Part A Report

Introduction

In Part A, the objective is to fill in the missing values in the incomplete data set by using the given values to create a regression model and approximating the expected values of the missing values. We also performed a variety of tests to help us understand the relationship between the independent and dependent variables better - this is known as **regression analysis.** Some of the research questions are: How strong is the relationship between the independent and dependent variables? What is the probability that the relationship we observed is merely due to random chance? How accurately can we predict the missing values in our data set?

Methods

We used the program RStudio to perform our statistical analysis. Our files were merged by first creating two variables, IV and DV, with their corresponding values with the code IV<-read.csv("602237_IV.csv") and DV<-read.csv("602237_DV.csv"). We then merged IV (independent variable) and DV (dependent variable) with the line partA<-merge(IV,DV, by = 'ID'). We used the function mice to help us input the missing data. It is important to note that there were 472 complete data sets, 78 sets in which only the IV was missing, 72 sets in which only the DV was missing, and 31 sets in which both the IV and DV were missing. In our program, we had to drop the 31 sets that were missing the IV and DV, as they didn't give any information. My output was shown in figure 1 (in appendix).
We then used core summary statistics to estimate the slope and intercept of our data using ordinary least squares (OLS). I generated statistical data shown in figure 2.
I also generated an ANOVA table using the package "knitr", shown in figure 3.
I then generated the following scatter plot to help model the data as shown in figure 4.
Finally, I calculated the confidence level of the slope of the regression line and got the data shown in figure 5.

Results

In part A, I tested the null hypothesis that the slope of the regression line of the DV on the IV is zero against the alternative hypothesis that it is not. I can do this at various levels of significance. If I fail to reject the null hypothesis, that is essentially saying that at that particular significance level, I can't confidently claim that the results are not merely due to random chance. If we accept the alternative hypothesis, I can claim that the results are likely not due to chance and that there is a relationship between the independent and dependent variables. I tested the null hypothesis with two tests - the t-statistic test and f-statistic test. Note I do not know the value of σ, so I cannot perform the z-test. For the t-statistic test for both variables, I would reject the null hypothesis because it says the P(>|t|) is <2e-16 for both variables. This means that the probability of the relationship just being due to random chance is so negligibly small that we would reject the null hypothesis at every realistic α value. The same is true for the f-statistic test - the p-value is <2.2e-16, so I would reject the null hypothesis at every α value. The fitted function is $\hat{y} = 35.5603 + 3.1801x$. The 95% confidence interval for the intercept is (34.368832, 36.751863). The 95% confidence interval for the IV is (2.942413, 3.417738). The 99% confidence interval for the intercept is (33.992533, 37.128162, 3.492795). The 99% confidence interval for the IV is (2.867355,3.492795). This means that if you have 100 samples, a 99% percent confidence interval would mean you would expect the CI to include the mean 99 times. If you have 100 samples, a 95% percent confidence interval would mean you would expect the CI to include the mean 95 times. The fraction of the variation of the dependent variable that was explained is the r^2 value, or about 0.54. The r^2 value is represented by 1-RSS/TSS and measures correlation.

Conclusion and Discussion

Overall, in this portion, I got a better idea of the relationship between the independent and dependent variables. We found an association between the independent and dependent variables. The multiple r^2 value was found to be 0.5444, while the adjusted r^2 value was found to be 0.5436. Performing these experiments is key in identifying that there is a relationship between the two variables and ways we can model the relationship.

Part B Report

Introduction

In part B, I am performing a transformation on the given set of data, binning the data, and performing a lack of fit test. The objective is to maximize the p-value in the transformation to show that there isn't a statistically significant lack of fit in our model. I compared the before and after transformed models using a variety of tests, including the t-test, f-test, and r^2 analysis.Some of the research questions are: What transformation models our data best? How can a lack of fit test show how effective our transformation is?

Methods

In part B, I first tried a possible transformation, and kept the transformed data under "data_trans". I then used cut() to "cut" our scatter plot into groups. My groups are shown in figure 6.

For instance, there were 38 points between -∞ and 1.13. I calculated the average value of all the points at each interval. This step is key in data bining.

I then utilized a lack of fit test using the function **pureErrorAnova()** in the package **alr3**. I then tried more transformations until we found the "best" one. Our lack of fit test will be shown in the results section.

Results

I found that the best transformation was raising the y value to the (-2/3) power, while keeping the x value as it was. I tried many transformations on both x and y and found that this yielded the best p-value. I used the line: data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(-2/3)) to make my transformation.

My before test results are shown in figure 7, compared to the after shown in figure 8.

As we can clearly see, the multiple r^2 squared value after the transformation is greater than before. This indicates that the transformation resulted in a stronger fit. Moreover, the f-statistic is greater (809.6 compared to 640.7) on 1 and 576 DF after the transformation. A greater F-statistic indicates a greater chance of rejecting the null hypothesis. The null hypothesis here would be that there isn't a relationship, so these results indicate a greater chance of a relationship after the transformation. Additionally, the t-test also indicates a greater chance of rejecting the null after the transformation. The fitted equation before the transformation would be $\hat{y}$ = 0.552013-0.209258x. The fitted equation after the transformation would be $\hat{y}$ = 0.79926+1.20423x.

Figure 9 shows the lack of fit test results after the transformation:

As we can see, the Pr(>F) is 0.4388. This is considered high, so this indicates that the transformation is accurate. We would reject the null hypothesis at every significance level because 0.4388 is greater than 0.1, 0.05, and 0.01, which are the most commonly used significance values. This means there is no lack of fit.

Furthermore, my residual plot is shown in figure 10.

As you can see, the mean is constant among the lines, the outliers above and below are constant (not all above or below the mean), the variance is constant, and it appears normally distributed. This indicates that my model is a good fit for the given data.

Conclusion and Discussion

This part of the experiment had me generate the best transformation after trial-and-error with various different transformations. I then compared the before and after transformed data using a lack of fit test, residual plot, and the F-statistic test. I concluded that my transformation was effective and gave me a greater r^2 value (both multiple r^2 and adjusted r^2). The before r^2 value was approximately 0.53 and the after was approximately 0.58. This overall taught me how to better model a given set of data.

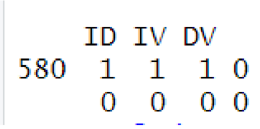
Appendix

Code (parts taken from One Predictor Linear Regression Handout)

```
#Part A
setwd ("C:/Users/arjun/OneDrive/Desktop/AMS 315 Project 1")
getwd()
IV<-read.csv("602237_IV.csv")
DV<-read.csv("602237_DV.csv")
partA<-merge(IV,DV, by = 'ID')
str(partA)
any(is.na(partA[,2]) == TRUE)
any(is.nan(partA[,2]) == TRUE)
partA_incomplete <- partA
#install.packages('mice')
library(mice)
md.pattern(partA_incomplete)
partA_imp <- partA[!is.na(partA$IV)==TRUE|!is.na(partA$DV)==TRUE,]
imp <- mice(partA_imp, method = "norm.boot", printFlag = FALSE)
partA_complete <- complete(imp)
md.pattern(partA_complete)
M <- lm(DV ~ IV, data=partA_complete)
summary(M)
#install.packages('knitr')
library(knitr)
kable(anova(M), caption='ANOVA Table')
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')
confint(M, level = 0.95)
confint(M, level = 0.99)
#Part B
plot(fitted(fit_b),resid(fit_b),)
data <- read.csv('602237_partB.csv', header = TRUE)
M2 <- lm(y ~ x, data=data)
summary(M2)
data_trans <- data.frame(xtrans=data$x, ytrans=data$y^(-2/3))
M3 <- lm(ytrans ~ xtrans, data=data_trans)
summary(M3)
groups <- cut(data_trans$xtrans,breaks=c(-Inf,seq(min(data_trans$xtrans)+0.03,
max(data_trans$xtrans)-0.03,by=0.03),Inf))
table(groups)
x <- ave(data_trans$xtrans, groups)
data_bin <- data.frame(x=x, y=data_trans$ytrans)
fit_b <- lm(y ~ x, data = data_bin)
#install.packages('remotes')
library(remotes)
#install_github("cran/alr3")
```

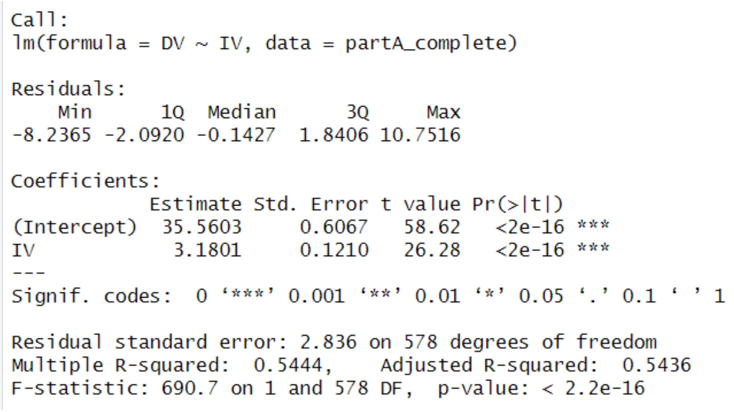
library(alr3)
pureErrorAnova(fit_b)

Figures



```
     ID IV DV
580  1  1  1 0
     0  0  0 0
```

Figure 1



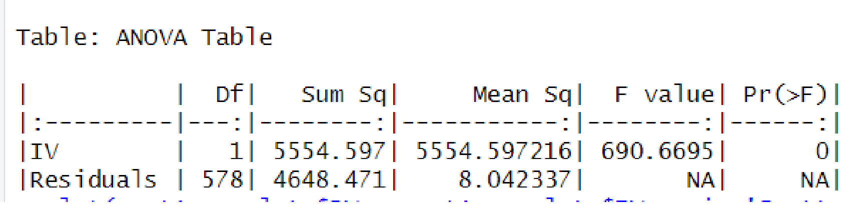
```
Call:
lm(formula = DV ~ IV, data = partA_complete)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2365 -2.0920 -0.1427  1.8406 10.7516 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  35.5603     0.6067   58.62   <2e-16 ***
IV            3.1801     0.1210   26.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.836 on 578 degrees of freedom
Multiple R-squared:  0.5444,	Adjusted R-squared:  0.5436 
F-statistic: 690.7 on 1 and 578 DF,  p-value: < 2.2e-16
```

Figure 2



```
Table: ANOVA Table

|          |  Df|   Sum Sq|     Mean Sq|  F value| Pr(>F)|
|:---------|---:|--------:|-----------:|--------:|------:|
|IV        |   1| 5554.597| 5554.597216| 690.6695|      0|
|Residuals | 578| 4648.471|    8.042337|       NA|     NA|
```

Figure 3

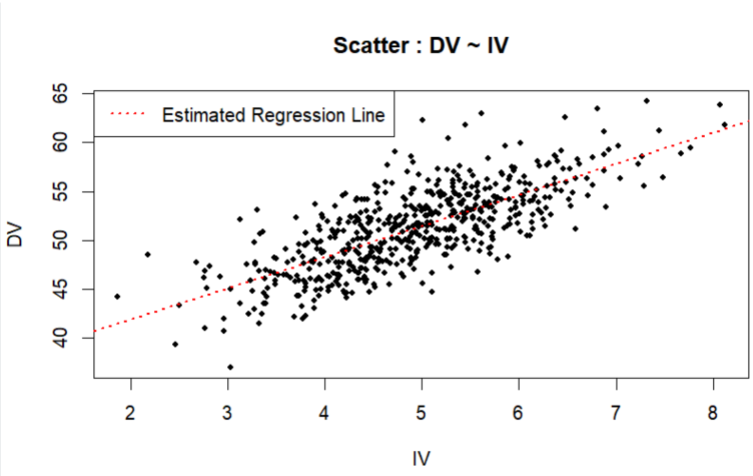
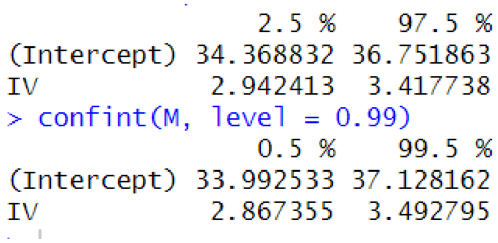
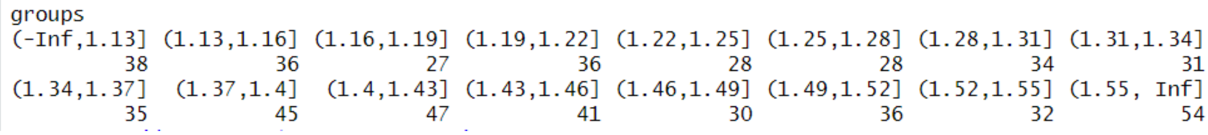


Figure 4



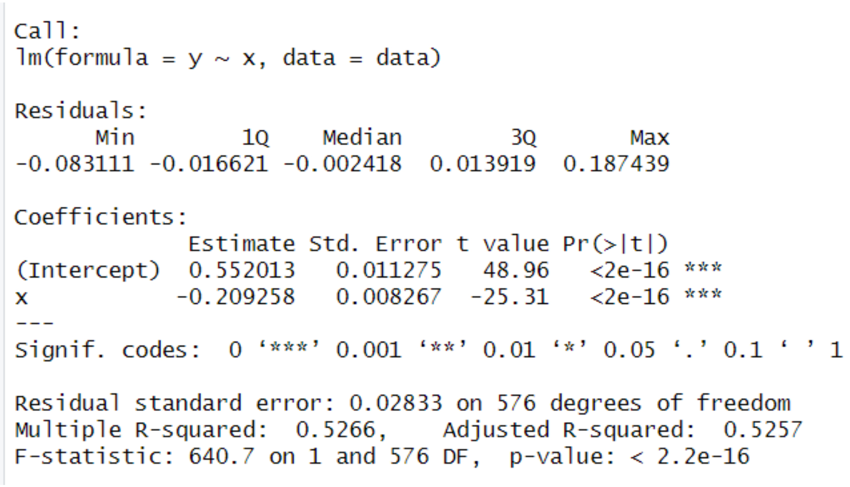
```
                 2.5 %    97.5 %
(Intercept) 34.368832 36.751863
IV           2.942413  3.417738
> confint(M, level = 0.99)
                 0.5 %    99.5 %
(Intercept) 33.992533 37.128162
IV           2.867355  3.492795
```

Figure 5



```
groups
(-Inf,1.13] (1.13,1.16] (1.16,1.19] (1.19,1.22] (1.22,1.25] (1.25,1.28] (1.28,1.31] (1.31,1.34] 
         38          36          27          36          28          28          34          31 
(1.34,1.37]  (1.37,1.4]  (1.4,1.43] (1.43,1.46] (1.46,1.49] (1.49,1.52] (1.52,1.55] (1.55, Inf] 
         35          45          47          41          30          36          32          54 
```

Figure 6



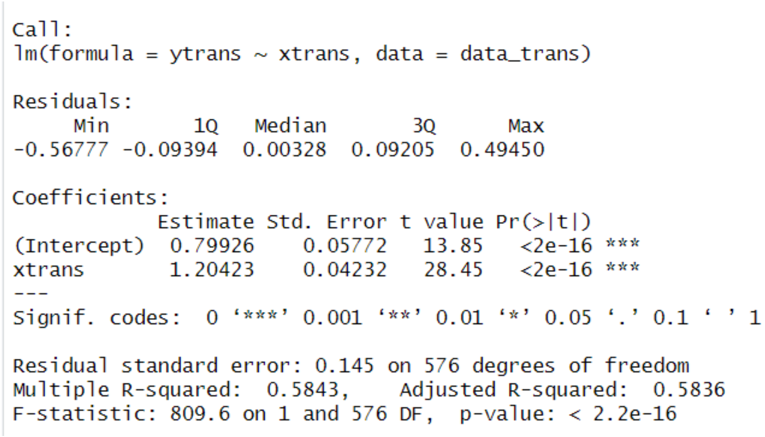
```
Call:
lm(formula = y ~ x, data = data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.083111 -0.016621 -0.002418  0.013919  0.187439 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.552013   0.011275   48.96   <2e-16 ***
x           -0.209258   0.008267  -25.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02833 on 576 degrees of freedom
Multiple R-squared:  0.5266,	Adjusted R-squared:  0.5257 
F-statistic: 640.7 on 1 and 576 DF,  p-value: < 2.2e-16
```

Figure 7



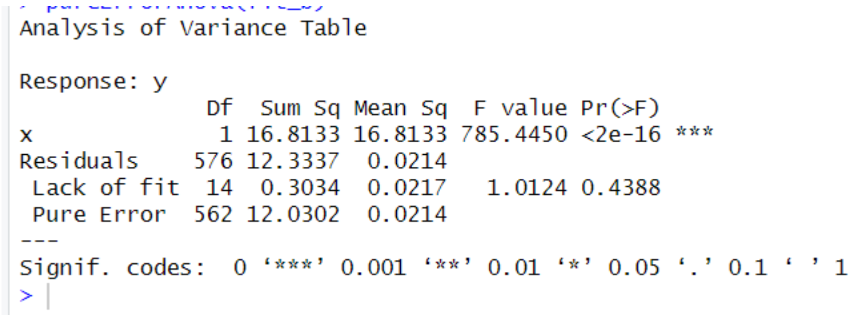
```
Call:
lm(formula = ytrans ~ xtrans, data = data_trans)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.56777 -0.09394  0.00328  0.09205  0.49450 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.79926    0.05772   13.85   <2e-16 ***
xtrans       1.20423    0.04232   28.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.145 on 576 degrees of freedom
Multiple R-squared:  0.5843,	Adjusted R-squared:  0.5836 
F-statistic: 809.6 on 1 and 576 DF,  p-value: < 2.2e-16
```

Figure 8



```
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq  F value Pr(>F)    
x              1 16.8133 16.8133 785.4450 <2e-16 ***
Residuals    576 12.3337  0.0214                    
 Lack of fit  14  0.3034  0.0217   1.0124 0.4388    
 Pure Error  562 12.0302  0.0214                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Figure 9

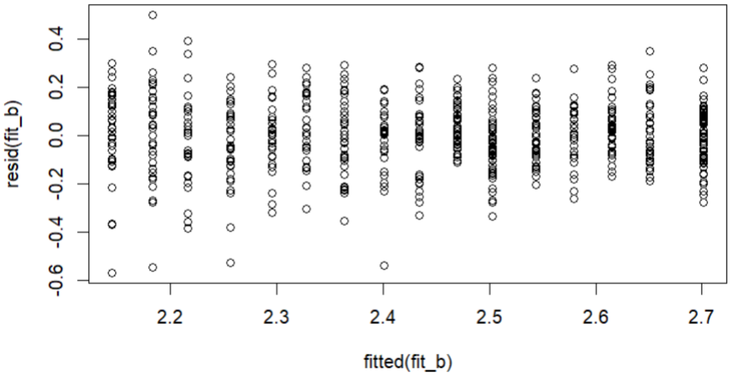


Figure 10