ISYE 6051 : Homework 5
2/24/2021

Table of Contents

8.1 Situations where linear regression analysis would be beneficial

7.2 Using Exponential Smoothing to determine the unofficial end of summer

8.1 Linear Regression Analysis

Linear Regression is the process of quantifying the relationship between a predictor variable with that of an outcome.

My analysis would center on Blood Pressure readings in women where their blood pressure is controlled either naturally or by prescribed medications. Therefore, the blood pressure readings captured over a 30-day period at regular intervals (8am, 12 noon and 4pm) are the outcome. The 5 predictors chosen are 1) number of minutes of cardio exercise each day where heart beats per minute is greater than 100 bpms, 2) number of grams of sodium ingested each day, 3) number of bananas or oranges eaten per day (high potassium foods), 4) number of hours of sleep per day and 5) number of alcoholic drinks per day. For simple linear regression, I would only evaluate one predictor against the outcome.

8.2 Evaluating crime with linear regression models

This assignment looks for the observed crime data (outcome) with a set of 15 possible predictors:

Predictor	*Value*
M	*14.0*
So	*0*

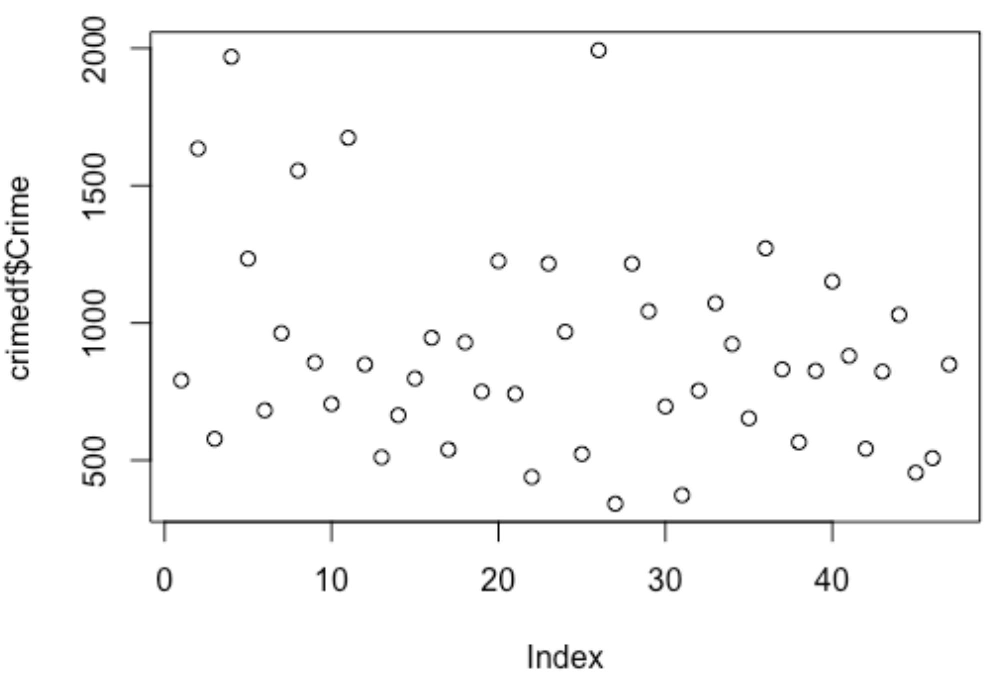
Ed	*10.0*
Po1	*12.0*
Po2	*15.5*
LF	*0.640*
M.F	*94.0*
Pop	*150*
NW	*1.1*
U1	*0.120*
U2	*3.6*
Wealth	*3200*
Ineq	*20.1*
Prob	*0.04*
Time	*39.0*

Reading the uscrime.txt data in and plotting it provides a graph that shows that there are perhaps 2 outliers that may be taken into consideration for possible exclusion during a later evaluation of the model.

```
rm(list = ls())
set.seed(17)
setwd("~/Documents/ISYE6501 Intro to Analytics
Modeling/FA_SP_hw5")

#library definition

# Read the data in
crimedf <- read.table("uscrime.txt", header = TRUE)
head(crimedf)
tail(crimedf)

plot(crimedf$Crime)
```

```
head(crimedf)
    M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq      Prob
1 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
2 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
3 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
4 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
5 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
6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
0.034201
    Time Crime
1 26.2011   791
2 25.2999  1635
3 24.3006   578
4 29.9012  1969
5 21.2998  1234
6 20.9995   682
> tail(crimedf)
     M So   Ed  Po1 Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq      Prob
42 14.1  0 10.9  5.6 5.4 0.523  96.8   4  0.2 0.107 3.7   4890 17.0
0.088904
43 16.2  1  9.9  7.5 7.0 0.522  99.6  40 20.8 0.073 2.7   4960 22.4
0.054902
44 13.6  0 12.1  9.5 9.6 0.574 101.2  29  3.6 0.111 3.7   6220 16.2
0.028100
45 13.9  1  8.8  4.6 4.1 0.480  96.8  19  4.9 0.135 5.3   4570 24.9
0.056202
```

```
46 12.6  0 10.4 10.6 9.7 0.599  98.9  40  2.4 0.078 2.5   5930 17.1
0.046598
47 13.0  0 12.1  9.0 9.1 0.623 104.9   3  2.2 0.113 4.0   5880 16.0
0.052802
     Time Crime
42 12.1996   542
43 31.9989   823
44 30.0001  1030
45 32.5996   455
46 16.6999   508
47 16.0997   849
```



For the first evaluation all 15 predictors are used to determine the outcome. The lm() – linear model function is used to evaluate the dataset fitting for a general linear model assuming that the errors have a normal distribution.

```
model15 <- lm(Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW +
U1 + U2 + Wealth +
            Ineq + Prob + Time, data = crimedf)
summary(model15)
```

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

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
```

The median is close to 0, the p-value is <0.05 and there are 4 predictors that are significantly different than 0.

The next approach tried was using the glm() function. GLM allowed for a generalized linear model where response variables can follow different distributions.

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

Deviance 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

(Dispersion parameter for gaussian family taken to be 43707.93)

    Null deviance: 6880928  on 46  degrees of freedom
Residual deviance: 1354946  on 31  degrees of freedom
AIC: 650.03

Number of Fisher Scoring iterations: 2
```

The median is the same as are the number of predictors that are different than 0 (4).

Since neither model showed differences, I decided to use the lm() function with the 5 predictors that showed significance (M, Ed, U2, Ineq and Prob) since these predictors have the closest value to 0.05.

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

Residuals:
   Min     1Q Median     3Q    Max 
-478.8 -233.6  -46.5  143.2  797.1 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3336.52    1435.26  -2.325  0.02512 *
M              85.33      54.39   1.569  0.12437
Ed            214.69      73.20   2.933  0.00547 **
U2            160.01      65.54   2.441  0.01903 *
Ineq           29.50      21.56   1.368  0.17880
Prob        -6897.24    2427.81  -2.841  0.00697 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 328.6 on 41 degrees of freedom
Multiple R-squared:  0.3565,	Adjusted R-squared:  0.278
F-statistic: 4.542 on 5 and 41 DF,  p-value: 0.002186
```

Excluding all but 5 predictors still does not show a fitted model since only 3 of the predictors have a significant difference than 0. Using the R^2 value of 0.278 I adjusted the number of predictors to 0.07 to now use 6 predictors.

```
model6 <- lm(Crime ~ M + Ed + Po1 + U2  +
               Ineq + Prob, data = crimedf)
summary(model6)
```

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

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
```

All 6 predictors are significantly different than 0. In this evaluation model6 appears to be the best model to utilize. I also tried an evaluation including U1 as a predictor and this p-value was not significantly different than 0.

Other evaluations tools that can be used to validate the best model that can be used for prediction are AIC and BIC.
AIC is the Akaike Information Criterion helps to avoid overfitting prderring comparison models with a smaller AIC. I will validate model15 (which contains all 15 predictors) against model6. BIC – Baysian Information Criterion can be used when there are more data points than parameters.

```
AIC(model15)
AIC(model6)
BIC(model15)
BIC(model6)
```

```
AIC(model15)
```

```
[1] 650.0291
> AIC(model6)
[1] 640.1661
> BIC(model15)
[1] 681.4816
> BIC(model6)
[1] 654.9673
```

Both the AIC and BIC criterions show that model6 is the best model for calculating Crime. Model6 is smaller with AIC (smaller is better with AIC comparisons) and there is a larger difference than 10 for BIC which is caegorized as a model that is very likely to be better.

I created a dataframe (predictvalue) to store the test values provided for the homework. Based on this information the estimate for Crime using model6 is 1304. This is reasonable looking at the raw data in the 47 data points for crime in the raw data uscrime.txt.

```
predictvalue <- 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 =
3200, Ineq = 20.1, Prob = 0.04, Time =
+                          39.0)
> predict(model6,predictvalue)
       1
1304.245
```