CHE384 Data to Decisions SOLUTION
Chris Mack, Fall 2018

**Exam 1 – Take home portion (75%)**

Notes:

- Bring your printed exam (with answers to the following questions, including graphs, etc.) to class, to be turned in at the end of class Thursday. Try to make this document stand-alone, so that it makes sense without the supporting spreadsheet or R code/results (e.g., insert graphs as appropriate, provide values for statistics and results, explain what you did to get your results, etc.).
- Please name the file with supporting calculations using this format: Exam1_yourname.xlsx or Exam1_yourname.R, and email the supporting documents by the start of class on Thursday to: chris@lithoguru.com.

In Data_Sets_2.xlsx, the "Water BP" tab contains measurements of the boiling point of water (T) versus barometric pressure (P). For an ideal gas, these are related by the Clausius-Clapeyron equation:

$$\ln P = -\frac{L}{R}\left(\frac{1}{T}\right) + c$$

where L = specific latent heat of vaporization, and R = universal gas constant.

1. (30 pts) Using OLS, create a model to predict the boiling point of water given the barometric pressure. Report the best fit model coefficients and their confidence intervals. Provide an explanation for what this model means in the context of this problem (i.e., what does the slope tell you? what does $R^2$ tell you?)

2. (35 pts) Test the assumptions of OLS for this model. Provide the results of these tests and your conclusions.

3. (5 pts) Perform an Overall F-test on the model. What do you conclude?

4. (5 pts) There are other options for regression besides OLS. Based on the regression approaches we have discussed in class so far, describe which one(s) might be a candidate approach for this data set and why. (You do not actually need to perform the regression!)

1. (30 pts) Using OLS, create a model to predict the boiling point of water given the barometric pressure. Report the best fit model coefficients and their confidence intervals. Provide an explanation for what this model means in the context of this problem (i.e., what does the slope tell you? what does $R^2$ tell you?)

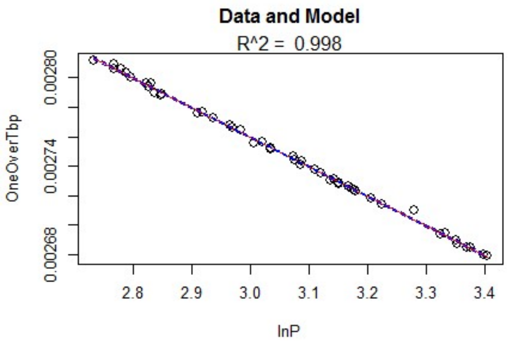
Model: $1/T = b_1 \ln(P) + b_0$ (T in Kelvin; units of P only affect the intercept)

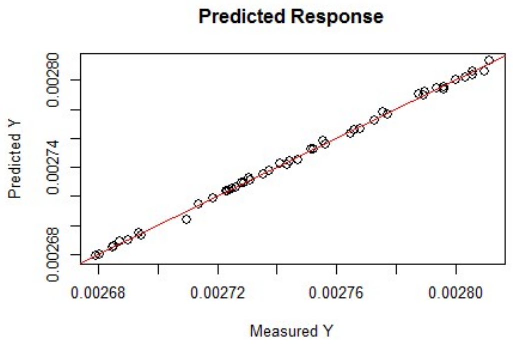
The coefficient of ln(P) is –(R/L) with units of 1/K. The model fit provides a value of this coefficient of -2.003e-04 $K^{-1}$ with a standard error of 1.21e-06 $K^{-1}$. The 95% confidence interval of the slope is (-0.0002028, -0.0001979). Using R = 8.3145 J/K-mol, this means that from the data, L = 41.5 ± 0.5 kJ/mol. The literature value for the specific latent heat of vaporization of water is 40.7 kJ/mol.

The $R^2$ was 0.9983, meaning that 99.83% of the variance in 1/T is explained by the model.

R Results:

```
Call:
lm(formula = OneOverTbp ~ lnP, data = Water)

Residuals:
       Min         1Q     Median         3Q        Max 
-2.983e-06 -8.369e-07 -8.390e-08  6.918e-07  6.187e-06 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.361e-03  3.703e-06   907.7   <2e-16
lnP         -2.003e-04  1.206e-06  -166.1   <2e-16

Residual standard error: 1.631e-06 on 46 degrees of freedom
Multiple R-squared:  0.9983,   Adjusted R-squared:  0.9983 
F-statistic: 2.759e+04 on 1 and 46 DF,  p-value: < 2.2e-16

                    2.5 %        97.5 %
(Intercept)  0.0033532060  0.0033681116
lnP         -0.0002027777 -0.0001979217
```

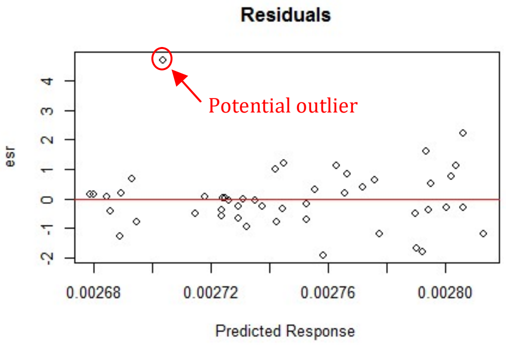


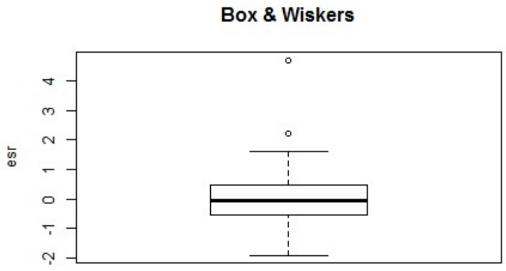


2. (35 pts) Test the assumptions of OLS for this model. Provide the results of these tests and your conclusions.

**Step 1: Plot and study residuals.**

Plotting esr versus predicted response, no obvious model error was observed.





**Step 2: Look for outliers**

The Grubb’s Test using α = 0.01 found one data point (at 1/T = 0.0271, ln(P) = 3.28, data point number 12, with esr = 4.69) that tested positive as an outlier (p-value = 3e-5).

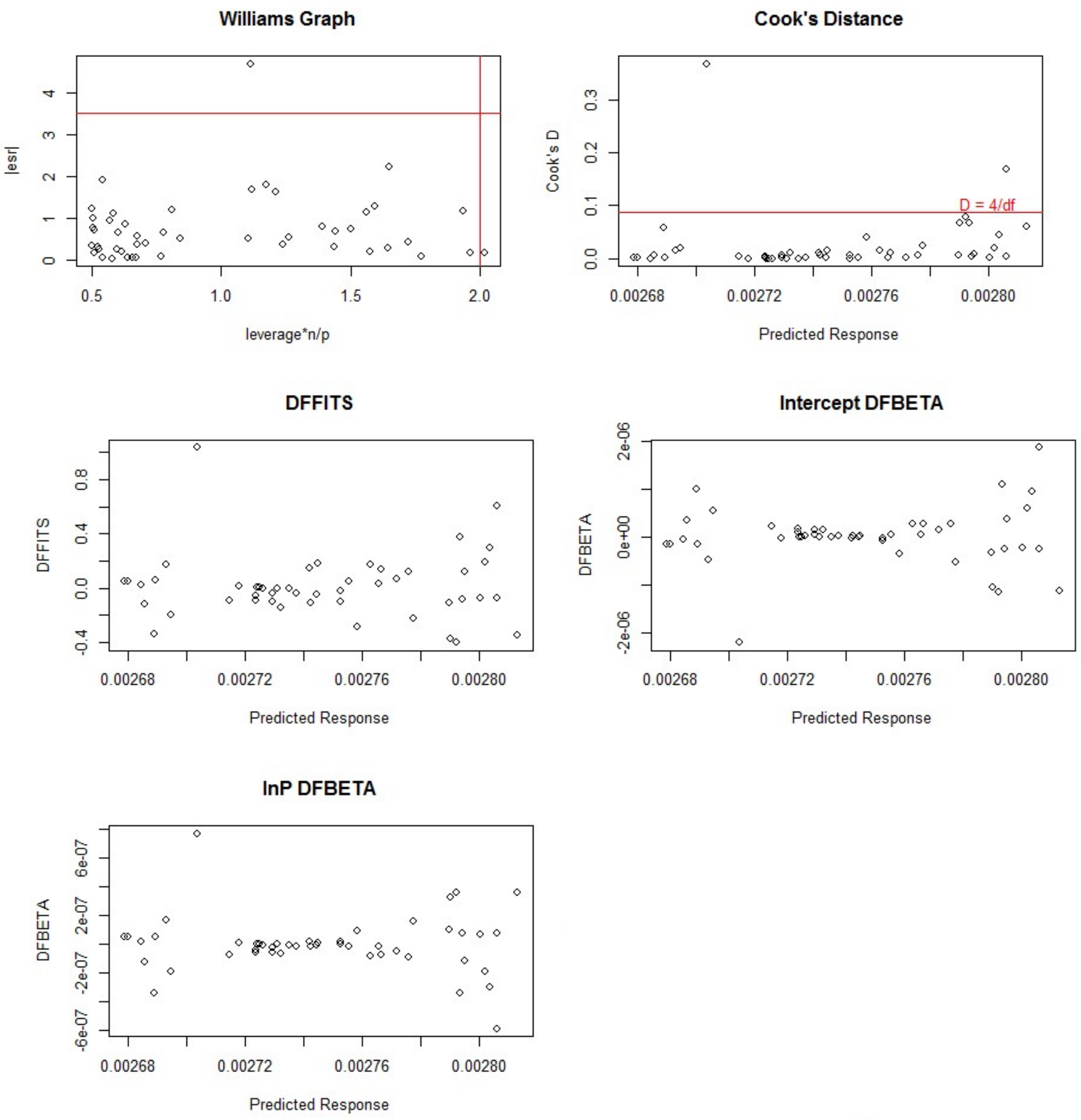
```
Testing for Outliers:
Grubbs' Critical Value (alpha = 0.01) = 3.497499
Grubbs' test for one outlier using outlierTest from the car package:
   rstudent unadjusted p-value Bonferonni p
12 4.686659          2.5909e-05    0.0012436
```

**Step 3: Look for influential data points**

The Cook’s Distance, DFFITS, and DFBETA tests all identified this same point (#12) as being the most influential, though its leverage was not high (normalized leverage just over 1.0, the average leverage for all points). The Cook’s distance for this point (0.37) was bigger than the cut-off guideline (4/n = 0.87) but less than 1, and the DFFITS value (1.03) was bigger than it’s cut-off (sqrt(4p/n) = 0.42). The DFBETAs for the slope and intercept, however, were much smaller than their cut-offs, indicating that this data point is not influential in determining the model coefficients. For example, the slope of the model fit changed from -2.003e-04 to -2.011e-04 when point #12 was removed, within the confidence interval of that parameter. In conclusion, while one data point was found to be a statistical outlier, it was not influential and so will not be removed.

```
Using Externally Studentized Residuals (esr) to Look for Influence:
Cook's Distance cut-off (4/df) =  0.08695652
Maximum Cook's Distance =  0.3675636
  which occurs at index =  12
Maximum DFFITS =  1.034488
```

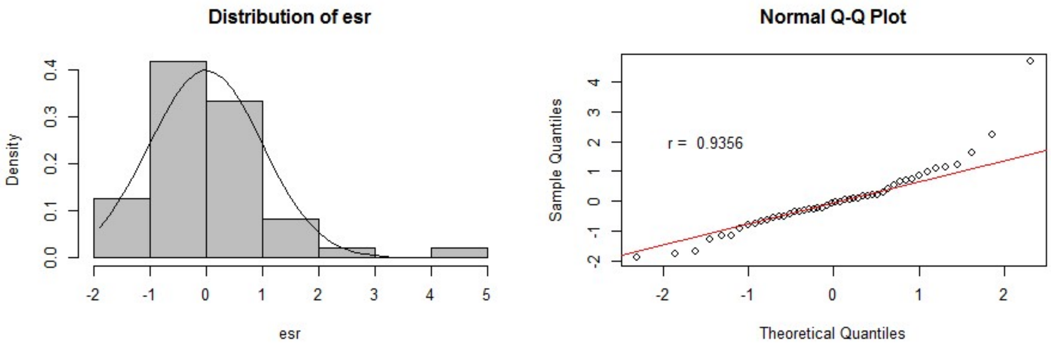
```
  which occurs at index =  12
Intercept Maximum DFBETA =  2.217721e-06
  which occurs at index =  12
lnP Maximum DFBETA =  7.680857e-07
  which occurs at index =  12
```



## Step 4: Test for normality

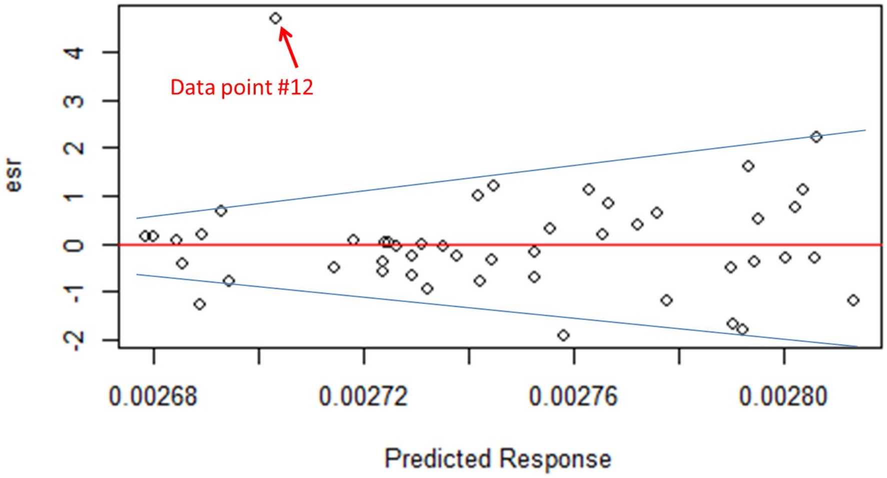
The Grubb's Test (α = 0.01) found one data point that tested positive as an outlier (p-value = 3e-5). Testing for skewness and kurtosis (α = 0.05) found that both skewness (p-value = 1e-6) and kurtosis (p-value ~ 0) exist in the esr, but of course that could be due to the detected outlier. The Shapiro-Wilk normality test (α = 0.05) also rejected the null hypothesis of a normal distribution of esr (p-value = 0.0002). As confirmation, the Grubb's identified outlier (data point #12) was removed and these tests were repeated. Without this one data point, the skewness, kurtosis, and Shapiro-Wilk tests failed to reject the null hypothesis that the esr distribution is normal. Thus, it appears that we have a set of residuals that is following an approximately normal distribution with the exception of one data point (#12).

```
Testing Externally Studentized Residuals (esr) for Normality (using all
data points):
(Small p-values means we can reject the assumption of esr normality)
Skewness Z =  4.825 ,   p-value =  1.400371e-06
Kurtosis Z =  9.264 ,   p-value =  0

        Shapiro-Wilk normality test

data:  esr
W = 0.88536, p-value = 0.0002189
```



**Step 5: Test for heteroscedasticity**

Several tests for heteroscedasticity (α = 0.05) did not allow the assumption of homoscedasticity to be rejected when the outlier point #12 was included (see R results below). When that data point was excluded, however, all tests for heteroscedasticity rejected the null hypothesis that the esr distribution was homoscedastic (Breusch-Pagan test p-value = 0.0036, Barlett test p-value = 0.00078, Brown-Forsythe test p-value = 0.0014). A plot of the externally studentized residuals versus predicted value indicates an increase in variance as a function of y, but the outlier data point #12 masks this heteroscedasticity. Roughly speaking, the standard deviation of the residuals is about doubling over the range of the predicted response. This level of variation in the variance is probably too small to be worrisome, though it will mean that estimated confidence intervals of the OLS model parameters are too small. No attempt was made to address this heteroscedasticity (see answer to question #4).



```
Testing for Homoscedasticity (using all data points):
(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 = 0.045956, df = 1, p-value = 0.8303

Barlett test from barletett.test, stats package:

        Bartlett test of homogeneity of variances

data:  esr_sorted and as.factor(group)
Bartlett's K-squared = 0.011532, df = 1, p-value = 0.9145

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 = 1.8262, p-value = 0.1832
```

3. (5 pts) Perform an Overall F-test on the model. What do you conclude?

> The overall F-test produced a very small p-value (2e-16), indicating that the probability of getting the observed trend randomly is extremely small. The model is statistically significant. Note that the Overall F-test is a blunt instrument, and only tells whether this model is better than nothing. It is only a useful test when it is difficult to discern any trend in the data.

```
F-statistic: 2.759e+04 on 1 and 46 DF,  p-value: < 2.2e-16
```

4. (5 pts) There are other options for regression besides OLS. Based on the regression approaches we have discussed in class so far, describe which one(s) might be a candidate approach for this data set and why. (You do not actually need to perform the regression!)

> Since both measurement of T and measurement of P include experimental error, added knowledge of the measurement errors for each of these terms could be used to perform a total regression. Since there is no obvious “right” variable to use as the regressor versus the regression output, a geometric regression is also an option.
>
> Additionally, the presence of heterscedasticity (found without the outlier point) indicates that wieghted regression might be useful.