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.

In Organizational Psychology, linear regression models are widely used to research on how employeesÕ psychological states influence their behavior in company. For example, to explore what factors will influence employeesÕ KPI performance, we can measure their job satisfaction, organization commitment, fairness perception as predictors, and then use the KPI performance in a later time point as dependent varible.

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

Considering that we donÕt have any theoretical knowledge to decide which predictors we should choose, our first step is selecting all the predictors and using ÒsimultaneousÓ 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, Ineq, Prob* are statistically significantly associated to *Crime*; and *Po1, U2* are marginally significant. ItÕs 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)
> crime = read.table("uscrime.txt", header=TRUE)   # import data
> model_1 = lm(Crime~., data=crime)
> summary(model_1)
```

```

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

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
```

```
> 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 ÒMASSÓ 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=671.13
Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
    U2 + Wealth + Ineq + Prob + Time

         Df Sum of Sq     RSS    AIC
- So      1        29 1354974 512.65
- LF      1      8917 1363862 512.96
- Time    1     10304 1365250 513.00
- Pop     1     14122 1369068 513.14
- NW      1     18395 1373341 513.28
- M.F     1     31967 1386913 513.74
- Wealth  1     37613 1392558 513.94
- Po2     1     37919 1392865 513.95
<none>                1354946 514.65
- U1      1     83722 1438668 515.47
- Po1     1    144306 1499252 517.41
- U2      1    181536 1536482 518.56
- M       1    193770 1548716 518.93
- Prob    1    199538 1554484 519.11
- Ed      1    402117 1757063 524.86
- Ineq    1    423031 1777977 525.42

Step:  AIC=666.16
Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
    Wealth + Ineq + Prob + Time

         Df Sum of Sq     RSS    AIC
- Time    1     10341 1365315 511.01
- LF      1     10878 1365852 511.03
- Pop     1     14127 1369101 511.14
- NW      1     21626 1376600 511.39
- M.F     1     32449 1387423 511.76
- Po2     1     37954 1392929 511.95
```

```
- Wealth  1     39223 1394197 511.99
<none>                1354974 512.65
- U1      1     96420 1451395 513.88
+ So      1        29 1354946 514.65
- Po1     1    144302 1499277 515.41
- U2      1    189859 1544834 516.81
- M       1    195084 1550059 516.97
- Prob    1    204463 1559437 517.26
- Ed      1    403140 1758114 522.89
- Ineq    1    488834 1843808 525.13

Step:  AIC=661.87
Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
    Wealth + Ineq + Prob

         Df Sum of Sq     RSS    AIC
- LF      1     10533 1375848 509.37
- NW      1     15482 1380797 509.54
- Pop     1     21846 1387161 509.75
- Po2     1     28932 1394247 509.99
- Wealth  1     36070 1401385 510.23
- M.F     1     41784 1407099 510.42
<none>                1365315 511.01
- U1      1     91420 1456735 512.05
+ Time    1     10341 1354974 512.65
+ So      1        65 1365250 513.00
- Po1     1    134137 1499452 513.41
- U2      1    184143 1549458 514.95
- M       1    186110 1551425 515.01
- Prob    1    237493 1602808 516.54
- Ed      1    409448 1774763 521.33
- Ineq    1    502909 1868224 523.75

Step:  AIC=657.87
Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + Wealth +
    Ineq + Prob

         Df Sum of Sq     RSS    AIC
- NW      1     11675 1387523 507.77
- Po2     1     21418 1397266 508.09
- Pop     1     27803 1403651 508.31
- M.F     1     31252 1407100 508.42
- Wealth  1     35035 1410883 508.55
<none>                1375848 509.37
- U1      1     80954 1456802 510.06
+ LF      1     10533 1365315 511.01
+ Time    1      9996 1365852 511.03
+ So      1      3046 1372802 511.26
- Po1     1    123896 1499744 511.42
- U2      1    190746 1566594 513.47
- M       1    217716 1593564 514.27
- Prob    1    226971 1602819 514.54
- Ed      1    413254 1789103 519.71
- Ineq    1    500944 1876792 521.96
```

```
Step:  AIC=654.18
Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth + Ineq + 
    Prob

         Df Sum of Sq     RSS    AIC
- Po2     1     16706 1404229 506.33
- Pop     1     25793 1413315 506.63
- M.F     1     26785 1414308 506.66
- Wealth  1     31551 1419073 506.82
<none>                1387523 507.77
- U1      1     83881 1471404 508.52
+ NW      1     11675 1375848 509.37
+ So      1      7207 1380316 509.52
+ LF      1      6726 1380797 509.54
+ Time    1      4534 1382989 509.61
- Po1     1    118348 1505871 509.61
- U2      1    201453 1588976 512.14
- Prob    1    216760 1604282 512.59
- M       1    309214 1696737 515.22
- Ed      1    402754 1790276 517.74
- Ineq    1    589736 1977259 522.41

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
+ Po2     1     16706 1387523 507.77
+ 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
```

```
+ Po2     1     13259 1413315 506.63
+ 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
+ Po2     1     14148 1438919 505.47
+ 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           U1
   -6426.10        93.32       180.12       102.65        22.34     -6086.63
         U2         Ineq         Prob
     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                                            residual standard error
[1,] "m1: Crime~."                                    "209.064"
[2,] "m2: Crime~M+Ed+Po1+U2+Ineq+Prob"                "200.69"
[3,] "m3: Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob"         "195.547"
[4,] "m4: Crime~M+Ed+Po1+M.F+U1+U2+Wealth+Ineq+Prob"  "196.357"
     AICc
[1,] "671.133"
[2,] "643.956"
[3,] "645.426"
[4,] "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
```

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
```