ISyE 6501-HOMEWORK 5

Group Members:
Jinyu, Huang | jhuang472@gatech.edu | GTID: 903522245
Chengqi, Huang | chengqihuang@gatech.edu | GTID: 903534690
Yuefan, Hu | yuefanhu@gatech.edu | GTID:903543027
Jingyu, Li | alanli@gatech.edu | GTID: 903520148

Qusetion 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.

Answer

An instructor of a course would like to predict the score that an average student will recieve for mid-term examination based on the data from previous students who regiesterd this course. A linear regression model will provide some help. The predictors that the instructor is suggested to use will include but not be limited to:

1. students' overall GPA: measures the students' overall academic performance (positive correlation)
2. the number of courses a student takes: measures to what extent students could devote to this exam (if they took too many courses, may not have enough time to prepare for this one,negative correlation)
3. students' major: to see whether it is related to this exam (if related, probably get higher score);related=1, non-related=0
4. the level of the exam: the ratio of easy questions to hard ones , (positive correlation, more easy questions indicates higher score)

Qusetion 8.2

Using crime data, use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data

Answer

To explore the data preliminary, we run a pairwise correlation between all variables. Firstly, we find the corelations between predictors and *crime* are various. Secondly, *Po1* and *Po2* is highly correlated($r = 0.99$). We check the description of the data, and it shows *Po1* means per capita expenditure on police protection in 1960, while *Po2* means per capita expenditure on police protection in 1959. So there's no doubt these to predictors are correlated. To avoid multicollinearity, we remove *Po2*.

```
> data = read.table("uscrime.txt", header=TRUE)  # import data
> round(cor(data),2)
```

```
        M    So    Ed   Po1   Po2    LF   M.F   Pop    NW    U1    U2
M    1.00  0.58 -0.53 -0.51 -0.51 -0.16 -0.03 -0.28  0.59 -0.22 -0.24
So   0.58  1.00 -0.70 -0.37 -0.38 -0.51 -0.31 -0.05  0.77 -0.17  0.07
Ed  -0.53 -0.70  1.00  0.48  0.50  0.56  0.44 -0.02 -0.66  0.02 -0.22
Po1 -0.51 -0.37  0.48  1.00  0.99  0.12  0.03  0.53 -0.21 -0.04  0.19
Po2 -0.51 -0.38  0.50  0.99  1.00  0.11  0.02  0.51 -0.22 -0.05  0.17
LF  -0.16 -0.51  0.56  0.12  0.11  1.00  0.51 -0.12 -0.34 -0.23 -0.42
M.F -0.03 -0.31  0.44  0.03  0.02  0.51  1.00 -0.41 -0.33  0.35 -0.02
Pop -0.28 -0.05 -0.02  0.53  0.51 -0.12 -0.41  1.00  0.10 -0.04  0.27
NW   0.59  0.77 -0.66 -0.21 -0.22 -0.34 -0.33  0.10  1.00 -0.16  0.08
U1  -0.22 -0.17  0.02 -0.04 -0.05 -0.23  0.35 -0.04 -0.16  1.00  0.75
```

```
U2     -0.24  0.07 -0.22  0.19  0.17 -0.42 -0.02  0.27  0.08  0.75  1.00
Wealth -0.67 -0.64  0.74  0.79  0.79  0.29  0.18  0.31 -0.59  0.04  0.09
Ineq    0.64  0.74 -0.77 -0.63 -0.65 -0.27 -0.17 -0.13  0.68 -0.06  0.02
Prob    0.36  0.53 -0.39 -0.47 -0.47 -0.25 -0.05 -0.35  0.43 -0.01 -0.06
Time    0.11  0.07 -0.25  0.10  0.08 -0.12 -0.43  0.46  0.23 -0.17  0.10
Crime  -0.09 -0.09  0.32  0.69  0.67  0.19  0.21  0.34  0.03 -0.05  0.18
       Wealth  Ineq  Prob  Time Crime
M       -0.67  0.64  0.36  0.11 -0.09
So      -0.64  0.74  0.53  0.07 -0.09
Ed       0.74 -0.77 -0.39 -0.25  0.32
Po1      0.79 -0.63 -0.47  0.10  0.69
Po2      0.79 -0.65 -0.47  0.08  0.67
LF       0.29 -0.27 -0.25 -0.12  0.19
M.F      0.18 -0.17 -0.05 -0.43  0.21
Pop      0.31 -0.13 -0.35  0.46  0.34
NW      -0.59  0.68  0.43  0.23  0.03
U1       0.04 -0.06 -0.01 -0.17 -0.05
U2       0.09  0.02 -0.06  0.10  0.18
Wealth   1.00 -0.88 -0.56  0.00  0.44
Ineq    -0.88  1.00  0.47  0.10 -0.18
Prob    -0.56  0.47  1.00 -0.44 -0.43
Time     0.00  0.10 -0.44  1.00  0.15
Crime    0.44 -0.18 -0.43  0.15  1.00
```

```
> crime = data[,-5]
```

Considering that we don't have any theoretical knowledge to decide which predictors we should choose, our first step is selecting all the predictors and using "simultaneous" enter method. Because the sample size is relatively small, we get a good fitting model with adjusted $R^2 = 0.709$ as expected. Given all other predictors in the model, *M, Ed, Po1, U2, Ineq* are statistically significantly associated to *Crime*; and *Prob* are marginally significant. It's hard to do validation to select among different models in such a small data set. As our goal is to make prediction based on given datas on the independent value, we choose to record the residual standard error and AICc of different models for further comparison.

```
> library(MuMIn)
> model_1 = lm(Crime~., data=crime)
> summary(model_1)
```

```

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

Residuals:
    Min      1Q  Median      3Q     Max
-442.55 -116.46    8.86  118.26  473.49

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.379e+03  1.569e+03  -4.066 0.000291 ***
M            8.986e+01  4.157e+01   2.162 0.038232 *
So           5.669e+00  1.481e+02   0.038 0.969705
Ed           1.773e+02  6.082e+01   2.915 0.006445 **
Po1          9.653e+01  2.392e+01   4.035 0.000317 ***
LF          -2.801e+02  1.408e+03  -0.199 0.843538
M.F          1.822e+01  2.029e+01   0.898 0.376026
```

```
Pop         -7.836e-01  1.286e+00  -0.609 0.546523
NW           2.446e+00  6.187e+00   0.395 0.695239
U1          -5.416e+03  4.178e+03  -1.296 0.204164
U2           1.694e+02  8.215e+01   2.062 0.047441 *
Wealth       9.072e-02  1.033e-01   0.878 0.386292
Ineq         7.271e+01  2.256e+01   3.222 0.002921 **
Prob        -4.285e+03  2.184e+03  -1.962 0.058484 .
Time        -1.128e+00  6.692e+00  -0.168 0.867251
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 208.6 on 32 degrees of freedom
Multiple R-squared:  0.7976,    Adjusted R-squared:  0.709
F-statistic: 9.006 on 14 and 32 DF,  p-value: 1.673e-07
```

```
> quality = matrix(nrow=4, ncol=3)  # store the quality matrix
> colnames(quality) = c("model", "residual standard error", "AICc")
> quality[1,1] = "m1: Crime~."
> quality[1,2] = round(summary(model_1)$sigma,3)
> quality[1,3] = round(AICc(model_1),3)
```

The second model we try only includes the significant predictors in model_1, which are *M, Ed, Po1, U2, Ineq, Prob*. The summary table shows all the predictors in this model are significant and the adjusted R^2 of the overall model equals to 0.731.

```
> model_2 = lm(Crime~M+Ed+Po1+U2+Ineq+Prob, data=crime)
> summary(model_2)
```

```

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

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
```

```
> quality[2,1] = "m2: Crime~M+Ed+Po1+U2+Ineq+Prob"
> quality[2,2] = round(summary(model_2)$sigma,3)
> quality[2,3] = round(AICc(model_2),3)
```

Furthermore, we try the stepwise method, which iteratively adds and removes predictors from the model to find a subset of variables resulting in the lowest predicting error. The general function of stepwise method for regression is in library "MASS" and called *stepAIC*. Because our sample size is small and AICc would be a better indicator, we used a modified version called *stepAICc*(https://stat.ethz.ch/pipermail/r-help/2009-April/389888.html). We use the full model (including all predictors) as initial model and choose stepwise method. Results show some top models according to AICc values.

```
> library(MASS)
> full.model = lm(Crime~., data=crime)
> stepAICc(full.model, direction = "both", steps=2000)
```

```
Start:  AIC=667.46
Crime ~ M + So + Ed + Po1 + LF + M.F + Pop + NW + U1 + U2 + Wealth +
    Ineq + Prob + Time

         Df Sum of Sq     RSS    AIC
- So      1        64 1392929 511.95
- Time    1      1236 1394101 511.99
- LF      1      1723 1394588 512.00
- NW      1      6802 1399667 512.18
- Pop     1     16168 1409033 512.49
- Wealth  1     33582 1426447 513.07
- M.F     1     35080 1427945 513.12
<none>                1392865 513.95
- U1      1     73136 1466001 514.35
- Prob    1    167590 1560455 517.29
- U2      1    185009 1577874 517.81
- M       1    203389 1596254 518.35
- Ed      1    369864 1762729 523.01
- Ineq    1    451937 1844802 525.15
- Po1     1    708738 2101603 531.28

Step:  AIC=662.81
Crime ~ M + Ed + Po1 + LF + M.F + Pop + NW + U1 + U2 + Wealth +
    Ineq + Prob + Time

         Df Sum of Sq     RSS    AIC
- Time    1      1319 1394247 509.99
- LF      1      2646 1395574 510.04
- NW      1      8949 1401878 510.25
- Pop     1     16166 1409095 510.49
- Wealth  1     36125 1429054 511.15
- M.F     1     36467 1429396 511.16
<none>                1392929 511.95
- U1      1     86999 1479928 512.80
+ So      1        64 1392865 513.95
- Prob    1    171381 1564310 515.40
- U2      1    196372 1589301 516.15
- M       1    206121 1599050 516.43
- Ed      1    371159 1764088 521.05
- Ineq    1    534611 1927540 525.22
- Po1     1    728570 2121499 529.72

Step:  AIC=658.5
Crime ~ M + Ed + Po1 + LF + M.F + Pop + NW + U1 + U2 + Wealth +
```

```
    Ineq + Prob

         Df Sum of Sq     RSS    AIC
- LF      1      3019 1397266 508.09
- NW      1      7996 1402243 508.26
- Pop     1     19634 1413881 508.65
- Wealth  1     35276 1429524 509.17
- M.F     1     40680 1434928 509.34
<none>                1394247 509.99
- U1      1     85946 1480194 510.80
+ Time    1      1319 1392929 511.95
+ So      1       147 1394101 511.99
- U2      1    195095 1589343 514.15
- M       1    206909 1601157 514.50
- Prob    1    223309 1617557 514.98
- Ed      1    381593 1775840 519.36
- Ineq    1    537046 1931294 523.31
- Po1     1    764536 2158784 528.54

Step:  AIC=654.5
Crime ~ M + Ed + Po1 + M.F + Pop + NW + U1 + U2 + Wealth + Ineq +
    Prob

         Df Sum of Sq     RSS    AIC
- NW      1      6963 1404229 506.33
- Pop     1     23381 1420648 506.87
- Wealth  1     34787 1432053 507.25
- M.F     1     41289 1438555 507.46
<none>                1397266 508.09
- U1      1     84385 1481652 508.85
+ LF      1      3019 1394247 509.99
+ Time    1      1692 1395574 510.04
+ So      1      1369 1395898 510.05
- U2      1    197849 1595115 512.32
- Prob    1    221812 1619078 513.02
- M       1    226201 1623468 513.15
- Ed      1    395884 1793150 517.82
- Ineq    1    534370 1931637 521.32
- Po1     1    834362 2231628 528.10

Step:  AIC=650.88
Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Wealth + Ineq +
    Prob

         Df Sum of Sq     RSS    AIC
- Pop     1     22345 1426575 505.07
- Wealth  1     32142 1436371 505.39
- M.F     1     36808 1441037 505.54
<none>                1404229 506.33
- U1      1     86373 1490602 507.13
+ NW      1      6963 1397266 508.09
+ So      1      3807 1400422 508.20
+ LF      1      1986 1402243 508.26
+ Time    1       575 1403654 508.31
```

```
- U2      1    205814 1610043 510.76
- Prob    1    218607 1622836 511.13
- M       1    307001 1711230 513.62
- Ed      1    389502 1793731 515.83
- Ineq    1    608627 2012856 521.25
- Po1     1   1050202 2454432 530.57

Step:  AIC=647.99
Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob

         Df Sum of Sq     RSS    AIC
- Wealth  1     26493 1453068 503.93
<none>                1426575 505.07
- M.F     1     84491 1511065 505.77
- U1      1     99463 1526037 506.24
+ Pop     1     22345 1404229 506.33
+ NW      1      5927 1420648 506.87
+ So      1      5724 1420851 506.88
+ LF      1      5176 1421398 506.90
+ Time    1      3913 1422661 506.94
- Prob    1    198571 1625145 509.20
- U2      1    208880 1635455 509.49
- M       1    320926 1747501 512.61
- Ed      1    386773 1813348 514.35
- Ineq    1    594779 2021354 519.45
- Po1     1   1127277 2553852 530.44

Step:  AIC=645.43
Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob

         Df Sum of Sq     RSS    AIC
<none>                1453068 503.93
+ Wealth  1     26493 1426575 505.07
- M.F     1    103159 1556227 505.16
+ Pop     1     16697 1436371 505.39
+ So      1      9329 1443739 505.63
+ LF      1      4374 1448694 505.79
+ NW      1      3799 1449269 505.81
+ Time    1      2293 1450775 505.86
- U1      1    127044 1580112 505.87
- Prob    1    247978 1701046 509.34
- U2      1    255443 1708511 509.55
- M       1    296790 1749858 510.67
- Ed      1    445788 1898855 514.51
- Ineq    1    738244 2191312 521.24
- Po1     1   1672038 3125105 537.93


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

Coefficients:
(Intercept)            M           Ed          Po1          M.F
```

```
  -6426.10        93.32       180.12       102.65        22.34
        U1           U2         Ineq         Prob
  -6086.63       187.35        61.33     -3796.03
```

So we choose the top 2 models from stepwise method and add them into comparison. Based on residual standard error and AICc, *model_2* and *model_3* are two better ones. Because AICc contains penalty term to balance likelihood with simplicity, we can see the more simple model(*model_2*) has a lower AICc but higher residual standard error. We propose that lower residual standard error is more important in this question for doing prediction and the complexity of *model_3* and *model_2* differs not much. So we finally choose *model_3*.

```
> model_3 = lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob, data=crime)
> quality[3,1] = "m3: Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob"
> quality[3,2] = round(summary(model_3)$sigma,3)
> quality[3,3] = round(AICc(model_3),3)
>
> model_4 = lm(Crime~M+Ed+Po1+M.F+U1+U2+Wealth+Ineq+Prob, data=crime)
> quality[4,1] = "m4: Crime~M+Ed+Po1+M.F+U1+U2+Wealth+Ineq+Prob"
> quality[4,2] = round(summary(model_4)$sigma,3)
> quality[4,3] = round(AICc(model_4),3)
>
>
> quality
```

```
     model
[1,] "m1: Crime~."
[2,] "m2: Crime~M+Ed+Po1+U2+Ineq+Prob"
[3,] "m3: Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob"
[4,] "m4: Crime~M+Ed+Po1+M.F+U1+U2+Wealth+Ineq+Prob"
     residual standard error AICc
[1,] "208.631"               "667.46"
[2,] "200.69"                "643.956"
[3,] "195.547"               "645.426"
[4,] "196.357"               "647.993"
```

As for model_3, the adjusted R^2 is 0.744 and overall F-value is 17.74 with $p \approx 0$. The regression model is: $Crime = -6426.10 + 93.32M + 180.12Ed + 102.65PO1 + 22.34M.F - 6086.63U1 + 187.35U2 + 61.33Ineq + -3796.03Prob$

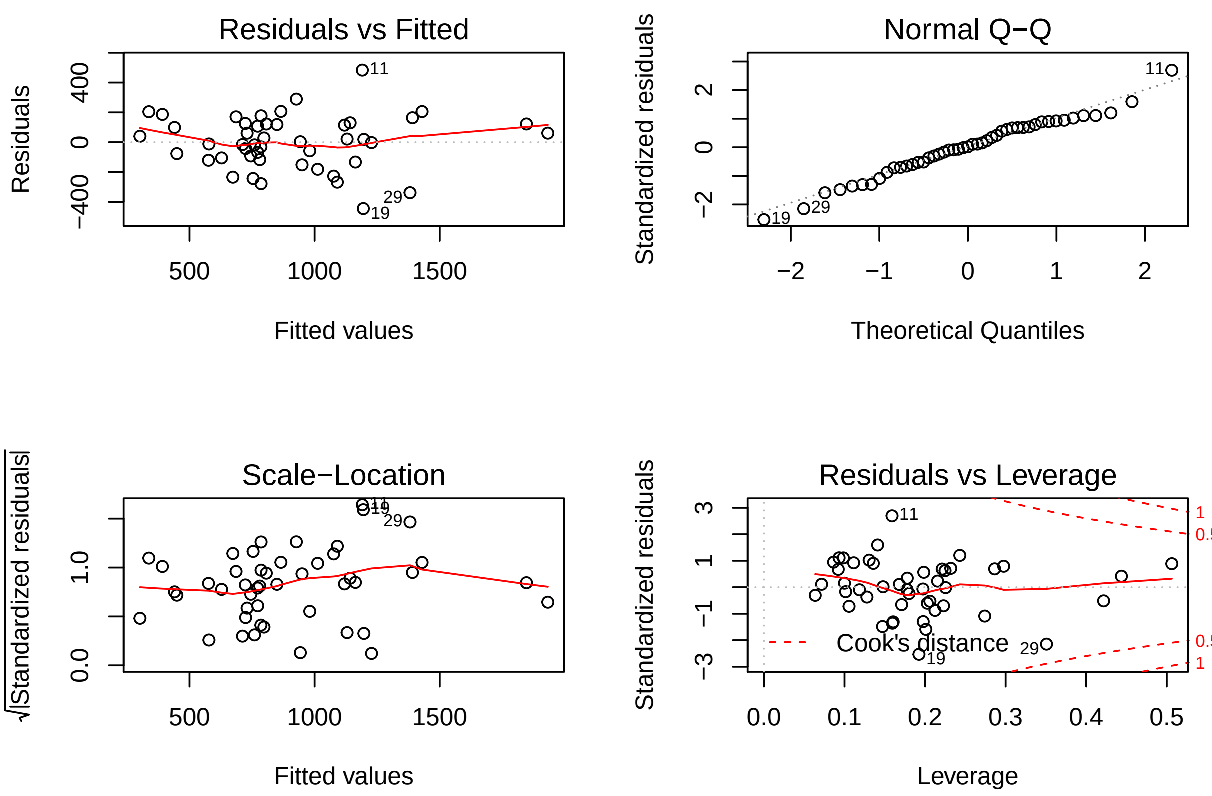
```
> summary(model_3)
```

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

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
```

Based on the residual analysis, with a small sample size, we propose that the assumptions for linear regression generally hold (*Lineartiy, Constant Variance, Independence, Normal Distribution*).

```
> par(mfrow=c(2,2))
> plot(model_3)
```



Then we use model_3 to do prediction based on the given data. We apply the *predict* function and use 0.99 confidence level. With the new given data, the predicted crime is 1038 and the 99% confidence interval is $(376.41, 1700.41)$.

```
> target_data = data.frame(M=14,
+                          Ed=10,
+                          Po1=12,
+                          M.F=94,
```

```
+                             U1=0.12,
+                             U2=3.6,
+                             Ineq=20.1,
+                             Prob=0.04)
> predict(model_3, target_data, interval="predict", level=0.99)
```

```
       fit      lwr      upr
1 1038.413 376.4115 1700.415
```