**AMS 315 Data Analysis Project 1- Part A**

Young Seok Seo(109803869)

April 1st , 2019

**Introduction**

In this project, Part A, I was asked to have a linear regression model function with the given data. In this project, some programs can be used to generate the linear function. In my case, I used Excel and R program. The independent variable and dependent variable were given. Each data should be sorted by the ID number in ascending order.

**Methodology**

I received two data sets, one had the ID# and the independent variable column, and the other had the ID# and the dependent variable column, to import the data and merged with ID number by merge() function. The independent and dependent variable data file had a total of 500 pairs of data. The ID number was ranging from 1 to 500. The independent variable data set had 50 missing data value, NA. The dependent variable data set had 100 missing data value, NA. Moreover, I got total 361 pairs of data set from independent and dependent variable. Then, by using Amelia Function, I imputed missing value. As a result, linear relationship was predicted for the data and simple linear regression function was carried out with the lm() function.

**Results**

After importing my data into R program, I was able to get a result of fitted function for the model, Y=B0+B1\*X. With the 0.4408 or 44.08% of coefficient of determination, I got the function, Y = 63.9230 +12.4844\*X. Since significance level is 0.01, Regression coefficients were statistically significant. Moreover, this model is also signiciant because of low value of

p-value(p<0.001). The 95% confidence interval for the intercept was [11.24913, 13.71964] and F = 394.3 on 498 degrees of freedom.

> summary(l)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 63.9230 1.9167 33.35 <2e-16 \*\*\*

x 12.4844 0.6287 19.86 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.785 on 498 degrees of freedom

Multiple R-squared: 0.4419, Adjusted R-squared: 0.4408

F-statistic: 394.3 on 1 and 498 DF, p-value: < 2.2e-16

> anova(l)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 37756 37756 394.3 < 2.2e-16 \*\*\*

Residuals 498 47685 96

---

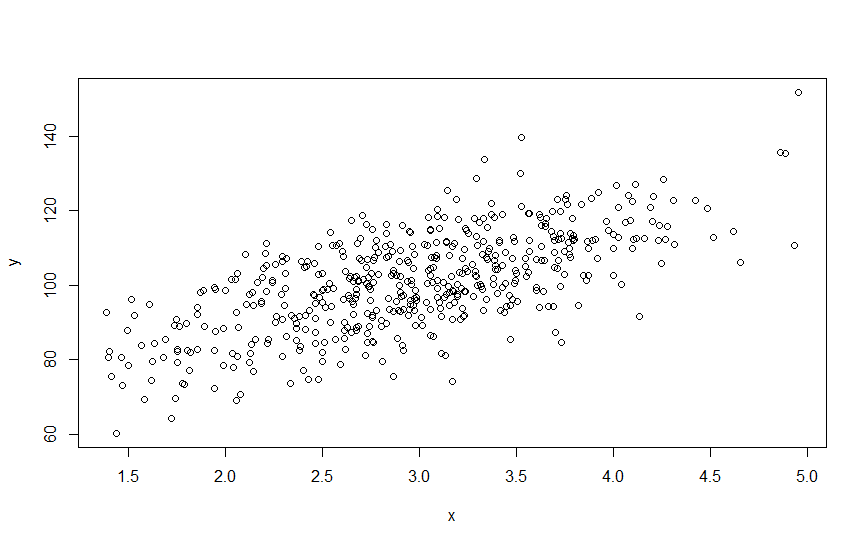
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> confint(l)

2.5 % 97.5 %

(Intercept) 60.15731 67.68878

x 11.24913 13.71964



**Conclusion**

The IV and DV in this problem had a statistically significant positive relationships, with 44.08% of the DV variation. Furthermore, the plot of regression residual versus expected value goes to a positive linear line, then it is reasonable to assume that a linear model describes the data.

**End of Report**

**AMS 315 Data Analysis Project 1- Part B**

Young Seok Seo(109803869)

April 1st, 2019

**Introduction**

For Part B, I was given the dependent variable and independent variable data set that had already been merged. What I was asked in this part is to analyze and generate the function in the given data as I did in Part A, and to determine the best fitted linear model due to the transformations of those data. Besides, I need to find repeated (or close) independent variable values if there is any, and apply a lack-of-fit test on both equations.

**Methodology**

The given data contains each value in independent variable and dependent variable matches with each other. There were total 800 observations with no missing data. Observations were grouped by x in order to have repeated value of independent variable to find the LOF test. Then calculated group means and rounded upto second decimal points. Therefore, I used x\_bin as new variable in order to attach to the original dataset. Then, I started transforming the given independent variables through R into sqrt(X), X^2, ln(X), 1/X. I tried to test all the combinations for determining the best fitted linear model. The dependent variable y was transformed into y^2, sqrt(y) and independent variable x into x^2, sqrt(x), 1/x, and ln(x).

**Results**

The fitted function of the original IV and DV for the model Y=B+B1\*Xwas DV=1.17162+0.71632\*X with Adjusted R-squared value 0.4862. The 95% confidence interval for the slope was [0.660862,0.7717836]. The 95% confidence interval for the intercept was [1.057363,1.2858866]. The test statistics for the null hypothesis is F=642.8 with 798 Degree of Freedom.

The fitted function of the transformed IV and original DV was sqrt(DV)=1.83334 – 0.96853(X\_sqrt) + 0.56538(X) with Adjusted R-squared value 0.4786. Adjusted R-sqaured value represents that 47.86% fraction of variance was explained by the IV. The 95% confidence interval for the slope was [0.398546,0.7322136]. The 95% confidence interval for the intercept was [1.520059,2.1466266]. The test statistics for the null hypothesis is F=718.197 with

The ANOVA table is shown below and the association between the independent and dependent variable is highly significant (p=0.000). When applied the lack-of-fit test, it showed nothing which means I have a good model.

Table and Plot

**Y=B+B1\*X**

> summary(l1)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.17162 0.05821 20.13 <2e-16 \*\*\*

x 0.71632 0.02825 25.35 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4034 on 798 degrees of freedom

Multiple R-squared: 0.4461, Adjusted R-squared: 0.4454

F-statistic: 642.8 on 1 and 798 DF, p-value: < 2.2e-16

> anova(step(l1,direction = 'both'))

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 104.61 104.614 642.78 < 2.2e-16 \*\*\*

Residuals 798 129.88 0.163

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> confint(step(l1,direction = 'both'))

Start: AIC=-1450.42

y ~ x

Df Sum of Sq RSS AIC

<none> 129.88 -1450.42

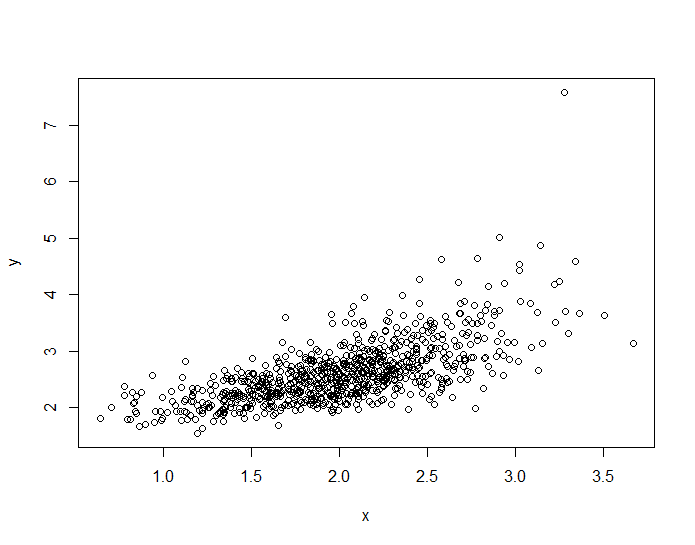
- x 1 104.61 234.49 -979.75

2.5 % 97.5 %

(Intercept) 1.057363 1.2858866

x 0.660862 0.7717836

**GGraph**



**Y= B1(X)+B2\*sqrt(X) + B**

> summary(step(l3,direction = 'both'))

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.83334 0.15960 11.487 < 2e-16 \*\*\*

x 0.56538 0.08499 6.652 5.36e-11 \*\*\*

x\_sqrt -0.96853 0.23388 -4.141 3.82e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1145 on 797 degrees of freedom

Multiple R-squared: 0.4799, Adjusted R-squared: 0.4786

F-statistic: 367.7 on 2 and 797 DF, p-value: < 2.2e-16

> anova(step(l3,direction = 'both'))

Analysis of Variance Table

Response: y\_sqrt

Df Sum Sq Mean Sq F value Pr(>F)

x 1 9.4233 9.4233 718.197 < 2.2e-16 \*\*\*

x\_sqrt 1 0.2250 0.2250 17.149 3.824e-05 \*\*\*

Residuals 797 10.4572 0.0131

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> confint(step(l3,direction = 'both'))

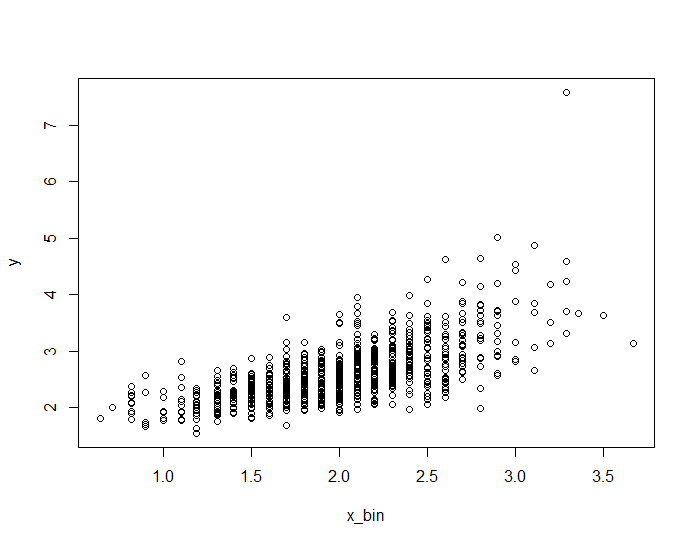
2.5 % 97.5 %

(Intercept) 1.520059 2.1466266

x 0.398546 0.7322136

x\_sqrt -1.427628 -0.5094399

**GGraph**



**Conclusion**

Adjusted R-squared value should be closer to 1 in order to have more effective and valid function. In the result that I presented, Adjusted R-squared value is the highest value among the other combinations. According to result, I determined that sqrt(DV)=1.83334 – 0.96853(X\_sqrt) + 0.56538(X) is the best fitted function. In the final model, IV were all statistically significant and explained about 47.86% of the DV variation. The result of ANOVA analysis confirmed the validity of this model.

**End of Report**