DATA1046 Part A

Introduction

The objective is to find the appropriate model describing the data in Problem A. A simulation program using an unknown linear function was used to generate the data.

Methodology

We used R to solve the problem. The original data files were supplied with two data sheets. One data sheet had the ID of an object and its associated independent variable value, and the other had the ID and its associated dependent variable value. Both independent variable data file and dependent variable data file had total values with ID# ranging from 1 to 777, but with some data missing. The first data set has 79 missing values and the second data set has 120 missing values. We first sorted data in both files in ascending ID# order using Excel and then used R's merge function to merge the files. We next used complete cases function of R to remove 207 missing values that were missing either the independent variable value or the dependent variable value. After preprocessing the data set, there were 590 entries with both values, with ID# ranging from 1 to 590. After plotting the data (see appendix), we see a good linear relation between independent variable values and dependent variable values. Therefore, we use linear transformation of the data.

Results

The fitted function for the model $Y= B+B_1 X$ which was DV=7.9739V+88.1249with 39.7% fraction of variance was explained. The 95% confidence interval for the slope was [7.177996 , 8.769857]. The 95% confidence interval for the intercept was [86.46178,89.788055]. The analysis of variance table is given below and the association between the independent variable and dependent variable was highly significant (p=0.000).

Table 1
Analysis of Variance Table
DV regressed on IV
(n=670)

ANOVA

Model	Sum of Squares	Df	Mean Square	F	Pr(>F)
Regression	9151.308	1	9151.30835	387.1509	0
Residual	13898.893	588	23.63757		
Total	23050.201	589			

Conclusion

For problem A, the association between independent variables and dependent variables was highly significant (p=0.000), with 39.7% fraction of variance was explained. The plot of residual versus predicted value confirmed the validity of this model.

Appendix

Table 2

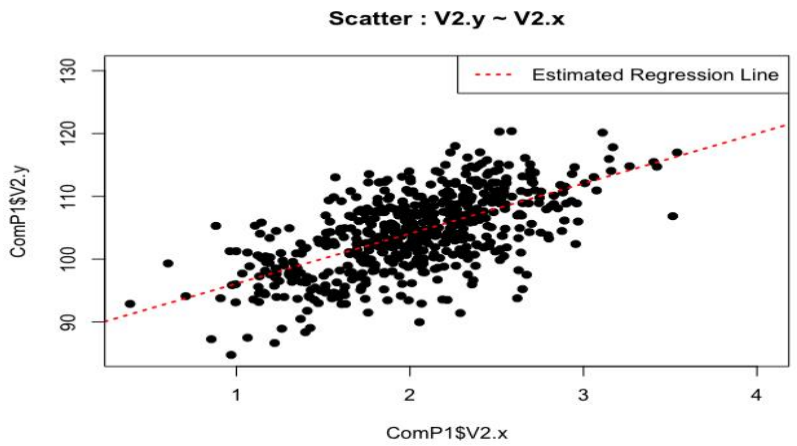

Table 3

	Estimate	*Standard Error*	*t value*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	88.1249	0.8468	104.07	<2e-16	86.461788	89.788055	85.936591	90.313253
V2. x	7.9739	0.4053	19.68	<2e-16	7.177996	8.769857	6.926651	9.021202

Table 4

Regression Statistics	
Multiple R Square	0.397
Adjusted R Square	0.396
Standard Error	387.2
Observations	590

End of Report