DATA1646 Project 1 Problem B

Introduction

This report was designed to show the method of solving Problem B. The aim of problem B was to find the relation of the dependent variables (DV) and the independent variables (IV). We also need to bin the data together and apply lack of fit tests to decide the best fit function.

Methodology

We imported the data set in Excel and noted that it contains 800 entries. Then we sorted the data from smallest to largest based on the values of x using excel. We then bin nearly repeated data together and then get 40 groups with 20 data in each group. In order to find the best fitting model, we apply several transformations on IV or DV such as IV^2, exp (IV), sqrt (DV), log (DV). For each of the model, we applied a lack of fit test. We first calculate the Sum of Squares of Pure Error (SSPE)and divide them by the degree of freedom to get the Mean Square of Pure Error (MSPE). The next step is to use ktable function in R to work out the Analysis of Variance Table in order to calculate the lack of fit (LOF) Sum of Squares and the degree of freedom. Now we get the Mean Square of LOF and divide it by the Mean Square of Pure Error. Next, we get the F-value. Finally, we compared all the model statistics, and choose the best fitting model by comparing the F value with 1.

Results

After testing the linear regression of 5 models, which are (IV, DV), (IV^2, DV), (IV, sqrt (DV)), (exp (IV), DV) and (IV, log (DV)). The analysis of variance tables is shown below. The F value of model (IV, log (DV)) is 96.2893805, the F value of model (exp (IV), DV) is 28.909432 and the F-value of model (DV ~ IV^2) is 6.23102258, which indicates that we have to conclude a lack of fit. For the other 2 models with reasonable F values, we choose the one (IV, DV) with F value 1.92071229.

Conclusion

The best fitting model is DV = a*IV + b, which is DV=0.0110430IV+4.7187063. We do not conclude that there is a lack of fit after we compared the F value with 1 in the lack of fit test. The association between IV and DV is highly significant, since the p value of t test is almost zero, and 55.71% fraction of variance is explained.

Appendix for Problem B

Analysis of Variance Table
Linear Regression and Lack of Fit Sums of Squares

Model 1: DV ~ IV

Source	Sum of Squares	Degree of freedom	Mean Square	F test
Regression	0.8686407	1		1003.956
Error	0.6904441	798		
Lack of (linear) fit	0.06049757	38	0.00159204	1.92071229
Pure Error	0.62994653	760	0.00082888	
Total	1.5590848	799		

Model 2: DV ~ IV^2

Source	Sum of Squares	Degree of freedom	Mean Square	F test
Regression	0.7328770	1		
Error	0.8262079	798		
Lack of (linear) fit	0.19626137	38	0.00516477	6.23102258
Pure Error	0.62994653	760	0.00082888	
Total	1.5590848	799		

Model 3: sqrt (DV) ~ IV

Source	Sum of Squares	Degree of freedom	Mean Square	F test
Regression	0.0452645	1		
Error	0.0359682	798		
Lack of (linear) fit	0.00318485	38	0.00008381	1.94278725
Pure Error	0.03278355	760	0.00004314	
Total	0.0812327	799		

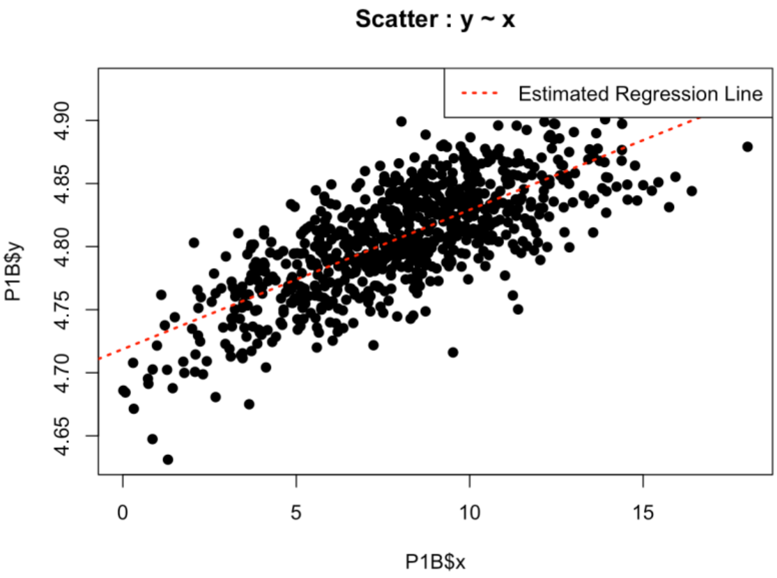
Model 4: DV ~ exp (IV)

Source	Sum of Squares	Degree of freedom	Mean Square	F test
Regression	0.0185651	1		0.0019958
Error	1.5405197	798		
Lack of (linear) fit	0.91057317	38	0.02396245	
Pure Error	0.62994653	760	0.00082888	28.909432
Total	1.5590848	799		

Model 5: log (DV) ~ IV

Source	Sum of Squares	Degree of freedom	Mean Square	F test
Regression	0.0377411	1		
Error	0.0299829	798		
Lack of (linear) fit	0.02483379	38	0.00065352	96.2893805
Pure Error	0.00514911	760	0.00000678	
Total	0.067724	799		

Model 1 Scatter (DV, IV)



	Estimate	*Standard Error*	*t value*	*P-value*
Intercept	4.7187063	0.0029988	1573.54	<2e-16
V2. x	0.0110430	0.0003485	31.68	<2e-16

Regression Statistics	
Multiple R Square	0.5571
Adjusted R Square	0.5566
Standard Error	0.02941
Observations	800