IE 360 – Assignment 2

Spring'19

Due Date: 10.04.2019

1. Our aim is to build a simple linear regression model where we explain the response variable Y with the explanatory (prediction) variable X. For this, you are supposed to use the following data:

```
x<-scan()
186 119 180 137 152 171 169 131 137 105 191 195 118 117 150 186 183
176 160 138 109 108 160 150 127 193 182 114 118 129 170 175 115 166
136 144 145 185 146 132 117 148 136 154 157 183 189 133 180 141
y<-scan()
2397 156 1572 339 600 1364 1317 265 288 121 2101 2038 214 132 601 1729
1986 1738 757 280 106 112 637 559 195 2227 1689 93 144 273 849 1364 159
953 322 357 391 2273 516 229 162 571 338 594 614 2102 2635 278 1960 374
```

Using "lm" function in R,

a) Construct the standard model $Y = \beta_0 + \beta_1 X + \varepsilon$ and check the model assumptions. Which of them are not met?

b) To fulfill the model assumptions, suggest a better model and re-check the model assumptions.

c) Interpret the estimated regression coefficient for X and construct a 95% confidence interval for it.

d) Using "predict" command, make forecasts for observed values x = 125 and x = 250 of the explanatory variable. Discuss the reliability of these forecasts.

e) Construct a 95% confidence interval and prediction interval for x = 150.

f) How are the confidence and prediction intervals different than each other? Explain the reason of the difference between them.

2. The file "salesperson.txt" contains a sample data to forecast the SALES (per month) of a person. Using the following variables, we are trying to forecast if a particular applicant will be a good salesperson or not.

APT: Selling aptitude test score

AGE: Age (in years)

ANX: Anxiety test score

EXP: Experience (in years)

GPA: High school GPA

Here, all of these variables may not be needed to forecast sales of a person, so you need to implement stepwise regression to reach a sensible final model.

a) Calculate the correlation matrix of all 6 variables and look at all scatter plots between the variables. Which variables do you think are needed to forecast sales values?

b) Implement stepwise regression by following the steps below and obtain a final regression model.

c) Write down your estimates for the intercept, coefficient(s) for the variables and residual variance.

d) Test if high school GPA of a person has an influence on sales value (Use $\alpha = 0.05$). State H0, H1 and the p-value of the test.

Step 1: Choose the variable having the highest absolute correlation value. Construct an initial simple linear regression model using this variable and the response.

Step 2: Out of the variables that are not in the model, build a new model by adding one variable into your current model. Use the command anova(currentmodel,newmodel) to test the significance of this new variable with an F-test. Do this for all variables which are not in the current model. Choose the variable that corresponds to largest F-statistic (smallest p-value) and update your current model by adding this variable.

Step 3: Once a new variable is added into your current model, build a reduced model by removing one of the variables which was already in your current model (except the last one added in the previous step). Use the command anova(currentmodel,reducedmodel) to test the significance of the removed variable with an F-test. If the p-value of this test is larger than a sensible significance level (if F-statistic is small then critical F-value), then update your current equation by removing this variable. Otherwise, do not touch that variable. Do this for all variables in your current model, except the last variable added in the second step.

Step 4: Repeat step 2 and 3 until all possible additions are nonsignificant and all possible deletions are significant. (For this question, do not focus on the model assumptions.)

While writing your asssignment report, please try to follow the guidelines given in “Assignment Format.pdf" file, as much as possible. Write down all of your code and the necessary output.

3. The file “salesdata.txt" contains quarterly PROFIT of a company (in thousand dollars) with the quarterly SALES of their product (in tons) starting from the first quarter of 1988.

a) Build a linear regression model that explains the variability in the PROFIT with the observed information SALES. Use any dummy variables if necessary.

b) Check if the model assumptions are fulfilled or not.

c) Find a way to meet all model assumptions. Build a new model. Again, use any dummy variables if necessary. Check the model assumptions for this new model. Is this new model reliable?

d) Your expected sales in the first quarter of 2013 is 30 tons. According to your model in (c), what is your forecast for the profit in this quarter?

While writing your asssignment report, please try to follow the guidelines given in “Assignment Format.pdf" file, as much as possible. Write down all of your code and the necessary output.