Stat 6021: Homework Set 1

1. (R required) R provides some data sets though its datasets package that you can readily use. The list of the data sets, as well as their descriptions, can be found here.

   We will look at a classic data set concerning the Old Faithful geyser at Yellowstone National Park. To load this data set, just type `faithful`. You may attach the data by typing `attach(faithful)`. The data set contains 272 observations on two variables: *eruptions* is the eruption time in minutes, and *waiting* is the waiting time to the next eruption. It is hypothesized that the waiting time to the next eruption can be predicted using the eruption time of the current eruption.

   During the Washburn Expedition of 1870, members noticed the geyser erupted at regular intervals, thus leading to its name Old Faithful.

   (a) What is the response variable in this analysis? What is predictor in this analysis?

   (b) Produce a scatterplot of the two variables. How would you describe the relationship between the two variables?

   (c) What is the correlation between the eruption times and waiting times for the next eruption? Interpret this correlation contextually. How reliable is this interpretation?

   For parts 1d to 1i, assume the regression assumptions are all met.

   (d) Use the `lm()` function to fit a linear regression for the two variables. Where are the values of $\hat{\beta}_1$, $\hat{\beta}_0$, R^2, and $\hat{\sigma}^2$ for this linear regression?

   (e) Interpret the values of $\hat{\beta}_1$, $\hat{\beta}_0$ contextually. Does the value of $\hat{\beta}_0$ make sense in this context?

   (f) Use the `anova()` function to produce the ANOVA table for this linear regression. What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic?

   (g) Obtain the 95% confidence interval for the slope, β_1. Is this confidence interval consistent with your conclusion from part 1f? Briefly explain.

   (h) The latest eruption at Old Faithful lasted for 3.5 minutes. Obtain an appropriate 95% interval that predicts the waiting time for the next eruption.

(i) What is the 95% interval for the average waiting time for the next eruption among current eruptions that last 3.5 minutes?

(j) Create a residual plot, an ACF plot of the residuals, and the QQ plot of the residuals. Based on these plots, are the regression assumptions met? Is your answer surprising, given the context of this data set?

2. (R required) For this question, we will use the `cornnit` data set from the `faraway` package. Be sure to install and load the `faraway` package first, and then load the data set. The data explore the relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application in a study carried out in Wisconsin.

   (a) What is the response variable and predictor for this study? Create a scatterplot of the data, and interpret the scatterplot.

   (b) Fit a linear regression without any transformations. Create the corresponding residual plot. Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.

   (c) Create a Box Cox plot for the profile loglikelihoods. How does this plot aid in your data transformation?

   (d) Perform the necessary transformation to the data. Re fit the regression with the transformed variable(s) and assess the regression assumptions. You may have to apply transformations a number of times. Be sure to explain the reason behind each of your transformations. Perform the needed transformations until the regression assumptions are met. What is the regression equation that you will use?

   Note: in part 2d, there are a number of solutions that will work. You must clearly document your reasons for each of your transformations.

3. (No R required) A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The data consist of 10 shipments. The variables are number of times the carton was transferred from one aircraft to another during the shipment route (*transfer*), and the number of ampules found to be broken upon arrival (*broken*). We want to fit a simple linear regression. A simple linear regression model is fitted using R. You may assume all the regression assumptions are met. The corresponding output from R is shown next, with some values missing.

```
Call:
lm(formula = broken ~ transfer)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2000     0.6633  ______ ________ ***
transfer      4.0000     0.4690  ______ ________ ***

Residual standard error: 1.483 on 8 degrees of freedom
```

```
...

Analysis of Variance Table

Response: broken
          Df Sum Sq Mean Sq F value    Pr(>F)
transfer   1  160.0   160.0  ______ ________ ***
Residuals  8   17.6     2.2
```

The following values are also provided for you, and may be used for the rest of this question: $\bar{x} = 1$, $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 10$.

(a) Calculate the value of R^2, and interpret this value in context.

(b) Carry out a hypothesis test to assess if there is a linear relationship between the variables of interest.

(c) Calculate a 95% confidence interval that estimates the unknown value of the population slope.

(d) A consultant believes the mean number of broken ampules when no transfers are made is different from 9. Conduct an appropriate hypothesis test (state the hypotheses statements, calculate the test statistic, and write the corresponding conclusion in context, in response to his belief).

4. (No R required) A chemist studied the concentration of a solution, y, over time, x, by fitting a simple linear regression. The scatterplot of the dataset, and the residual plot from the regression model are shown in Figure 1.

   (a) The profile log-likelihoods for the parameter, λ, of the Box-Cox power transformation, is shown in Figure 2. Your classmate says that you should apply a log transformation to the response variable first. Do you agree with your classmate? Be sure to justify your answer.

   (b) Your classmate is adament on applying the log transformation to the response variable, and fits the regression model. The R output is shown in Figure 3. Write down the estimated regression equation for this model. How do we interpret the regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_0$ in context?

5. Reminder: please complete the Module 1 to 4 Guided Question Set Participation Self- and Peer-Evaluation Questions via Test & Quizzes by July 21.

Figure 1: Scatterplot of Concentration of Solution against Time (left). Residual Plot from SLR (right)

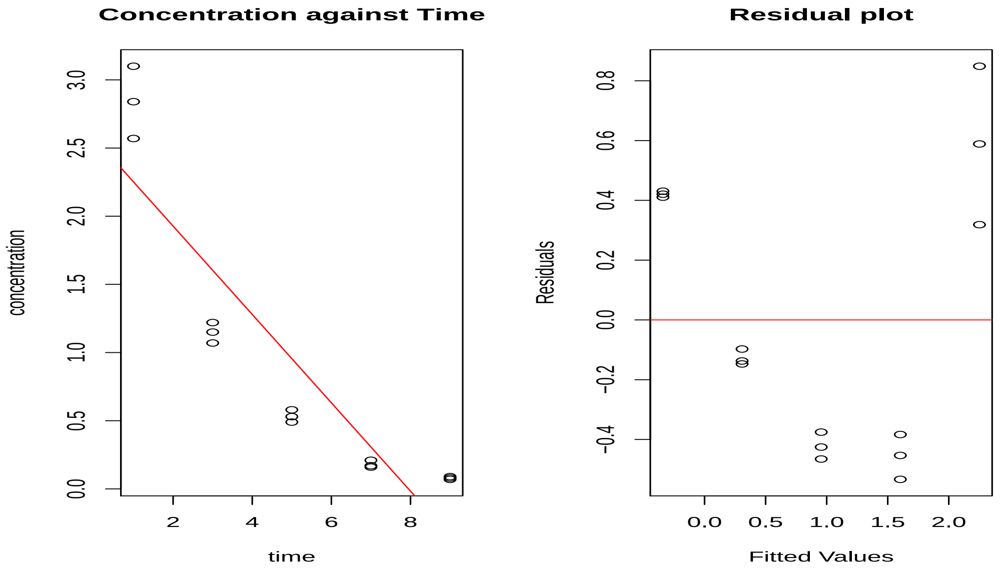


Figure 2: Profile Log-likelihoods for λ.

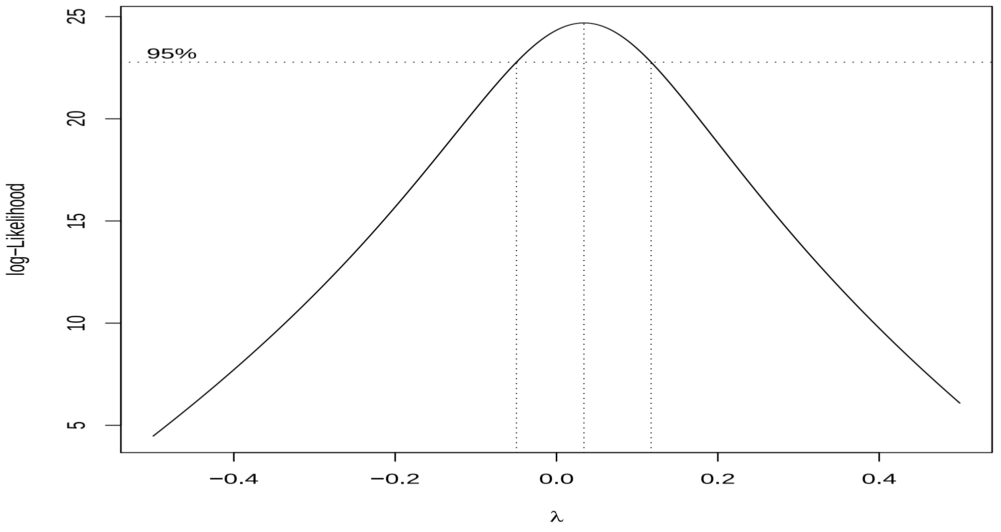
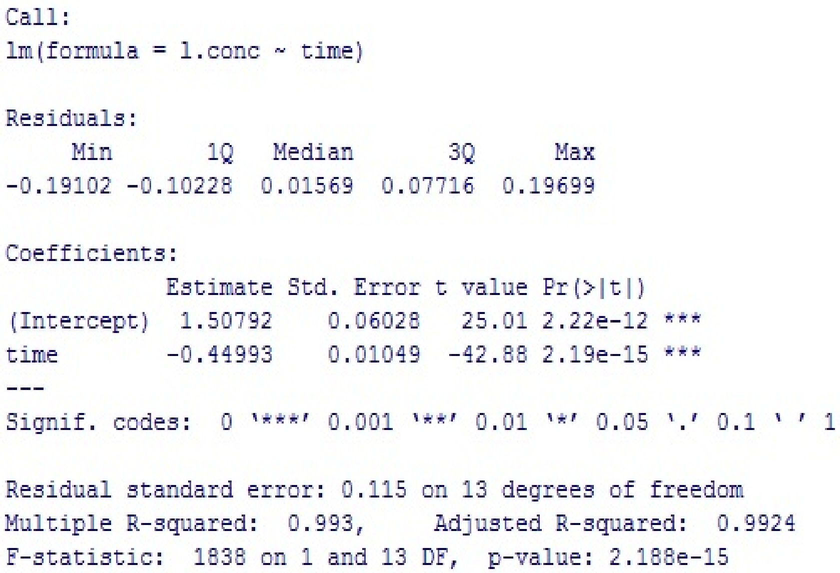


Figure 3: R Output after Transforming Response Variable.



```
Call:
lm(formula = l.conc ~ time)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.19102 -0.10228  0.01569  0.07716  0.19699 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.50792    0.06028   25.01 2.22e-12 ***
time        -0.44993    0.01049  -42.88 2.19e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.115 on 13 degrees of freedom
Multiple R-squared:  0.993,     Adjusted R-squared:  0.9924 
F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15
```