**Homework 5**

**SHAFAGH YAZDANI**

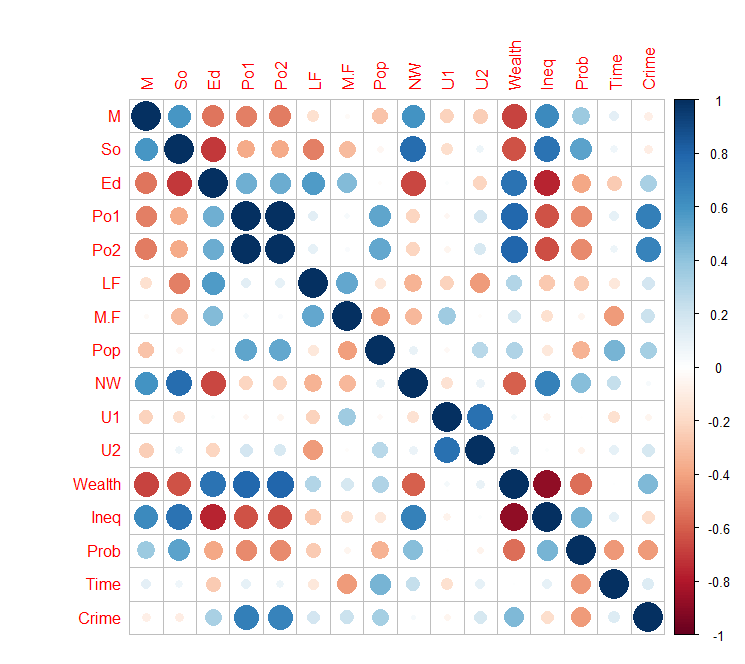
**Q 8.1**

**Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.**

One of the application of the linear regression model in the healthcare field is that we can use this model for analysis some of the human data. This is very beneficial in health centers and hospital to estimate how much is overall health of the patient. for example we can apply regression model in using multiple predictors like BMI ( body mass index) and waist- hip ratio , diet score(number of healthy item were consumed) , exercise score(total time exercise per week) to predict total cholesterol. This linear regression model can be used to predict the total cholesterol in patients. The total cholesterol are very crucial in patient overall health and also especially is critical in developing arteriosclerotic cardiovascular disease and coronary artery disease.

**Q 8.2**

We can observe our variables in the dataset and check how they are correlated with the response.



Summary of model with 15 predictor:

We know that null hypothesis for each predictor is that the predictor is not significant which means its coefficient =0

Our threshold is 0.05, Therefore the variables with the P value <0.05 are significant predictor.

Here are the colored variables.

**Model1**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.984e+03 1.628e+03 -3.675 0.000893 \*\*\*

M 8.783e+01 4.171e+01 2.106 0.043443 \*

So -3.803e+00 1.488e+02 -0.026 0.979765

Ed 1.883e+02 6.209e+01 3.033 0.004861 \*\*

Po1 1.928e+02 1.061e+02 1.817 0.078892 .

Po2 -1.094e+02 1.175e+02 -0.931 0.358830

LF -6.638e+02 1.470e+03 -0.452 0.654654

M.F 1.741e+01 2.035e+01 0.855 0.398995

Pop -7.330e-01 1.290e+00 -0.568 0.573845

NW 4.204e+00 6.481e+00 0.649 0.521279

U1 -5.827e+03 4.210e+03 -1.384 0.176238

U2 1.678e+02 8.234e+01 2.038 0.050161 .

Wealth 9.617e-02 1.037e-01 0.928 0.360754

Ineq 7.067e+01 2.272e+01 3.111 0.003983 \*\*

Prob -4.855e+03 2.272e+03 -2.137 0.040627 \*

Time -3.479e+00 7.165e+00 -0.486 0.630708

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.1 on 31 degrees of freedom

Multiple R-squared: 0.8031, Adjusted R-squared: 0.7078

F-statistic: 8.429 on 15 and 31 DF, p-value: 3.539e-07

Since our model is overfit, and we have variables with p values > 0.05 (not significant)

Therefore, we need to redo the regression model and rerun the model with the new chosen variables.

We can see that the M, Ed, Ineq, and Prob are the significant variable since they have P value <0.05.

It looks like 4-6 variables are important for our ﬁt: \* M \* Ed \* Po1 (maybe, p-value is between 0.05~0.10) \* U2 (maybe, p-value is between 0.05~0.10) \* Ineq \* Prob

**Model 2 with selected variables**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5040.5 899.8 -5.60 1.7e-06

M 105.0 33.3 3.15 0.0031

Ed 196.5 44.8 4.39 8.1e-05

Po1 115.0 13.8 8.36 2.6e-10

U2 89.4 40.9 2.18 0.0348

Ineq 67.7 13.9 4.85 1.9e-05

Prob -3801.8 1528.1 -2.49 0.0171

(Intercept) \*\*\*

M \*\*

Ed \*\*\*

Po1 \*\*\*

U2 \*

Ineq \*\*\*

Prob \*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 201 on 40 degrees of freedom

Multiple R-squared: 0.766, Adjusted R-squared: 0.731

F-statistic: 21.8 on 6 and 40 DF, p-value: 3.42e-11

Although the Multiple R-squared went down from 0.80 to 0.77, the Adjusted R-squared increased from 0.71 to 0.73! Furthermore, after reﬁtting the model, all the parameters have p-values of less than 0.05.

AIC value for Model 1 is : **650**

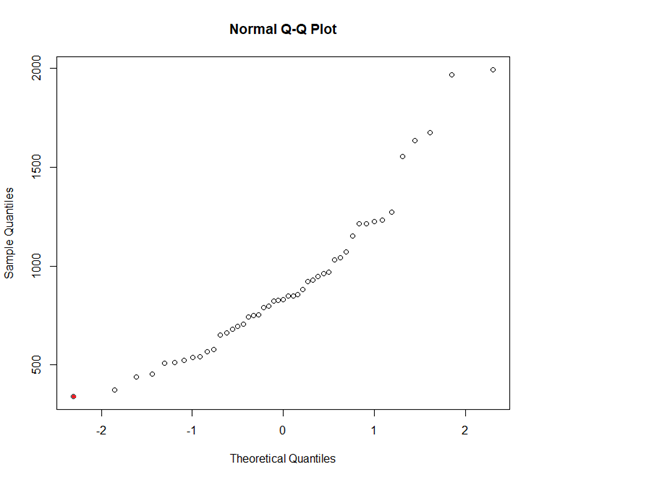
AIC value for Model 2 is : **640**

So our model got improved , therefore model 2 is better fitted model than model 1.

Our prediction with the model 1:

Predicted value for data point is **155**

This is the Q-Q plot for the prediction value in Crime data, We can observe that predicted value is the minimum value in left low corner of graph.



|  |
| --- |
| Prediction of test data point with model 2 is **1304** |
| |  | | --- | |  | |

We run the cross validation for our model 2 regression, and we got better R2 squared and adjusted R squared value :

Here is the summary of our **CV model 2 with 5 fold**:

|  |
| --- |
| Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -5040.5 899.8 -5.60 1.7e-06  M 105.0 33.3 3.15 0.0031  Ed 196.5 44.8 4.39 8.1e-05  Po1 115.0 13.8 8.36 2.6e-10  U2 89.4 40.9 2.18 0.0348  Ineq 67.7 13.9 4.85 1.9e-05  Prob -3801.8 1528.1 -2.49 0.0171    (Intercept) \*\*\*  M \*\*  Ed \*\*\*  Po1 \*\*\*  U2 \*  Ineq \*\*\*  Prob \*  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 201 on 40 degrees of freedom  Multiple R-squared: 0.766, Adjusted R-squared: 0.731  F-statistic: 21.8 on 6 and 40 DF, p-value: 3.42e-11  Here Our R-squared: 0.766, Adjusted R-squared: 0.731 is the same as model2. |
|  |
| Our predicted value for the test data point with CV-model2 is : |