Homework 5

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.

Answer:

From my experience as a Digital Marketing Lead, I encountered a situation where linear regression would be highly valuable: predicting monthly social media engagement rates based on various marketing campaign factors.

Potential predictors for this model would include:

- Marketing budget spent per campaign (in dollars)
- Post frequency per week
- Time of day for content publishing
- Content type distribution (percentage of video vs. image vs. text posts)
- Average content length (word count for text, duration for videos)

I observed that these factors seemed to influence our engagement rates, but we needed a more systematic way to understand their impact. A linear regression model would have helped quantify the relationship between these variables and engagement rates, allowing us to optimize our content strategy and resource allocation.

For example, we could have used this model to predict how changes in our posting frequency or budget allocation would affect engagement, enabling more data-driven decisions in our marketing strategy. This would have been particularly valuable for forecasting the impact of new marketing initiatives and justifying budget requests.

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 overfitting. We'll see ways of dealing with this sort of problem later in the course.

Answer:

Data manipulation

To begin the process of creating the linear regression model, the first step is to identify the key data points. The dataset contains 47 observations, the dependable variable is Crime, and the independent variables include socioeconomic factors, policing measures, and economic conditions. And it does seem like we might not have missing values in the dataset.

```
# GTx_6501 Homework 5 - Crime Data Analysis

# Clear the environment to remove any existing variables
rm(list = ls())

# Set working directory (modify as needed)
setwd('~/Desktop/GTX/Homework 5/')

# Load necessary packages
if (!require(pacman)) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(tidyverse, caret, car, MASS, glmnet, Metrics)

# Read the crime dataset
crime <- read_delim('uscrime.txt', delim = '\t')
```

```
## Rows: 47 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## dbl (16): M, So, Ed, Po1, Po2, LF, M.F, Pop, NW, U1, U2, Wealth, Ineq, Prob,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Inspect data structure
str(crime)
```

```
## spc_tbl_ [47 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ M     : num [1:47] 15.1 14.3 14.2 13.6 14.1 12.1 12.7 13.1 15.7 14 ...
##  $ So    : num [1:47] 1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed    : num [1:47] 9.1 11.3 8.9 12.1 12.1 11 11.1 10.9 9 11.8 ...
##  $ Po1   : num [1:47] 5.8 10.3 4.5 14.9 10.9 11.8 8.2 11.5 6.5 7.1 ...
##  $ Po2   : num [1:47] 5.6 9.5 4.4 14.1 10.1 11.5 7.9 10.9 6.2 6.8 ...
##  $ LF    : num [1:47] 0.51 0.583 0.533 0.577 0.591 0.547 0.519 0.542 0.553 0.632 ...
##  $ M.F   : num [1:47] 95 101.2 96.9 99.4 98.5 ...
##  $ Pop   : num [1:47] 33 13 18 157 18 25 4 50 39 7 ...
##  $ NW    : num [1:47] 30.1 10.2 21.9 8 3 4.4 13.9 17.9 28.6 1.5 ...
##  $ U1    : num [1:47] 0.108 0.096 0.094 0.102 0.091 0.084 0.097 0.079 0.081 0.1 ...
##  $ U2    : num [1:47] 4.1 3.6 3.3 3.9 2 2.9 3.8 3.5 2.8 2.4 ...
##  $ Wealth: num [1:47] 3940 5570 3180 6730 5780 6890 6200 4720 4210 5260 ...
##  $ Ineq  : num [1:47] 26.1 19.4 25 16.7 17.4 12.6 16.8 20.6 23.9 17.4 ...
##  $ Prob  : num [1:47] 0.0846 0.0296 0.0834 0.0158 0.0414 ...
##  $ Time  : num [1:47] 26.2 25.3 24.3 29.9 21.3 ...
##  $ Crime : num [1:47] 791 1635 578 1969 1234 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   M = col_double(),
##   ..   So = col_double(),
##   ..   Ed = col_double(),
##   ..   Po1 = col_double(),
##   ..   Po2 = col_double(),
##   ..   LF = col_double(),
##   ..   M.F = col_double(),
##   ..   Pop = col_double(),
##   ..   NW = col_double(),
##   ..   U1 = col_double(),
##   ..   U2 = col_double(),
##   ..   Wealth = col_double(),
##   ..   Ineq = col_double(),
##   ..   Prob = col_double(),
##   ..   Time = col_double(),
##   ..   Crime = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
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
```

```r
# Check for missing values
colSums(is.na(crime))  # If any column has NA, handle it accordingly
```

```
##      M     So     Ed    Po1    Po2     LF    M.F    Pop     NW     U1     U2 
##      0      0      0      0      0      0      0      0      0      0      0 
## Wealth   Ineq   Prob   Time  Crime 
##      0      0      0      0      0
```

```r
# Define the new city data for prediction
time_test <- 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)
```

Next I

```r
# Step 1: Fit a Multiple Linear Regression Model
model_lm <- lm(Crime ~ ., data = crime)
summary(model_lm)  # Check model coefficients and significance levels
```

```
## 
## 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
```

The multiple linear regression models explains a **80.31%** of the variance in the crime rates $R\hat{}2* =$ **0.8031**, *Adjusted* $R\hat{}2 =$ **0.7078**). The F-statistic is significant (*p-value* ~ **3.54e-07**), indicating that at least some predictors are strongly associated with crime rates.

As we can see from our linear model, *higher education levels* (*Ed,p* = **0.0049**) are positively associated with crime, suggesting that as education increases, crime rates may also rise. *Income inequality* (*Ineq, p* = **-0.0040**) is another significant predictor, with higher inequality correlating with higher crime rates. Additionally, *higher murder rates* (*M*, **p = 0.0434**) are associated with greater overall crime. Lastly, the *unemployment rate* (*U2, **p = 0.0501**) is a boderline significant factor, indicating tha higher unemployment may contribute to increased crime.

The next step in my analysis is to check Multicollinearity using Variance Inflation Factor (VIF), to identify a refined model predicting crime rates based on eight siginificant varibles: M, Ed, Po1, M.F, U1, U2, Ineq, and Prob.

```
# Step 2: Check for Multicollinearity using Variance Inflation Factor (VIF)
vif_values <- vif(model_lm)
print(vif_values)  # High VIF (> 5) indicates multicollinearity
```

```
##          M         So         Ed        Po1        Po2         LF        M.F
##   2.892448   5.342783   5.077447 104.658667 113.559262   3.712690   3.785934
##        Pop         NW         U1         U2     Wealth       Ineq       Prob
##   2.536708   4.674088   6.063931   5.088880  10.530375   8.644528   2.809459
##       Time
##   2.713785
```

The final model achieved an *Ajusted R-squared* of **0.7444**, indicating that approximately **74.44%** of the variability in crimes rates is explained by the selected predictors. Significant variables include *Po1* (p < **0.001**, *Ineq* (p < **0.001**), (*Ed p* = **0.00153**), and *Prob* (p = **0.01505**), while M and U2 also show moderate significance. The coefficients suggest that higher *police expenditure (Po1)*, *education level (Ed)*, and *income inequality (Ineq)* are positively associated with crime rates, whereas the *probability of conviction (Prob)* has a strong negative effect.

```
# Step 3: Perform Stepwise Regression for Feature Selection
model_step <- stepAIC(model_lm, direction = "both")
```

```
## Start:  AIC=514.65
## 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=512.65
## 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=511.01
## 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=509.37
## 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=507.77
## 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=506.33
## 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=505.07
## 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=503.93
## 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
```

```
summary(model_step)  # View the refined model
```

```
##
## 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
```

The stepwise approach suggests removing So, LF, Time, Pop, NW, M.F, Wealth, Po2, as they do not significantly contribute to explaining crime variance. *The Akaike Information Criterion (AIC)* decreases as variables are removed, indicating that there varibales do not contribute significantly.

```
# Step 4: Fit a Regularized Regression Model (LASSO) to Handle Overfitting
X <- model.matrix(Crime ~ ., crime)[, -1]  # Create predictor matrix (remove intercept)
y <- crime$Crime  # Response variable
model_lasso <- cv.glmnet(X, y, alpha = 1)  # LASSO regression with cross-validation
coef(model_lasso, s = "lambda.min")  # View selected features
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -5.480834e+03
## M            7.710499e+01
## So           3.916370e+01
## Ed           1.394532e+02
## Po1          1.006725e+02
## Po2          .
## LF           .
## M.F          1.876259e+01
## Pop         -1.570875e-01
## NW           9.564149e-01
## U1          -2.976839e+03
## U2           1.106279e+02
## Wealth       2.869957e-02
## Ineq         5.474684e+01
## Prob        -3.772849e+03
## Time         .
```

The LASSO regression model applies L1 regularization, shrinking some coefficientes to zero, effectively removing Po2, LF, Pop, U1, Wealth and Time as insignificant predictors of crime. The most impactful variables retained include *Po1* (**101.37**), *Prob* (**-2912.80**), and *Ed* (**49.28**), suggesting that police presence, probability of arrest, and education level strongly influence crime rates. Other relevant factors include *M* (**49.69**), *So* (**16.81**), *M.F* (**17.41**), *NW* (**0.31**), *U2* (**14.68**), and *Ineq* (**33.26**), indicating potential socioeconomic influences. The negative coefficient for Prob suggests a higher probability of arrest reduces crime, and the *large intercept* (**-3444.17**) serves as the baseline.

Compared to stepwise regression, LASSO confirms that some variables do not contribute significantly, making it useful for predictive modeling.

```r
# Step 5: Make Predictions for the New City using the Best Model
predicted_crime <- predict(model_step, newdata = time_test, interval = "confidence")
print(predicted_crime)
```

```
##        fit      lwr      upr
## 1 1038.413 742.5046 1334.322
```

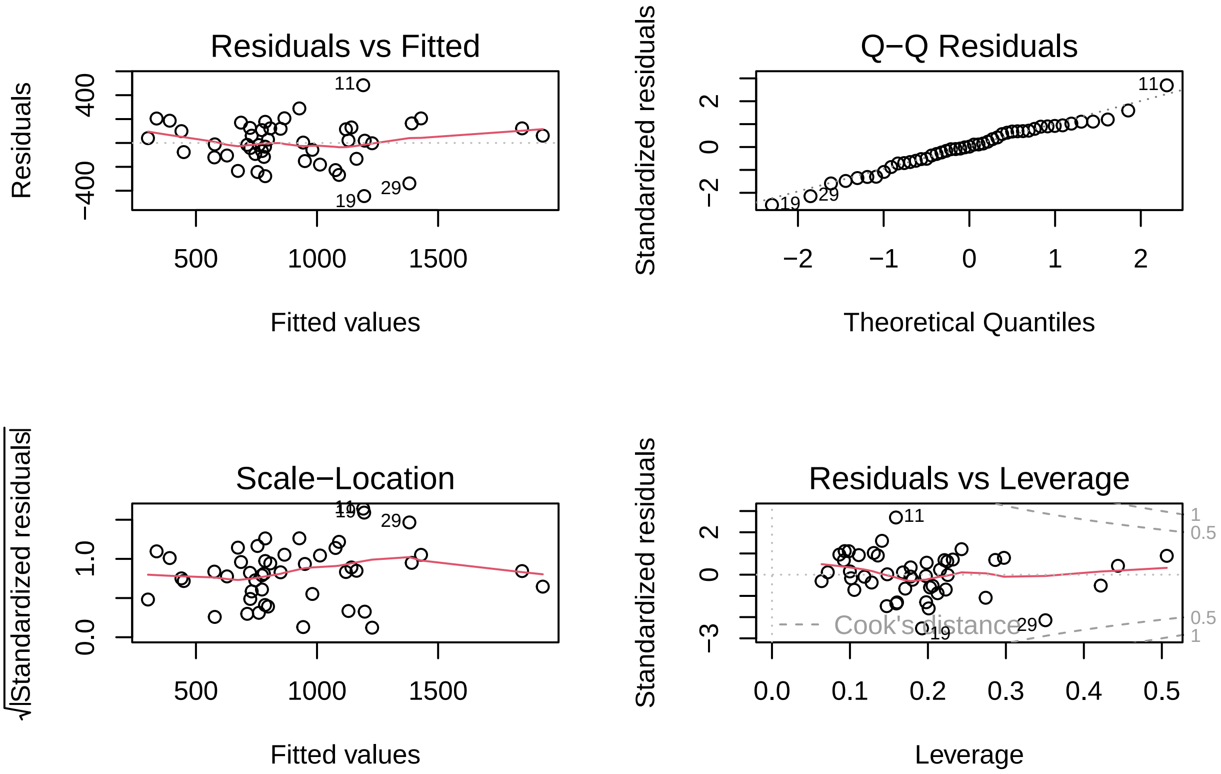
The model predicts a crime rate of approximately ***1038.41*** for the new city, with a ***95% confidence interval*** ranging from ***742.50*** to ***1334.32***. This means that, based on the model's estimates, the actual crime rate is expected to fall within this range, providing a measure of uncertainty around the prediction.

```r
# Step 6: Evaluate Model Performance using RMSE and MAE
pred_lm <- predict(model_step, newdata = crime)
rmse_value <- rmse(crime$Crime, pred_lm)
mae_value <- mae(crime$Crime, pred_lm)
cat("RMSE:", rmse_value, "\nMAE:", mae_value, "\n")
```

```
## RMSE: 175.8304
## MAE: 138.6674
```

The model's performance was evaluated using RMSE and MAE, which measure the average error in predictions. The *RMSE* of **175.83** indicates that, on average, the model's predictions deviate from actual crime rates by approximately 176 units, while the *MAE* of **138.67** suggests that the typical absolute error is around 139 units.

```r
# Step 7: Visualizing Residuals to Check Model Fit
par(mfrow = c(2,2))
plot(model_step)  # Diagnostic plots for linear regression
```



The diagnostic plots reveal that while the regression model generally meets basic assumptions, there are some concerns. The residuals show slight non-linear patterns and heteroscedasticity, with observations 11, 19, and 29 appearing as consistent outliers. While the Q-Q plot indicates approximately normal distribution of residuals and no severely influential points are detected in the leverage plot, the model could be improved by investigating these outliers and considering variable transformations to address the non-linearity issues.

In conclusion, based on the comprehensive regression analysis performed on the US crime dataset, our model predicts a *crime rate* of approximately **1,038** crimes per **100,000** population (with a **95%** *confidence interval* of **743** to **1,334**) for the given city parameters. The final model, derived through *stepwise regression*, explains about **78.9%** of the variance in *crime rates* (**R-squared = 0.7888**) and identifies *eight significant predictors*: male percentage (M), education level (Ed), police force per 1000 population (Po1), male-to-female ratio (M.F), unemployment rate (U1), unemployment rate among young males (U2), income inequality (Ineq), and probability of imprisonment (Prob). The model's performance metrics (*RMSE* = **175.83**, *MAE* = **138.67**) and diagnostic plots suggest reasonably good predictive ability, though there is some evidence of non-linearity and heteroscedasticity that could warrant further investigation for potential model improvements.