Data Analysis - One Predictor Linear Regression
Bushra Paracha

I. Part A

Introduction: The main goal of this report is to observe the number of observations and attain data for the independent and dependent variables presented. Any missing data in the files will be imputed appropriately to facilitate the identification of the fitted function.

Methods: Inorder to do the analysis, programming language R was used. First using R, the merge() function was used to merge both the files together. After examining them, using str() the missing values presented in the data set that need to be addressed. The imputation of the missing values will be solved by using package mice. Afterwards, we use the bootstrap method and complete() to impute the rest of the missing values. Using R, it is also capable of performing OLS regression by using the lm() function, which provides a summary of the completed data set. The ANOVA table can also be provided by using the package knitr. A scatter plot is then created by using plot() with its useful estimated regression line (Fig. 5). After this, the confidence intervals of the slope can be calculated by using confint().

Results: After using R it was found that two files contain 630 observations of 3 variables. One file contains ID numbers as integers with respective IV numbers and the other file has its ID with its respective DV numbers. There are 630 observations in this file and three columns in the data set, called 'ID', 'IV' and 'DV'. Fig.1 attached below shows the merged data set of the two files provided. After merging the data it was observed that the file has 499 complete data sets,138 cases missing either IV OR DV, 79 cases missing IV, 45 cases missing and 7 cases missing both IV AND DV. This can be seen in Fig.2. The 7 cases missing both IV and DV were dropped because they did not provide any information. Next we used the bootstrap method to impute values. It was found that there are 623 complete data sets. After finding the complete data set it was found that estimated regression equation for this set is DV = 49.0220 + 4.8722*(IV). The equation was found using simple linear regression as shown in Fig.3. For the following set the p-value,2.2e-16,provided in the summary is significantly less than 0.5 which means that we can reject the null hypothesis that the slope is 0 and that there is a significant relationship between the DV and IV. Using knitr package in R we found the ANOVA table for the following set. It was found that the F value ANOVA table further supports that we can reject the null hypothesis. This can be seen in Fig.4. After the ANOVA table we created a scatter plot for the following data set. We found the 95% and 99% confidence intervals of the slope. For the 95% confidence interval the slope is between (4.64365 5.100699) and for the 99% confidence interval the slope is between (4.571504 5.172845). Fig.5 shows the scatter plot for this data set.

Conclusion and Discussion: By looking at the p-value we see that there is a significant relationship between the DV and IV. Based on the summary of our data, the R-squared value that was found was 0.7834 or 78.34%. This typically means that 78.34% of the DVs can be analyzed by the IVs. Our estimated regression equation of DV= DV = 49.0220 + 4.8722*(IV) can be used to predict the values about 78.34% of the time, which is quite significant.

II. Part B

Introduction: For this part, the main goal is to apply a transformation of either IV or DV or both if required to find a fitted model. An appropriate LOF test will be applied. We will find repeated independent variables and bin near repeated data into one level.

Methods: For this project I chose to use R inorder to perform data analysis. First step was to get a visual representation of the data. This can be done using plot() function or View() to examine the values in a table. Next inorder to calculate the correlation coefficient needed to discern the relationship between x and y variables cor() function was used. For the transformation of the data, the function data.frame() is used and set to a different variable to keep original and transformed data separate. I then used the lm() function to view a summary of both the original and transformed data. The next step for this project was to create groups for the transformed data, for that I used the cut() function. Then using the ave() function the average values of x were calculated after binning the data. At the end inorder to perform the Lack of Fit text, I used the alr3 package remotes to access the library, where to find analysis of the variable table pureErrorAnova() function is used.

Results: The data set in part B file contains 530 observations of 3 variables “ID”, “x”, and “y”. Using the data and plot() function we view the graph to see if we can locate the transformation in the data. The graph can be seen in Fig.6. From the figure we can note that it is an exponential decay. After that using cor() function the correlation coefficient was found between x and y which came out to be -0.6422736. Looking at Fig.6 it can be seen that there is an obvious outlier. After some trial and error the transformation found was IV and log (DV). After plotting transformed data we can see the changed graph in fig.7. The correlation coefficient for the following transformed data comes out to be 0.7797685. The summary of the transformed data shows that the R squared value is 0.608 which can be seen as 8.2. This means that the 60.8% of the variance can be explained. In contrast, in the original data, summary revealed only 41.25% of the variance could be explained; this can be seen in fig 8.1. After binning the data cut() function was used to create groups as seen in fig.9. These groups were later on used for the LOF test. As seen in fig 9 we have a total of 3 groups with each group being separated by 0.3. After this we fit the bins and perform LOF test using pureErrorAnova() function. The analysis showed that the F value is 2.2178 and that the p value of the data is 0.137 within the Lack of Fit row. The high p-value within the test indicates that there is not a significant LOF in the fitted regression model after we transformed our data as seen in fig 10. It is also apparent that the p-value of x is less than 0.05, which shows that we can reject the null hypothesis and say that there is a significant relationship between x and y.

Conclusions and Discussions: After looking at the correlation coefficient in the original and transformed data, the original graph has a negative correlation while the transformed data has a positive correlation. Furthermore, we can see that R seems to improve after the data has transformed and F-value in comparison to the p-value indicates there is not a significant lack of fit in the regression model.

Part A Appendix

Fig.1. Data received using R by using str()

```
> str(PartA)
'data.frame':    630 obs. of  3 variables:
 $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
 $ IV: num  5.14 5.08 4.91 6.69 4.71 ...
 $ DV: num  69.6 73.4 NA 79 69.3 …
```

Fig.2. Pattern of missing values in merged set

```
>md.pattern(PartA_incomplete)
```

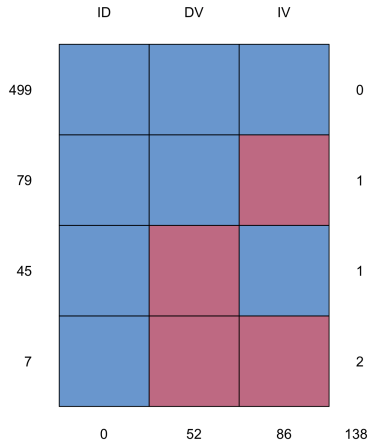


Fig.3. Summary of completed set M

```
> summary(M)
Call:
lm(formula = DV ~ IV, data = PartA_complete)
Residuals:
    Min      1Q  Median      3Q     Max
-7.9296 -2.1151 -0.1427  1.9961 11.4261
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.0220     0.5920   82.80   <2e-16 ***
IV            4.8722     0.1164   41.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.943 on 621 degrees of freedom
Multiple R-squared:  0.7384,  Adjusted R-squared:  0.738
F-statistic:  1753 on 1 and 621 DF,  p-value: < 2.2e-16
```

Fig.4. Table of the ANOVA table

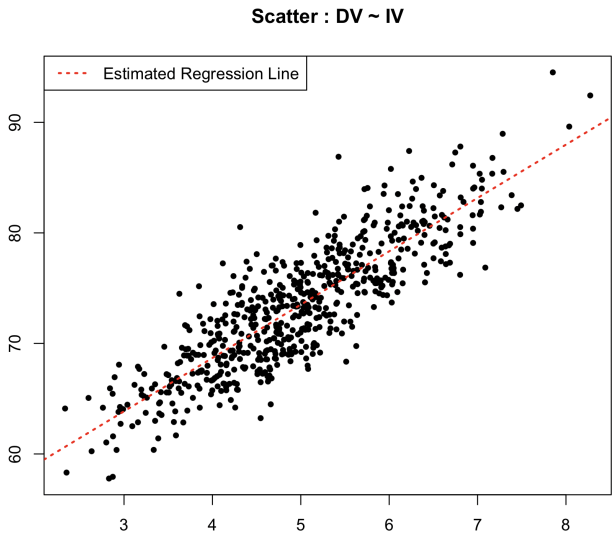
```
> kable(anova(M), caption='ANOVA Table')
```

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	15186.555	15186.554551	1752.957	0
Residuals	621	5379.968	8.663394	NA	NA

Fig.5. Scatter plot for data set

```
> plot(PartA_complete$DV ~ PartA_complete$IV, main='Scatter : DV ~ IV', xlab='IV',
ylab='DV', pch=20)
>  abline(M, col='red', lty=3, lwd=2)
>  legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
```



Part B Appendix

FIG.6. Original Data Set Plot

> plot(data$y ~ data$x, main='Scatter: y ~ x', xlab='x', ylab='y', pch=20)

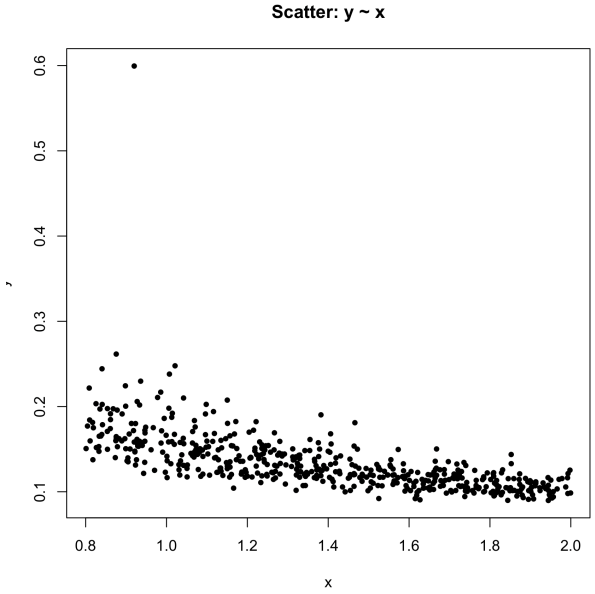


FIG.7. Transformed Data Set Plot

> plot(data_trans$y ~ data_trans$x, main='Scatter: y ~ x', xlab='x', ylab='y', pch=20)

> model <- lm(data_trans$y~data_trans$x)

> abline(model,col='red',lty=3,lwd=2)

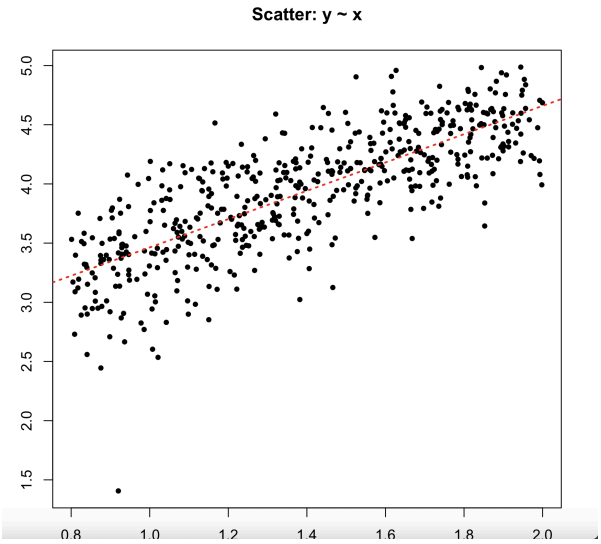


Fig.8.1. Original Data Summary

```
> model <- lm(data$y~data$x)
> summary(model)
Call:
lm(formula = data$y ~ data$x)
Residuals:
    Min       1Q   Median       3Q      Max
-0.04419 -0.01231 -0.00281  0.00829  0.43465
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.226204   0.004968   45.53   <2e-16 ***
data$x      -0.066700   0.003464  -19.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.02695 on 528 degrees of freedom
Multiple R-squared:  0.4125,  Adjusted R-squared:  0.4114
F-statistic: 370.7 on 1 and 528 DF,  p-value: < 2.2e-16
```

Fig.8.2. Transformed data summary

```
> model <- lm(data_trans$ytrans~data_trans$xtrans)
> summary(model)
Call:
lm(formula = data_trans$ytrans ~ data_trans$xtrans)
Residuals:
    Min       1Q   Median       3Q      Max
-1.96144 -0.19586  0.02713  0.22390  0.85275
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.26741    0.05994   37.83   <2e-16 ***
data_trans$xtrans  1.19612    0.04179   28.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.3252 on 528 degrees of freedom
Multiple R-squared:  0.608,    Adjusted R-squared:  0.6073
F-statistic: 819.1 on 1 and 528 DF,  p-value: < 2.2e-16
```

Fig.9. Transformed Data groups

```
> groups <- cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.3,
max(data_trans$xtrans)-0.3,by=0.3),Inf))
> table(groups)
groups
(-Inf,1.1]  (1.1,1.4] (1.4, Inf]
       131        143        256
```

Fig.10. Variance Table using pureAnova() function

```
> corpureErrorAnova(fit_b)
Analysis of Variance Table
Response: y
             Df Sum Sq Mean Sq  F value Pr(>F)
x             1 78.331  78.331 646.6600 <2e-16 ***
Residuals    528 64.105   0.121
 Lack of fit   1  0.269   0.269   2.2178  0.137
 Pure Error  527 63.836   0.121
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```