ISYE 6501

Homework #5

8.1]

In the current market where the new housing developments have been stunned due to supply chain issues. We can measure the impact of overall raw material suppliers output numbers to the delivery dates of a newly building house. The raw materials used as parameter input can be as follows:

1) Wood
2) Labor
3) Drywall
4) Cement

8.2]

With the given dataset of crime rate in use for 47states , the first thing I did was to check for any correlation between the dataset parameters. The Cor function was used for that. Based on the plotted parameters in Fig[1], it look like the Po1 and Po2 are closely corelated along with U1 and U2 which are fairly correlated as well. It is worth noting how wealth and Ineq are completely unrelated to eachother.

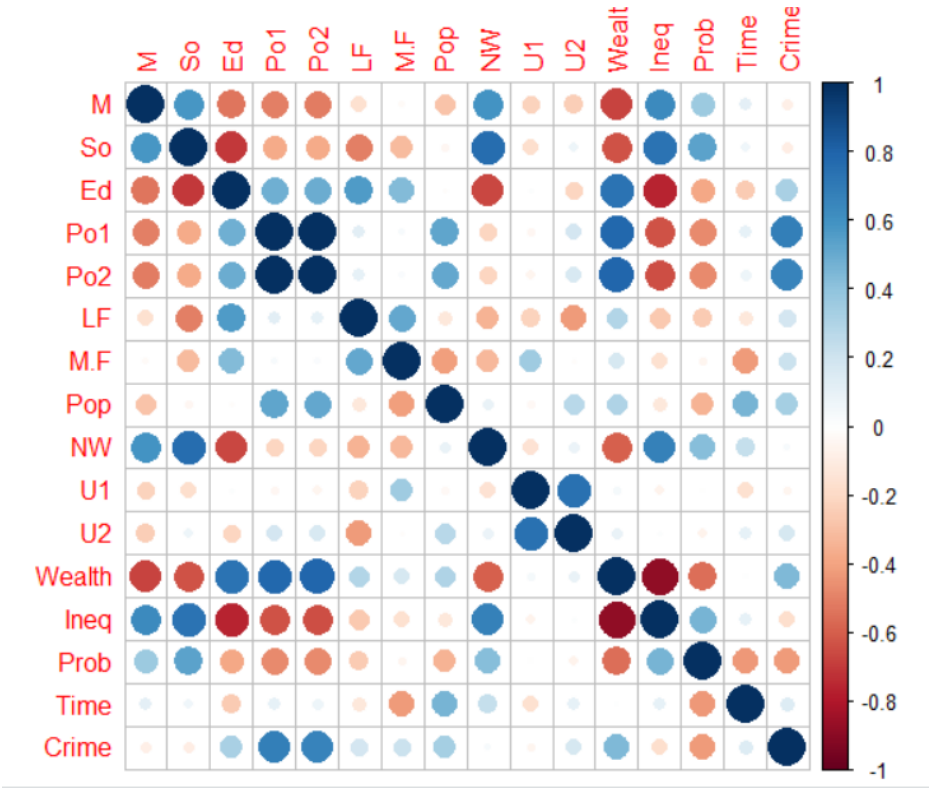


Fig [1]

To check multiple regression models, I decided to leave the parameters that are correlated as seen in Fig[1] in the first model fit. Using the lm() function , the first regression model was fitted on crime_table. The summary of the model was as shown in Fig[2]

```
> summary(model_1)

Call:
lm(formula = Crime ~ ., data = crime_table)

Residuals:
    Min      1Q  Median      3Q     Max
-395.74  -98.09   -6.69  112.99  512.67

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
```

Fig [2]

Based on the summary table, the following parameters seem the most relevant with a Pr score of less than or close to 0.05: M, Ed, Po1, U2, Ineq and Prob. To compare multiple models the output of model_1 on given set of parameters for the homework was tested. Fig[3] shows the predicted value of the crime rate based on the following parameter inputs :M = 14.00, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640, M.F = 94.0, Pop = 150, NW = 1.1, U1 = 0.120, U2 = 3.6, Wealth = 3200, Ineq = 20.1, Prob = 0.04, Time = 39.0 :

```
> model_1_fit
       1
155.4349
```

Fig[3]

I will compare the model fits towards the end of the analysis, currently the focus will be on creating a new model and noting the predicted value for the homework parameters.

As mentioned above from Fig[1], I decided to setup a new model by including one of the two correlated values between Po1 and Po2 / U1 and U2. Based on Fig[2] since Po1 and U2 are more relevant I decided to include those in the model_2. Fig[4] shows the summary of model_2 fit. It is interesting to notice that the parameters that are more relevant to the regression fit have changed to So, Pol and Prob. My understanding here is since the model is based on the parameter inputs, the dependency / relevant parameters in the model will change. As a result, I decided to go with the model that has better prediction, which I will assess towards the end.

```
> summary(model_2)

Call:
lm(formula = Crime ~ M + So + Ed + Po1 + LF + M.F + Pop + NW +
    U1 + Wealth + Prob + Time, data = crime_table)

Residuals:
    Min      1Q  Median      3Q     Max
-422.51 -154.42    2.12  130.52  550.16

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.319e+03  1.757e+03  -2.458 0.019221 *
M            6.416e+01  4.853e+01   1.322 0.195005
So           2.812e+02  1.560e+02   1.803 0.080289 .
Ed           8.261e+01  6.593e+01   1.253 0.218736
Po1          1.066e+02  2.726e+01   3.911 0.000418 ***
LF           1.007e+03  1.626e+03   0.619 0.539717
M.F          2.422e+01  2.330e+01   1.040 0.305782
Pop          1.965e-01  1.481e+00   0.133 0.895277
NW           3.120e+00  7.344e+00   0.425 0.673639
U1           1.531e+03  3.120e+03   0.491 0.626741
Wealth      -8.586e-02  9.710e-02  -0.884 0.382778
Prob        -5.267e+03  2.576e+03  -2.045 0.048666 *
Time         1.081e+00  7.898e+00   0.137 0.891959
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 248.1 on 34 degrees of freedom
Multiple R-squared:  0.6959,    Adjusted R-squared:  0.5886
F-statistic: 6.484 on 12 and 34 DF,  p-value: 7.922e-06
```

Fig[4]

Model_2 was then used to predict value for the same parameters as on model_1. The predicted value is as shown in Fig[5]

```
> model_2_fit
       1
1379.001
```

Fig[5]

A third model was also created using just the relevant parameters as observed from Fig[2] which are : M,Ed, Po1, U2, Ineq and Prob. Fig[6] shows the summary of model_3, which shows all parameters are relevant in the regression model based on the Pr value.

```
> summary(model_3)

Call:
lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_table)

Residuals:
    Min      1Q  Median      3Q     Max
-470.68  -78.41  -19.68  133.12  556.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
M             105.02      33.30   3.154  0.00305 **
Ed            196.47      44.75   4.390 8.07e-05 ***
Po1           115.02      13.75   8.363 2.56e-10 ***
U2             89.37      40.91   2.185  0.03483 *
Ineq           67.65      13.94   4.855 1.88e-05 ***
Prob        -3801.84    1528.10  -2.488  0.01711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 200.7 on 40 degrees of freedom
Multiple R-squared:  0.7659,    Adjusted R-squared:  0.7307
F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

Fig[6]

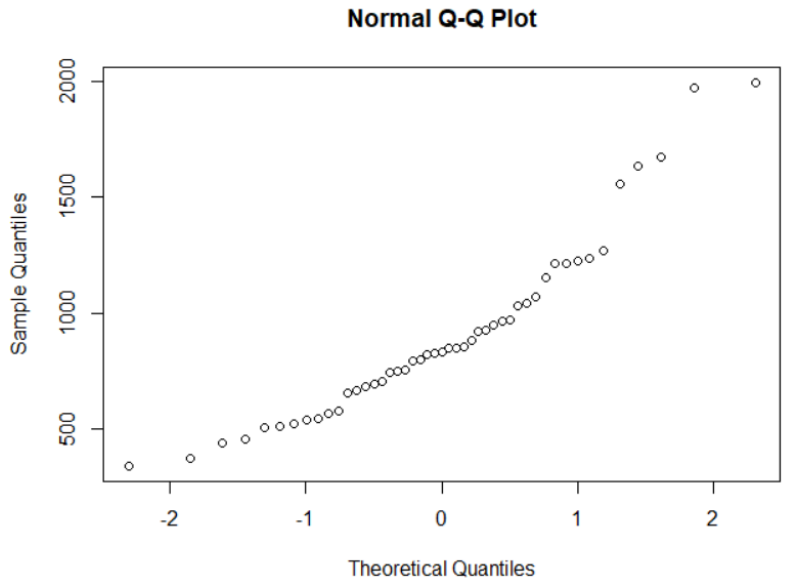
This model was then used to predict the value of test parameters as before. Fig[7] shows the predicted value for model_3

```
> model_3_fit
       1
1304.245
```

Fig[7]

I decided to do multiple checks to see which model fits the best for predicting the crime rate for the test parameter set.

1) First check to select the best fit model was to see where the predicted values fit with respect to the current crime rate values in our primary dataset. Fig[8] shows the plot:



Fig[8]

From a usual check of values Model_fit_1 = 155 seems too low given the input parameters in the test data are closely similar to our primary dataset. Model_fit_2 = 1379 and Model_fit_3=1304 values are close to the densely populated area of our primary dataset which is more likely based on the input parameters.

2) Adjusted R squared Values:
   Model_1_fit = 0.7078
   Model_2_fit= 0.5886
   Model_3_fit= 0.7307

   Based on the R squared adjusted values it seems like Model_3 with the least amount of parameters is the best regression model

   The equation of the regression model is as follows:

   Crime=-5040.50+105.02M+196.47Ed+115.02Po1+89.37U2+67.65Ineq-3801.84Prob

R Code (for Reference):

```
rm(list=ls())
set.seed(100)


#Read data to R
crime_table=read.table("M:/OMSA/ISYE6501/HW5/crime.txt",header=TRUE)
head(crime_table)

#Check data points for coorelation
library(corrplot)
corrplot(cor(crime_table))

#First regression model fit
model_1=lm(Crime~.,data=crime_table)
summary(model_1)


#input required hw data set to test
test_data=data.frame(M = 14.00,So = 0,Ed = 10.0,Po1 = 12.0,Po2 = 15.5,LF = 0.640,M.F = 94.0,Pop = 1
                     NW = 1.1,U1 = 0.120,U2 = 3.6,Wealth = 3200,Ineq = 20.1,Prob = 0.04,Time = 39.0

#Test test data on regression model_1
model_1_fit=predict(model_1,test_data)
model_1_fit


#Regression model_2
model_2=lm(Crime~M+So+Ed+Po1+LF+M.F+Pop+NW+U1+Wealth+Prob+Time,data=crime_table)
summary(model_2)

#Test test data on regression model_2
model_2_fit=predict(model_2,test_data)
model_2_fit

#Regression model_3
model_3=lm(Crime~M+Ed+Po1+U2+Ineq+Prob,data=crime_table)
summary(model_3)

#Test test data on regression model_3
model_3_fit=predict(model_3,test_data)
model_3_fit

#Crime data plot
qqnorm(crime_table$Crime)
```