Biostatistics 200B: Project II

Chad Pickering

3/26/2018

Introduction.

We are provided a dataset of county demographic information (CDI) for the 440 most populous counties in the United States pertaining to the years 1990. Our objective is to develop a model to predict the number of serious crimes per capita, transformed to a rate per 1000 persons. We will carry out several different variable selection methods in order to find the model that results in the smallest prediction error, and discuss some of the advantages, disadvantages, and results of each method, and indicate some shortcomings and limitations.

Methods and Discussion.

For each county, the state, region of the U.S., land area in square miles, total population in 1990, and percent of the population aged 18-34 and 65+ are given. Total number of active physicians, hospital beds, and serious crimes are also indicated, as well as high school graduation rate, percentage of adults with a bachelor's degree, per capita income, total personal income, percent of population below the poverty line, and unemployment rate. While formulating the body of covariates to consider in the full model, population density for each county is calculated, as well as doctors and hospital beds per 1000 people to match the units of the response variable, serious crimes per 1000 people (Table 1).

When looking at the downward curvature and increasing behavior of the simple linear regression (SLR) plots, and seeing that the residuals of the response versus fitted values for each individual predictor are irregular/curved downward, we see that population, poverty, population density, and total personal income should be log-transformed. Looking at other diagnostic plots for the SLRs (residuals vs. leverage), we see that there are two overarching influential points/outliers. Kings County, NY (observation 6) is an extreme outlier in the design space for every covariate (and has a very large Cook's distance for around half), and is therefore removed. Los Angeles County, CA (observation 1) is a huge outlier in the design space with respect to the predictors population, number of physicians per 1000 people, and total personal income, but is not removed.

In order to generate models for validation purposes, the data is split into a training set (329 observations, 75%) and a test set (110 observations, 25%). The training and test sets only include the variables that should be considered as potential predictors - all of the log-transformed variables mentioned in addition to the other variables listed in Table 1 except for state.

Population percentage under 18, per capita income, region, and population density are included in all models selected (poverty and total personal income included in this list for all algorithm-generated models). The first model chosen includes only covariates thought to be good predictors of serious crimes - percent of population under 18, high school graduation rate, unemployment rate, per capita income, region, physicians per 1000 people, and population density (Table 2). Thus, if we have concerns about our model being the most interpretable, this is the one we would choose. All of the models selected using the other methods mentioned below have more variables and more complex interpretations, specifically because of the large number of transformed covariates in them. The testing root mean squared error (RMSE) is 18.65, the highest of all the models created.

Best subset selection and backward selection yield the same model, with an RMSE of 17.06 (Table 2). Best subset selection chose the five best models per number of variables based on Mallow's C_p measure, while backward selection using AIC does not penalize complex models as much as, say, BIC would, so the resultant model has all of the covariates from the full model except for three, the most complex chosen by any algorithm (tied with LASSO). Forward selection and stepwise selection choose a model with three less covariates than that of best subset and backward selections (elderly population rate, unemployment rate, and number of physicians are excluded), with an RMSE of 17.47 (Table 2). Since they both start with the null model in their process, we expect the models to be quite similar; in this case, the “backward look” of stepwise selection did not find any insignificant covariates to remove once initially included. The LASSO method found a complex most predictive model identical to that of the best subset and backwards methods, except it excludes population as a predictor of serious crimes and includes high school graduation rate, for an RMSE of 16.91 (Table 2). By optimizing the shrinkage coefficient in order to render a handful of beta coefficients negligible, the resultant model was quite complex, suggesting that most predictors included in our overall set are at least mildly significant in and of themselves. This postulate is given more evidence with the fact that the bivariate p-value method with a significance threshold of 0.15 (0.1 yields the same model) outputs a model suggesting that only four covariates are insignificant, with an RMSE of 16.50 (Table 2). If we are to trust the one random split and not take an average of several, this model should be chosen to predict future observations. Such a controversial and theoretically problematic method was not expected to give the most predictive model; some likely flaws with our current procedures could take the blame.

Limitations and Conclusion.

Instead of merely using one validation step, 5-fold cross validation (CV) (with k carefully chosen) should be used to confirm the best model - an average of the RMSEs for each selection method is always more telling than the reliance of one random split. The performance of the LASSO, based solely on its shrinkage properties, should do better overall most of the time - an increased number of steps could be attempted to see if any model variablility results, or if the same model is output every time. Additionally, the presence of outliers and influential points in the dataset should be more carefully dealt with even though the most egregious data point was removed in this process. There was some evidence that there were a few other observations, such as Los Angeles County, that could have considerably affected the significance of the beta coefficients in a few covariates. Lastly, there are some variables such as population and population density that, although transformed, do depend on the same values (in this case, population counts); we should take some effort to include only one variable for consideration that tells the same story.

In conclusion, in order to predict the number of serious crimes (at least from 1990, but assuming the same trends continue today), we should choose the model generated by the bivariate p-value method in Table 2, with reservations (see above). Depending on purpose, a simpler model should be chosen (see discussion), but all models indicated are fairly predictive - there are no glaring differences between them. However, as discussed, we should re-do this analysis using 5-fold CV to confirm both method performance and model choice.

Tables and figures.

Table 1. Characteristics of the cohort.

Characteristic	Full cohort (n=440)	Training set (n=330)	Test set (n=110)
State			
California	34 (0.077)	25 (0.076)	9 (0.082)
Florida	29 (0.066)	18 (0.055)	11 (0.1)
Pennsylvania	29 (0.066)	22 (0.067)	7 (0.064)
Texas	28 (0.064)	22 (0.067)	6 (0.055)
Ohio	24 (0.055)	18 (0.055)	6 (0.055)
New York	22 (0.05)	21 (0.064)	1 (0.009)
Area (sq. mi.)	656.5 (451.2-946.8)	640.0 (431.5-933.5)	715.0 (526.0-1039.0)
Population	217280 (139027-436064)	224833 (140368-467118)	209828 (137062-347138)
Population Pct. 18-34	28.57 ± 4.19	28.78 ± 4.01	27.94 ± 4.65
Population Pct. 65+	12.17 ± 3.99	11.89 ± 3.38	13.00 ± 5.37
Number of active physicians	401.0 (182.8-1036.0)	421.5 (198.2-1155.2)	359.5 (173.5-791.8)
Number of hospital beds	755.0 (390.8-1575.8)	809.0 (411.5-1633.8)	658.5 (356.2-1402.2)
Number of serious crimes	11820 (6220-26280)	12188 (6126-28444)	10365 (6887-19723)
Pct. adults 12+ yrs. school	77.56 ± 7.02	77.73 ± 6.93	77.04 ± 7.26
Pct. adults with BS degree	21.08 ± 7.65	21.29 ± 7.54	20.45 ± 7.98
Pct. population in poverty	7.90 ± 4.66	8.51 ± 4.46	9.36 ± 5.17
Pct. population unemployed	6.6 ± 2.3	6.5 ± 2.1	7.0 ± 2.8
Per capita income	18561 ± 4059	18718 ± 4043	18093 ± 4091
Total personal income per county (mil. dollars)	3857 (2311-8654)	4025 (2351-9346)	3510 (2270-6918)
Region			
Northeast	103 (0.234)	85 (0.276)	18 (0.164)
North central	108 (0.245)	77 (0.233)	31 (0.282)
South	152 (0.345)	113 (0.342)	39 (0.355)
West	77 (0.175)	55 (0.167)	22 (0.200)

Statistics are in the form mean ± SD, median (IQR), or discrete count (percentage). Point 6, Kings County, NY, excluded from analysis, is included in these summaries.

Table 2. Models Generated from Variable Selection Methods.

Note: Best subset used C_p criterion. All others used AIC.

Var.	A priori	Best subset	Forward	Backward	Stepwise	Lasso	Bivar. p-val.
area		X	X	X	X	X	
pop18	X	X	X	X	X	X	X
pop65		X		X		X	
hsgrad	X					X	X
bagrad							
unemp	X	X		X		X	
pcincome	X	X	X	X	X	X	X
region	X	X	X	X	X	X	X
docsper1000	X						X
bedsper1000		X		X		X	X
ln(pop)		X	X	X	X		X
ln(poverty)		X	X	X	X	X	X
ln(pop_dens)	X	X	X	X	X	X	X
ln(totalinc)		X	X	X	X	X	X
RMSE	18.65	17.06	17.47	17.06	17.47	16.91	**16.50**

AIC: Akaike's information criterion; RMSE: root mean squared error.