Data Project 1: Part A

Introduction

This report details the process and results of recovering the function that was used to generate the dependent variable values based on the values of the independent variable for the assigned data. This work incorporates simple linear regression. Research questions include: Is there a relationship between the two variables? How significant is this relationship? And finally, what is the fitted function?

Methods

For this work, the R programming language was used via RStudio. The methods and code herein largely follow those shown in "One Predictor Linear Regression Handout" by Benjy Hetchman and Jin Huang (2021). The two files provided for this part, one containing the independent variable (IV) and the other containing the dependent variable (DV), were merged into one data set by the subject ID of each observation with the native R function "merge". Using the R package "mice", this list of 554 observations was checked for missing data using function "md.pattern". In total, there were 534 subject IDs that had either an IV value or a DV value, with 470 having both, 500 having an IV value, and 504 having a DV value (see Figure 1 in Appendix). This left 20 observations to be dropped, as they offered no information at all. For the 534 IDs that had data, missing values were imputed using the "norm.boot" (linear regression using bootstrap) method of imputation provided in the mice package (see Figure 2 in Appendix). With the now complete data set, a linear regression model was made using the function "lm". A model summary, 95% slope confidence interval, and ANOVA table were computed using this model, and a scatter plot with the fitted line was produced.

Results

The fitted function recovered is $\hat{y}_i = 31.2067 + 4.9658x_i$. The null hypothesis that slope was equal to 0 was rejected with $p < 2e\text{-}16$ (see summary in Figure 3). The slope of this fitted line has a 95% confidence interval of (4.5857, 5.3459), meaning there is a 95% probability the true slope is between these values. There exists a strong association between the independent and dependent variables, with 55.32% of the variance in DV explained by IV ($r^2 = 0.5532$, see Figure 3 in Appendix). See Figure 3 in the appendix for the analysis of variance table.

Conclusions and Discussion

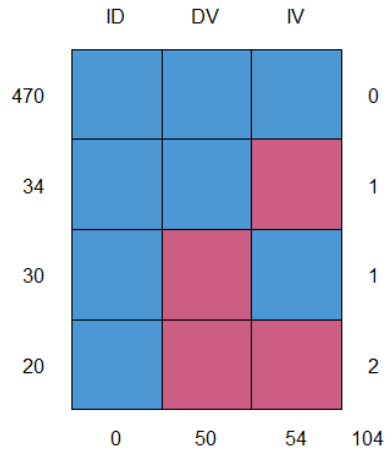
As mentioned in the results, there is indeed a significant relationship between the variables with an r^2 of 0.5532. The function recovered above thus accurately models this relationship to a large degree and could perhaps be used for prediction of future values. Figure 4 in the appendix depicts the data and model addressed in this report.

Appendix

Figures and Computer Code

Figure 1

Diagram Showing Data Completeness After Merge



Note: Blue = present, red = missing, with left side counting amount per type.

Figure 2

Code Showing Merge, Data Checking, And Imputation

```
> PartA_IV <- read.csv('327651_IV.csv', header=TRUE)
> PartA_DV <- read.csv('327651_DV.csv', header=TRUE)
> PartA <- merge(PartA_IV, PartA_DV, by ='ID')
> library(mice)

Attaching package: 'mice'

The following object is masked from 'package:stats':

    filter

The following objects are masked from 'package:base':

    cbind, rbind

Warning message:
package 'mice' was built under R version 4.0.5
> md.pattern(PartA)
    ID DV IV
470  1  1  1   0
34   1  1  0   1
30   1  0  1   1
20   1  0  0   2
     0 50 54 104
> 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)
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'
```

Merge

Checking Completeness

Imputation

Note: Code largely adopted from handout mentioned.

Figure 3

R Outputs for Model Summary, ANOVA Table, And 95% C.I.

```
> model <- lm(DV ~ IV, data=PartA_complete)
> summary(model)

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

Residuals:
     Min       1Q   Median       3Q      Max
-23.1223  -5.9952   0.3128   5.9417  29.5954

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2067     1.0208   30.57   <2e-16 ***
IV            4.9658     0.1935   25.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.725 on 532 degrees of freedom
Multiple R-squared:  0.5532,    Adjusted R-squared:  0.5524
F-statistic: 658.7 on 1 and 532 DF,  p-value: < 2.2e-16

> library(knitr)
Warning message:
package 'knitr' was built under R version 4.0.5
> kable(anova(model), caption='ANOVA Table')

Table: ANOVA Table

|          | Df|   Sum Sq|     Mean Sq| F value| Pr(>F)|
|:---------|---:|--------:|-----------:|-------:|------:|
|IV        |  1| 50149.08| 50149.08443| 658.704|      0|
|Residuals | 532| 40502.73|    76.13296|      NA|     NA|
> plot(PartA_complete$DV ~ PartA_complete$IV, main='Scatter Plot: DV ~ IV', xlab='IV', ylab='DV', pch=20)
> abline(model, col='red', lty=3, lwd =3)
> legend('topleft', legend = 'Estimated Regression Line', lty=3, lwd=2, col='red')
> legend('topleft', legend = 'Estimated Regression Line', lty=3, lwd=3, col='red')
> confint(model, level=0.95)
                2.5 %    97.5 %
(Intercept) 29.201413 33.212025
IV           4.585754  5.345932
```

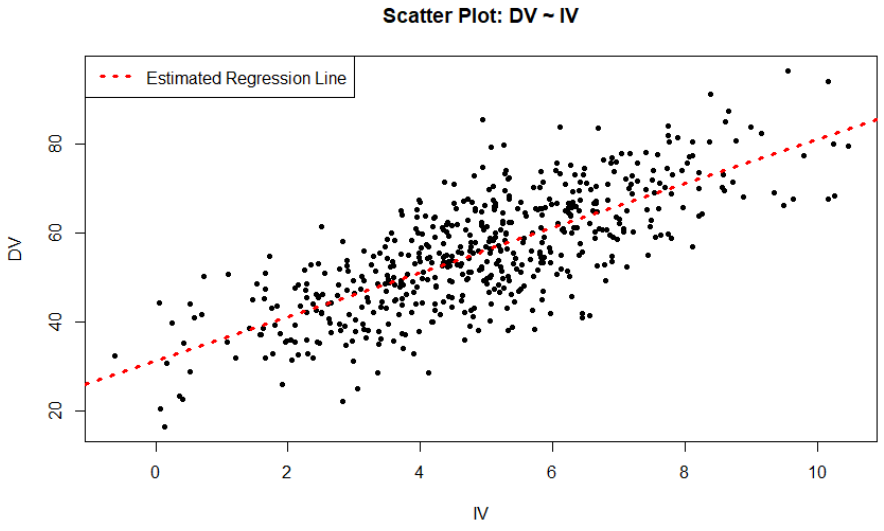
Model Summary

ANOVA

95% C.I. For Slope

Figure 4

Scatter Plot Showing Estimated Fitted Line and Data Points



Data Project 1: Part B

Introduction

The goal of this report is to find the original function used to generate the dependent variable values from the values of the independent variable for the assigned dataset. This will require knowledge of simple linear regression, specifically dealing with transforming variables and testing for lack of fit. Some questions here are: What is the original function? What transformations will work best? To what degree can these estimates be considered accurate?

Methods

Here, the R programming language is used to generate statistics and plots. Adopting code from "One Predictor Linear Regression Handout" facilitated this use (Hetchman & Haung, 2021). First, the assigned file containing 589 observations, each with a subject ID, independent variable x value, and dependent variable y value, was read into RStudio. Using function "plot" for $y \sim x$, it was immediately evident that there was increasing variability as both x and y increased (see Figure 1 in Appendix). To make this plot more linear, several transformations of both x and y were tested in trial and error while observing the changes they made to r^2 and the residual vs fitted values plot, until eventually arriving at $x^{\frac{5}{3}}$ and $y^{\frac{1}{3}}$, as these maximized desirable properties (see Figures 2, 3, & 4 for observed changes). With these transformations applied, the new data set was binned in terms of x into intervals of $\frac{3}{2}$ using the native R "cut" function. A linear regression model for the relationship of x and y was computed using "lm", and an approximate lack of fit test was applied. A summary, analysis of variance table, and scatter plot with estimated regression line were produced (see Appendix).

Results

Whilst the original data had an r^2 of 0.3219, the transformations increased r^2 to 0.3777 and visibly "fixed" the residual vs fitted values and scatter plots (Figures 2, 3 and 4). R's summary for the linear regression model on the transformed, binned data shows that the fitted line function is estimated to be $\hat{y}_i = 67.2371 + 0.5546x_i$, with a 95% confidence interval for slope of (0.4969, 0.6124). The result of the approximate lack of fit test on this model was insignificant, with $p = 0.1587$, meaning insufficient evidence exists to suggest this model is inappropriate (Figure 5). Computing corr(x,y) for this data gave a value of 0.6146 (see Figure 5).

Conclusions and Discussion

With a low r^2 and similarly fair correlation coefficient, the association of x and y even after their transformations seems to only be moderate at best. Going by r^2, merely 37.77% of the variance in y is explained. Nonetheless, this was the best model found (depicted in Figure 6).

Appendix

Figures and Computer Code

Figure 1

Scatter Plot of Original Data

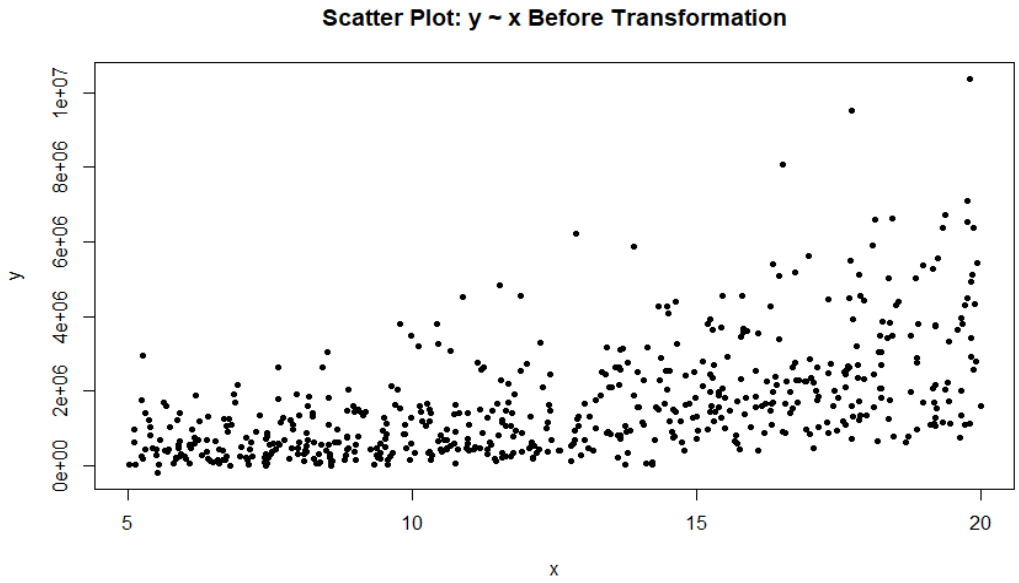
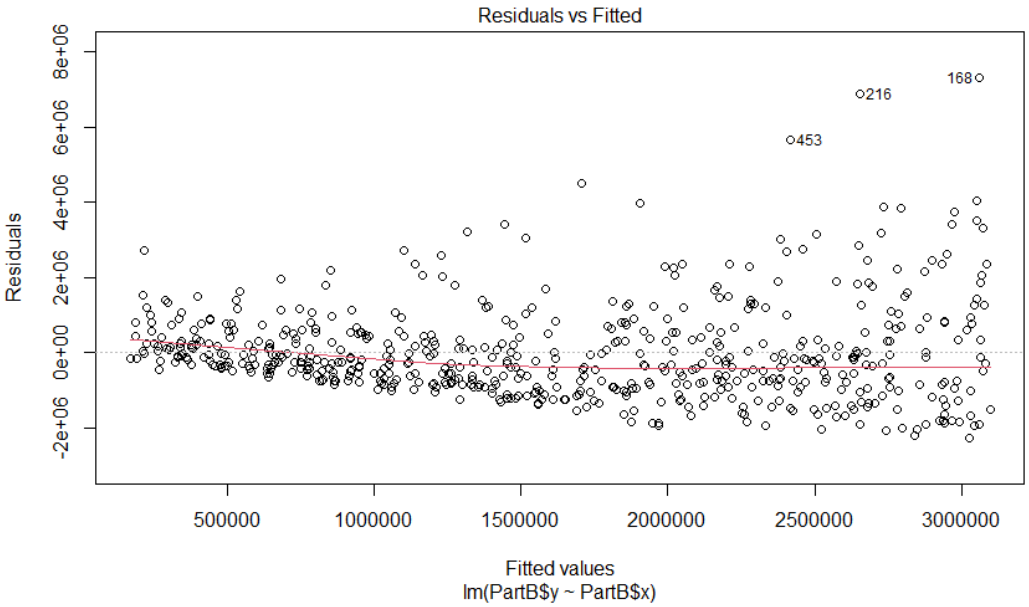


Figure 2

Residuals vs Fitted Plot for Original Data



Note: This plot has a clear bias. Should instead be more evenly spread and elliptical in shape.

Figure 3

Residuals vs Fitted Plot for Transformed x and y

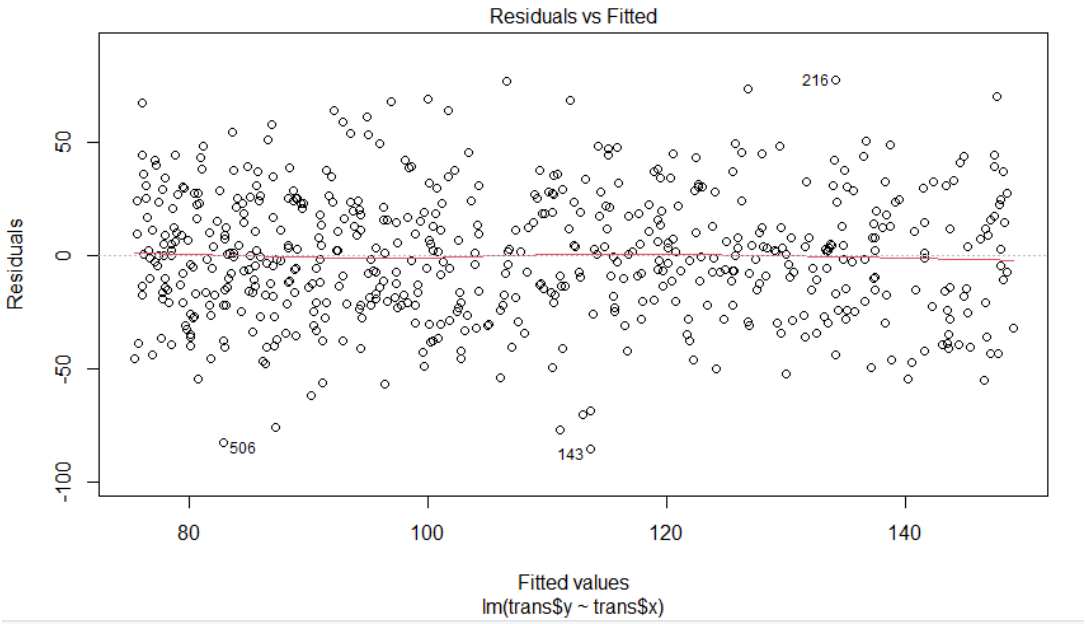


Figure 4

Increase in R^2 by Transformations

```
> summary(M)

Call:
lm(formula = PartB$y ~ PartB$x)

Residuals:
     Min       1Q   Median       3Q      Max
-2266484  -773104  -234743   529658  7320939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -814021     155624  -5.231 2.35e-07 ***
PartB$x       195571      11716  16.692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1236000 on 587 degrees of freedom
Multiple R-squared:  0.3219,    Adjusted R-squared:  0.3207
F-statistic: 278.6 on 1 and 587 DF,  p-value: < 2.2e-16

> summary(fit_trans)

Call:
lm(formula = trans$y ~ trans$x)

Residuals:
    Min      1Q  Median      3Q     Max
-85.584 -19.524   0.261  19.186  77.855

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.21160    2.41864   27.79   <2e-16 ***
trans$x      0.55499    0.02939   18.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.82 on 586 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3782,    Adjusted R-squared:  0.3772
F-statistic: 356.5 on 1 and 586 DF,  p-value: < 2.2e-16
```

R^2 for original data

R^2 for transformed data

Note: This r^2 very slightly differs from that in the report because fit_trans was not yet binned.

Figure 5

LOF Test Result, Summary of Binned Fit Line, Corr(x,y) and Slope 95% C.I.

- Lack of Fit Test

```
> pureErrorAnova(fit_b)
Analysis of Variance Table

Response: y
             Df Sum Sq Mean Sq  F value Pr(>F)
x             1 275608  275608 364.4797 <2e-16 ***
Residuals   586 454088     775
 Lack of fit  86  76004     884   1.1687 0.1587
 Pure Error  500 378084     756
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Summary of Fit_B

```
> fit_b <- lm(y ~ x, data = data_bin)
> summary(fit_b)

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

Residuals:
    Min      1Q  Median      3Q     Max
-85.341 -19.443   0.292  19.188  77.925

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.23710    2.41983   27.79   <2e-16 ***
x            0.55464    0.02941   18.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.84 on 586 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3777,    Adjusted R-squared:  0.3766
F-statistic: 355.7 on 1 and 586 DF,  p-value: < 2.2e-16
```

- 95% Confidence Interval for Slope + Corr(x,y)

```
> confint(fit_b, level =0.95)
                  2.5 %    97.5 %
(Intercept) 62.4845052 71.989692
x            0.4968771  0.612398
> cor(data_bin$x, data_bin$y)
[1] NA
> cor(data_bin$x, data_bin$y, use = "complete.obs")
[1] 0.614575
```

Figure 6

Scatter Plot With Estimated Regression Line (Final Model)