Part B

Introduction

The objective of this assignment is to recover the function that was used to generate the dependent variable value based on the value of the independent variable and perform a lack of fit test to evaluate how well the regression model fits the data. The given contains one line for each subject ID alongside the values of the respective independent and dependent variables.

Methods

First, we determine an appropriate transformation for our dataset. To do so, we create a scatter plot and residual plot of our data to observe relations between the independent and dependent variables. After applying the transformation, we bin near repeated data into common levels. For example, suppose that x_1=1.01, x_2=1.02, x_3=1.03 and y_1=2, y_2=3, y_3=4. These can be grouped into one bin, and define their x-values replaced with the bin average. The cut() function was used to define bins with width 0.1 and the ave() function to compute the average x value within each bin. Use lm() to make a linear regression model of the binned data. To evaluate how well the model fits the binned data, we use the pureErrorAnova() function from the alr3 package to get the P-value for the lack of fit. In addition, we use summary() to get data about the linear regression models including the r_square values and slope. We can get the confidence interval of the slope using the confint() function.

Results

Figure 1 and 2 give the scatterplot and residual plot of the original data. We can see that the plots indicate a relation that is increasing at an increasing rate and the variability around the curve increases as the predicted y value increases. As such we transform our dataset by taking the natural log of the dependent variable. After binning the transformed data we plot a new scatterplot and residual plot for the data as shown in Figures 4 and 5 respectively. The figures show that a more linear relation with consistent variance throughout. The linear regression model of the original data explained 43.8% of the variation in the dependent variable as R^2 = .438 (as seen in Figure 3). The linear regression model of the transformed and binned data explains 50.1% of data as R^2 = .5013 (as seen in Figure 7). The fitted function of the original slope is $\hat{y}$ = -685.85 + 292.63x and for the binned model it is $\hat{y}$ = 1.19 + 1.43x as shown in Figures 3 and 7 respectively. The function of The 99% confidence interval for the original slope is [249.86, 335.40], and for the binned model it is [1.04, 1.35]. Figure 6 shows the ANOVA table for the binned dataset, with a lack of fit p-value of 0.5805. Since this exceeds 0.05, we conclude that there is no significant lack of fit in our regression model. In addition, we reject the null hypothesis that the slope was zero.

Conclusion and Discussion

The goal of this assignment was to recover the function that was used to generate the dependent variable value based on the value of the independent variable value. We do so by taking the natural log of the dependent variable to transform the data and then binning the data. The linear regression model obtained from the transformed dataset has an R^2 = .5013, indicating that 50.1% of dependent variables could be explained by the model. In addition, the lack of fit test gives a p value of .5805 indicating a not significant lack of fit. This shows that the linear regression model we produce adequately fits the data and the model can be used for further analysis.

Appendix

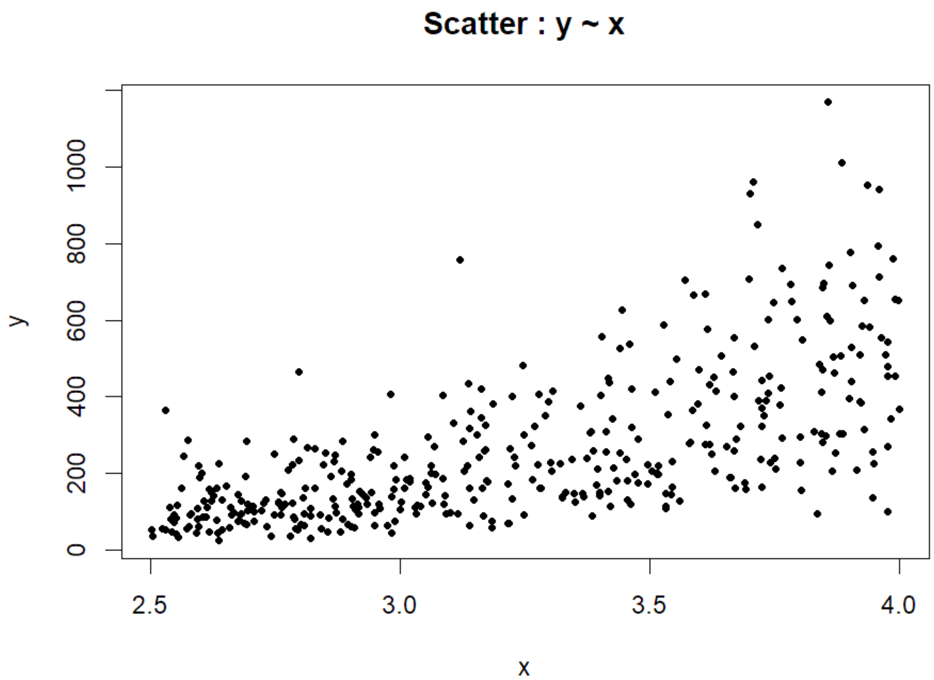

Figure 1: Scatter plot of original data showing a non-linear trend

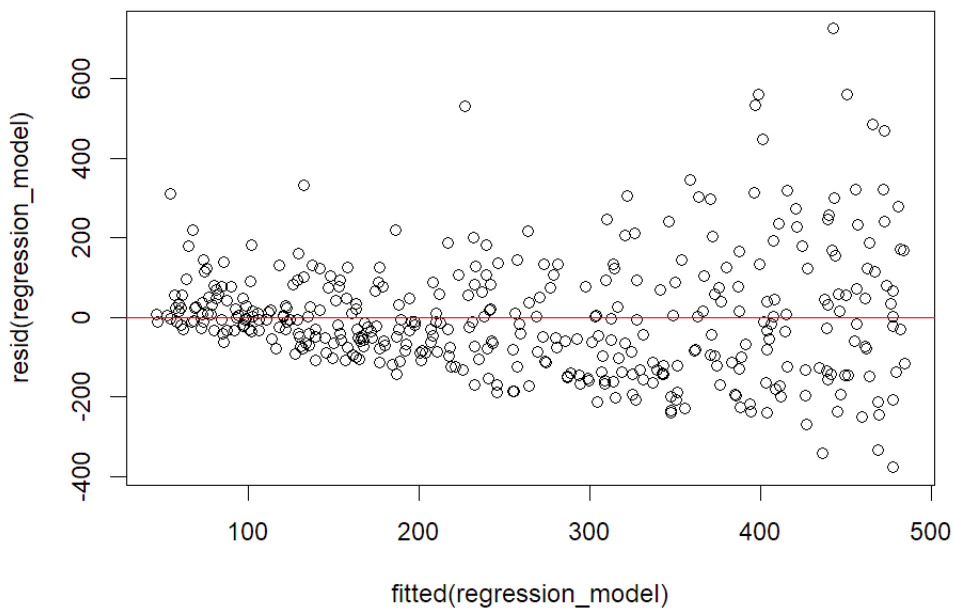

Figure 2: Residual plot of original model indicating increasing variance as y increases

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

Residuals:
    Min      1Q  Median      3Q     Max
-377.09  -91.55  -12.93   70.89  727.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -685.85      53.94  -12.71   <2e-16 ***
x             292.63      16.53   17.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147.9 on 402 degrees of freedom
Multiple R-squared:  0.4382,    Adjusted R-squared:  0.4368
F-statistic: 313.6 on 1 and 402 DF,  p-value: < 2.2e-16
```

Figure 3: Summary of original regression model

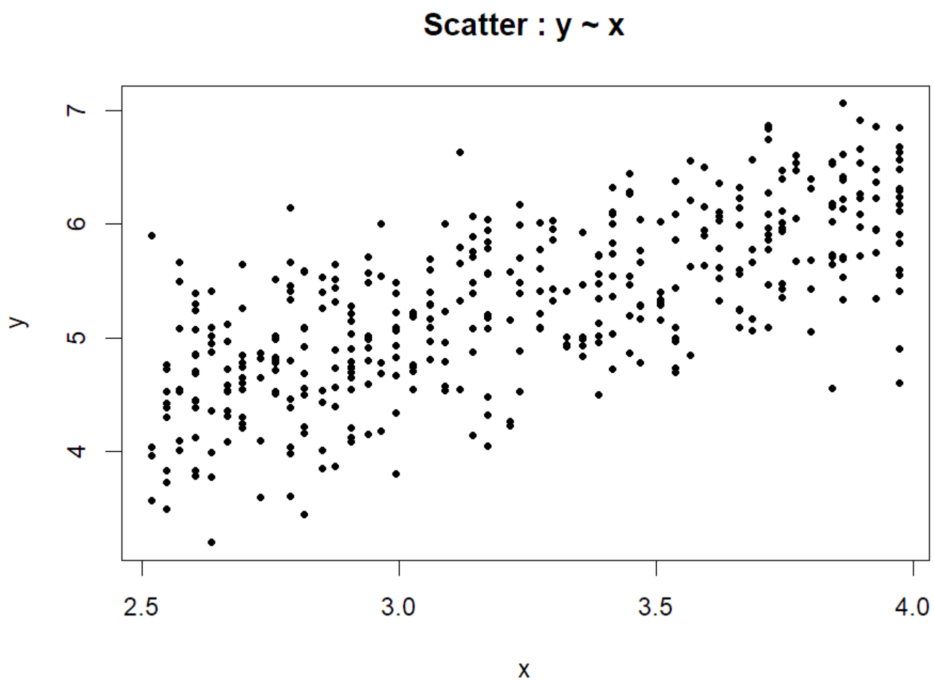

Figure 4: Scatter plot after transforming and binning data

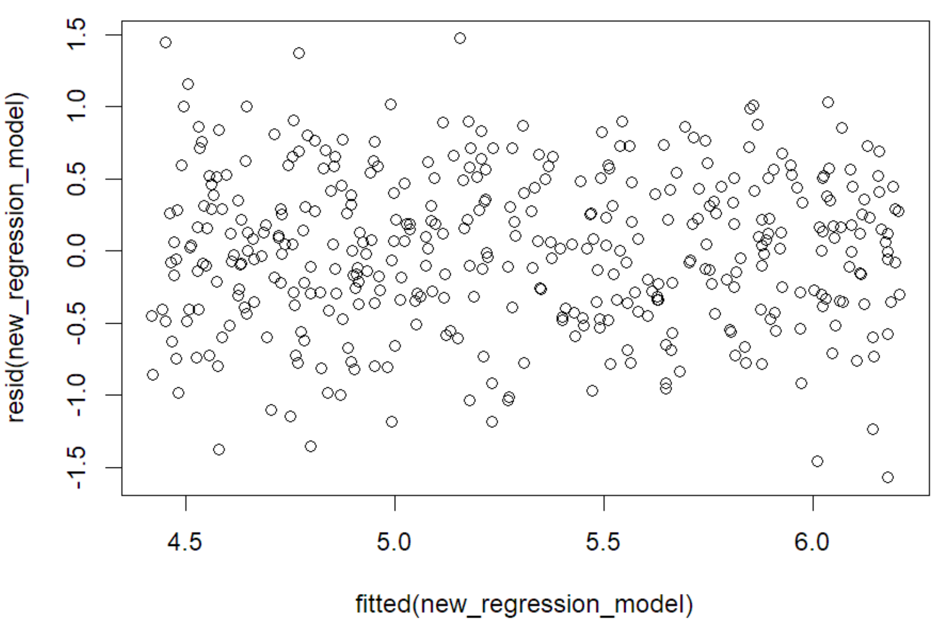

Figure 5: Residual plot of model after transforming and binning data

```
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq  F value Pr(>F)
x              1 114.258 114.258 401.5220 <2e-16 ***
Residuals    402 113.647   0.283
 Lack of fit  47  12.628   0.269   0.9442 0.5805
 Pure Error  355 101.019   0.285
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: ANOVA table of transformed and binned model

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.56741 -0.35104  0.01123  0.38314  1.47487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42944    0.19402   7.367 9.97e-13 ***
x            1.19493    0.05944  20.104  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5317 on 402 degrees of freedom
Multiple R-squared:  0.5013,    Adjusted R-squared:  0.5001
F-statistic: 404.2 on 1 and 402 DF,  p-value: < 2.2e-16
```

Figure 7: Summary of linear regression model of transformed and binned data

```r
library(knitr)
library(remotes)
library(alr3)

setwd("~/AMS 315 Project 1")
PartB <- read.csv('Part B/422817_partB.csv', header = TRUE)
plot(PartB$y ~ PartB$x, main='Scatter : y ~ x', xlab='x', ylab='y', pch=20)
regression_model <- lm(y ~ x, data=PartB)
summary(regression_model)
confint(regression_model, level = 0.99)


kable(anova(regression_model), caption='ANOVA Table')
plot(fitted(regression_model), resid(regression_model))
abline(0, 0, col='red')

PartB_trans <- data.frame(x=PartB$x, log_y=log(PartB$y))
plot(PartB_trans$log_y ~ PartB_trans$x, main='Scatter : y ~ x', xlab='x', ylab='y', pch=20)
new_regression_model <- lm(log_y ~ x, data=PartB_trans)

groups <- cut(PartB_trans$x,breaks=c(-Inf,seq(min(PartB_trans$x)+0.03, max(PartB_trans$x)-0.03,by=0.03),Inf))
table(groups)

x <- ave(PartB_trans$x, groups)
data_bin <- data.frame(x=x, y=PartB_trans$log_y)
plot(data_bin$y ~ data_bin$x, main='Scatter : y ~ x', xlab='x', ylab='y', pch=20)
fit_b <- lm(y ~ x, data = data_bin)
plot(fitted(fit_b), resid(fit_b))
pureErrorAnova(fit_b)
summary(fit_b)
confint(fit_b, level = 0.99)
```

Figure 8: R code used to perform analysis