# Variable Selection to Predict Crimes per Capita

*Prepared by Emilie Campos*

*March 26, 2018*

## Introduction

The county demographic information (CDI) data set contains data from 440 of the most populous counties in the United States. Our objective is to develop a model to predict the number of serious crimes per capita, transformed to a rate per 1000 persons, i.e. crimesper1000 = 1000*crimes/pop.

## Methods

The CDI data set consits of 440 of the most populous counties in the United States. The information generally pertains to the years 1990 and 1992. The 17 variables are given in Table 1. In order to create a predictive model, we split this data set into two parts: 75% training and 25% test. The summary statistics for each are provided in Table 2. The mean and standard error are given for continuous variables and the count and percentage are given for categorical variables.

Table 1: Variable Descriptions

| Variable | Description |
|---|---|
| Identification Number | 1-440 |
| County | County name |
| State | Two-letter state abbreviation |
| Land area | Land area (square miles) |
| Total Population | Estimated 1990 population |
| Percent of population aged 18-34 | Percent 1990 CDI popluation aged 18-34 |
| Percent of population 65 or older | Percent 1990 CDI population aged 65 years old or older |
| Number of active physicians | Number of professionally active nonfederal physicians during 1990 |
| Number of hospital beds | Total number of beds, cribs, and bassinets during 1990 |
| Total serious crimes | Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies |
| Percent high school graduates | Percent of adult population (persons 25 years old or older) who completed 12 or more years of school |
| Percent bachelor's degrees | Percent of adult population (persons 25 years old or older with bachelor's degree |
| Percent below poverty level | Percent of 1990 CDI popluation with income below poverty level |
| Percent unemployment | Percent of 1990 CDI labor force that is unemployed |
| Per capita income | Per capita income of 1990 CDI popluation (dollars) |
| Total personal income | Total personal income of 1990 CDI population (in millions of dollars) |
| Geographic region | Geographic region classification that is used by the US Bureau of the Census, where 1 = NE, 2 = NC, 3 = S, 4 = W |

### Transformations

Upon inspection of the univariate and bivariate distributions, and component residual plots, we determined that we need to log-transform the variables: crimes per 1000 persons, number of beds, number of physicians, total population, total income, percent below the poverty line, percent of unemployment, and per capita income.

## Model Selection

Beginning with a base model which includes all of our possible predictors:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logdocs + \beta_3 logpop + \beta_4 logtotalinc + \beta_5 logpoverty \\ &+ \beta_6 logunemp + \beta_7 logpcincome + \beta_8 logarea + \beta_9 pop18 + \beta_{10} pop65 \\ &+ \beta_{11} hsgrad + \beta_{12} bagrad + \beta_{13} ncregion + \beta_{14} sregion + \beta_{15} wregion, \end{aligned}$$

we can then perform several variable selection procedures to find an optimal prediction model. For each variable selection procedure, the form of the model is given and the summary of regression coefficient values is given in Table 3. All analyses were performed in R 3.4.3 using packages: `stats`, `olsrr`, and `glmnet`.

i. Best a priori judgment:
Based on prior knowledge, we chose the following variables: total area, total population, percent of population aged 18-34, percent of high school graduates, percent below the poverty line, percent of unemployment, and per capita income. This is the model:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logpop + \beta_2 logpoverty + \beta_3 logunemp + \beta_4 logpcincome + \beta_5 logarea \\ &+ \beta_6 pop18 + \beta_7 hsgrad \end{aligned}$$

ii. Best subset selection:
We used the function `ols_best_subset()`. Based off a combination of factors including Mallow's Cp and AIC, this resulted in the following model with 5 predictors:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logpoverty + \beta_3 logpcincome + \beta_4 pop65 + \beta_5 ncregion \\ &+ \beta_6 sregion + \beta_7 wregion \end{aligned}$$

iii. Forward selection:
Using the criterion AIC and the function `ols_stepaic_forward()`, this resulted in the following model:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logpoverty + \beta_3 logpcincome + \beta_4 pop65 + \beta_5 ncregion \\ &+ \beta_6 sregion + \beta_7 wregion \end{aligned}$$

iv. Backward elimination:
Using the criterion AIC and the function `ols_stepaic_backward()`, this resulted in the following model:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logpop + \beta_3 logtotalinc + \beta_4 logpoverty + \beta_5 pop65 \\ &+ \beta_6 ncregion + \beta_7 sregion + \beta_8 wregion \end{aligned}$$

v. Stepwise:
Using the criterion AIC and the function `ols_stepaic_both()`, this resulted in the following model:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logpoverty + \beta_3 logpcincome + \beta_4 pop65 + \beta_5 ncregion \\ &+ \beta_6 sregion + \beta_7 wregion \end{aligned}$$

vi. Lasso:
Using techniques from Introduction to Statistical Learning in R and the function `cv.glmnet()`, LASSO regression resulted in the following model:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logdocs + \beta_3 logpoverty + \beta_4 pop65 + \beta_5 ncregion \\ &+ \beta_6 sregion + \beta_7 wregion \end{aligned}$$

vii. Bivariate p-value selection:
By regressing each predictor singly and using a threshold of 0.25, we ended with the following model:

$$\begin{aligned} logcrimesper1000 = \beta_0 &+ \beta_1 logbeds + \beta_2 logpop + \beta_3 logtotalinc + \beta_4 logpoverty + \beta_5 logpcincome \\ &+ \beta_6 logarea + \beta_7 pop18 + \beta_8 pop65 + \beta_9 hsgrad + \beta_{10} bagrad \\ &+ \beta_{11} ncregion + \beta_{12} sregion + \beta_{13} wregion \end{aligned}$$

# Discussion

The Best Subset, Forward and Stepwise Selection methods resulted in the same exact model. It appears that region was a very important factor in modeling the crime rate as it was included in each of the algorithmic variable selection methods. The Bivariate p-value approach included a lot more variables because it does not have a penalty for less parsimonious models. The model based on a priori judgement was very different from the other models, including the variables that most of the variable selection algorithms found unimportant. This highlights the fact that our perceptions about crime in the United States may be incorrect.

Each of the models had a fairly small root mean test error, taking on values between 0.44-0.47. In terms of intpretability, the model chosen by the best subset, forward, and stepwise methods would be the most parsimonious. If we are more concerned with prediction, the bivariate p-value approach has the smallest root mean test error. However, the difference is minimial.

One concern after training the model is that an outlier was randomly included in the training set. Had this value been excluded from the training set, the regression coefficients would have been different and the root mean test error could have been improved. In addition to this, the functions used in R were using AIC as the criterion for choosing variables and there was not an equivalent function for using BIC or Mallow's Cp. It would have been interesting to see if these would have led to the same models chosen based on AIC. This could be a potential future project.

# Tables and Figures

Table 2: Summary Statistics

| | Training Set | Test Set |
|---|---|---|
| Land area, in square miles | 1,049 (1,624.9) | 1,019 (1,624.9) |
| Total Population | 360,721 (364,166.1) | 489,881 (364,166.1) |
| Percent of population aged 18-34 | 29 (4.31) | 28 (4.31) |
| Percent of population 65 or older | 12 (4.07) | 12 (4.07) |
| Number of active physicians | 912 (1,231.01) | 1,216 (1,231.01) |
| Number of hospital beds | 1,343 (1,626.65) | 1,806 (1,626.65) |
| Total serious crimes | 25,172 (47,480.29) | 32,932 (47,480.29) |
| Percent high school graduates | 78 (7.25) | 77 (7.25) |
| Percent bachelor's degrees | 21 (7.88) | 20 (7.88) |
| Percent below poverty level | 9 (4.99) | 9 (4.99) |
| Percent unemployment | 7 (2.42) | 7 (2.42) |
| Per capita income | 18,611 (4,278.65) | 18,414 (4,278.65) |
| Total personal income, in millions | 7,217 (8,120.44) | 9,825 (8,120.44) |
| Geographic region | | |
| North East region | 78 (24%) | 25 (23%) |
| North Central region | 83 (25%) | 25 (23%) |
| South region | 117 (35%) | 35 (32%) |
| West region | 52 (16%) | 25 (23%) |

Table 3: Estimated Coefficients by Selection Method

| | A priori | Best Subset | Forward | Backward | Stepwise | Lasso | Bivariate |
|---|---|---|---|---|---|---|---|
| No. of hospital beds, 1% increase | | 0.181 | 0.181 | 0.16 | 0.181 | 0.142 | 0.143 |
| No. of active physicians, 1% increase | | | | | | 0.088 | |
| Total Population, 1% increase | 0.189 | | | -0.565 | | | -39.964 |
| Total personal income, 1% increase | | | | 0.595 | | | 39.991 |
| Percent below poverty, 1% increase | 0.688 | 0.395 | 0.395 | 0.407 | 0.395 | 0.247 | 0.417 |
| Percent unemployment, 1% increase | -0.227 | | | | | | |
| Per capita income, 1% increase | 0.656 | 0.589 | 0.589 | | 0.589 | | -39.352 |
| Land area, 1% increase | -0.006 | | | | | | |
| Percent of population aged 18-34 | 0.009 | | | | | | 0.005 |
| Percent of population aged 65 or older | | -0.017 | -0.017 | -0.015 | -0.017 | -0.014 | -0.014 |
| Percent high school graduates | 0.002 | | | | | | 0.002 |
| Percent bachelor's degrees | | | | | | | -0.003 |
| Geographic Region | | | | | | | |
| North East | | -3.968 | -3.968 | 3.933 | -3.968 | 1.792 | 547.574 |
| North Central | | 0.232 | 0.232 | 0.236 | 0.232 | 0.242 | 0.235 |
| South | | 0.549 | 0.549 | 0.546 | 0.549 | 0.546 | 0.55 |
| West | | 0.469 | 0.469 | 0.459 | 0.469 | 0.456 | 0.459 |
| Root Mean Test Error | 0.4659 | 0.4499 | 0.4499 | 0.4501 | 0.4499 | 0.4516 | 0.4470 |