Homework 5: Linear Regression

2025-02-11

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.

A linear regression model would be useful for predicting monthly revenue at a gym based on different factors that influence it. Since revenue isn't random, I can use historical data to identify key variables that impact how much the gym earns each month.

Some important predictors I would include are:

1. **Number of active members** – More memberships generally mean higher revenue.
2. **Personal training sessions booked** – Since personal training is a major revenue source, tracking these bookings helps with forecasting.
3. **Average check in frequency per member** – Members who visit more often are more likely to renew and make additional purchases.
4. **Marketing spend** – Promotions and ads influence new sign ups, which affect overall revenue.
5. **Time of year (seasonality)** – Trends show that revenue fluctuates, with peaks in January and slower months in summer.

By using a linear regression model, I can see how these factors impact revenue and make better decisions, like adjusting marketing strategies, scheduling staff efficiently, and planning promotions to boost engagement during slow periods.

How Linear Regression Stands Out:

- **What it's used for:** Linear regression is good for when you're trying to predict a continuous number, like monthly gym revenue. KNN and SVM are usually better for categorizing things (like predicting whether a member will renew or not).
- **How it predicts:** Linear regression finds a straight line relationship between different factors (like memberships, check ins, and revenue). KNN, on the other hand, looks at similar past cases to guess the outcome, while SVM tries to find a boundary that best separates the data.

Which One Should I Use for Gym Revenue?

Since I'm trying to predict a number, and I suspect revenue has a mostly straightforward relationship with things like memberships and marketing spend, linear regression is my best bet. It's simple, clear, and gives me useful insights.

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.

```
rm(list = ls())
#load the uscrime data
data <- read.table('uscrime.txt',stringsAsFactors = FALSE, header = TRUE)

library(corrplot)
```

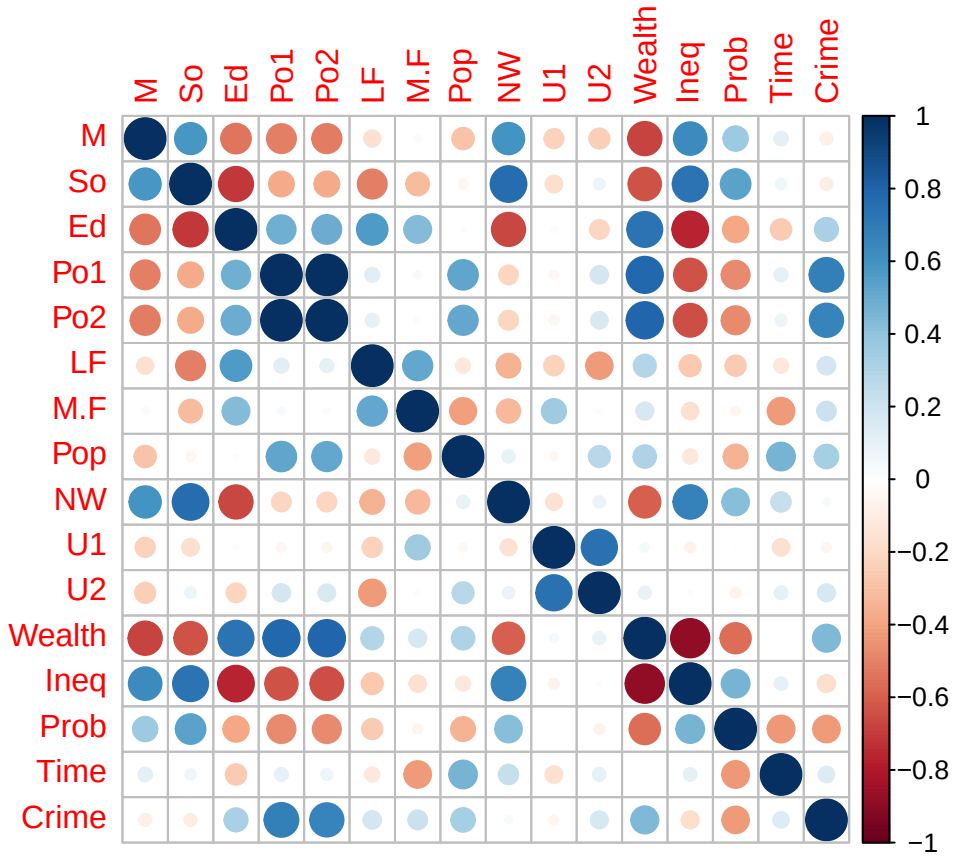
```
## corrplot 0.95 loaded
```

```
cor(data)
```

```
##                  M          So          Ed         Po1         Po2         LF
## M       1.00000000  0.58435534 -0.53023964 -0.50573690 -0.51317336 -0.1609488
## So      0.58435534  1.00000000 -0.70274132 -0.37263633 -0.37616753 -0.5054695
## Ed     -0.53023964 -0.70274132  1.00000000  0.48295213  0.49940958  0.5611780
## Po1    -0.50573690 -0.37263633  0.48295213  1.00000000  0.99358648  0.1214932
## Po2    -0.51317336 -0.37616753  0.49940958  0.99358648  1.00000000  0.1063496
## LF     -0.16094882 -0.50546948  0.56117795  0.12149320  0.10634960  1.0000000
## M.F    -0.02867993 -0.31473291  0.43691492  0.03376027  0.02284250  0.5135588
## Pop    -0.28063762 -0.04991832 -0.01722740  0.52628358  0.51378940 -0.1236722
## NW      0.59319826  0.76710262 -0.66488190 -0.21370878 -0.21876821 -0.3412144
## U1     -0.22438060 -0.17241931  0.01810345 -0.04369761 -0.05171199 -0.2293997
## U2     -0.24484339  0.07169289 -0.21568155  0.18509304  0.16922422 -0.4207625
## Wealth -0.67005506 -0.63694543  0.73599704  0.78722528  0.79426205  0.2946323
## Ineq    0.63921138  0.73718106 -0.76865789 -0.63050025 -0.64815183 -0.2698865
## Prob    0.36111641  0.53086199 -0.38992286 -0.47324704 -0.47302729 -0.2500861
## Time    0.11451072  0.06681283 -0.25397355  0.10335774  0.07562665 -0.1236404
## Crime  -0.08947240 -0.09063696  0.32283487  0.68760446  0.66671414  0.1888663
##                M.F         Pop          NW           U1          U2
## M      -0.02867993 -0.28063762  0.59319826 -0.224380599 -0.24484339
## So     -0.31473291 -0.04991832  0.76710262 -0.172419305  0.07169289
## Ed      0.43691492 -0.01722740 -0.66488190  0.018103454 -0.21568155
## Po1     0.03376027  0.52628358 -0.21370878 -0.043697608  0.18509304
## Po2     0.02284250  0.51378940 -0.21876821 -0.051711989  0.16922422
## LF      0.51355879 -0.12367222 -0.34121444 -0.229399684 -0.42076249
## M.F     1.00000000 -0.41062750 -0.32730454  0.351891900 -0.01869169
## Pop    -0.41062750  1.00000000  0.09515301 -0.038119948  0.27042159
## NW     -0.32730454  0.09515301  1.00000000 -0.156450020  0.08090829
## U1      0.35189190 -0.03811995 -0.15645002  1.000000000  0.74592482
## U2     -0.01869169  0.27042159  0.08090829  0.745924815  1.00000000
## Wealth  0.17960864  0.30826271 -0.59010707  0.044857202  0.09207166
## Ineq   -0.16708869 -0.12629357  0.67731286 -0.063832178  0.01567818
## Prob   -0.05085826 -0.34728906  0.42805915 -0.007469032 -0.06159247
## Time   -0.42769738  0.46421046  0.23039841 -0.169852838  0.10135833
## Crime   0.21391426  0.33747406  0.03259884 -0.050477918  0.17732065
##              Wealth        Ineq         Prob          Time       Crime
## M      -0.6700550558  0.63921138  0.361116408  0.1145107190 -0.08947240
```

```
## So      -0.6369454328  0.73718106  0.530861993  0.0668128312 -0.09063696
## Ed       0.7359970363 -0.76865789 -0.389922862 -0.2539735471  0.32283487
## Po1      0.7872252807 -0.63050025 -0.473247036  0.1033577449  0.68760446
## Po2      0.7942620503 -0.64815183 -0.473027293  0.0756266536  0.66671414
## LF       0.2946323090 -0.26988646 -0.250086098 -0.1236404364  0.18886635
## M.F      0.1796086363 -0.16708869 -0.050858258 -0.4276973791  0.21391426
## Pop      0.3082627091 -0.12629357 -0.347289063  0.4642104596  0.33747406
## NW      -0.5901070652  0.67731286  0.428059153  0.2303984071  0.03259884
## U1       0.0448572017 -0.06383218 -0.007469032 -0.1698528383 -0.05047792
## U2       0.0920716601  0.01567818 -0.061592474  0.1013583270  0.17732065
## Wealth   1.0000000000 -0.88399728 -0.555334708  0.0006485587  0.44131995
## Ineq    -0.8839972758  1.00000000  0.465321920  0.1018228182 -0.17902373
## Prob    -0.5553347075  0.46532192  1.000000000 -0.4362462614 -0.42742219
## Time     0.0006485587  0.10182282 -0.436246261  1.0000000000  0.14986606
## Crime    0.4413199490 -0.17902373 -0.427422188  0.1498660617  1.00000000
```

```
# check each colume's correlation with other column
corrplot(cor(data))
```

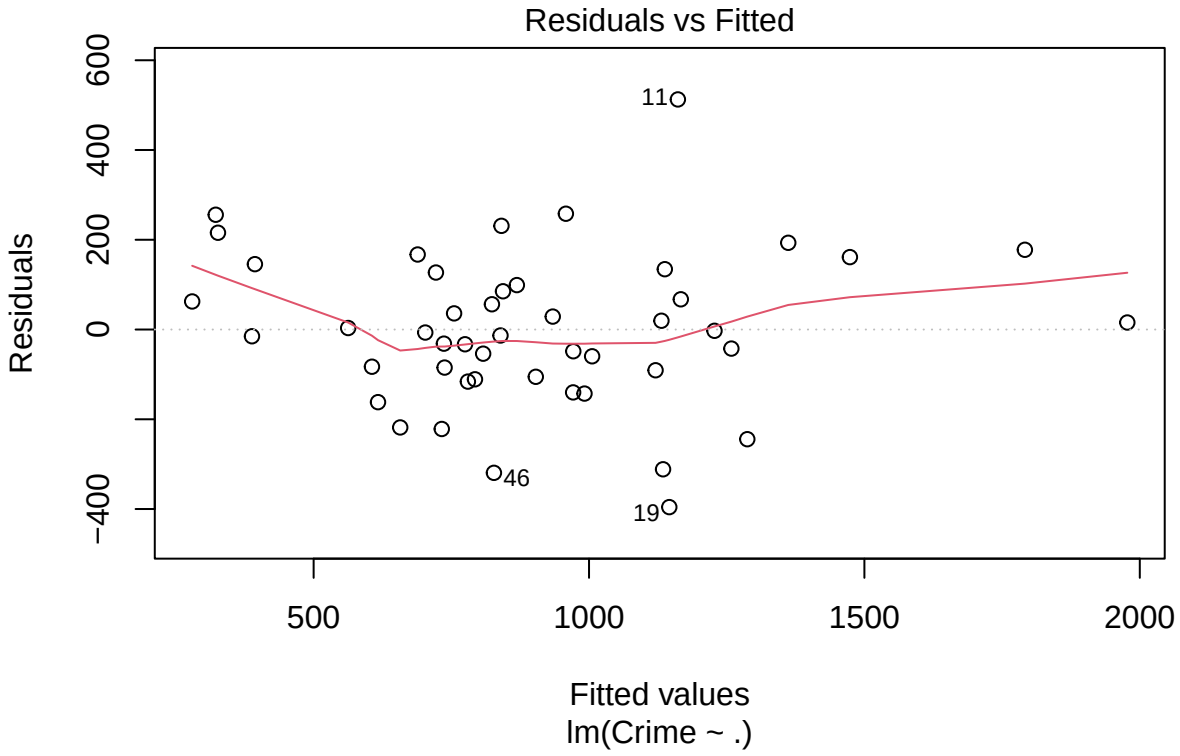


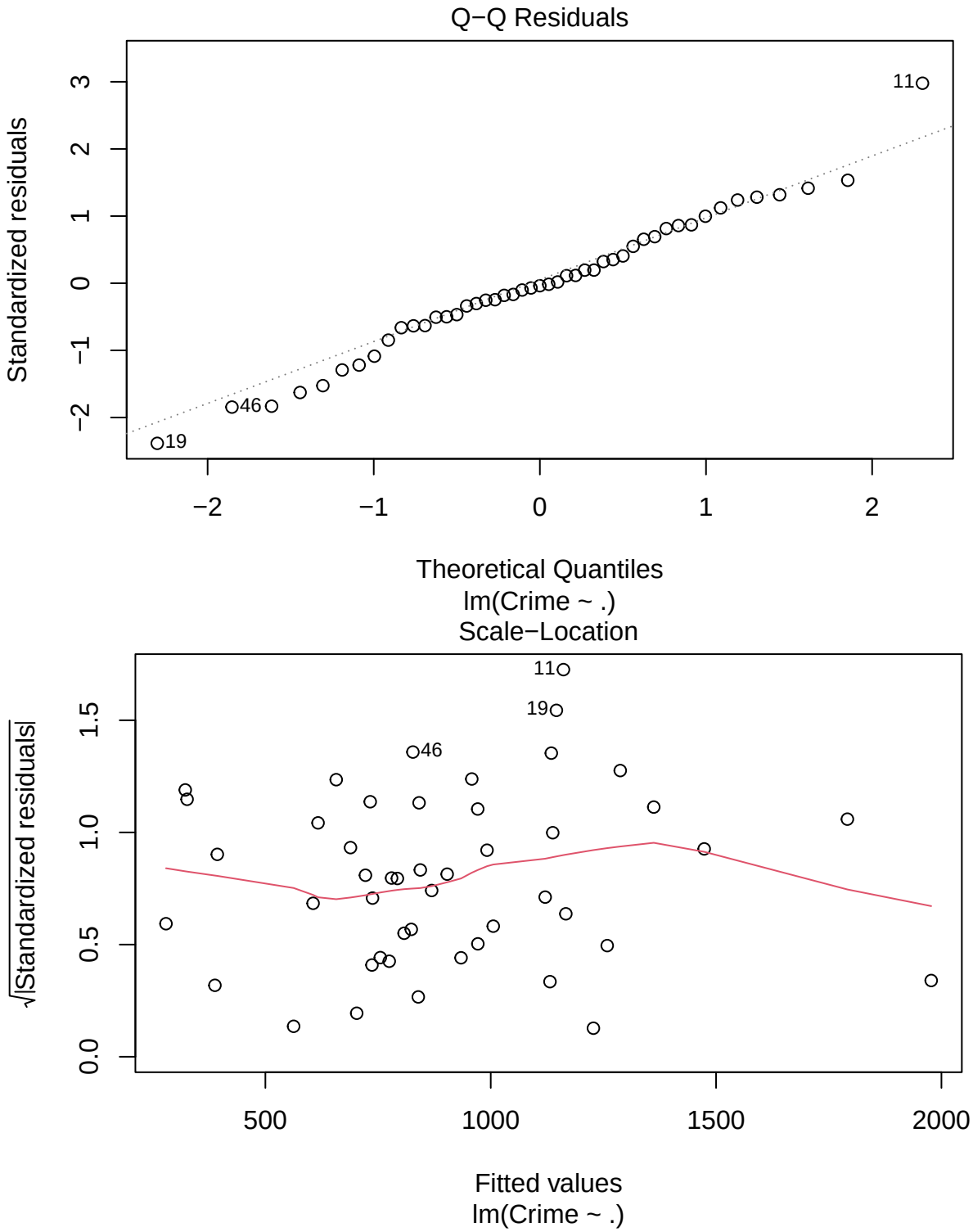

```
#linear regression model by using all crime data
data_lm <- lm(Crime~., data = data)
summary(data_lm)
```

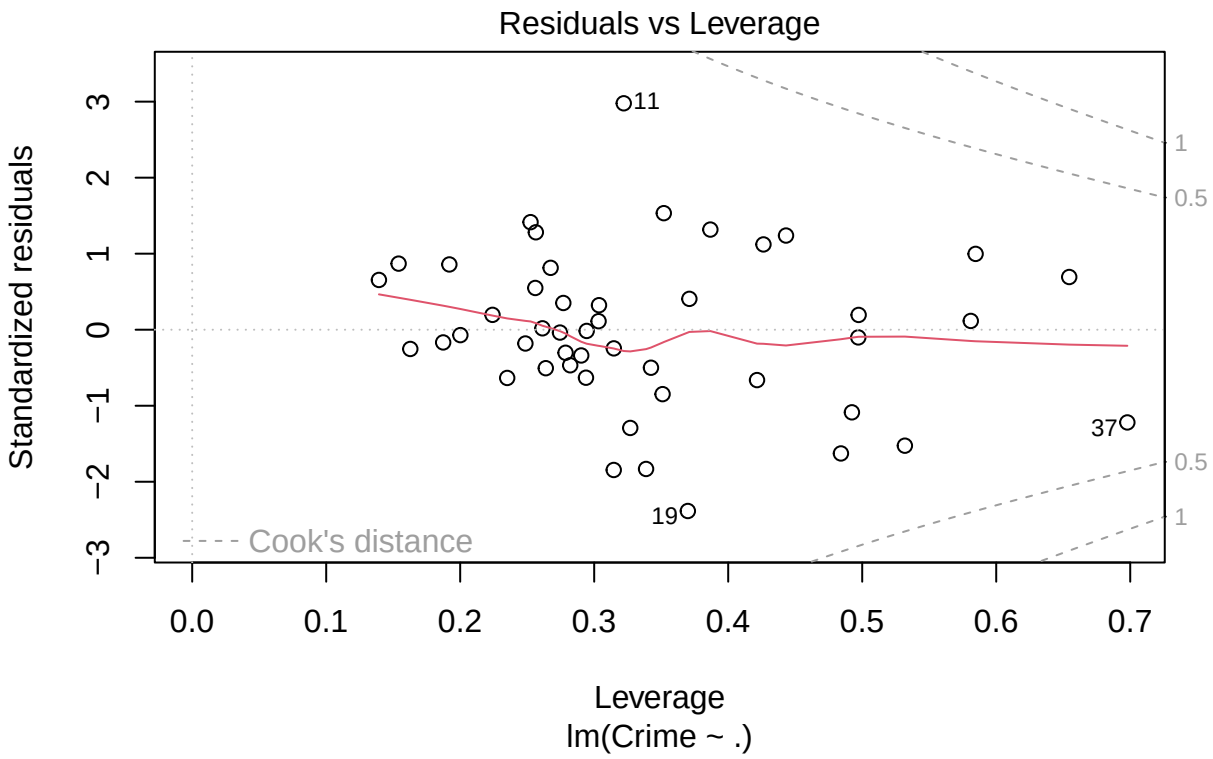
```
##
## Call:
## lm(formula = Crime ~ ., data = data)
##
## Residuals:
```

```
##     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
```

```
plot(data_lm)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Crime ~ .)
Scale−Location
√|Standardized residuals|
Fitted values
lm(Crime ~ .)
11
19
46




```
#create a data frame by using given data.
data_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, U
#predict crime rate by using given data
model_predict <- predict(data_lm, data_test)
model_predict
```

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

```
# the output number is quite lower than the lowest city's, we will construct
#another model and calculate its output and compare.
# construct model2 by using columns which are more positively correlated to
#crime rate(shown above)
model2 <- lm(Crime ~ Ed+Po1+Po2+LF+M.F+Pop+U2+Wealth+Time, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = Crime ~ Ed + Po1 + Po2 + LF + M.F + Pop + U2 + Wealth +
##     Time, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -571.18 -149.15   13.72  143.96  480.51
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.570e+03  1.734e+03  -2.059   0.0466 *
## Ed           5.020e+01  7.239e+01   0.693   0.4924
## Po1          1.618e+02  1.341e+02   1.207   0.2350
## Po2         -4.100e+01  1.438e+02  -0.285   0.7771
```

```
## LF           5.666e+02  1.545e+03   0.367   0.7160
## M.F          3.217e+01  2.080e+01   1.546   0.1306
## Pop         -2.348e-01  1.600e+00  -0.147   0.8841
## U2           4.032e+01  6.241e+01   0.646   0.5222
## Wealth      -1.890e-01  9.438e-02  -2.003   0.0526 .
## Time         1.064e+01  7.184e+00   1.480   0.1472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.8 on 37 degrees of freedom
## Multiple R-squared:  0.5968, Adjusted R-squared:  0.4987
## F-statistic: 6.085 on 9 and 37 DF,  p-value: 3.264e-05
```

```
model2_predict <-predict(model2, data_test)
model2_predict
```

```
##        1
## 1544.587
```

Predicting Crime Rates Using Regression Models

Crime is influenced by a mix of social, economic, and demographic factors. In this analysis, we used regression models to see if we could predict crime rates based on data from 47 U.S. cities. We tested two models: one that included all available factors and another that focused only on the strongest predictors.

Building the Models

We started with a **full model** that included all **15 factors**, like education levels, income inequality, and arrest probability. The model explained about **80.3% of the variation** in crime rates ($\mathbf{R^2 = 0.8031}$), which is pretty good. But when we looked closer, some of the factors had little to no impact (P value of $>$ 0.05), meaning the model might be overfitting, picking up on noise rather than real trends.

A few notes from the full model:

1. **Higher education levels (Ed) were linked to higher crime rates**
   - **Check the coefficient for Ed**: In the first model output, the estimate for **Ed** was **196.47**, which is positive. This suggests that as education levels increase, crime rates also tend to increase.
   - **Is it statistically significant?** The p-value for **Ed** was **0.00008** (very low, $p < 0.05$), meaning this relationship is statistically significant.
   - **True**: Higher education levels were seemingly linked to higher crime rates in this model (even though logically that may not be true).
2. **Greater income inequality (Ineq) was strongly tied to higher crime rates**
   - **Check the coefficient for Ineq**: The estimate was **67.65**, which is positive, indicating that more income inequality is associated with higher crime rates.
   - **Is it statistically significant?** The p-value for **Ineq** was **0.0000188**, which is very low ($p < 0.05$), meaning the effect is statistically significant.
   - **True**: Greater income inequality was strongly tied to higher crime rates.
3. **A higher probability of arrest (Prob) was associated with lower crime rates**
   - **Check the coefficient for Prob**: The estimate was **-3801.84**, which is **negative**, suggesting that a higher probability of arrest is linked to lower crime rates.

- **Is it statistically significant?** The p-value for **Prob** was **0.01711**, which is below 0.05, making it statistically significant.

- **True**: A higher probability of arrest was associated with lower crime rates.

Using this model, we predicted the crime rate for a new city, which came out to **155.43**.

Key Takeaways from the Second Model

After refining the full model, we built a second model using only the most relevant predictors. While it doesn't explain as much variation as the full model ($\mathbf{R^2 = 0.5968}$ vs. **0.8031**), it is likely less prone to overfitting.

What Stood Out?

- **Wealth and Crime:** Wealth had a small but negative relationship with crime rates, meaning wealthier areas tend to have slightly lower crime, though the effect wasn't very strong.

- **Arrest Probability (Prob):** Higher chances of getting arrested were still linked to lower crime, consistent with the first model.

- **Education (Ed):** The relationship between education and crime remained unclear, likely due to other factors.

Performance & Accuracy

- The model simplifies things by removing less meaningful predictors, but this comes at the cost of **predictive accuracy**.

- The crime rate prediction for our test city jumped to **1544.59**, significantly higher than the first model's estimate, showing how different models can produce very different results.

- The adjusted R^2 is lower, meaning it doesn't capture the data as well, but it might work better when predicting new data.

Conclusion

The second model is simpler and less likely to overfit, but it struggles with accuracy. It still highlights important factors like income inequality and the probability of arrest, but some relationships like education don't fully align with expectations. This shows that crime is influenced by a mix of social factors that a basic linear model might not fully capture. Plus, things like data limitations could be affecting the results.