Report

Carlos Espino, Xavier Gonzalez, Diego Llarrull, Woojin Kim

December 14, 2015

Contents

Introduction

Understanding the factors behind criminal behaviour is one of the most crucial task for preventing and controlling future crime. In this report, we explore the potential factors affecting crime rates based on the demographics and econometrics data gathered from 197 counties in North Carolina from 1981 to 1987. Using various statistical methods and modeling techniques, we analyze and identify the most important factors and metrics tied to crime rates. We also present a predictive model capable of estimating the crime rate with under 25% error using the selected parameters.

Dataset

Predictor	Description
county	county identifier
year	year from 1981 to 1987
crmrte	crimes committed per person
prbarr	'probability' of arrest
prbconv	'probability' of conviction
prbpris	'probability' of prison sentence
avgsen	average sentence, days
polpc	police per capita
density	people per square mile
taxpc	tax revenue per capita
region	one of 'other', 'west' or 'central'
smsa	'yes' or 'no' if in SMSA
pctmin	percentage minority in 1980
wcon	weekly wage in construction
wtuc	weekly wage in trns, util, commun
wtrd	weekly wage in whole sales and retail trade
wfir	weekly wage in finance, insurance and real estate
wser	weekly wage in service industry
wmfg	weekly wage in manufacturing
wfed	weekly wage of federal emplyees
wsta	weekly wage of state employees
wloc	weekly wage of local governments employees mix offence mix: face-to-face/other
pctymle	percentage of young males

Table 1: Description of the predictors in the dataset

Non-linear modeling

From the pairwise plot we generated shown in (), we identified a predictor `pctmin`, corresponding to the proportion of minorities in the region, showing a clear non-linear relationship with the crime rate that can benefit from higher order polynomial regression/splines and also help with predicting the overall crime rate.

Linear model

First we evaluated the regression model generated using only using a linear model:

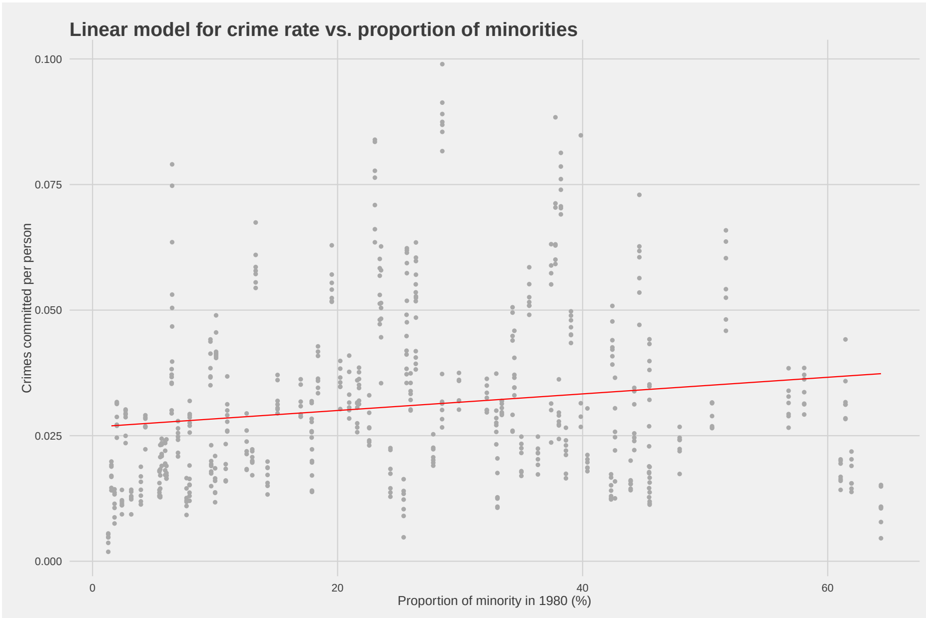

Figure 1: Linear model for crime rate vs. proportion of minorities

This naïve model results in an error rate of 0.4808 for the testing set. From the plot it is very clear that the relationship between the crime rate and the proportion of minorities in the area is not linear.

Polynomial model

Next we obtained a degree-4 polynomial fit for a smooth fit over the `pctmin` data:

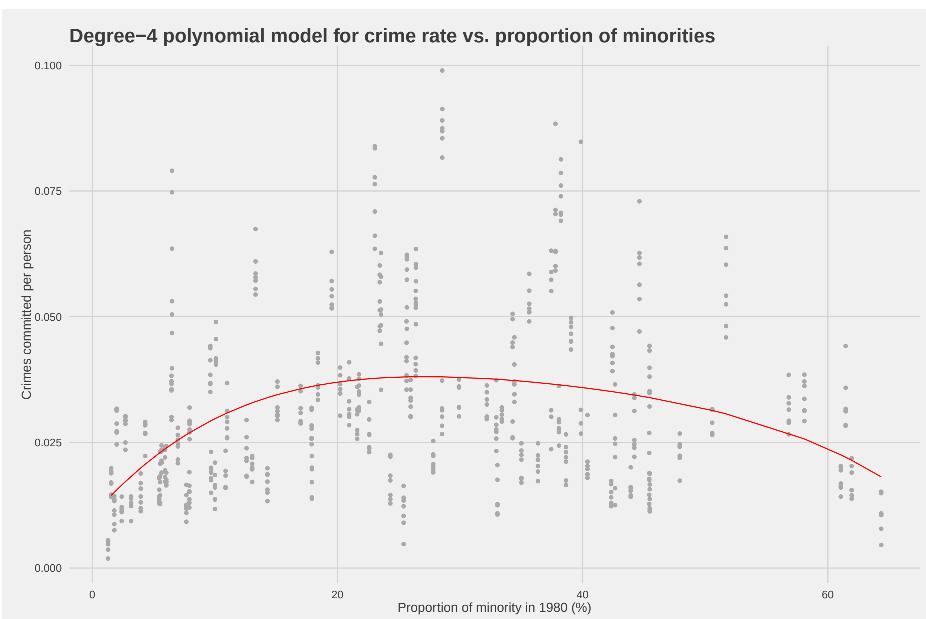

Figure 2: Degree-4 polynomial model for crime rate vs. proportion of minorities

The error rate was improved by reducing the bias of the model. This fit resulted in an error rate of 0.4755.

Splines

We further attempt to reduce the bias, introducing a more flexible piecewise polynomial by using knots. Using a cubic spline, the fitted curves are constrained to be continuous. As splines often leads to high variance at the outern ranges of the predictors, we fit natural cubic spline, which forces the function to be linear at the boundary. `ns()` function was used to generate natural cubic knots with 6 degrees of freedom, with matrix of basis functions for splines and knots at 6.03%, 14.28%, 24.31%, 34.28%, and 42.64% of `pctmin`.

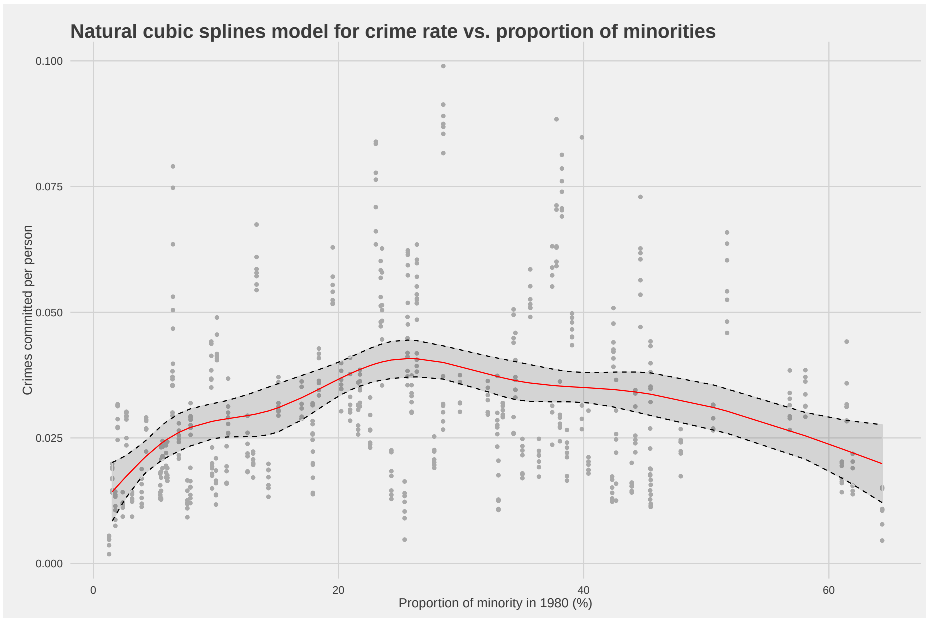

Figure 3: Natural cubic splines model for crime rate vs. proportion of minorities

Consequently, we see a modest improvement in the mean error: 0.468.

We also attempted to fit a smoothing spline with a value of λ chosen cross-validation. This resulted in a model very similar to the polynomial fit and failed to improve the mean k-folds cross-validation error.

Combined models

Combining the predictors from the previous section with just the linear `pctmin` results in a mean error of 0.2385.

Finally, we included the splined version of `pctmin` predictor alongside the selected predictors, which resulted in a slight improvement in mean error to 0.2355.