ISYE 6501 HW5

February 2021

1 Question 8.1

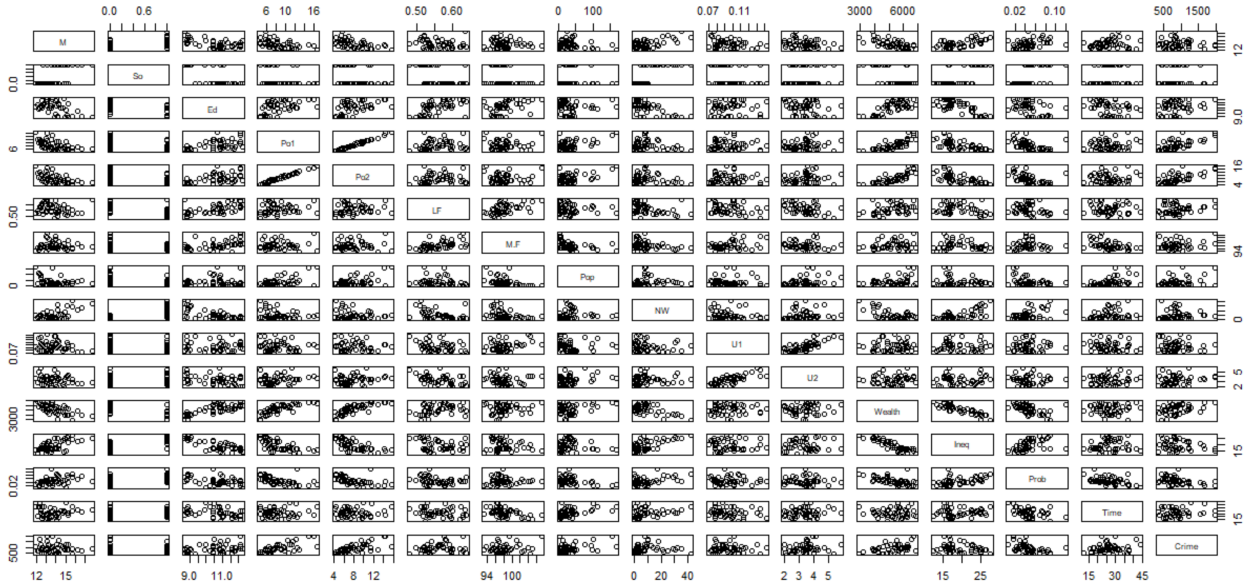
The Total sales of a grocery store can be predicted using regression. Predictors can be which day in a month, which day in a week, weather forecast of the day, a dummy variable for holidays, location of the store, even crime rate in the surrounding area of the store.

2 Question 8.2

First, let's review the data. Looks like "So" is a dummy variable, and "Wealth" contains much higher values which needs scaling (I will use logarithm transformation on this predictor).

M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
15.1	1	9.1	5.8	5.6	0.510	95.0	33	30.1	0.108	4.1	3940	26.1	0.084602	26.2011	791
14.3	0	11.3	10.3	9.5	0.583	101.2	13	10.2	0.096	3.6	5570	19.4	0.029599	25.2999	1635
14.2	1	8.9	4.5	4.4	0.533	96.9	18	21.9	0.094	3.3	3180	25.0	0.083401	24.3006	578
13.6	0	12.1	14.9	14.1	0.577	99.4	157	8.0	0.102	3.9	6730	16.7	0.015801	29.9012	1969
14.1	0	12.1	10.9	10.1	0.591	98.5	18	3.0	0.091	2.0	5780	17.4	0.041399	21.2998	1234

Second, review the correlation between predictors. It seems that "Po1" & "Po2" are highly correlated. "U1" & "U2" are correlated. "Wealth" are "Ineq" are each correlated with multiple predictors. But the correlation between predictors are not effect how we predict a value of crime rate, so we don't need to worry about it in this case.



Then to the model selection step. I have fitted all 15 predictors to the model with "Wealth" log transformed and got the results listed below. We can see that there are a lot of coefficients that have p-values much larger than the threshold 0.05. This means they might highly possible do not have impact on the model thus can be removed. The model has a fairly high R squared value of 0.8037 and adjusted R squared value of 0.7087 which suggests the model is not a bad fit.

```
Call:
lm(formula = Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop +
    NW + U1 + U2 + log(Wealth) + Ineq + Prob + Time, data = df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9508.3318  4242.1298  -2.241  0.03230 *
M              84.7257    41.2458   2.054  0.04847 *
So            -10.6676   149.8029  -0.071  0.94369
Ed            187.5393    62.0378   3.023  0.00499 **
Po1           188.5560   106.1743   1.776  0.08557 .
Po2          -103.9264   117.1066  -0.887  0.38167
LF           -694.5372  1470.3506  -0.472  0.63998
M.F            18.6132    20.2189   0.921  0.36438
Pop            -0.7396     1.2876  -0.574  0.56983
NW              5.0435     6.6740   0.756  0.45554
U1          -5901.9619  4196.7085  -1.406  0.16957
U2            168.0669    82.0589   2.048  0.04910 *
log(Wealth)   463.6256   472.7260   0.981  0.33431
Ineq           69.9587    21.8556   3.201  0.00316 **
Prob        -4688.6914  2290.0989  -2.047  0.04918 *
Time           -2.9428     7.1107  -0.414  0.68183

Multiple R-squared:  0.8037,        Adjusted R-squared:  0.7087
F-statistic: 8.462 on 15 and 31 DF,  p-value: 3.384e-07
```

Now we want to select a model that is the "best" fit, which means it's simple enough and not over fitting. We can use the stepAIC function in MASS library to complete this process. Let direction = "both" to use both forward and backward selection strategy, the result is the following model, which keeps predictor "M", "Ed", "Po1", "M.F", "U1", "U2", "Ineq", & "Prob". Model R-squared value is 0.7888, Adjusted R-squared value is 0.7444. Suggested this is a good fit as well. The higher Adjusted R-squared value indicates the AIC selected model is better than the original all 15 predictor model.

```
Call:
lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-444.70 -111.07    3.03  122.15  483.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6426.10    1194.61  -5.379 4.04e-06 ***
M              93.32      33.50   2.786  0.00828 **
Ed            180.12      52.75   3.414  0.00153 **
Po1           102.65      15.52   6.613 8.26e-08 ***
```

```
M.F            22.34      13.60   1.642  0.10874
U1          -6086.63    3339.27  -1.823  0.07622 .
U2            187.35      72.48   2.585  0.01371 *
Ineq           61.33      13.96   4.394 8.63e-05 ***
Prob        -3796.03    1490.65  -2.547  0.01505 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 195.5 on 38 degrees of freedom
Multiple R-squared:  0.7888,        Adjusted R-squared:  0.7444
F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10
```

Using the selected model to predict the Crime rate of the given data point, the predicted crime rate is 1038.413.

M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time
14	0	10	12	15.5	0.64	94	150	1.1	0.12	3.6	8.070906	20.1	0.04	39

2.1 My R code is:

```
library(MASS)
df <- read.table("uscrime.txt", sep = '\t', stringsAsFactors = FALSE, header = TRUE)
str(df["Crime"])
head(df)
summary(df)
plot(df)

df[,12] <- log(df[,12])

#fit
fit <- lm(Crime~., data=df)
summary(fit)
# Stepwise regression model
AIC_fit <- stepAIC(fit, direction = "both", trace = FALSE)
summary(AIC_fit)

# predict
newdf <- data.frame (M = 14.0,
                     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 = log(3200),
                     Ineq = 20.1,
                     Prob = 0.04,
                     Time = 39.0)
newdf
crime_hat <- predict(AIC_fit, newdata=newdf)
crime_hat
```