reduces the octane from its intercept value of about 93.9, but higher condition increases the octane. Thus, let's pick low values for materials 1 and 2, then find the condition value required to make the octane 95. Setting Material 1 to 20 and Material 2 to 5, a Condition of 1.52 produces an octane of 95. Of course, many other solutions are possible.

Note that using the subset model is roughly equivalent to setting the Material 3 amount to be zero in the full model.