Statistics 101A
Homework Two Due Friday Jan. 31st 2020 @ 5 pm

Question One: Problem one from chapter three 3.4 Exercises (The data file airfares.txt)

Question Two: Problem eight from chapter three 3.4 exercises (The Diamond stones data file)

Question Three:

a) Using the stress echo UCLA data (see week four), fit a linear model to predict basal blood pressure from systolic blood pressure. Report the equation for the model. Report a residual plot and comment what it tells us about the assumption of linearity.

b) Report the ANOVA table. Show how you can find the F value reported in the ANOVA table using R^2. What is the null hypothesis that you are testing through ANOVA? Compare the F value that you calculate with value that you find from the F table and decide whether you are going to reject or fail to reject the null hypothesis). Check if this equation is true: $(Se)^2$ is approximately equal to $var(Y) * (1 - r^2)$.

c) Calculate R^2 adjusted and compare it to R^2. Comment on the difference.

d) Check the diagnostic plots and comment on each one of them.

e) Create two new variables: one for the leverage of a point and one for the standardized residuals. Create a table from both variables to identify the following:

Leverage/Outliers	Yes	No
Yes		
No		

f) Use ggplot2 library to create a plot of Leverage Vs Standardizes residuals divided into regions to help you identify bad and good leverage points, outliers and not leverage points and all the ordinary points.

Question Four:

Use the Echo data from question three to transform the data and compare the results to the SLR created in question three:

a) Use the inverse response plot to find the best λ to transform the y variable to minimize the SSE. Construct a SLR of the transformed y variable and systolic blood pressure. Check diagnostics. Is this one better than the SLR in question three.

b) Use the power transform function to find the best λ(s) to transform both the y variable and the x variable to make the densities of these two variables as close as possible to normal. Construct a SLR of the transformed variables. Check diagnostics. Is this one better than the SLR in question three.

Question Five:

Consider the following R output predicting Marine water growth from Freshwater growth in Salmon:

```
> SL1<- lm(salmon$Marine~salmon$Freshwater)
> summary(SL1)
Call:
lm(formula = salmon$Marine ~ salmon$Freshwater)

Residuals:
    Min      1Q  Median      3Q     Max
-88.222 -27.382  -3.406  24.784  89.977
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       511.3656    18.2547   28.01  < 2e-16 ***
salmon$Freshwater  -0.9602     0.1512   -6.35 6.75e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 39.12 on 98 degrees of freedom
Multiple R-squared:  0.2915,   Adjusted R-squared:  0.2843
F-statistic: 40.32 on 1 and 98 DF,  p-value: 6.747e-09
> summary(salmon$Freshwater)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   53.0    99.0   117.5   117.9   140.0   179.0
> var(salmon$Freshwater)
[1] 676.0541
> summary(salmon$Marine)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  301.0   367.0   396.5   398.1   428.2   511.0
> var(salmon$Marine)
[1] 2138.142
```

a) Construct ANOVA table based on the given output.

Consider the three observations: 4, 41 and 53

	Observation	SalmonOrigin	Freshwater	Marine
1	4	Alaska	86	506
2	41	Alaska	84	511
3	53	Canada	179	407

b) Which of these three points is(are) a leverage point?

c)Which of these three points is (are) an outlier?

d) Based on your answers of part b and c, classify these points as one of the following:

i) A bad leverage point

ii) An outlier but Not a leverage point.

iii) A good leverage point

iv) Not a leverage point nor an outlier (ordinary)