Homework 5: Linear Regression Modeling and Crime Rate Prediction

Georgia Institute of Technology, Business Analytics

Introduction to Analytics Modeling

Professor Joel Sokol

September 24, 2025

Files submitted: homework5_answers.pdf (this doc), homework5.R

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.

Situation

At my job, we track daily active users (DAU) on our data platform. Leadership often asks what factors influence usage spikes or drops. A linear regression model would be appropriate to quantify how different factors affect DAU and to estimate future usage patterns.

Predictors

Some predictors I might use are:

1. Day of Week: DAU is typically higher on weekdays than weekends.
2. Number of Deployments Released: new features or fixes may drive engagement.
3. Marketing Campaign Activity: presence/absence of campaigns that drive traffic.
4. System Uptime (%): downtime or slow performance can reduce usage.
5. Average Session Length: longer sessions may correlate with higher DAU.

Why Linear Regression Fits

- DAU is a continuous outcome variable.
- Predictors are a mix of categorical (day of week, campaign activity) and continuous (uptime, session length).
- A linear regression model can help identify which predictors are statistically significant drivers of DAU and quantify their effect sizes.

This model would help leadership understand which operational or behavioral factors drive usage. For example, if session length strongly correlates with DAU, UX improvements could be prioritized. Overfitting may occur if too many predictors are used without regularization. Future work could include interaction terms or time series modeling.

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

Methodology

I used R to fit a linear regression model (lm) on the uscrime.txt dataset. The dataset contains 47 observations and 15 predictors. After skipping the header row and converting all columns to numeric, I fit the model using all predictors to estimate the crime rate (Crime). I then used predict() to estimate the crime rate for a new city profile.

```
>
> # Load the data, skipping the first row (which contains headers as data)
> crime_raw <- read.table("uscrime.txt", header = TRUE)
>
> # Convert all columns to numeric
> crime <- as.data.frame(lapply(crime_raw, as.numeric))
>
> #  View structure of the dataset
> str(crime)
'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    : num  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   : num  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: num  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 : num  791 1635 578 1969 1234 ...
>
```

Results

```
> summary(model)

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

This shows:

- Coefficients for each predictor
- Significance levels (e.g., Ed, Ineq, Prob, M)
- $R^2 = 0.8031$, Adjusted $R^2 = 0.7078$
- Residual standard error = 209.1

```
>
> # Predict crime rate
> predicted_crime <- predict(model, newdata = new_city)
> print(predicted_crime)
       1
155.4349
>
```

Predicted crime rate for the new city: 155.43 crimes per 100,000 population

Discussion of Results

The model explains ~70% of the variance in crime rates, which is strong given the small sample size. Education level, income inequality, and probability of arrest were significant predictors. Some variables (e.g., LF, Pop, Time) were not significant, possibly due to multicollinearity or noise. Overfitting is likely due to the high predictor-to-observation ratio. Future improvements could include:

- Regularization (e.g., Ridge or Lasso)
- Feature selection
- Cross-validation

Ethically, care must be taken when modeling crime data to avoid reinforcing biases or misinterpreting correlations as causation.

REFERENCES

https://www.geeksforgeeks.org/machine-learning/ml-linear-regression/