Bilal Haider (109217390)
Yun Joon Soh (108256259)
Shengyuan Luo (110598918)

Report for Data Set A

Introduction

This is a report on part A of the computer project. The main goal of this part is to merge the two data files giving the independent and dependent variable, remove any patients with missing information, and find a linear generating function that best represents the data.

Methodology

In order to solve problem A, we used R to make all our calculations. We first set a working directory into a folder on our Macintosh computers. We then used the read.csv() function to place our independent variable and dependent variable into an x and y data frame respectively. When calling this function, we set the na.string to the string "NULL" so that the dataframe will automatically place NA where NULL is presented in the data. We then used the merge function provided by R to merge the data based on the "Patientno" and then put this in another dataframe called 'total'. The total dataframe still has the NA data, so we called the na.omit() function to remove rows where the NA exists (in either the row, or the column). We then used the lm() function to find the fitted linear model and placed the contents in the variable linearAnalysis. The summary() function allows us to see the coefficients with statistical significance. We then used the anova() function to get the ANOVA table from the linearAnalysis variable.

Results

The linear model has a $\beta_0$value of 139246.19 and $\beta_1$value of 323.9 (y = 323.9x + 139246.19). The explained fraction of variation of DV was determined by the the adjusted r^2 value, 0.2989. The IV-DV association p-value was less than .001. 99% confidence interval for β_0 is [134869.025, 143623.365] and [278.863, 368.968] for slope (i.e., β_1). In the list-wise deletion, 301 rows were incomplete and thus taken out.

Table 1 - Analysis of Variance Table

ANOVA[a]

Model		Df	Sum of Squares	Mean Square	F	Sig.
1	Regression	1	8.2193e+10	8.2193e+10	344.6	.000[b]
	Residual	805	1.9200e+11	2.3851e+08		
	Total	806				

a. Dependent Variable: DV b. Predictors: (Constant), IV (<2e-16)

Conclusions

We can reject our null hypothesis, for no association at multiple conventional alpha values (.05, .01, .005), that is the slope is non-zero. We will make the conclusion that there is a highly significant association between IV and DV and we can model it by an equation.

Bilal Haider (109217390)
Yun Joon Soh (108256259)
Shengyuan Luo (110598918)

Report for Data Set B

Introduction

This is a report on part B of the computer project. The main goal of this part was to transform either or both the IV and DV to find the best fitted model. We conducted a lack of fit test to see whether this fit was adequate.

Methodology

We used Minitab and R for our calculations. We tried a variety of different procedures at first. We first tried to get rid of a few data points points where the IV value was much lower than the rest of the data point. We realized this was not the best approach. We had trouble justifying them as outliers since they were within three standard deviations from the mean when considering the IV sample size. Additionally, eliminating data points should be erred with extreme caution, as it removes data that potentially is important. In the end, we did not remove any data points. We tried a multitude of functions to transform the IV, but in the end, we transformed the data by applying ln(IV) as it made the best fit. After finding the fitted model, we analyzed the results to see the lack of fit conclusion to see whether it was valid or not, since our data had multiple DV values per IV value.

The resulted for the linear fit and lack of fit values from Minitab was verified with R, using anova method on normal linear model and factorized linear model where IV was factorized and treated as classes instead of numerical value. (See Appendix II for details)

Results

The explained fraction of variance was determined by the adjusted r^2 value, which was .9348. The lack of fit test gave a F-statistic of 1.33 and a p-value of .123. The regression equation is 13303.6 + 237.19 ln(IV), where β_0 was 13303.6 and β_1 was 237.19. The association between IV and DV p-value was less than .001. Confidence interval at 99% for β_0 and β_1 are [13172.0848, 13435.0352] and [231.7528, 242.6214] respectively.

Table 2 - Analysis of Variable Table (With Lack of Fit)

Model		DF	SS	MS	F-Value	P-value
Reg. Model	Source					
DV ~ ln(IV)	Regression	1	131280796	131280796	12694.26	0.000
	Residual	884	9142101	10342		
	Lack of Fit	26	355272	13664	1.33	0.123
	Pure Error	858	8786829	10241		
	Total	885	140422897			

Conclusion

Based on the p-value of our data in the lack of fit test, we do not reject the null hypothesis, and say there is not enough evidence to support a lack of fit, so our fit is adequate. We also conclude the slope is non-zero.

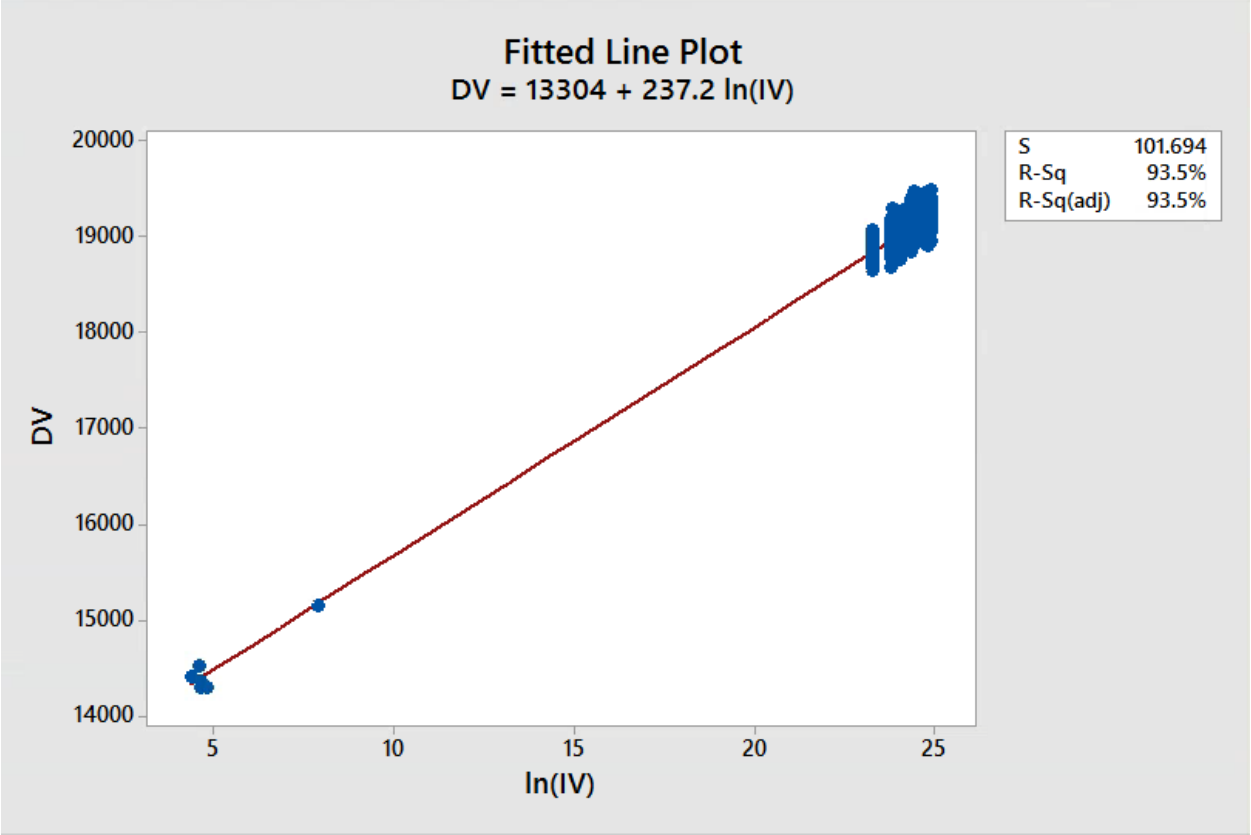
Appendix I

By Bilal Haider

Part A

```
setwd("~/Documents/Stony Brook/Yr 4 - Spring/AMS 315 Project 1")
x = read.csv("A 1 1 .csv", header = T , na.string = "NULL")
y = read.csv("A 1 2 .csv", header = T, na.string = "NULL")
total = merge(x,y,by="Patientno")
total_modified = na.omit(total)
linearAnalysis = lm(formula = total_modified$DV ~ total_modified$IV)
summary(linearAnalysis)

Call:
lm(formula = total_modified$DV ~ total_modified$IV)

Residuals:
   Min     1Q Median         3Q    Max
-38679  -9421   -260   9662  51524

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         139246.19    1695.30   82.14   <2e-16 ***
total_modified$IV    323.92      17.45   18.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15440 on 805 degrees of freedom
Multiple R-squared:  0.2998,        Adjusted R-squared:  0.2989
F-statistic: 344.6 on 1 and 805 DF,  p-value: < 2.2e-16

anova(linearAnalysis)
Analysis of Variance Table

Response: total_modified$DV
                     Df         Sum Sq    Mean Sq F value            Pr(>F)
total_modified$IV   1 8.2193e+10 8.2193e+10   344.6 < 2.2e-16 ***
Residuals           805 1.9200e+11 2.3851e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result of Minitab: Part B



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	131280796	131280796	12694.26	0.000
ln(IV)	1	131280796	131280796	12694.26	0.000
Error	884	9142101	10342		
Lack-of-Fit	26	355272	13664	1.33	0.123
Pure Error	858	8786829	10241		
Total	885	140422897			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
101.694	93.49%	93.48%	93.46%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	13303.6	50.9	261.20	0.000	
ln(IV)	237.19	2.11	112.67	0.000	1.00

Appendix II

By Yun Joon Soh

R Code: Part A

```
################################################################################
###
# Part A

# Read files
x = read.csv("A 1 1 .csv", header = T , na.string = "NULL")
y = read.csv("A 1 2 .csv", header = T, na.string = "NULL")
A = merge(x,y,by="Patientno")

# List wise deletion
A = na.omit(A)

# Linear Model
fitA = lm(formula = A$DV ~ A$IV)

# Output results
summary(fitA)
anova(fitA)

# Calculate confidence interval
confint(fitA, level = 0.99)

################################################################################
###
```

Result of R Code: Part A

```
> summary(fitA)
Call:
lm(formula = A$DV ~ A$IV)

Residuals:
   Min     1Q Median     3Q    Max
-38679  -9421   -260   9662  51524

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 139246.19    1695.30   82.14   <2e-16 ***
A$IV           323.92      17.45   18.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15440 on 805 degrees of freedom
Multiple R-squared:  0.2998,     Adjusted R-squared:  0.2989
F-statistic: 344.6 on 1 and 805 DF,  p-value: < 2.2e-16


> anova(fitA)
Analysis of Variance Table

Response: A$DV
```

```
            Df     Sum Sq    Mean Sq F value    Pr(>F)
A$IV         1 8.2193e+10 8.2193e+10   344.6 < 2.2e-16 ***
Residuals  805 1.9200e+11 2.3851e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> confint(fitA, level = 0.99)
                  0.5 %     99.5 %
(Intercept) 134869.025 143623.365
A$IV           278.863    368.968
```

R Code: Part B

```
###############################################################################################
###
# Part B

# Read the file
B = read.csv("B 1 .csv")  # read csv file

# Plot
plot(B$IV, B$DV)

# Linear model for log and factor
lnfit1 = lm(DV ~ log(IV), data = B)
lnfit2 = lm(DV ~ factor(IV), data = B)

# Analyze
aovB = aov(DV ~ log(IV), data = B)
anova(lnfit1, lnfit2)

# Output results
summary(aovB)

# Calculate confidence interval
confint(lnfit1, level = 0.99)
###############################################################################################
###
```

Result of R Code: Part B

```
> anova(lnfit1, lnfit2)
Analysis of Variance Table

Model 1: DV ~ log(IV)
Model 2: DV ~ factor(IV)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    884 9142101
2    858 8786829 26    355272 1.3343 0.1233

> # Output results
> summary(aovB)
             Df    Sum Sq   Mean Sq F value Pr(>F)
log(IV)       1 131280796 131280796   12694 <2e-16 ***
Residuals   884   9142101     10342
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Calculate confidence interval

> confint(lnfit1, level = 0.99)
                 0.5 %     99.5 %
(Intercept) 13172.0848 13435.0352
log(IV)       231.7528   242.6214
```