AMS 315 Project 1

Group 13

108855476 Tang Guo

109553043 Rong Bo

107968074 Chen Xinhe

108854857Kung Preston

Problem A

Introduction:

For problem A, we used two sets of data, which were listed with identification numbers. The aim of the problem is to find the function of the line, which is based on the linear regression. The data used in this program is statistically analyzed through a simulated program.

Methodology:

In order to solve problem A, we use the statistics package R and Microsoft Excel. The original data files were supplied with two data sheets in Excel, with one data sheet had ID and associated independent variables and the other one had ID and associated dependent variables. The raw data files had total 1882 independent variables with ID# ranging from 1 to 1882, and it also had total 1685 dependent variables with ID # ranging from 1 to 1882. We first sorted data in both files in ascending orders and then merged it by using Excel. From this merge we can see that there is missing data and we used list-wide delete to delete these missing values. Finally, we merged two files into one file with three columns: ID, IV and DV, with total 1685 entries and ID# ranging from 1 to 1882. Therefore, with the computational statistics package R, the data can be analyzed and show the diverse results.

Results:

The fitted function for the model Y= B+B1. The linear regression fitted function found was: y = 5.97387x-79.51276. The 95% confidence interval for the slope was [5.94086 6.006887]. The 95% confidence interval for the intercept was [-83.60556 -75.419961]. The analysis of variance table is shown below and the association between the independent variable and dependent variable is highly significant (p=2.2e-16).

Conclusions:

From the linear regression results, we can conclude that the null hypothesis, where the slope equals zero, is rejected by the data. Using the statistical simulated program, we were able to manipulate our data to generate a straight line that is well fitted. The 99% of the variance explained is almost perfect. This makes it certain that the slope is accurate of the data set provided.

> setwd("~/Desktop/315")

> dataA = read.csv("dataA.csv", header=T)

> IV = dataA[,1]

> DV = dataA[,2]

> fit = lm(DV~IV)

> summary(fit)

Call:

lm(formula = DV ~ IV)

Residuals:

Min 1Q Median 3Q Max

-300.057 -58.893 -1.202 57.943 278.087

Coefficients:

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

(Intercept) -79.51276 2.08670 -38.1 <2e-16 \*\*\*

IV .97387 0.01683 354.9 <2e-16 \*\*\*

---

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

Residual standard error: 85.65 on 1683 degrees of freedom

Multiple R-squared: 0.9868, Adjusted R-squared: 0.9868

F-statistic: 1.26e+05 on 1 and 1683 DF, p-value: < 2.2e-16

> fit$coef

(Intercept) IV

-79.512762 5.973873

> # (Intercept) IV

> # -79.512762 5.973873

> # fitted linear model: DV = -79.51276 + 5.97387\*IV

> anova(fit)

Analysis of Variance Table

Response: DV

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

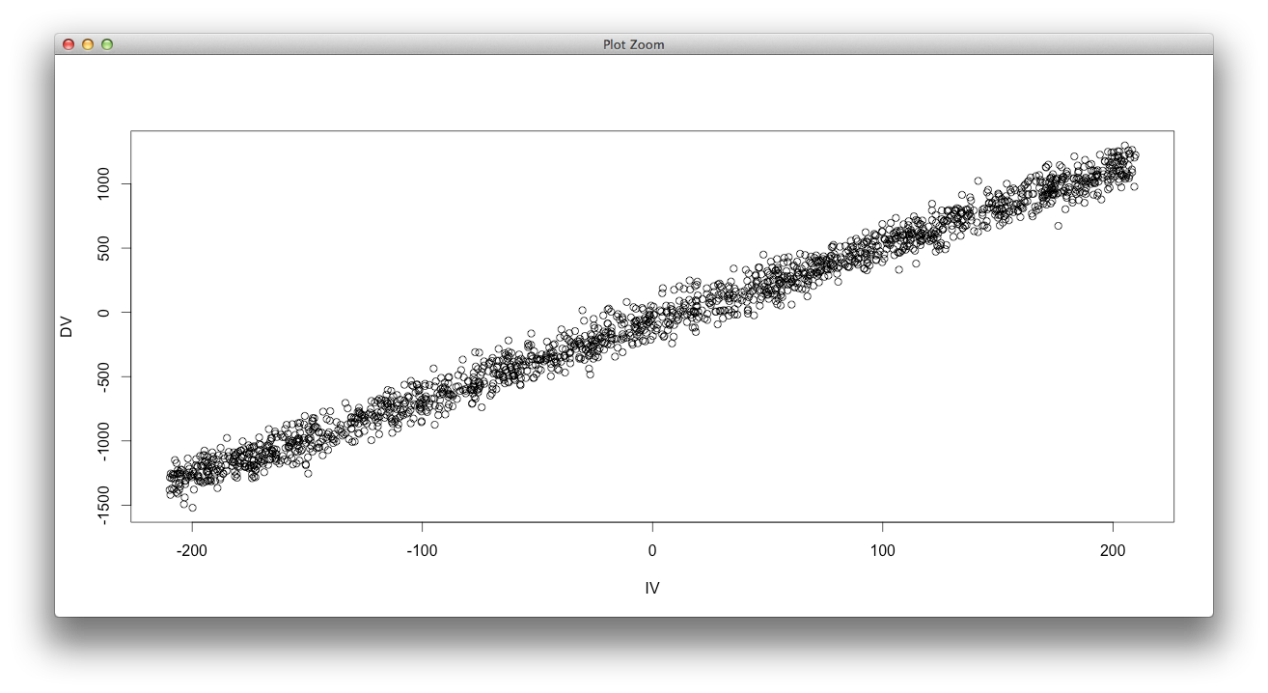
IV 1 924154072 924154072 125967 < 2.2e-16 \*\*\*

Residuals 1682 12347339 7337

---

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

> plot(DV~IV)



> confint(fit, level = 0.95)

2.5 % 97.5 %

(Intercept) -83.60556 -75.419961

IV 5.94086 6.006887

> confint(fit, level = 0.99)

0.5 % 99.5 %

(Intercept) -84.893848 -74.131675

IV 5.930468 6.017278

Problem B:

Introduction:

In part B, we use two files of data, each has dependent variables or independent valuables. The goal of problem B is to use the transformation on independent variables or independent variables to make data more approach to linear regression model.

Methodology:

In order to solve problem B, we uses the statistics package R and Microsoft Excel. The original data files were supplied with two data sheets in Excel, with one data sheet had ID and associated independent variables and the other one had ID and associated dependent variables. The raw data files had total 1734 independent variables and dependent variables with ID# ranging from 1 to 1734. The data was then imported into R. We sorted the data and saw that no data was missing so nothing had to be deleted. We used R to analyze the data and given the plot we could see that the data was not linear. We squared the given dependent variables, which gave us a linear representation of the data. We then plugged the new dependent variable (DV^2) and the same dependent variable back into R to get the results below.

Results:

The linear regression fitted function found was: y = 760.57x-239102. The 95% confidence interval for the slope is [-960.42302 -912.00731]. The 95% confidence interval for the slope was [5.94086 6.006887].At a 0.05 significance level, the null hypothesis was tested and rejected.

Conclusions:

The observation of the data such as the fraction of variance is almost too perfect and therefore, is a strong evidence to accept a linear regression model. According to the data, the function’s computed fit is perfect. The 95% confidence interval calculated is extremely narrow and thus, the data indicates a linear relationship between variables exists.

> dataB = read.csv("dataB.csv", header = T)

> IV = dataB[,2]

> DV = dataB[,3]

> avg = mean(DV, na.rm=TRUE)

> DV[is.na(DV)] = avg

> #check if all the NA's have been replaced with mean

> n = sum(is.na(DV))

> n

[1] 0

> fit2 = lm(DV~IV)

> summary(fit2)

Call:

lm(formula = DV ~ IV)

Residuals:

Min 1Q Median 3Q Max

-6.7651 -1.3987 -0.4340 0.9773 7.9520

Coefficients:

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

(Intercept) 24.1597 0.1253 192.88 <2e-16 \*\*\*

IV -936.2152 12.3425 -75.85 <2e-16 \*\*\*

---

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

Residual standard error: 2.167 on 1732 degrees of freedom

Multiple R-squared: 0.7686, Adjusted R-squared: 0.7685

F-statistic: 5754 on 1 and 1732 DF, p-value: < 2.2e-16

> fit2$coef

(Intercept) IV

24.15972 -936.21517

> # (Intercept) IV

> # 24.15972 -936.21517

> # fitted linear model: DV = 24.15972 + -936.21517 \*IV

> anova(fit2)

Analysis of Variance Table

Response: DV

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

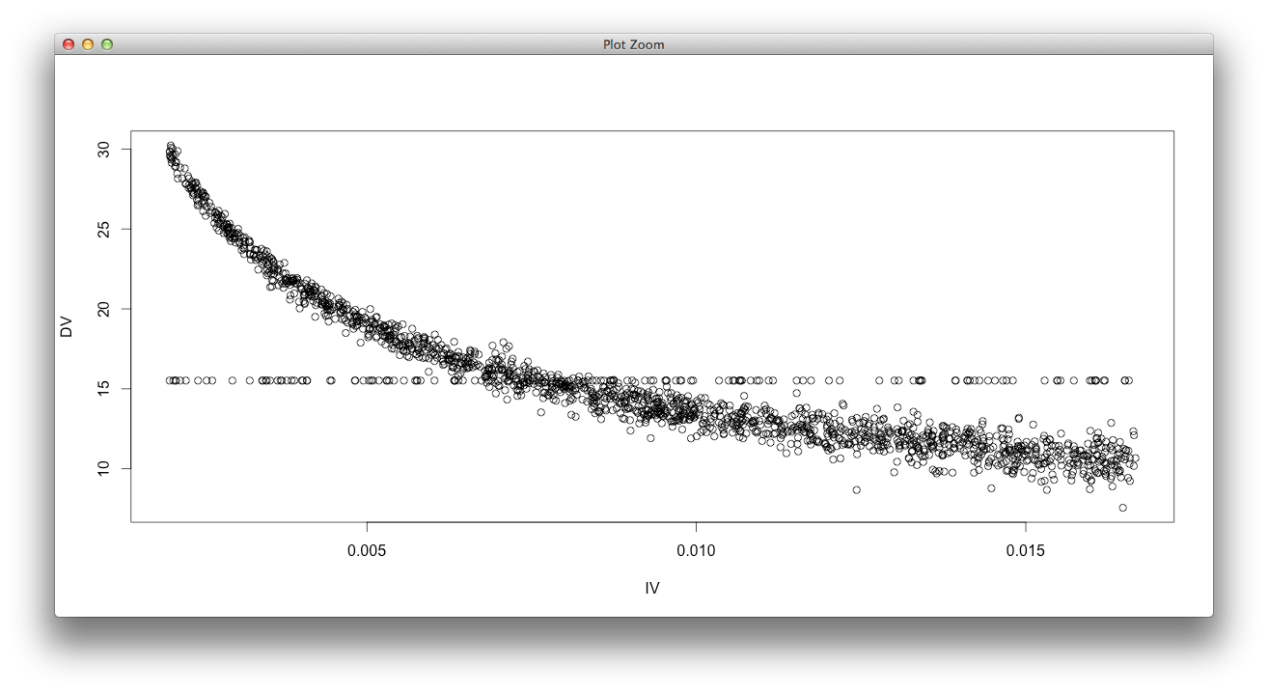
IV 1 27027.2 27027.2 5753.6 < 2.2e-16 \*\*\*

Residuals 1734 8135.9 4.7

---

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

> plot(DV~IV)



> confint(fit2, level = 0.95)

2.5 % 97.5 %

(Intercept) 23.91406 24.40539

IV -960.42302 -912.00731

> confint(fit2, level = 0.99)

0.5 % 99.5 %

(Intercept) 23.83673 24.48271

IV -968.04252 -904.38782