Report of Simple Linear Regression - Part A

Group 50: Claire Pang, Kunlang Li, Yutong Guo

Introduction

For part A, we are given two separate excel files containing two sets of data. Our goal is to merge this two file which sort by Patient ID into one data set and fit the data set to a linear model.

Methodology

Before recovering the linear regression function, we need to sort the two files by Patient ID and merge them. In order to do so, we used the statistical software R. We would also use R to solve all the following problem. The two files we were given were originally in .xls files, we first changed them into .csv files and imported them into R. Then we merged the two data sets into one using “merge” function in R. We tested if there were missing variable in the new data set using function “is.na” and found out the R deal with the missing variable automatically while we merged the two data set. After executiing “total” command we noticed now we had 481 data left while data set X had 497 data and data set Y had 496 data. Then we found the linear relation between X(IV) and Y(DV).The entire code would be attached in the appendix section.

Result

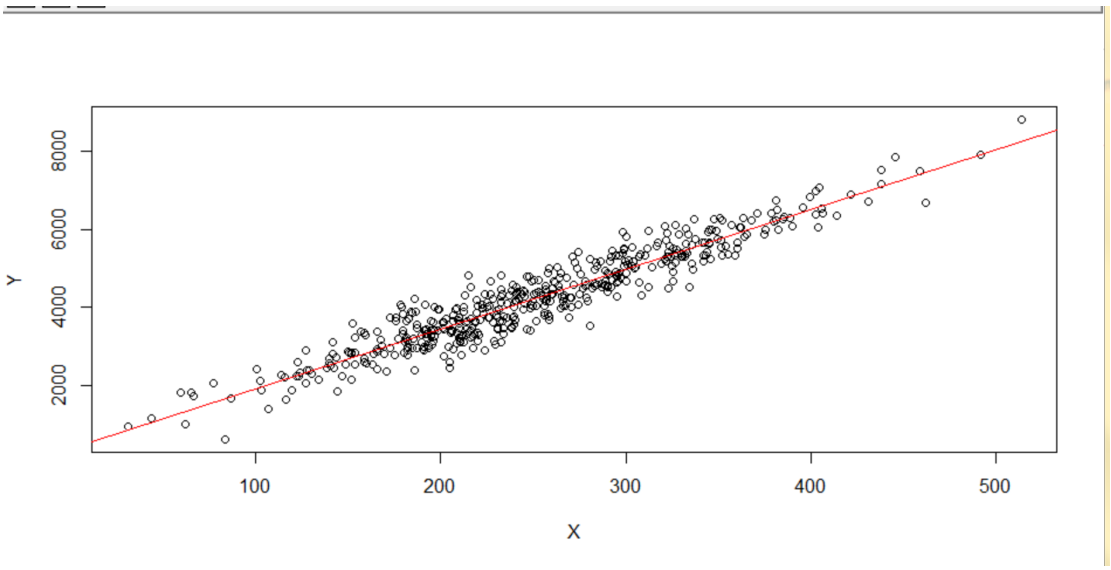
The fitted function we found for the model $Y= \beta_0 + \beta_1 X$ was

Y=394.67+15.29X. SInce R^2=0.901 we assumed that the fitted function is valid. The 95% confidence interval for the slope [14.83,15.74]. We did the correlation test with a result of p<2.2e-16 which is closed to 0, thus the association between the IV and DV is significant.

lm(formula = Y ~ X, data = Data)

Residuals:				
Min	1Q	Median	3Q	Max
-1141.54	-260.79	-21.59	263.61	1138.2
Coefficients:				
	Estimate	Std.Error	t value	Pr(>|t|)
Intercept:	3.95E+02	6.07E+01	6.499	<2.03e-10 ***
x:	1.53E+01	2.31E-01	66.107	< 2e-16 ***
Signif. Codes :0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 391.4 on 479 degrees of freedom				
Multiple R-squared: 0.9012, Adjusted R-squared: 0.901				
F-statistic: 4370 on 1 and 479 DF, p-value:<2.2e-16				

data: Data$X and Data$Y
t = 66.107, df = 479, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9396744 0.9574662
sample estimates:
0.9493254



Anova Table

Response: Y					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
X1	1	6.70E+08	6.70E+08	4370.1	< 2.2e-16 ***
Residuals	479	7.34E+07	1.53E+05		
Signif. Codes :0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Conclusion

For the fitted function in part A, 90.1% fraction of variance is explained. From the scatter plot and the correlation test, we can also tell that the model is valid.

Appendix

-R studio

```
> Data1=read.csv("C:/Download/Group_50_X.csv",header=TRUE)
> Data2=read.csv("C:/Download/Group_50_Y.csv",header=TRUE)
> Data=merge(Data1,Data2,by="Patient_ID")
> fit<-lm(Y~X,Data)
> summary(fit)
> plot(Y~X,Data)
abline(lm(Y~X,Data),col="red")
```

```
> anova(fit)
> cor.test(Data$X,Data$Y)
> confint(fit,'X',level=0.95)
```

Report of Simple Linear Regression - Part B

Introduction

In part B, we were given one set of data containing a independent variable (X) and dependent variable (Y). Our goal is to find the best fitted linear regression model by performing the transformation on the independent variables and/or the dependent variables.

Methodology

First, we add columns in the excel file we were given and named each column with (X1,X2,X3,X4,Y1,Y2) which correspond to (X^2,X-squared,lnX,1/X,Y^2,Y-squared). We use excel formula to get the data we need. Then we imported the data in R using the same method as in part A. We computed "lm" command to get the linear regression function of each combination of X and Y. By comparing the R^2 value, F-statistic value and the Std.Error, we determined that the best fitted linear regression equation exist for (IV^2,DV)=(X^2,Y). Then we apply the lack of fit test. All the code will be attached in the Appendix section.

Result

The fitted function of model $Y=\beta_0+\beta_1X$ was Y=-3164.74+157.03X and the fraction of variance is 0.906. The 95% confidence interval of slope is [154.01,160.04]. The fitted function for (IV^2,DV)=(X^2,Y) is $Y=1490+1.197X^2$ with fraction of variance 0.906. The 95% confident interval of slope is [1.1743,1.2201].

lm(formula = Y ~ X, data = PB)

Residuals:				
Min	1Q	Median	3Q	Max
-2916.3	-671.4	-29.6	702.8	3945.9

Coefficients:				
	Estimate	Std.Error	t value	Pr(>|t|)
Intercept:	-3164.743	104.265	-30.35	<2e-16 ***
x:	157.031	1.537	102.14	<2e-16 ***
Signif. Codes :0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 1028 on 1082 degrees of freedom				
Multiple R-squared: 0.906, Adjusted R-squared: 0.906				
F-statistic: 1.043e+04 on 1 and 1082 DF, p-value:< 2.2e-16				

lm(formula = Y ~ X1, data = PB)

Residuals:				
Min	1Q	Median	3Q	Max
-3375.8	-735.6	-17.4	716.7	3525

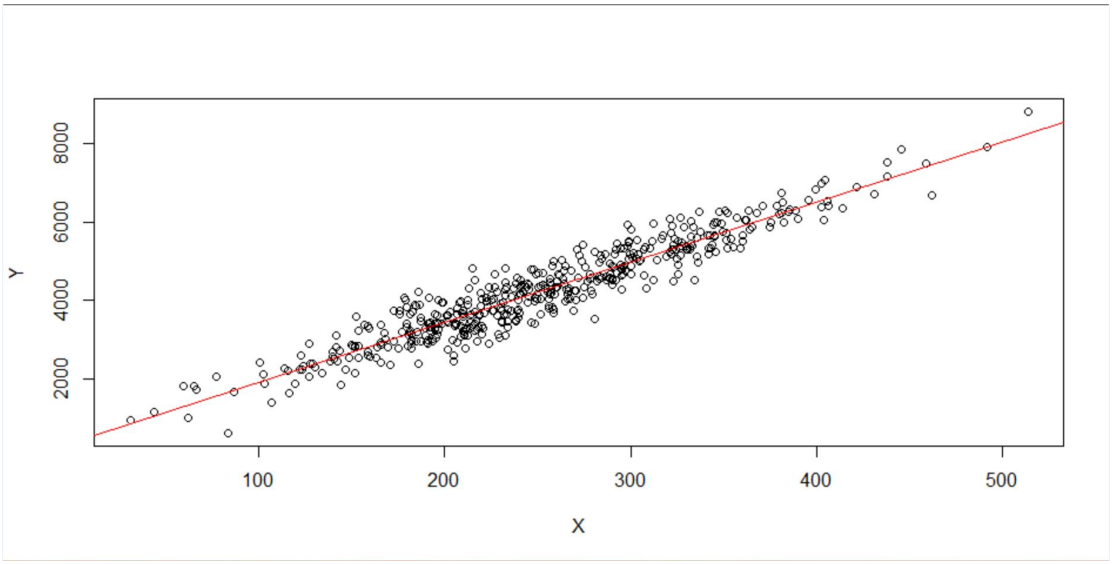
Coefficients:				
	Estimate	Std.Error	t value	Pr(>|t|)
Intercept:	1.49E+03	6.20E+01	24.04	<2e-16 ***
x:	1.20E+00	1.17E-02	102.67	<2e-16 ***
Signif. Codes :0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 1023 on 1082 degrees of freedom				
Multiple R-squared: 0.9069, Adjusted R-squared: 0.9068				
F-statistic: 1.054e+04 on 1 and 1082 DF, p-value: < 2.2e-16				

Anova Table

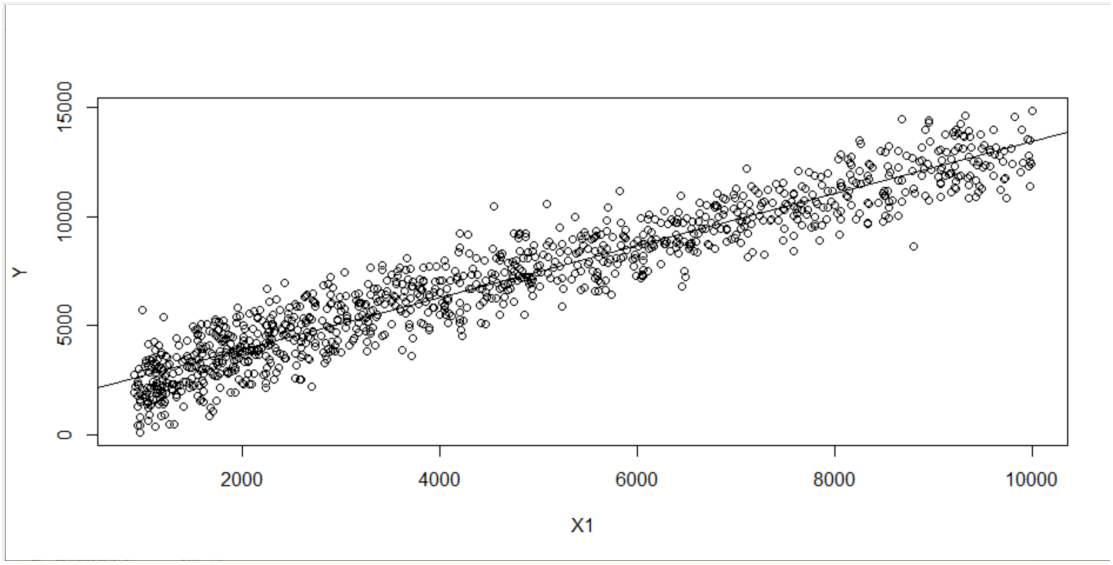
Response: Y					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
X1	1	1.10E+10	1.10E+10	10542	<2e-16 ***
Residuals	1082	1.13E+09	1.05E+06		
Signif. Codes :0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Model 1: Y ~ X1						
Model 2: Y ~ 0 + as.factor(X1)						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1082	1132158107				
2	29	34384085	1053	1097774022	0.8793	0.7178

Y vs. X



Y vs. X^2



Conclusion

The fraction of variance we got is formidable enough to accept our linear regression model. And accourding to the F-statistic value and R^2 value we accept that the relationship between Y and X^2 is our best fit linear regression equation.

Appendix

-R-studio

```
> PB=read.csv("C:/Download/Group_50_PartB.csv",header=TRUE)
> fit<-lm(Y~X,PB)
> summary(fit)
> plot(Y~X,Data)
> abline(lm(Y~X,Data),col="red")
> fit<-lm(Y~X1,PB)
> summary(fit)
> plot(Y~X1,PB)
> abline(lm(Y~X1,PB))
> anova(fit)
> reduced=lm(Y~X1,PB)
> full=lm(Y~0+as.factor(X1),PB）
> anova(reduced,full)
95 IC for origin
> confint(fit,"X",level=0.95)
        2.5 %      97.5 %
X 154.0142 160.0472
95 IC for new> fit<-lm(Y~X1,PB)
> confint(fit,"X1",level=0.95)
          2.5 %      97.5 %
X1 1.174329 1.220087
```