Name: ____SOLUTION_____

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

Exam #1 – In Class Portion (25%)

open book, open notes, calculators and laptops allowed

1) Use your OLS model to predict the boiling point of water when the barometric pressure is 15 inches of mercury. Provide a 95% confidence interval for this prediction (the formula below can be used).

$$s_{\hat{y}_i}^2 = \frac{[SE(e)]^2}{n} + [SE(b_1)]^2(x_i - \bar{x})^2$$

x_i = ln(15) = 2.70805, , $\bar{x}$ = 3.06332, $SE(b_1)$ = 1.206e-06, $SE(e)$ = 1.631e-06, DF = 46

SE(prediction) = 4.9e-7

Prediction (1/K) = 0.002818102 with 95%CI = (0.002817117, 0.002819088)

Prediction (K) = 354.85 with 95% CI = (354.73, 354.97)

R code to accomplish this task:

```
new <- data.frame(lnP = log(15))
predict(model, new, se.fit = TRUE)
predict(model, new, interval = "confidence")
```

2) Explain the difference between an internally studentized residual (isr) and an externally studentized residual (esr). Why would one choose to use an esr instead of an isr for analysis?

Internally studentized residual (isr) is the residual value divided by the standard error of that residual.

$$isr_i = \frac{e_i}{SE(e_i)} = \frac{e_i}{s_e\sqrt{1 - h_{ii}}}$$

The externally studentized residual uses the same formula but excludes the data point under consideration from the calculation of s_e. Thus, if that data point happens to be bad (an outlier, for example), then the esr will not be lowered by the fact that the s_e is high due to the outlier. Also, the esr is t-distributed and so is easy to test (the isr has a very complicated distribution). For these reasons, the esr is preferred for plotting, testing, etc.

3) Why is it important to deal with outliers before testing for normality (using skewness and kurtosis tests, for example)?

One outlier is enough to cause every test for normality to fail. Thus, if you plan to remove an outlier from your analysis, the normality tests should be done on the residuals of the model generated from the data set that excludes the residuals.

4) What two properties do you look for in a "best fit" estimate of a model parameter?

We want our model estimates to be unbiased and to have low standard errors (low uncertainty in the parameter estimates).

5) Explain what the hat matrix is used for.

The hat matrix is often used in the calculation of the OLS solution, though that is an internal aspect of the OLS algorithms used. From a user perspective, the diagonal of the hat matrix provideds the leverage of each data point.

6) Explain what is meant by a regression result that is "fragile".

A regression result is fragile if that result changes significantly by the removal of one or a very few data points from the data set.

7) In this problem, one could use 1/T as the regressor (x) variable and ln(P) as the output (y) variable, or the other way around (modeling 1/T as a function of ln(P)). Is there any difference between these two approaches? Would you expect to get different results? Why?

Yes, there is a difference. OLS assumes that all uncertainty is in y (the predictor) and that the regressors have no uncertainty. This results in differences in the regression results. For example, a regression of 1/T as a function of ln(P) produces a slope of -2.003e-4, while a regression of ln(P) versus 1/T produces 1/slope = -2.007e-4. In this case, the difference is quite small. In other cases, the difference is larger. This is why geometric regression produces a different result than OLS.