**Homework 8: “Modeling Crime Rates with Stepwise, Lasso, and Elastic Net Regression**

Georgia Institute of Technology, Business Analytics

Introduction to Analytics Modeling

Professor Joel Sokol

October 16, 2025

Files submitted: homework8\_answers.pdf (this doc), homework8.R

**Question 11.1**

**Using the crime data set uscrime.txt from Questions 8.2, 9.1, and 10.1, build a regression model using:**

# Stepwise regression

1. **Lasso**
2. **Elastic net**

**For Parts 2 and 3, remember to scale the data first – otherwise, the regression coefficients will be on different scales and the constraint won’t have the desired effect.**

**For Parts 2 and 3, use the glmnet function in R.**

**Notes on R:**

* **For the elastic net model, what we called λ in the videos, glmnet calls “alpha”; you can get a range of results by varying alpha from 1 (lasso) to 0 (ridge regression) [and, of course, other values of alpha in between].**
* **In a function call like glmnet(x,y,family=”mgaussian”,alpha=1) the predictors x need to be in R’s matrix format, rather than data frame format. You can convert a data frame to a matrix using as.matrix – for example, x <- as.matrix(data[,1:n-1])**
* **Rather than specifying a value of T, glmnet returns models for a variety of values of T.**

## Methodology

This question required building three regression models using the uscrime.txt dataset:

1. **Stepwise Regression**

* **Tool Used**: stepAIC() from the MASS package in R.
* **Approach**: Started with a full linear model using all predictors. Stepwise selection was performed in both directions (forward and backward) using AIC as the criterion.
* **Evaluation**: Final model selected based on lowest AIC value.

2. **Lasso Regression**

* **Tool Used**: glmnet() with alpha = 1.
* **Preprocessing**: Predictors were standardized using scale() to ensure consistent penalization across features.
* **Evaluation**: Used cv.glmnet() to perform cross-validation and select the optimal λ (lambda.min). Coefficients were extracted at this value.

3. **Elastic Net Regression**

* **Tool Used**: glmnet() with alpha = 0.5.
* **Preprocessing**: Same standardized predictors as Lasso.
* **Evaluation**: Used cv.glmnet() to select optimal λ. Coefficients were extracted at lambda.min.

**Results**

**Stepwise Regression Output:**

**Final Model**: Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob

**Adjusted R²**: 0.7444 **Significant Predictors**:

* Po1 (p < 0.001)
* Ed, Ineq, Prob, M, U2 (p < 0.05)
* U1 (marginal, p ≈ 0.076)
* M.F (not significant)

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 1: Summary output of stepwise\_model showing coefficients, p-values, R², and AIC.**

**Lasso Regression Output:**

**Selected Predictors** (non-zero coefficients at lambda.min):

* M, So, Ed, Po1, LF, M.F, NW, U2, Ineq, Prob **Zeroed Out**:
* Po2, Pop, U1, Wealth, Time

A graph with lines drawn on it

AI-generated content may be incorrect.

**Figure 2: Lasso coefficient path plot — the one with colored lines and “-Log(λ)” on the x-axis.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Figure 3: Output of coef(cv\_lasso, s = "lambda.min") — showing which variables were retained.**

**Elastic Net Regression Output:**

**Selected Predictors**:

* M, So, Ed, Po1, Po2, LF, M.F, Pop, NW, U1, U2, Wealth, Ineq, Prob

**Zeroed Out**:

* Time

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 4: Output of coef(cv\_enet, s = "lambda.min") — showing coefficients at optimal λ.**

## Discussion of Results

## Insights:

* **Po1**, **Ed**, and **Ineq** consistently emerged as strong predictors across all models, suggesting a robust relationship with crime rates.
* **Stepwise regression** offered a clean, interpretable model with high explanatory power.
* **Lasso** effectively performed variable selection, shrinking less informative coefficients to zero.
* **Elastic Net** retained more variables, useful when predictors are correlated (e.g., Po1 and Po2).

**Surprises:**

* **U1** was dropped in Lasso but retained in Stepwise and Elastic Net, indicating sensitivity to penalization.
* **Time** was consistently dropped, suggesting limited predictive value in this dataset.

**Improvements:**

* Future models could explore interaction terms or nonlinear relationships.
* Consider using external socioeconomic data to enrich the feature set.

**Ethical Considerations:**

* Predictive models on crime data must be interpreted cautiously to avoid reinforcing biases.
* Variables like race, income, and education should be contextualized to avoid misrepresentation.

**REFERENCES**

https://ieeexplore.ieee.org/document/10937580