AMS 315 Project 1 Part A
Professor: Stephen Finch
Group 87: Ivan Tinov, Lipeng Feng

Introduction:

In part A of this project, our goal is to merge two sets of data in the form of .csv files with one file containing the independent variable (X), while the other file contains the dependent variable (Y). After the two files are merged, any data set that contains missing data is removed using a list-wise deletion. Finally, we used both R and Minitab to find the fitted linear model described below.

Methodology:

In order to solve this part of the project, we used RStudio and the R programming language to create the tables and models below. We also used Minitab to check our data with another statistical package and to just experiment more with these kinds of software. In RStudio, we first set a working directory where all of our files will be located. After this, we loaded the independent and dependent variables into a data set within R using the read.csv() function. Likewise, we set the na.string to "NA" so that the data frame that is storing our data will replace every missing value with the "NA" string. After this we merged our two independent and dependent data sets using the merge(x, y, by = "ID") function which merges together the data sets based on their respective ID. We assigned the merge function to M, a new variable in which the merged data is now stored in. However, this new data frame still has the missing data, so we used the na.omit(M) function, and removed all of the rows/columns that contained the "NA" string. Then, we found the fitted linear model using the lm(formula = M$Y ~ M$X) function and assigned it to the variable fitM. We used three other important functions which are the summary(fitM) function to find the coefficients with statistical significance as well as the anova(fitM) function to find the ANOVA table, and the confint(fitM, level = 0.99) to find the confidence interval. All the tables can be viewed below.

Result:

The fitted linear model is observed to have a β_0 value of 9.994e+05, and a β_1 value of -3.883e+01. Therefore, the linear fitted model, $Y = \beta_0 + \beta_1 X$ can be described as Y = 9.994e+05 - 3.883e+01X with an R^2 value of 77.1% (0.771). The 99% confidence interval for the intercept (β0) is (997563.54010, 1001234.98173), and the 99% confidence interval for the slope (β_1) is (-40.09369, -37.56025). After using the list-wise deletion, 141 rows contained missing values and thus were removed from the data frame.

The ANOVA table is:

Table 1

Analysis of Variance (ANOVA) Table

Regression Analysis: DV vs. IV

(n = 1858)

ANOVA*						
Model		**Sum of Squares**	**Df**	**Mean Square**	**F**	**Sig**
1	**Regression**	5.8800e+12	1	5.8800e+12	6246.4	0.000[b]
	Residual	1.7481e+12	1857	9.4133e+08		
	Total	7.6281e+12	1858			
a) Dependent Variable: DV						
b) Predictors: (Constant), IV						

The Coefficients table is:

Table 2

Coefficients[a]

		Unstandardized Coefficients		**Standardized Coefficients**			**95.0% Confidence Interval for B**	
Model		**Estimate**	**Std. Error**	**Beta**	**T-Value**	**Sig.**	**Lower Bound**	**Upper Bound**
1	(Const.)	9.994e+05	7.119e+02		1403.77	.000	997563.540	1001234.982
	X	-3.883e+01	4.913e-01	-0.878	-79.03	.000	-40.094	-37.560
a) Dependent Variable: Y								

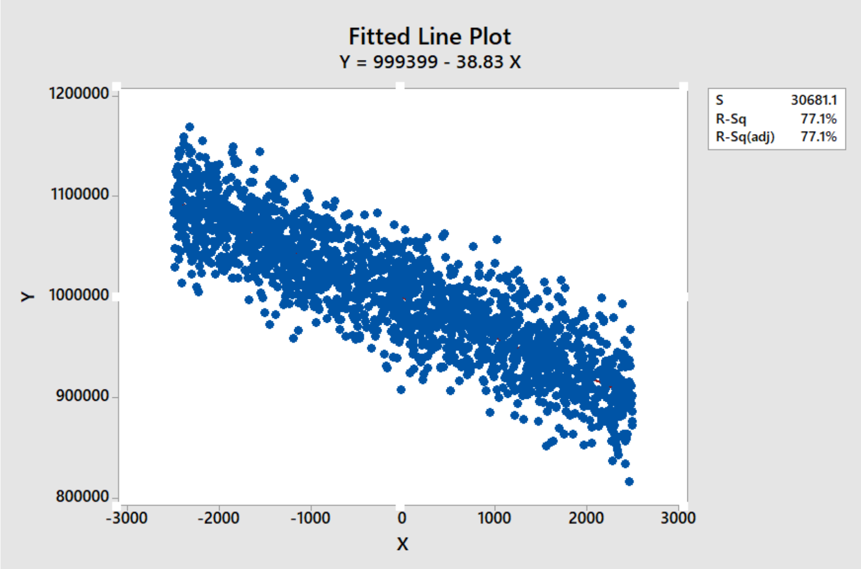
The Model Summary table is:

Table 3

Model Summary[b]

						Change Statistics			
Model	**R**	**R^2**	**Adjusted R Square**	**Std. Error of the Estimate**	**R^2 Change**	**F Change**	**df1**	**df2**	**Sig. F Change**
1	0.878	0.771	0.771	30680	0.878	6246	1	1857	0.000
a) Predictors: (Constant), X									
b) Dependent Variable: Y									

Graph for the Fitted Line Plot:



The plot has an R^2 value of 77.1% as described above in the results.

Conclusion:

We concluded that for problem A, the dependent variable and the independent variables are highly associated with a p – value of 0.000. This we can reject the null hypothesis. The residual plots can also confirm this conclusion. Similarly, the fitted linear model of Y = 9.994e+05 - 3.883e+01X showcases the relationship between the independent and dependent variables.

End of Report