**DSC 424: Advanced Data Analysis and Regression**

**Assignment 01**

**Name: Sidhant Thakur**

**Student ID: 2020181**

**Problem 1**

**By hand calculation**

**Text, letter

Description automatically generated**

**Input**

**Text

Description automatically generated**

**Output**

**Graphical user interface, text, application

Description automatically generated**

I get the same result when I calculate by hand and when I program in R.

**Problem 2**

**Hand Calculation**

**Text, letter

Description automatically generated**

**Input**

**Text

Description automatically generated**

**Output**

**Text

Description automatically generated**

The calculation yields the same result. i.e., the output matrix by hand and by r is the same.

**Problem 3**

Graphical user interface, text, application, email

Description automatically generated

Output:

Table

Description automatically generated

Yes, we have the same coefficients in the output.

**Problem 4**

**a)**

**Graphical user interface, text

Description automatically generated**

**Text

Description automatically generated**

**b)**

**For F test,** p-value: < 2.2e-16.

So, we reject the null hypothesis i.e. (All the betas are equal to zero) and accept the alternative hypothesis that at least one beta is not equal to zero.

For the **ZN,** we have significant p-value; so, we reject the null hypothesis(B1=0) and accept the alternative hypothesis(B1=/=0). And we are going to use prediction of 0.030200.

For the **DIS**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 0.000848.

For the **RAD**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 4.95e-10

For the **INDUS, CHAS, NOX, RM, AGE, TAX, PTRATIO, LSTAT,** we do not have significant p-value. so, we fail to reject the null hypothesis (B1=0)

The **adjusted R-square** for the model is 0.4069(40.69%), which mean that 40.69% of the variability in CRIM is explained by our regression model.

In my opinion, the model is not performing well enough because we have a low p-value for the F test but the p-value for T test of the variables such as INDUS, CHAS, NOX, RM, AGE, TAX, PTRATIO, LSTAT are greater than 0.05 or 5%.

**c)**

**Backward Elimination**

**Graphical user interface, text

Description automatically generated**

**Output:**

**Text

Description automatically generated**

**For F test,** p-value: < 2.2e-16.

So, we reject the null hypothesis i.e. (All the betas are equal to zero) and accept the alternative hypothesis that at least one beta is not equal to zero.

For the **ZN,** we have significant p-value; so, we reject the null hypothesis(B1=0) and accept the alternative hypothesis(B1=/=0). And we are going to use prediction of 0.03195.

For the **NOX**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 0.02399.

For the **DIS**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 0.00045.

For the **RAD**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 2e-16

For the **PTRATIO**, we do not have significant p-value; so, we fail to reject the null hypothesis (B1=0)

For the **MEDV**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 1.11e-06.

The adjusted R-square for the model is 0.4109(41.09%), which mean that 41.09% of the variability in CRIM is explained by our regression model.

**Forward Elimination**

**A picture containing text

Description automatically generated**

**Output**

**Text

Description automatically generated**

**For F test,** p-value: < 2.2e-16.

So, we reject the null hypothesis i.e. (All the betas are equal to zero) and accept the alternative hypothesis that at least one beta is not equal to zero.

For the **RAD**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 2e-16

For the **MEDV**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 1.11e-06.

For the **PTRATIO**, we do not have significant p-value; so, we fail to reject the null hypothesis (B1=0)

For the **DIS**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 0.00045.

For the **NOX**, we have significant p-value; so, we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0). And we are going to use prediction of 0.02399.

For the **ZN,** we have significant p-value; so, we reject the null hypothesis(B1=0) and accept the alternative hypothesis(B1=/=0). And we are going to use prediction of 0.03195.

The adjusted R-square for the model is 0.4109(41.09%), which mean that 41.09% of the variability in CRIM is explained by our regression model.

**d)**

Backward Elimination output

**Text

Description automatically generated**

Forward Selection output

**Text

Description automatically generated**

On comparing the variables for both the model, it is clearly seen that both the model has same variables. (RAD, MEDV, PTRATIO, DIS, NOX, ZN), but their arrangement is not same.

I think this is because backward elimination starts with the **full model** and remove variables one at a time and record r-square, **whereas** forward selection starts with **null** **model** and fit all one-variable model in turn and pick up the model with best r square

For both the model, the P value of F Test is under significant value which mean that we reject the null hypothesis i.e. (All the betas are equal to zero) and accept the alternative hypothesis that at least one beta is not equal to zero.

And the p value of T test is also under significant values which mean that we reject the null hypothesis (B1=0) and accept the alternative hypothesis (B1=/=0).

No, variables that were not deemed statistically significant by the model in part a were retained, because all variables with values greater than 0.05 or 5% are removed in both forward selection and back elimination.

**e)**

When comparing the performance of the three models, I believe that forward selection and backward elimination have done a better job in predicting the crime rate because both models have the same variable and have significant p values for F test and T test, but the variables are not in the same arrangement, but this has no effect on the model's performance.

In addition, the adjusted r -square is 41.09 percent.

Whereas in linear regression model, we have a lot of non-significant variables, moreover the adjusted r-square is 40.69%.

But based on adjusted r-square we can’t say that backward elimination and forward elimination is a good model because it’s just 1% increase in adjusted-r square.

We are saying it because forward and backward elimination has fewer variables as compared to linear regression model and model with fewer model are easy to interpret and understand.

The benefits of having a simpler model justify accepting slightly worse performance than a more complex model is that with fewer model are easy to interpret and understand and can be easily explained to the person, who have limited or zero knowledge about the data analysis.

In, other term it is called as a parsimonious model