HW 5

2023-09-21

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 work in asset management where we design various investment strategies based on client risk/return profile and investment needs. One strategy is a large cap core strategy which focuses on large cap companies. It would be helpful to predict the performance of this strategy in order to determine if it meets certain client expectations (return for given risk). Five variables that we can use to predict performance are:

S&P 500 Index - The overall return of the S&P 500 provides a benchmark for large cap performance. A strong positive correlation would be expected. Interest Rates - Declining interest rates tend to predict higher equity returns. Lower rates could indicate stronger performance. Inflation Rate - High inflation historically leads to lower P/E ratios and stock returns. So lower inflation may predict better performance. Market Volatility (Risk) - Measures like the VIX indicate general market volatility. Higher volatility has been associated with weaker returns. Sentiment Indicators - Metrics that quantify market sentiment or investor psychology like put/call ratios. High negative sentiment predicts better returns.

Question 8.2 Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at 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 over fitting. We'll see ways of dealing with this sort of problem later in the course.

```r
library(ggplot2)
```

```r
# Define the URL where the data is located
url <- "http://www.statsci.org/data/general/uscrime.txt"

# Use read.table to read the data from the URL into a data frame
data <- read.table(url, header = TRUE, sep = "\t")

# Display the first few rows of the data frame
head(data)
```

```
##      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
```

```
# Check dimensions (number of rows and columns)
cat("Dimensions of the data frame:\n")
```

```
## Dimensions of the data frame:
```

```
dim(data)
```

```
## [1] 47 16
```

```
# Check data types of columns
cat("\nData types of columns:\n")
```

```
##
## Data types of columns:
```

```
str(data)
```

```
## 'data.frame':    47 obs. of  16 variables:
##  $ M     : num  15.1 14.3 14.2 13.6 14.1 12.1 12.7 13.1 15.7 14 ...
##  $ So    : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed    : num  9.1 11.3 8.9 12.1 12.1 11 11.1 10.9 9 11.8 ...
##  $ Po1   : num  5.8 10.3 4.5 14.9 10.9 11.8 8.2 11.5 6.5 7.1 ...
##  $ Po2   : num  5.6 9.5 4.4 14.1 10.1 11.5 7.9 10.9 6.2 6.8 ...
##  $ LF    : num  0.51 0.583 0.533 0.577 0.591 0.547 0.519 0.542 0.553 0.632 ...
##  $ M.F   : num  95 101.2 96.9 99.4 98.5 ...
##  $ Pop   : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW    : num  30.1 10.2 21.9 8 3 4.4 13.9 17.9 28.6 1.5 ...
##  $ U1    : num  0.108 0.096 0.094 0.102 0.091 0.084 0.097 0.079 0.081 0.1 ...
##  $ U2    : num  4.1 3.6 3.3 3.9 2 2.9 3.8 3.5 2.8 2.4 ...
##  $ Wealth: int  3940 5570 3180 6730 5780 6890 6200 4720 4210 5260 ...
##  $ Ineq  : num  26.1 19.4 25 16.7 17.4 12.6 16.8 20.6 23.9 17.4 ...
##  $ Prob  : num  0.0846 0.0296 0.0834 0.0158 0.0414 ...
##  $ Time  : num  26.2 25.3 24.3 29.9 21.3 ...
##  $ Crime : int  791 1635 578 1969 1234 682 963 1555 856 705 ...
```

```
# Summary statistics
cat("\nSummary statistics:\n")
```

```
##
## Summary statistics:
```

```
summary(data)
```

```
##        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
```

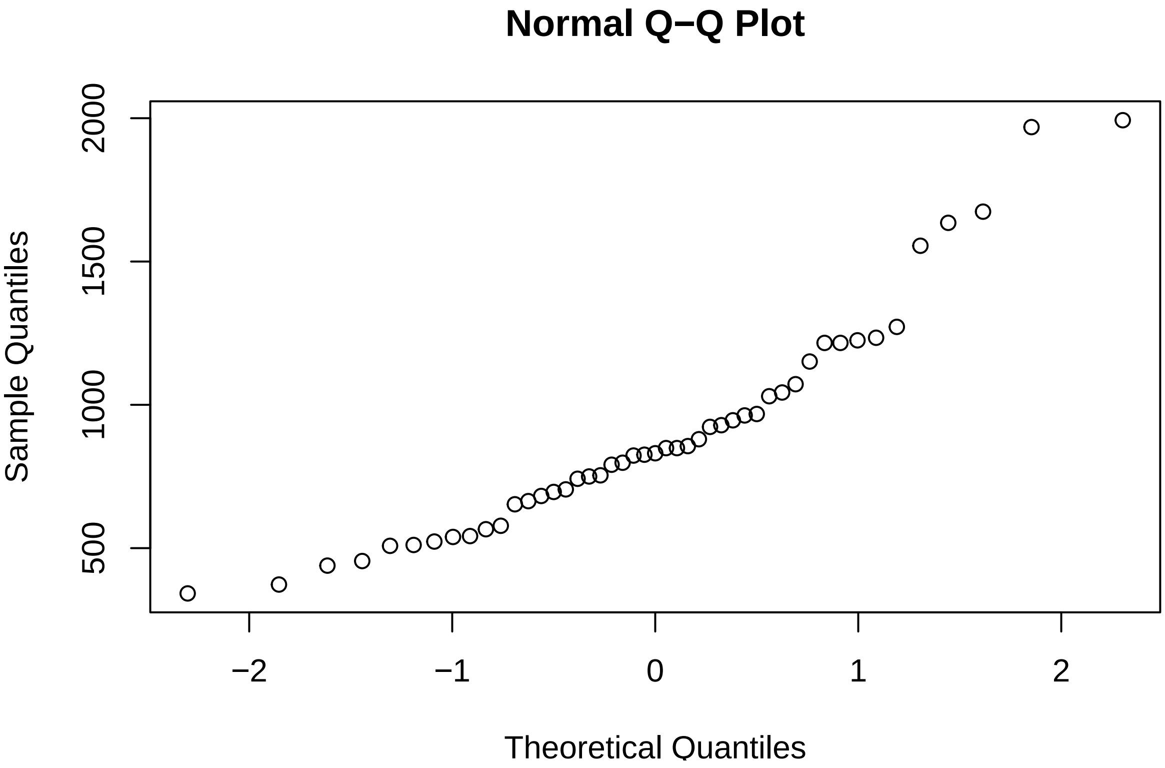
```
# Check for missing values
cat("\nMissing values:\n")
```

```
## 
## Missing values:
```

```
missing_values <- sum(is.na(data))
cat("Total missing values:", missing_values, "\n")
```

```
## Total missing values: 0
```

```
qqnorm(data$Crime)
```



The highest value in our data is 1993 and the lowest is 342.

```
# Extract predictor columns
predictors <- data[,c("M","So","Ed","Po1","Po2","LF","M.F","Pop","NW","U1","U2","Wealth","Ineq","Prob",
```

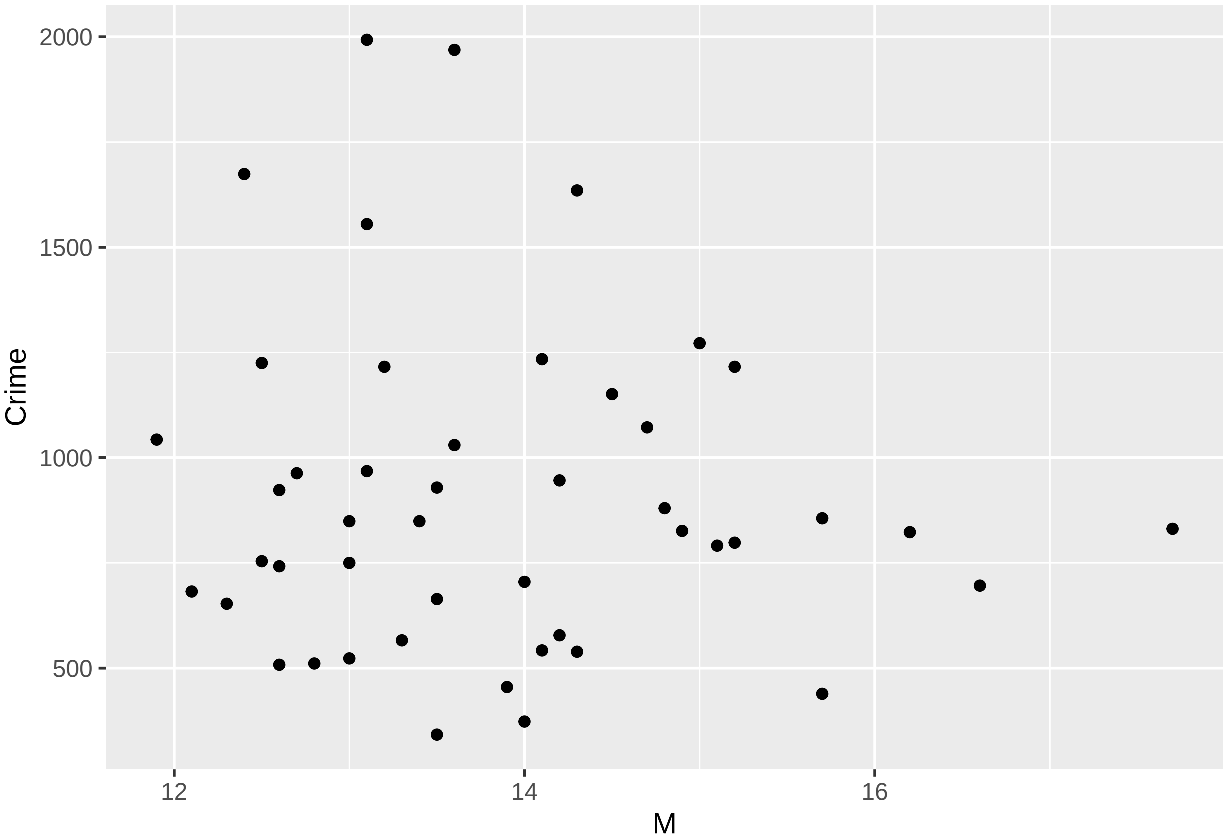
```r
# Calculate covariance matrix
corr_matrix <- cor(predictors)

# Round to 2 decimal places
corr_matrix <- round(corr_matrix, 2)

# Print covariance matrix
print(corr_matrix)
```

```
##            M    So    Ed   Po1   Po2    LF   M.F   Pop    NW    U1    U2 Wealth
## M       1.00  0.58 -0.53 -0.51 -0.51 -0.16 -0.03 -0.28  0.59 -0.22 -0.24  -0.67
## So      0.58  1.00 -0.70 -0.37 -0.38 -0.51 -0.31 -0.05  0.77 -0.17  0.07  -0.64
## Ed     -0.53 -0.70  1.00  0.48  0.50  0.56  0.44 -0.02 -0.66  0.02 -0.22   0.74
## Po1    -0.51 -0.37  0.48  1.00  0.99  0.12  0.03  0.53 -0.21 -0.04  0.19   0.79
## Po2    -0.51 -0.38  0.50  0.99  1.00  0.11  0.02  0.51 -0.22 -0.05  0.17   0.79
## LF     -0.16 -0.51  0.56  0.12  0.11  1.00  0.51 -0.12 -0.34 -0.23 -0.42   0.29
## M.F    -0.03 -0.31  0.44  0.03  0.02  0.51  1.00 -0.41 -0.33  0.35 -0.02   0.18
## Pop    -0.28 -0.05 -0.02  0.53  0.51 -0.12 -0.41  1.00  0.10 -0.04  0.27   0.31
## NW      0.59  0.77 -0.66 -0.21 -0.22 -0.34 -0.33  0.10  1.00 -0.16  0.08  -0.59
## U1     -0.22 -0.17  0.02 -0.04 -0.05 -0.23  0.35 -0.04 -0.16  1.00  0.75   0.04
## U2     -0.24  0.07 -0.22  0.19  0.17 -0.42 -0.02  0.27  0.08  0.75  1.00   0.09
## Wealth -0.67 -0.64  0.74  0.79  0.79  0.29  0.18  0.31 -0.59  0.04  0.09   1.00
## Ineq    0.64  0.74 -0.77 -0.63 -0.65 -0.27 -0.17 -0.13  0.68 -0.06  0.02  -0.88
## Prob    0.36  0.53 -0.39 -0.47 -0.47 -0.25 -0.05 -0.35  0.43 -0.01 -0.06  -0.56
## Time    0.11  0.07 -0.25  0.10  0.08 -0.12 -0.43  0.46  0.23 -0.17  0.10   0.00
## Crime  -0.09 -0.09  0.32  0.69  0.67  0.19  0.21  0.34  0.03 -0.05  0.18   0.44
##         Ineq  Prob  Time Crime
## M       0.64  0.36  0.11 -0.09
## So      0.74  0.53  0.07 -0.09
## Ed     -0.77 -0.39 -0.25  0.32
## Po1    -0.63 -0.47  0.10  0.69
## Po2    -0.65 -0.47  0.08  0.67
## LF     -0.27 -0.25 -0.12  0.19
## M.F    -0.17 -0.05 -0.43  0.21
## Pop    -0.13 -0.35  0.46  0.34
## NW      0.68  0.43  0.23  0.03
## U1     -0.06 -0.01 -0.17 -0.05
## U2      0.02 -0.06  0.10  0.18
## Wealth -0.88 -0.56  0.00  0.44
## Ineq    1.00  0.47  0.10 -0.18
## Prob    0.47  1.00 -0.44 -0.43
## Time    0.10 -0.44  1.00  0.15
## Crime  -0.18 -0.43  0.15  1.00
```

```r
# Create scatterplots
ggplot(data, aes(x=data[["M"]], y=data[["Crime"]])) +
  geom_point() +
  xlab("M") +
  ylab("Crime")
```

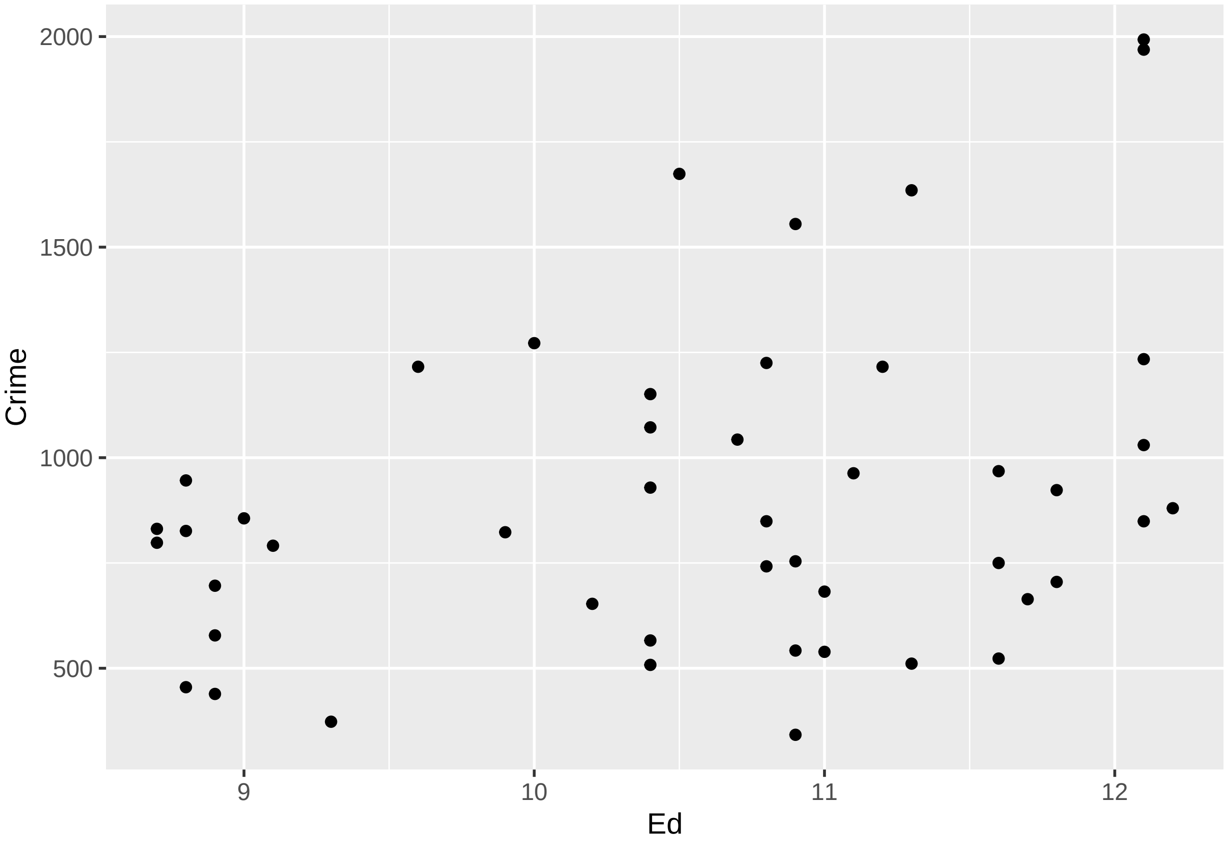
```
## Warning: Use of `data[["M"]]` is discouraged.
## i Use `.data[["M"]]` instead.
```

```
## Warning: Use of `data[["Crime"]]` is discouraged.
## i Use `.data[["Crime"]]` instead.
```



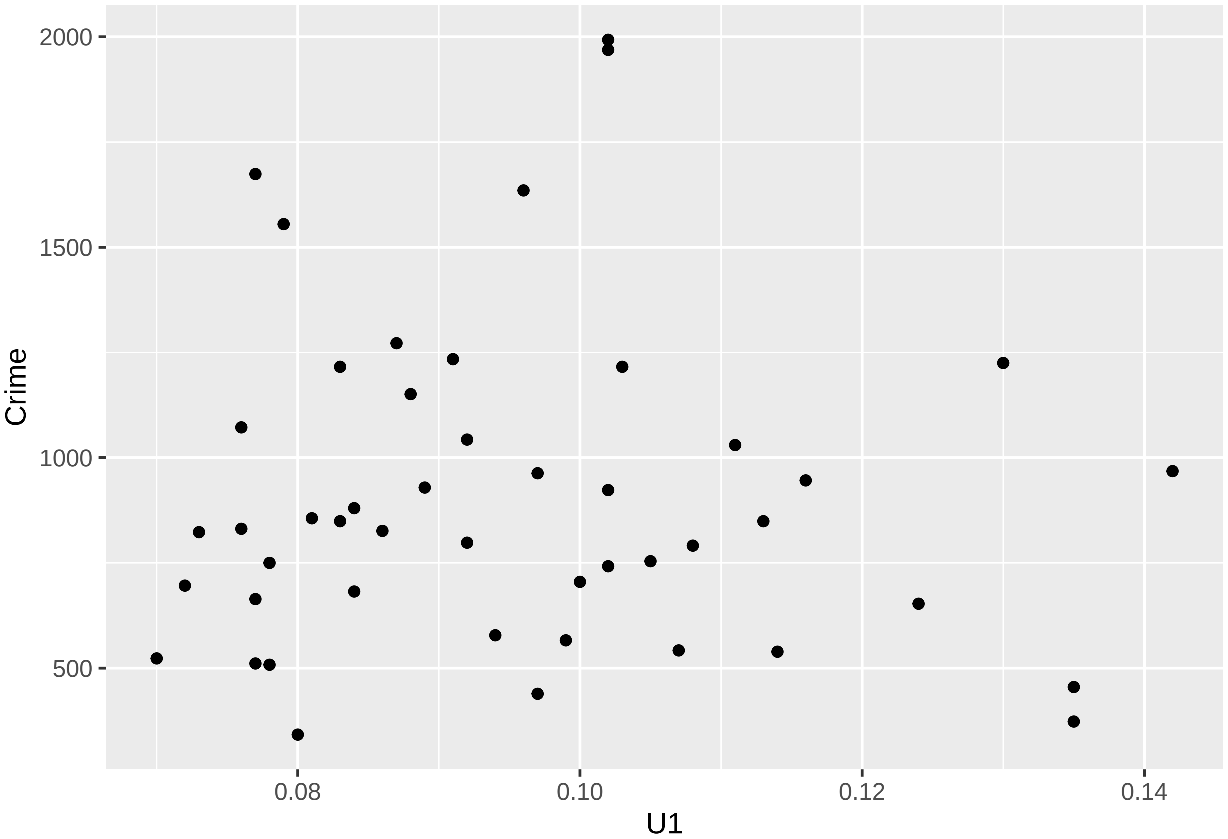
```
ggplot(data, aes(x=data[["Ed"]], y=data[["Crime"]])) +
  geom_point() +
  xlab("Ed") +
  ylab("Crime")
```

```
## Warning: Use of `data[["Ed"]]` is discouraged.
## i Use `.data[["Ed"]]` instead.
## Use of `data[["Crime"]]` is discouraged.
## i Use `.data[["Crime"]]` instead.
```



```r
ggplot(data, aes(x=data[["U1"]], y=data[["Crime"]])) +
  geom_point() +
  xlab("U1") +
  ylab("Crime")
```

```
## Warning: Use of `data[["U1"]]` is discouraged.
## i Use `.data[["U1"]]` instead.
## Use of `data[["Crime"]]` is discouraged.
## i Use `.data[["Crime"]]` instead.
```



After performing exploratory analysis and looking at the correlation results there are some key items to note. The proportion of young males and southern states seems to have a negative but immaterial relationship with crime rates. While the reverse may be what was actually expected, some assumptions could be based on the robustness of our data. What's noteworthy is that as unemployment among 35-39 year old males increases, crime increases, while crime decreases despite a fall in unemployment among youth 14-24. Other noteworthy occurrences in the data, are the positive relationship between crime and police and protection spending from 1959 to 1960, where possibly higher crime rates woudl require higher security spending.

```
# Load necessary libraries
options(repos = "https://cran.r-project.org")

install.packages("caret")
```

```
##
## The downloaded binary packages are in
##  /var/folders/6r/d7grt80x11v5vty52zj5cywh0000gn/T//RtmpRVZw1p/downloaded_packages
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
# Set a random seed for reproducibility
set.seed(123)

# Split the data into train (70%), validation (15%), and test (15%) sets
trainIndex <- createDataPartition(y = data$Crime, p = 0.7, list = FALSE)
validationIndex <- createDataPartition(y = data$Crime[-trainIndex], p = 0.5, list = FALSE)

trainData <- data[trainIndex, ]
validationData <- data[-c(trainIndex, validationIndex), ]
```

```r
testData <- data[-trainIndex, ]

# Standardize/normalize features in the training, validation, and test sets
# Using the "scale" function to standardize the features
trainData[, -1] <- scale(trainData[, -1])
validationData[, -1] <- scale(validationData[, -1])
testData[, -1] <- scale(testData[, -1])
```

```r
# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Train a full simple linear regression model using the full data
full_model_data <- lm(Crime ~ ., data = data)

#Predict given data using full model
predit_data <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640, M.F = 94.0, 

predict_model <- predict(full_model_data, predit_data)
predict_model
```

```
##        1
## 155.4349
```

```r
#full model on training data
full_model <- lm(Crime ~ ., data = trainData)

bfinal_model <- step(full_model, direction = "backward")
```

```
## Start:  AIC=-30.22
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq    RSS     AIC
## - Pop     1   0.02120 5.9363 -32.099
## - M.F     1   0.12375 6.0388 -31.500
## - Wealth  1   0.22564 6.1407 -30.914
## - LF      1   0.24916 6.1642 -30.780
## - Time    1   0.25440 6.1695 -30.751
## - So      1   0.32256 6.2376 -30.366
## <none>                5.9151 -30.224
## - U1      1   0.47335 6.3884 -29.530
## - Po2     1   0.49275 6.4078 -29.424
## - NW      1   0.60403 6.5191 -28.821
## - Po1     1   0.96445 6.8795 -26.938
## - M       1   1.01291 6.9280 -26.692
## - U2      1   1.06013 6.9752 -26.454
## - Prob    1   1.24604 7.1611 -25.534
## - Ineq    1   1.41096 7.3261 -24.737
## - Ed      1   1.58943 7.5045 -23.895
##
## Step:  AIC=-32.1
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + NW + U1 + U2 + Wealth + 
##     Ineq + Prob + Time
##
##          Df Sum of Sq    RSS     AIC
## - Wealth  1   0.20450 6.1408 -32.914
## - M.F     1   0.20502 6.1413 -32.911
## - LF      1   0.28165 6.2179 -32.477
## - Time    1   0.31951 6.2558 -32.264
## - So      1   0.33842 6.2747 -32.159
## <none>                5.9363 -32.099
## - Po2     1   0.53866 6.4749 -31.059
## - U1      1   0.56478 6.5011 -30.918
## - NW      1   0.63149 6.5678 -30.561
## - Po1     1   0.98759 6.9239 -28.713
## - M       1   1.01798 6.9543 -28.560
## - U2      1   1.16903 7.1053 -27.808
## - Prob    1   1.36479 7.3011 -26.856
## - Ineq    1   1.51330 7.4496 -26.152
## - Ed      1   1.59325 7.5295 -25.778
##
## Step:  AIC=-32.91
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + NW + U1 + U2 + Ineq + 
##     Prob + Time
##
##        Df Sum of Sq    RSS     AIC
## - LF    1   0.23375 6.3745 -33.606
## - M.F   1   0.24988 6.3907 -33.518
## - So    1   0.26395 6.4047 -33.441
## - Time  1   0.30657 6.4474 -33.209
## <none>              6.1408 -32.914
## - Po2   1   0.56982 6.7106 -31.808
## - NW    1   0.58033 6.7211 -31.753
## - U1    1   0.67425 6.8150 -31.268
## - M     1   0.83502 6.9758 -30.451
## - Po1   1   1.09608 7.2369 -29.166
## - Ineq  1   1.43221 7.5730 -27.577
## - U2    1   1.49987 7.6407 -27.265
## - Ed    1   1.70227 7.8431 -26.350
## - Prob  1   1.89704 8.0378 -25.492
##
## Step:  AIC=-33.61
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + NW + U1 + U2 + Ineq + 
##     Prob + Time
##
##        Df Sum of Sq    RSS     AIC
## - So    1   0.08974 6.4643 -35.117
## - M.F   1   0.09349 6.4680 -35.097
## - Time  1   0.23380 6.6083 -34.346
## <none>              6.3745 -33.606
## - NW    1   0.40052 6.7751 -33.474
## - Po2   1   0.42739 6.8019 -33.335
## - U1    1   0.47096 6.8455 -33.112
## - Po1   1   0.93265 7.3072 -30.827
## - M     1   0.94170 7.3162 -30.784
## - Ineq  1   1.23170 7.6062 -29.423
## - U2    1   1.47391 7.8484 -28.326
## - Ed    1   1.48936 7.8639 -28.257
## - Prob  1   1.66883 8.0434 -27.468
##
## Step:  AIC=-35.12
## Crime ~ M + Ed + Po1 + Po2 + M.F + NW + U1 + U2 + Ineq + Prob +
##     Time
##
##        Df Sum of Sq    RSS     AIC
## - M.F   1   0.11345 6.5777 -36.508
## - Time  1   0.18653 6.6508 -36.121
## - NW    1   0.31354 6.7778 -35.459
## <none>              6.4643 -35.117
## - U1    1   0.40334 6.8676 -34.999
## - Po2   1   0.40643 6.8707 -34.983
## - M     1   0.86293 7.3272 -32.731
## - Po1   1   0.91469 7.3790 -32.485
## - Ineq  1   1.14522 7.6095 -31.408
## - U2    1   1.40000 7.8643 -30.256
## - Ed    1   1.49206 7.9563 -29.848
## - Prob  1   1.60747 8.0717 -29.344
##
## Step:  AIC=-36.51
## Crime ~ M + Ed + Po1 + Po2 + NW + U1 + U2 + Ineq + Prob + Time
##
##        Df Sum of Sq    RSS     AIC
## - U1    1   0.34008 6.9178 -36.744
## - NW    1   0.35244 6.9302 -36.681
## <none>              6.5777 -36.508
## - Time  1   0.41236 6.9901 -36.380
## - Po2   1   0.73270 7.3104 -34.812
## - M     1   1.13193 7.7096 -32.951
## - Ineq  1   1.30974 7.8875 -32.153
## - U2    1   1.43802 8.0157 -31.588
## - Po1   1   1.47809 8.0558 -31.413
## - Prob  1   1.98050 8.5582 -29.296
## - Ed    1   2.35153 8.9292 -27.811
##
## Step:  AIC=-36.74
## Crime ~ M + Ed + Po1 + Po2 + NW + U2 + Ineq + Prob + Time
##
##        Df Sum of Sq    RSS     AIC
## - Time  1   0.27161 7.1894 -37.396
## - NW    1   0.33529 7.2531 -37.087
## <none>              6.9178 -36.744
## - Po2   1   0.58958 7.5074 -35.881
## - M     1   1.04284 7.9606 -33.829
## - Po1   1   1.34160 8.2594 -32.540
## - Ineq  1   1.40076 8.3186 -32.290
## - U2    1   1.55033 8.4681 -31.666
## - Prob  1   1.96793 8.8857 -29.982
## - Ed    1   2.01701 8.9348 -29.789
##
## Step:  AIC=-37.4
## Crime ~ M + Ed + Po1 + Po2 + NW + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS     AIC
## - NW    1   0.16380  7.3532 -38.607
## - Po2   1   0.40615  7.5955 -37.473
## <none>               7.1894 -37.396
## - Po1   1   1.10717  8.2966 -34.383
## - M     1   1.15774  8.3471 -34.170
## - U2    1   1.57974  8.7691 -32.444
## - Ineq  1   1.80251  8.9919 -31.566
## - Prob  1   1.84839  9.0378 -31.388
## - Ed    1   3.14353 10.3329 -26.700
##
## Step:  AIC=-38.61
## Crime ~ M + Ed + Po1 + Po2 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS     AIC
## - Po2   1   0.29699  7.6502 -39.222
## <none>               7.3532 -38.607
## - Po1   1   0.98383  8.3370 -36.212
## - U2    1   1.48569  8.8389 -34.167
## - Prob  1   1.69238  9.0456 -33.358
## - M     1   1.99652  9.3497 -32.200
## - Ineq  1   2.94473 10.2979 -28.819
## - Ed    1   2.98589 10.3391 -28.680
##
## Step:  AIC=-39.22
## Crime ~ M + Ed + Po1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS     AIC
## <none>               7.6502 -39.222
## - U2    1    1.5149  9.1651 -34.898
## - Prob  1    1.7214  9.3716 -34.118
## - M     1    2.1178  9.7680 -32.668
## - Ed    1    3.1549 10.8051 -29.137
## - Ineq  1    3.4791 11.1293 -28.102
## - Po1   1    8.4631 16.1133 -15.150
```

```r
# Print the final model summary
summary(bfinal_model)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20493 -0.28949  0.03786  0.27846  1.22823
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.70315    1.33303  -2.778  0.00965 **
## M            0.26587    0.09549   2.784  0.00951 **
## Ed           0.55755    0.16408   3.398  0.00205 **
## Po1          0.75272    0.13525   5.566 5.93e-06 ***
## U2           0.24050    0.10214   2.355  0.02578 *
## Ineq         0.61758    0.17307   3.568  0.00132 **
## Prob        -0.32704    0.13029  -2.510  0.01813 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5227 on 28 degrees of freedom
## Multiple R-squared:  0.775,  Adjusted R-squared:  0.7268
## F-statistic: 16.07 on 6 and 28 DF,  p-value: 6.387e-08
```

```
model.full <- lm(Crime ~ ., data = trainData)
model.null <- lm(Crime ~ 1, data = trainData)

MASS::stepAIC(model.null, direction = "forward", scope = list(lower = model.null,
                                                              upper = model.full))
```

```
## Start:  AIC=0.99
## Crime ~ 1
##
##          Df Sum of Sq    RSS      AIC
## + Po1     1   18.7432 15.257 -25.0614
## + Po2     1   17.5438 16.456 -22.4126
## + Prob    1   10.7509 23.249 -10.3179
## + Wealth  1    8.3821 25.618  -6.9220
## + Pop     1    4.8033 29.197  -2.3453
## + Ed      1    2.8305 31.169  -0.0568
## + M.F     1    2.3392 31.661   0.4906
## + U2      1    2.1609 31.839   0.6871
## <none>                34.000   0.9854
## + Ineq    1    1.6600 32.340   1.2335
## + Time    1    1.0723 32.928   1.8639
## + LF      1    0.8504 33.150   2.0989
## + M       1    0.7670 33.233   2.1868
## + So      1    0.4538 33.546   2.5151
## + NW      1    0.0186 33.981   2.9663
## + U1      1    0.0017 33.998   2.9837
##
## Step:  AIC=-25.06
## Crime ~ Po1
##
##          Df Sum of Sq    RSS     AIC
## + Ineq    1   2.28955 12.967 -28.752
## + M       1   1.98976 13.267 -27.952
## + M.F     1   1.39549 13.861 -26.419
## + NW      1   1.33175 13.925 -26.258
## + So      1   1.22507 14.032 -25.991
## + Po2     1   1.20356 14.053 -25.937
## <none>                15.257 -25.061
## + U2      1   0.55966 14.697 -24.369
## + Wealth  1   0.42052 14.836 -24.040
## + Prob    1   0.23814 15.019 -23.612
## + LF      1   0.17546 15.081 -23.466
## + Time    1   0.14481 15.112 -23.395
## + Ed      1   0.13894 15.118 -23.382
## + U1      1   0.06581 15.191 -23.213
## + Pop     1   0.04966 15.207 -23.175
##
## Step:  AIC=-28.75
## Crime ~ Po1 + Ineq
##
##          Df Sum of Sq    RSS     AIC
## + M.F     1   2.13432 10.833 -33.047
## + Ed      1   1.21092 11.756 -30.184
## + Prob    1   1.05379 11.913 -29.719
## + Wealth  1   0.95612 12.011 -29.433
## + Pop     1   0.72660 12.241 -28.771
## <none>                12.967 -28.752
## + M       1   0.64752 12.320 -28.545
## + LF      1   0.61741 12.350 -28.460
## + Po2     1   0.53145 12.436 -28.217
## + U1      1   0.15973 12.807 -27.186
## + U2      1   0.15516 12.812 -27.174
## + NW      1   0.00757 12.960 -26.773
## + Time    1   0.00198 12.965 -26.758
## + So      1   0.00113 12.966 -26.755
##
## Step:  AIC=-33.05
## Crime ~ Po1 + Ineq + M.F
##
##          Df Sum of Sq     RSS     AIC
## + Prob    1   1.41226  9.4206 -35.936
## + M       1   0.61142 10.2215 -33.080
## <none>                10.8329 -33.047
## + Wealth  1   0.57326 10.2596 -32.950
## + Time    1   0.56271 10.2702 -32.914
## + NW      1   0.15713 10.6758 -31.558
## + Ed      1   0.12745 10.7054 -31.461
## + U2      1   0.09473 10.7382 -31.354
## + Po2     1   0.09160 10.7413 -31.344
## + So      1   0.08386 10.7490 -31.319
## + U1      1   0.01778 10.8151 -31.104
## + Pop     1   0.01770 10.8152 -31.104
## + LF      1   0.00066 10.8322 -31.049
##
## Step:  AIC=-35.94
## Crime ~ Po1 + Ineq + M.F + Prob
##
##          Df Sum of Sq    RSS     AIC
## + M       1   0.54216 8.8785 -36.010
## <none>                9.4206 -35.936
## + NW      1   0.38451 9.0361 -35.394
## + U2      1   0.19374 9.2269 -34.663
## + Ed      1   0.18579 9.2348 -34.633
## + So      1   0.15777 9.2629 -34.527
## + Wealth  1   0.09815 9.3225 -34.302
## + LF      1   0.07683 9.3438 -34.222
## + Pop     1   0.06643 9.3542 -34.183
## + Po2     1   0.06202 9.3586 -34.167
## + Time    1   0.02180 9.3988 -34.017
## + U1      1   0.00787 9.4128 -33.965
##
## Step:  AIC=-36.01
## Crime ~ Po1 + Ineq + M.F + Prob + M
##
##          Df Sum of Sq    RSS     AIC
## + Ed      1   0.49800 8.3805 -36.031
## <none>                8.8785 -36.010
## + Wealth  1   0.46800 8.4105 -35.905
## + U2      1   0.40606 8.4724 -35.649
## + U1      1   0.08275 8.7957 -34.338
## + NW      1   0.07563 8.8028 -34.310
## + LF      1   0.05203 8.8264 -34.216
## + Po2     1   0.04645 8.8320 -34.194
## + So      1   0.02245 8.8560 -34.099
## + Pop     1   0.02132 8.8571 -34.094
## + Time    1   0.00442 8.8741 -34.028
##
## Step:  AIC=-36.03
## Crime ~ Po1 + Ineq + M.F + Prob + M + Ed
##
##          Df Sum of Sq    RSS     AIC
## + U2      1   1.03369 7.3468 -38.638
## + Wealth  1   0.49589 7.8846 -36.165
## <none>                8.3805 -36.031
## + LF      1   0.25396 8.1265 -35.108
## + U1      1   0.19058 8.1899 -34.836
## + Po2     1   0.09322 8.2873 -34.422
## + NW      1   0.08417 8.2963 -34.384
## + So      1   0.06733 8.3131 -34.313
## + Pop     1   0.04021 8.3403 -34.199
## + Time    1   0.00329 8.3772 -34.044
##
## Step:  AIC=-38.64
## Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2
##
##          Df Sum of Sq    RSS     AIC
## <none>                7.3468 -38.638
## + U1      1   0.35122 6.9956 -38.353
## + Wealth  1   0.26780 7.0790 -37.938
## + Po2     1   0.14939 7.1974 -37.357
## + NW      1   0.12013 7.2267 -37.215
## + Pop     1   0.05039 7.2964 -36.879
## + So      1   0.00085 7.3459 -36.642
## + LF      1   0.00070 7.3461 -36.641
## + Time    1   0.00000 7.3468 -36.638

##
```

```
## Call:
## lm(formula = Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2, data = trainData)
##
## Coefficients:
## (Intercept)          Po1         Ineq          M.F         Prob            M
##     -3.2862       0.7595       0.5647       0.1229      -0.3292       0.2359
##          Ed           U2
##      0.4225       0.2077
```

```
model_1 <- lm(Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2, data = trainData)

# Validate the model using the validation data
predicted_validation <- predict(model_1, newdata = validationData)
validation_rmse <- sqrt(mean((predicted_validation - validationData$Crime)^2))
cat("Validation RMSE:", validation_rmse, "\n")
```

```
## Validation RMSE: 0.532245
```

```
# Test the model using the test data
predicted_test <- predict(model_1, newdata = testData)
test_rmse <- sqrt(mean((predicted_test - testData$Crime)^2))
cat("Test RMSE:", test_rmse, "\n")
```

```
## Test RMSE: 0.5849555
```

```
# Calculate R-squared for validation and test sets
validation_r_squared <- 1 - (sum((validationData$Crime - predicted_validation)^2) / sum((validationData$
test_r_squared <- 1 - (sum((testData$Crime - predicted_test)^2) / sum((testData$Crime - mean(testData$Cr

# Adjusted R-squared
n <- nrow(trainData)
p <- length(coef(model_1)) - 1  # Number of predictors excluding intercept
adjusted_r_squared <- 1 - ((1 - validation_r_squared) * (n - 1) / (n - p - 1))

# AIC and BIC
aic <- AIC(model_1)
bic <- BIC(model_1)

# Print R-squared, Adjusted R-squared, AIC, and BIC
cat("Validation R-squared:", validation_r_squared, "\n")
```

```
## Validation R-squared: 0.6883868
```

```
cat("Test R-squared:", test_r_squared, "\n")
```

```
## Test R-squared: 0.6267205
```

```
cat("Adjusted R-squared:", adjusted_r_squared, "\n")
```

```
## Adjusted R-squared: 0.6075982
```

```
cat("AIC:", aic, "\n")
```

```
## AIC: 62.68767
```

```
cat("BIC:", bic, "\n")
```

```
## BIC: 76.6858
```

```
# Coefficients, t-stats, and p-values
coef_table <- summary(model_1)$coefficients
coef_table <- as.data.frame(coef_table)
cat("\nCoefficients, t-stats, and p-values:\n")
```

```
##
## Coefficients, t-stats, and p-values:
```

```
print(coef_table)
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -3.2862088  1.3876550 -2.368174 2.529356e-02
## Po1          0.7595246  0.1351230  5.620988 5.775352e-06
## Ineq         0.5647002  0.1798274  3.140235 4.062964e-03
## M.F          0.1229404  0.1164238  1.055973 3.003380e-01
## Prob        -0.3292050  0.1300392 -2.531584 1.748508e-02
## M            0.2359329  0.0994252  2.372969 2.502495e-02
## Ed           0.4225368  0.2077462  2.033909 5.189432e-02
## U2           0.2076927  0.1065594  1.949078 6.174568e-02
```

```
# Train a full simple linear regression model using the full data
model_1_data <- lm(Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2, data = data)

#Predict given data using full model
predit_data <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640, M.F = 94.0, 

new_predict_model <- predict(model_1_data, predit_data)
new_predict_model
```

```
##        1
## 1273.781
```

```
model_2 <- lm(Crime ~ M.F + Prob + M + Ed + U2, data = trainData)

# Validate the model using the validation data
predicted_validation <- predict(model_2, newdata = validationData)
validation_rmse <- sqrt(mean((predicted_validation - validationData$Crime)^2))
cat("Validation RMSE:", validation_rmse, "\n")
```

```
## Validation RMSE: 1.090326
```

```
# Test the model using the test data
predicted_test <- predict(model_2, newdata = testData)
test_rmse <- sqrt(mean((predicted_test - testData$Crime)^2))
cat("Test RMSE:", test_rmse, "\n")
```

```
## Test RMSE: 1.081525
```

```
# Calculate R-squared for validation and test sets
validation_r_squared <- 1 - (sum((validationData$Crime - predicted_validation)^2) / sum((validationData
test_r_squared <- 1 - (sum((testData$Crime - predicted_test)^2) / sum((testData$Crime - mean(testData$C

# Adjusted R-squared
n <- nrow(trainData)
p <- length(coef(model_2)) - 1  # Number of predictors excluding intercept
adjusted_r_squared <- 1 - ((1 - validation_r_squared) * (n - 1) / (n - p - 1))
```

```
# AIC and BIC
aic <- AIC(model_2)
bic <- BIC(model_2)

# Print R-squared, Adjusted R-squared, AIC, and BIC
cat("Validation R-squared:", validation_r_squared, "\n")
```

```
## Validation R-squared: -0.3076913
```

```
cat("Test R-squared:", test_r_squared, "\n")
```

```
## Test R-squared: -0.2760315
```

```
cat("Adjusted R-squared:", adjusted_r_squared, "\n")
```

```
## Adjusted R-squared: -0.5331554
```

```
cat("AIC:", aic, "\n")
```

```
## AIC: 89.28864
```

```
cat("BIC:", bic, "\n")
```

```
## BIC: 100.1761
```

```
# Coefficients, t-stats, and p-values
coef_table <- summary(model_2)$coefficients
coef_table <- as.data.frame(coef_table)
cat("\nCoefficients, t-stats, and p-values:\n")
```

```
##
## Coefficients, t-stats, and p-values:
```

```
print(coef_table)
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.9911024  2.0626998 -1.4500910 0.1577633276
## M.F          0.1742759  0.1664398  1.0470811 0.3037112281
## Prob        -0.5970307  0.1530939 -3.8997669 0.0005248512
## M            0.2147458  0.1477890  1.4530569 0.1569440232
## Ed           0.1955601  0.2321434  0.8424108 0.4064521991
## U2           0.3443343  0.1525585  2.2570646 0.0317098847
```

```
model_3 <- lm(Crime ~ ., data = trainData)

# Validate the model using the validation data
predicted_validation <- predict(model_3, newdata = validationData)
validation_rmse <- sqrt(mean((predicted_validation - validationData$Crime)^2))
cat("Validation RMSE:", validation_rmse, "\n")
```

```
## Validation RMSE: 0.5397676
```

```
# Test the model using the test data
predicted_test <- predict(model_3, newdata = testData)
test_rmse <- sqrt(mean((predicted_test - testData$Crime)^2))
cat("Test RMSE:", test_rmse, "\n")
```

```
## Test RMSE: 0.5735339
```

```
# Calculate R-squared for validation and test sets
validation_r_squared <- 1 - (sum((validationData$Crime - predicted_validation)^2) / sum((validationData$
test_r_squared <- 1 - (sum((testData$Crime - predicted_test)^2) / sum((testData$Crime - mean(testData$C

# Adjusted R-squared
n <- nrow(trainData)
p <- length(coef(model_2)) - 1  # Number of predictors excluding intercept
adjusted_r_squared <- 1 - ((1 - validation_r_squared) * (n - 1) / (n - p - 1))

# AIC and BIC
aic <- AIC(model_3)
bic <- BIC(model_3)

# Print R-squared, Adjusted R-squared, AIC, and BIC
cat("Validation R-squared:", validation_r_squared, "\n")
```

```
## Validation R-squared: 0.679516
```

```
cat("Test R-squared:", test_r_squared, "\n")
```

```
## Test R-squared: 0.6411551
```

```
cat("Adjusted R-squared:", adjusted_r_squared, "\n")
```

```
## Adjusted R-squared: 0.6242601
```

```
cat("AIC:", aic, "\n")
```

```
## AIC: 71.10122
```

```
cat("BIC:", bic, "\n")
```

```
## BIC: 97.54214
```

```
# Coefficients, t-stats, and p-values
coef_table <- summary(model_3)$coefficients
coef_table <- as.data.frame(coef_table)
cat("\nCoefficients, t-stats, and p-values:\n")
```

```
##
## Coefficients, t-stats, and p-values:
```

```
print(coef_table)
```

```
##                Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -3.26377899  1.8118747 -1.8013271 0.08754416
## M            0.23432259  0.1299070  1.8037724 0.08714701
## So          -0.23269497  0.2286065 -1.0178844 0.32151999
## Ed           0.54310135  0.2403610  2.2595238 0.03579301
## Po1          1.99470168  1.1332944  1.7600913 0.09448256
## Po2         -1.51751479  1.2062069 -1.2580883 0.22360181
## LF          -0.17072275  0.1908346 -0.8946110 0.38218833
## M.F          0.12166262  0.1929705  0.6304726 0.53589572
## Pop         -0.04255741  0.1630871 -0.2609489 0.79693937
## NW           0.31792117  0.2282411  1.3929181 0.17972996
## U1          -0.26173665  0.2122633 -1.2330752 0.23258662
## U2           0.39002350  0.2113564  1.8453359 0.08063300
## Wealth       0.29354293  0.3448020  0.8513375 0.40518437
## Ineq         0.67895660  0.3189243  2.1288958 0.04656432
## Prob        -0.41778283  0.2088279 -2.0006083 0.05993118
## Time        -0.16393080  0.1813440 -0.9039768 0.37732711
```

One of the most important steps in building the right linear prediction model is selecting the right variables. Some of the variables in the dataset may have positive correlation with crime rates while others may be negative. Before forming any assumptions, I first look at what the data presents to us.

I built a correlation matrix to first explore the data and then relate them to our assumptions.

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

M 1.00 0.58 -0.53 -0.51 -0.51 -0.16 -0.03 -0.28 0.59 -0.22 -0.24 -0.67 0.64 0.36 0.11 -0.09 So 0.58 1.00 -0.70 -0.37 -0.38 -0.51 -0.31 -0.05 0.77 -0.17 0.07 -0.64 0.74 0.53 0.07 -0.09 Ed -0.53 -0.70 1.00 0.48 0.50 0.56 0.44 -0.02 -0.66 0.02 -0.22 0.74 -0.77 -0.39 -0.25 0.32 Po1 -0.51 -0.37 0.48 1.00 0.99 0.12 0.03 0.53 -0.21 -0.04 0.19 0.79 -0.63 -0.47 0.10 0.69 Po2 -0.51 -0.38 0.50 0.99 1.00 0.11 0.02 0.51 -0.22 -0.05 0.17 0.79 -0.65 -0.47 0.08 0.67 LF -0.16 -0.51 0.56 0.12 0.11 1.00 0.51 -0.12 -0.34 -0.23 -0.42 0.29 -0.27 -0.25 -0.12 0.19 M.F -0.03 -0.31 0.44 0.03 0.02 0.51 1.00 -0.41 -0.33 0.35 -0.02 0.18 -0.17 -0.05 -0.43 0.21 Pop -0.28 -0.05 -0.02 0.53 0.51 -0.12 -0.41 1.00 0.10 -0.04 0.27 0.31 -0.13 -0.35 0.46 0.34 NW 0.59 0.77 -0.66 -0.21 -0.22 -0.34 -0.33 0.10 1.00 -0.16 0.08 -0.59 0.68 0.43 0.23 0.03 U1 -0.22 -0.17 0.02 -0.04 -0.05 -0.23 0.35 -0.04 -0.16 1.00 0.75 0.04 -0.06 -0.01 -0.17 -0.05 U2 -0.24 0.07 -0.22 0.19 0.17 -0.42 -0.02 0.27 0.08 0.75 1.00 0.09 0.02 -0.06 0.10 0.18 Wealth -0.67 -0.64 0.74 0.79 0.79 0.29 0.18 0.31 -0.59 0.04 0.09 1.00 -0.88 -0.56 0.00 0.44 Ineq 0.64 0.74 -0.77 -0.63 -0.65 -0.27 -0.17 -0.13 0.68 -0.06 0.02 -0.88 1.00 0.47 0.10 -0.18 Prob 0.36 0.53 -0.39 -0.47 -0.47 -0.25 -0.05 -0.35 0.43 -0.01 -0.06 -0.56 0.47 1.00 -0.44 -0.43 Time 0.11 0.07 -0.25 0.10 0.08 -0.12 -0.43 0.46 0.23 -0.17 0.10 0.00 0.10 -0.44 1.00 0.15 Crime -0.09 -0.09 0.32 0.69 0.67 0.19 0.21 0.34 0.03 -0.05 0.18 0.44 -0.18 -0.43 0.15 1.00

As discussed from lectures, correlation doesn't mean causation. A third variable may explain the interaction between these particular variables and the correlation results may only make sense for one sample. Moreover, not all the variables may be necessary.

Based on prior research, some assumptions that are held about the predictors/variables in the dataset are as follows:

M - Higher percentage of males aged 14-24 could indicate higher crime rates since this demographic has historically accounted for a disproportionate amount of crime So - Southern states may have higher crime rates due to socioeconomic factors Ed - Lower education levels may correlate with higher crime Po1, Po2 - Higher police expenditures likely indicate higher crime rates (more spending in response to crime) LF - Lower male labor force participation could be associated with more crime M.F - More males relative to females links to crime since males have higher criminality Pop - Higher population states have more crime in absolute terms NW - Higher non-white population has been associated with higher crime rates historically due to systemic socioeconomic factors U1, U2 - Higher unemployment rates may be tied to motive for crimes of economic necessity Wealth - Lower median wealth indicates poverty which is a known correlate of higher crime Ineq - More income inequality has been connected to higher crime rates Prob - Higher probability of imprisonment may deter crime (or higher crime leads to more imprisonment) Time - Longer prison times could deter crime but could also indicate high recidivism

Based on the results from the correlation matrix, these are some cases where there seems to be deviations from initial assumptions: a negative correlation between the % of 14-24 year old adults and crime, Southern states, which is also negative, a positive correlation between schooling + labor force participation and crime, lower unemployment and crime for 14-24 year old males shows a negative correlation but a positive correlation for males 35-39.

To observe the behavior of these predictors more closely, I decided to draft scatter plots of a few of them. And based on the results: M: there isn't a clear linear relationship between 14-24 year old and crime Ed: Even though the data is widely scattered I can still see a slight positive trend U1: Finally, I looked at the unemployment rate for these 14-24 year old males, and again the data is scattered but there is a slight negative trend.

While I do not expect this data to be the best representation of the population, since it only has 47 data points vs 15 variable (leading to potential overfitting), we can still perform hypothesis testing on each variable to determine their statistical significance and based on the q-q plot, we can see that the data meets the normality criteria. Also, please note that these are only assumptions and will only help us to better understand the data and inform the model that we are building.

Before proceeding to find our ideal model, we tested new data on the full model (including all the predictors) and found that the model predicted crime at 155 per 100,000, which is a significant deviation from 342 per 100,000, the lowest value in the given data. All the variables may not be necessary, which may explain why our model predicts values outside the expected range.
With these results and assumptions in mind, I decided to train, validate and test my model, using 70-15-15 splitting and cross validation. Since my dataset is really small, I think the use of a larger portion for training and allocating the rest for testing & validation would make the most sense. With a very limited dataset I decided to use cross-validation.

Model Selection:

Using the training set I conducted backward elimination and forward selection on the set of variables provided. Both methods produced the same result (Model below). The AIC of the model is 62.7. I also examined the adjusted R-Square, which showed that 60% of the variation is explained by the model and the F statistics of the model showed that at least one variable is significant. Both the unemployment rate of urban males 35–39 and the probability of imprisonment are statistically significant predictors. What is of particular importance, is whether the variables remaining in the model would make sense in predicting crime and in fact considering cases of multi collinearity, one could find that some of the variables that were removed may have been highly correlated with the remaining variables, thus we have a more simplified model with a better fit. For instance, including per capita expenditure on police protection in both 1959 and 1960 does not improve the model, these two variables are highly correlated with a value of 0.99… neither does including the two unemployment variables, which are highly correlated and could indicate that unemployment is uniform across the population.

Selected Model: lm(formula = Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2, data = trainData)

One final step, that I took was to reduce the model by two more variables, Po1, which is not a direct cause of crime, and secondly, Ineq, since we are already using U2 as the only economic variable. So using the "Selected model" as my baseline model, the reduced model and a full model (includes all the predictors), I compared the models (seen below) on the validation and test set to see how they fare against the baseline. I also recalculated the AIC along with the BIC to get a better idea of their fit and complexity:

model_1 <- lm(Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2, data = trainData) model_2 <- lm(Crime ~ M.F + Prob + M + Ed + U2, data = trainData) model_3 <- lm(Crime ~ ., data = trainData)

The results are as follows:

model 1 (Selected model): Validation RMSE: 0.532 Test RMSE: 0.585 Validation R-squared: 0.688 Test R-squared: 0.627 Adjusted R-squared: 0.608 AIC: 62.7 BIC: 76.7

model 2: Validation RMSE: 1.09 Test RMSE: 1.08 Validation R-squared: -0.308 Test R-squared: -0.276 Adjusted R-squared: -0.533 AIC: 89.3 BIC: 100

model 3: Validation RMSE: 0.54 Test RMSE: 0.574 Validation R-squared: 0.68 Test R-squared: 0.641 Adjusted R-squared: 0.624 AIC: 71.1 BIC: 97.5

The results confirmed that the "Selected model" (model_1) is the best model of the three models. The simpler model (model_2) produced significantly higher AIC and BIC results than the selected model (model 1). With a difference in BIC >10, model 1 (Crime ~ Po1 + Ineq + M.F + Prob + M + Ed + U2) is very likely better than model 2. Removing the variables did not improve our model, in fact it further exposed potential issues with the size of the dataset, overfitting (R-squared reduced from 0.775 on the training model to 0.627 on the test model) and the data quality. Model_3 however, showed valid results for r and adjusted r square, and which are very close to model_1. Still, model_1 is likely better, with a difference in BIC vs model_3 that is >6 but <10. And model_1 has the lowest RMSE of three models, indicating better predictive accuracy and precision.

Finally, I predicted the new data on the Selected model and the prediction is 1274, which falls within the range of the original data. This further proves than we have a more viable model than the full model where we used all the predictors.

To further our analysis we would have to address issues of multi collinearity, where too many of the predictor variables are highly correlated with each other (adding irrelevant or noisy variables to the model can introduce additional sources of error and reduce the model's predictive accuracy). A larger data set may also be a better representation of the true population, and could help to reduce overfitting.