ISYE6501x Homework 5

Done By: Joel Quek

Question 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.

I dabble in trading once in a while, and personally enjoy selling Bull Put Option Spreads on US Equities. My measure of success is when the options contracts expire Out Of The Money, which means I made the full premium as profit. When selling these Option Spreads there are many considerations, which I could use as predictors for my model. For example:

1. The Implied Volatility Percentile
2. The Delta Value of the Short Put Option
3. The Number of Days/Weeks to Expiration (A function of the Theta Value)
4. The overall trend of the market - Bullish, Bearish or Indiferrent (Dummy variables would need to be used)
5. The Probability of being Out-Of-The-Money (Using Black-Scholes Formula)

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (http://www.statsci.org/data/general/uscrime.txt) (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html (http://www.statsci.org/data/general/uscrime.html)), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

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

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course.

1. Importing Dataset and Libraries

In [1]:

```
library(stats)
```

In [26]:

```
library(dplyr)
```

...

In [2]:

```
crime <- read.table("uscrime.txt", header=TRUE)
```

In [3]:

```
head(crime)
```

A data.frame: 6 × 16

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>
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	26.2011	791
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	25.2999	1635
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	24.3006	578
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	29.9012	1969
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	21.2998	1234
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	20.9995	682

In [4]:

```
summary(crime)
```

```
       M               So               Ed              Po1       
 Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50  
 1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25  
 Median :13.60   Median :0.0000   Median :10.80   Median : 7.80  
 Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50  
 3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45  
 Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60  
      Po2              LF              M.F              Pop        
 Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00  
 1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00  
 Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00  
 Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62  
 3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50  
 Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00  
       NW              U1                U2            Wealth    
 Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880  
 1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595  
 Median : 7.60   Median :0.09200   Median :3.400   Median :5370  
 Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254  
 3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915  
 Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890  
      Ineq            Prob              Time           Crime       
 Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0  
 1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5  
 Median :17.60   Median :0.04210   Median :25.80   Median : 831.0  
 Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1  
 3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5  
 Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0  
```

In [5]:

```
names(crime)
```

'M' · 'So' · 'Ed' · 'Po1' · 'Po2' · 'LF' · 'M.F' · 'Pop' · 'NW' · 'U1' · 'U2' · 'Wealth' · 'Ineq' · 'Prob' · 'Time' · 'Crime'

2. Linear Regression Model 1 (Unscaled Data)

In [6]:

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

In [7]:

```
model
```

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

Coefficients:
(Intercept)            M           So           Ed          Po1          Po2  
 -5.984e+03    8.783e+01   -3.803e+00    1.883e+02    1.928e+02   -1.094e+02  
         LF          M.F          Pop           NW           U1           U2  
 -6.638e+02    1.741e+01   -7.330e-01    4.204e+00   -5.827e+03    1.678e+02  
     Wealth         Ineq         Prob         Time  
  9.617e-02    7.067e+01   -4.855e+03   -3.479e+00  
```

3. Linear Regression Model 2 (Scaled Data)

In [8]:

```
crime_scaled <- scale(crime[,1:15])
```

In [9]:

```
head(crime_scaled)
```

A matrix: 6 × 15 of type dbl

M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ir
0.9886930	1.3770536	-1.3085099	-0.9085105	-0.8666988	-1.2667456	-1.12060499	-0.09500679	1.943738564	0.69510600	0.8313680	-1.3616094	1.6793
0.3521372	-0.7107373	0.6580587	0.6056737	0.5280852	0.5396568	0.98341752	-0.62033844	0.008483424	0.02950365	0.2393332	0.3276683	0.0000
0.2725678	1.3770536	-1.4872888	-1.3459415	-1.2958632	-0.6976051	-0.47582390	-0.48900552	1.146296747	-0.08143007	-0.1158877	-2.1492481	1.4036
-0.2048491	-0.7107373	1.3731746	2.1535064	2.1732150	0.3911854	0.37257228	3.16204944	-0.205464381	0.36230482	0.5945541	1.5298536	-0.6767
0.1929983	-0.7107373	1.3731746	0.8075649	0.7426673	0.7376187	0.06714965	-0.48900552	-0.691709391	-0.24783066	-1.6551781	0.5453053	-0.5013
-1.3983912	-0.7107373	0.3898903	1.1104017	1.2433590	-0.3511718	-0.64550313	-0.30513945	-0.555560788	-0.63609870	-0.5895155	1.6956723	-1.7044

In [10]:

```
crime_scaled <- data.frame(crime_scaled)
```

In [11]:

```
head(crime_scaled)
```

A data.frame: 6 × 15

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	0.9886930	1.3770536	-1.3085099	-0.9085105	-0.8666988	-1.2667456	-1.12060499	-0.09500679	1.943738564	0.69510600	0.8313680	-1.3616094	1.67
2	0.3521372	-0.7107373	0.6580587	0.6056737	0.5280852	0.5396568	0.98341752	-0.62033844	0.008483424	0.02950365	0.2393332	0.3276683	0.0
3	0.2725678	1.3770536	-1.4872888	-1.3459415	-1.2958632	-0.6976051	-0.47582390	-0.48900552	1.146296747	-0.08143007	-0.1158877	-2.1492481	1.4
4	-0.2048491	-0.7107373	1.3731746	2.1535064	2.1732150	0.3911854	0.37257228	3.16204944	-0.205464381	0.36230482	0.5945541	1.5298536	-0.67
5	0.1929983	-0.7107373	1.3731746	0.8075649	0.7426673	0.7376187	0.06714965	-0.48900552	-0.691709391	-0.24783066	-1.6551781	0.5453053	-0.5
6	-1.3983912	-0.7107373	0.3898903	1.1104017	1.2433590	-0.3511718	-0.64550313	-0.30513945	-0.555560788	-0.63609870	-0.5895155	1.6956723	-1.7

In [12]:

```
model2 <- lm(crime$Crime ~., data=crime_scaled)
```

In [13]:

```
model2
```

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

Coefficients:
(Intercept)            M           So           Ed          Po1          Po2  
    905.085      110.382       -1.822      210.678      572.995     -305.958  
         LF          M.F          Pop           NW           U1           U2  
    -26.826       51.293      -27.906       43.234     -105.056      141.714  
     Wealth         Ineq         Prob         Time  
     92.792      281.954     -110.394      -24.655  
```

In [27]:

```
summary(model2)
```

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

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)  905.085     30.495  29.680  < 2e-16 ***
M            110.382     52.424   2.106  0.04344 *  
So            -1.822     71.250  -0.026  0.97977    
Ed           210.678     69.458   3.033  0.00486 ** 
Po1          572.995    315.347   1.817  0.07889 .  
Po2         -305.958    328.483  -0.931  0.35883    
LF           -26.826     59.394  -0.452  0.65465    
M.F           51.293     59.977   0.855  0.39900    
Pop          -27.906     49.095  -0.568  0.57385    
NW            43.234     66.642   0.649  0.52128    
U1          -105.056     75.906  -1.384  0.17624    
U2           141.714     69.536   2.038  0.05016 .  
Wealth        92.792    100.028   0.928  0.36075    
Ineq         281.954     90.630   3.111  0.00398 ** 
Prob        -110.394     51.667  -2.137  0.04063 *  
Time         -24.655     50.780  -0.486  0.63071    
---
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
```

4. Evaluation of Models

P-Values of Coefficients

A low P-value (< 0.05) means that the coefficient is likely not to equal zero. A high P-value (> 0.05) means that we cannot conclude that the explanatory variable affects the dependent variable

Source: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R (https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R)

I will remove coefficients with P-Value greater than 0.05 later

- so
- po1
- po2
- LF
- M.F.
- Pop
- NW
- U1
- U2
- wealth
- time

Train-Test Split Accuracy Test

Source: http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/ (http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/)

Source: https://www.statology.org/train-test-split-r/ (https://www.statology.org/train-test-split-r/)

Test on Model 1 (Unscaled Data)

In [15]:

```
library(caTools)
```

In [31]:

```
head(crime)
```

A data.frame: 6 × 16

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>
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	26.2011	791
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	25.2999	1635
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	24.3006	578
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	29.9012	1969
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	21.2998	1234
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	20.9995	682

In [17]:

```
set.seed(123)
#use 70% of dataset as training set and 30% as test set
split = sample.split(crime$Crime, SplitRatio = 0.7)
train  = subset(crime, split == TRUE)
test   = subset(crime, split == FALSE)
```

In [32]:

```
head(train)
```

A data.frame: 6 × 16

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>
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	26.2011	791
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	24.3006	578
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	20.9995	682
7	12.7	1	11.1	8.2	7.9	0.519	98.2	4	13.9	0.097	3.8	6200	16.8	0.042100	20.6993	963
9	15.7	1	9.0	6.5	6.2	0.553	95.5	39	28.6	0.081	2.8	4210	23.9	0.071697	29.4001	856
10	14.0	0	11.8	7.1	6.8	0.632	102.9	7	1.5	0.100	2.4	5260	17.4	0.044498	19.5994	705

In [19]:

```
names(train)
```

'M' · 'So' · 'Ed' · 'Po1' · 'Po2' · 'LF' · 'M.F' · 'Pop' · 'NW' · 'U1' · 'U2' · 'Wealth' · 'Ineq' · 'Prob' · 'Time' · 'Crime'

In [33]:

```
head(test)
```

A data.frame: 6 × 16

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>
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	25.2999	1635
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	29.9012	1969
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	21.2998	1234
8	13.1	1	10.9	11.5	10.9	0.542	96.9	50	17.9	0.079	3.5	4720	20.6	0.040099	24.5988	1555
11	12.4	0	10.5	12.1	11.6	0.580	96.6	101	10.6	0.077	3.5	6570	17.0	0.016201	41.6000	1674
16	14.2	1	8.8	8.1	7.7	0.497	95.6	33	32.1	0.116	4.7	4270	24.7	0.052099	26.0991	946

In [21]:

```
names(test)
```

'M' · 'So' · 'Ed' · 'Po1' · 'Po2' · 'LF' · 'M.F' · 'Pop' · 'NW' · 'U1' · 'U2' · 'Wealth' · 'Ineq' · 'Prob' · 'Time' · 'Crime'

In [34]:

```
# Train Model using Unscaled Data
model <- lm(Crime ~ M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data=train)
```

In [35]:

```
model
```

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

Coefficients:
(Intercept)            M           So           Ed          Po1          Po2  
 -3.094e+03    1.218e+02    6.252e+01    1.158e+02    1.505e+02   -7.226e+01  
         LF          M.F          Pop           NW           U1           U2  
  1.427e+03   -1.846e+01   -2.043e+00    6.782e+00    1.492e+03    6.862e+01  
     Wealth         Ineq         Prob         Time  
  7.033e-02    4.973e+01   -4.958e+03   -4.693e+00  
```

In [36]:

```
summary(model)
```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-287.128  -65.042   -4.796  103.893  298.405 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -3.094e+03  1.892e+03  -1.636   0.1214  
M            1.218e+02  5.851e+01   2.082   0.0538 .
So           6.252e+01  1.670e+02   0.374   0.7131  
Ed           1.158e+02  9.058e+01   1.278   0.2193  
Po1          1.505e+02  1.359e+02   1.107   0.2845  
Po2         -7.226e+01  1.376e+02  -0.525   0.6068  
LF           1.427e+03  2.346e+03   0.608   0.5515  
M.F         -1.846e+01  2.848e+01  -0.648   0.5260  
Pop         -2.043e+00  1.748e+00  -1.169   0.2595  
NW           6.782e+00  8.065e+00   0.841   0.4128  
U1           1.492e+03  5.299e+03   0.282   0.7819  
U2           6.862e+01  1.052e+02   0.652   0.5235  
Wealth       7.033e-02  1.300e-01   0.541   0.5958  
Ineq         4.973e+01  2.408e+01   2.065   0.0555 .
Prob        -4.958e+03  2.484e+03  -1.996   0.0632 .
Time        -4.694e+00  8.425e+00  -0.557   0.5852  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 186.3 on 16 degrees of freedom
Multiple R-squared:  0.6955,	Adjusted R-squared:   0.41 
F-statistic: 2.436 on 15 and 16 DF,  p-value: 0.04361
```

In [44]:

```
pred <- predict(model, test)
```

In [45]:

```
pred
```

2: 1307.43637908047 **4:** 1413.86347241968 **5:** 1183.18623573892 **8:** 1173.29695832952 **11:** 911.744484497506 **16:** 1117.75858148835 **20:** 1077.31969386968 **21:** 752.574620263444 **22:** 832.68523190931 **24:** 870.408783693826 **26:** 1599.58579046693 **31:** 439.77242807706 **32:** 713.3862501965 **34:** 817.194137524441 **37:** 1364.52664269435

In [46]:

```
model %>% predict(test)
```

2: 1307.43637908047 **4:** 1413.86347241968 **5:** 1183.18623573892 **8:** 1173.29695832952 **11:** 911.744484497506 **16:** 1117.75858148835 **20:** 1077.31969386968 **21:** 752.574620263444 **22:** 832.68523190931 **24:** 870.408783693826 **26:** 1599.58579046693 **31:** 439.77242807706 **32:** 713.3862501965 **34:** 817.194137524441 **37:** 1364.52664269435

https://www.statology.org/error-in-evalpredvars-data-env-object-not-found/#:~:text=This%20error%20occurs%20when%20you,used%20to%20fit%20the%20model (https://www.statology.org/error-in-evalpredvars-data-env-object-not-found/#:~:text=This%20error%20occurs%20when%20you,used%20to%20fit%20the%20model).

In [47]:

```
sum(pred == test[,16]) / nrow(test) # Incorrect to use this as it is not a classification model
```

0

In [49]:

```
# Calculate Sum of Squared Error For Predictions
sum((pred-test[,16])^2)
```

1817449.95771349

In [84]:

```
# Calculate Root Mean Squared Error For Predictions
sqrt((sum((pred-test[,16])^2)/nrow(test)))
```

348.085234553597

Test on Model 2 (Scaled Data)

In [56]:

```
crime_scaled <- crime
```

In [57]:

```
crime_scaled[,1:15]<-scale(crime_scaled[,1:15])
```

In [58]:

```
head(crime_scaled)
```

Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	C
:dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<
5099	-0.9085105	-0.8666988	-1.2667456	-1.12060499	-0.09500679	1.943738564	0.69510600	0.8313680	-1.3616094	1.6793638	1.6497631	-0.05599367	
0587	0.6056737	0.5280852	0.5396568	0.98341752	-0.62033844	0.008483424	0.02950365	0.2393332	0.3276683	0.0000000	-0.7693365	-0.18315796	
2888	-1.3459415	-1.2958632	-0.6976051	-0.47582390	-0.48900552	1.146296747	-0.08143007	-0.1158877	-2.1492481	1.4036474	1.5969416	-0.32416470	
1746	2.1535064	2.1732150	0.3911854	0.37257228	3.16204944	-0.205464381	0.36230482	0.5945541	1.5298536	-0.6767585	-1.3761895	0.46611085	
1746	0.8075649	0.7426673	0.7376187	0.06714965	-0.48900552	-0.691709391	-0.24783066	-1.6551781	0.5453053	-0.5013026	-0.2503580	-0.74759413	
8903	1.1104017	1.2433590	-0.3511718	-0.64550313	-0.30513945	-0.555560788	-0.63609870	-0.5895155	1.6956723	-1.7044289	-0.5669349	-0.78996812	

In [59]:

```
set.seed(123)
#use 70% of dataset as training set and 30% as test set
scale_split = sample.split(crime_scaled$Crime, SplitRatio = 0.7)
scale_train  = subset(crime_scaled, split == TRUE)
scale_test   = subset(crime_scaled, split == FALSE)
```

In [61]:

```
head(scale_train)
```

A data.frame: 6 × 16

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	0.9886930	1.3770536	-1.3085099	-0.9085105	-0.86669885	-1.2667456	-1.12060499	-0.09500679	1.9437386	0.69510600	0.8313680	-1.361609422	1
3	0.2725678	1.3770536	-1.4872888	-1.3459415	-1.29586316	-0.6976051	-0.47582390	-0.48900552	1.1462967	-0.08143007	-0.1158877	-2.149248102	1
6	-1.3983912	-0.7107373	0.3898903	1.1104017	1.24335901	-0.3511718	-0.64550313	-0.30513945	-0.5555608	-0.63609870	-0.5895155	1.695672300	-1
7	-0.9209743	1.3770536	0.4792798	-0.1009456	-0.04413392	-1.0440385	-0.03465789	-0.85673768	0.3683047	0.08497051	0.4761471	0.980579287	-0
9	1.4661099	1.3770536	-1.3978993	-0.6729708	-0.65211669	-0.2027004	-0.95092575	0.06259271	1.7978651	-0.80249928	-0.7079224	-1.081790417	1
10	0.1134288	-0.7107373	1.1050061	-0.4710795	-0.43753454	1.7521735	1.56032692	-0.77793793	-0.8375829	0.25137110	-1.1815502	0.006394603	-0

In [62]:

```
head(scale_test)
```

A data.frame: 6 × 16

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
2	0.3521372	-0.7107373	0.65805874	0.6056737	0.5280852	0.5396568	0.98341752	-0.62033844	0.008483424	0.02950365	0.2393332	0.3276683	0.
4	-0.2048491	-0.7107373	1.37317459	2.1535064	2.1732150	0.3911854	0.37257228	3.16204944	-0.205464381	0.36230482	0.5945541	1.5298536	-0.
5	0.1929983	-0.7107373	1.37317459	0.8075649	0.7426673	0.7376187	0.06714965	-0.48900552	-0.691709391	-0.24783066	-1.6551781	0.5453053	-0.
8	-0.6026964	1.3770536	0.30050081	1.0094561	1.0287769	-0.4748980	-0.47582390	0.35152511	0.757300739	-0.91343301	0.1209263	-0.5532434	0.
11	-1.1596827	-0.7107373	-0.05705712	1.2113474	1.2791227	0.4654211	-0.57763144	1.69112082	0.047383024	-1.02436673	0.1209263	1.3640350	-0.
16	0.2725678	1.3770536	-1.57667830	-0.1345942	-0.1156613	-1.5884338	-0.91698990	-0.09500679	2.138236568	1.13884089	1.5418097	-1.0196084	1.

In [63]:

```
# Train Model using Unscaled Data
model_scaled <- lm(Crime ~ M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data=scale_train)
```

In [64]:

```
model_scaled
```

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

Coefficients:
(Intercept)            M           So           Ed          Po1          Po2
     869.21       153.08        29.94       129.55       447.23      -202.06
         LF          M.F          Pop           NW           U1           U2
      57.67       -54.40       -77.78        69.74        26.90        57.96
     Wealth         Ineq         Prob         Time
      67.86       198.41      -112.73       -33.26
```

In [65]:

```
summary(model_scaled)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-287.128  -65.042   -4.796  103.893  298.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   869.21      40.72  21.345  3.5e-13 ***
M             153.08      73.54   2.082   0.0538 .
So             29.94      80.01   0.374   0.7131
Ed            129.55     101.34   1.278   0.2193
Po1           447.23     403.82   1.107   0.2845
Po2          -202.06     384.86  -0.525   0.6068
LF             57.67      94.80   0.608   0.5515
M.F           -54.40      83.91  -0.648   0.5260
Pop           -77.78      66.53  -1.169   0.2595
NW             69.74      82.93   0.841   0.4128
U1             26.90      95.53   0.282   0.7819
U2             57.96      88.86   0.652   0.5235
Wealth         67.86     125.39   0.541   0.5958
Ineq          198.41      96.08   2.065   0.0555 .
Prob         -112.73      56.47  -1.996   0.0632 .
Time          -33.26      59.71  -0.557   0.5852
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 186.3 on 16 degrees of freedom
Multiple R-squared:  0.6955,    Adjusted R-squared:   0.41
F-statistic: 2.436 on 15 and 16 DF,  p-value: 0.04361
```

In [68]:

```
scale_pred <- predict(model_scaled, scale_test)
```

In [69]:

```
scale_pred
```

2: 1307.43637908046 **4:** 1413.86347241967 **5:** 1183.18623573892 **8:** 1173.29695832952 **11:** 911.744484497504 **16:** 1117.75858148834 **20:** 1077.31969386968 **21:** 752.574620263441 **22:** 832.685231909299 **24:** 870.408783693821 **26:** 1599.58579046693 **31:** 439.772428077062 **32:** 713.386250196495 **34:** 817.194137524438 **37:** 1364.52664269433

In [67]:

```
# Calculate R Squared Error For Predictions
sum((scale_pred-scale_test[,16])^2)
```

1817449.95771349

In [85]:

```
# Calculate Root Mean Squared Error For Predictions
sqrt((sum((scale_pred-scale_test[,16])^2)/nrow(test)))
```

348.085234553597

The unscaled data also had a $RMSE$ score of 348.085234553597

5. Predictions

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

In [71]:

```
x <- c(14.0, 0, 10.0, 12.0, 15.5, 0.640, 94.0, 150, 1.1, 0.120, 3.6, 3200, 20.1, 0.04,39.0)
```

In [73]:

```
x<-list("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,
```

In [76]:

```
crime_pred_scaled <- predict(model_scaled, x)
```

In [77]:

```
crime_pred_scaled
```

1: 209924.369411726

In [78]:

```
crime_pred<-predict(model,x)
```

In [79]:

```
crime_pred
```

1: 603.326307807169

The Model Coefficients

In [80]:

```
model
```

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

Coefficients:
(Intercept)            M           So           Ed          Po1          Po2
 -3.094e+03    1.218e+02    6.252e+01    1.158e+02    1.505e+02   -7.226e+01
         LF          M.F          Pop           NW           U1           U2
  1.427e+03   -1.846e+01   -2.043e+00    6.782e+00    1.492e+03    6.862e+01
     Wealth         Ineq         Prob         Time
  7.033e-02    4.973e+01   -4.958e+03   -4.693e+00
```

Prediction of Crime Level

I have to use the unscaled model to predict because the required input is unscaled.

The predicted crime level is *603.326*.

The Quality of Fit

The $RMSE$ of the Model is 348.085234553597

In [82]:

```
sd(crime$Crime)
```

386.762697146186

In [83]:

```
summary(crime$Crime)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  342.0   658.5   831.0   905.1  1057.5  1993.0 
```

The RMSE is similar to the data Standard Deviation, and the value of the RMSE is close to the minimum value of the data, which is a good sign.To make the fit better, I can also remove the coefficients with high P-Values.