Model Building2

2023-12-08

1) The first question I'm trying to answer is how well Unemployment, labor force statistics, and population can predict certain types of crime in Maryland. For this PCA I removed all crime statistics because those are my dependent values. Last time I included them because I misunderstood the process. I also scaled the data for this new PCA to get more accurate results. Furthermore based on the suggestions from the last model building assignment I added a bi plot to display the variable relationships. As well as creating a regression model for seven different types of crime statistics. In the end I discovered that the pca could predict total crime better than all other crimes besides larceny theft. Larceny theft was the only crime that the model could predict better than total crime. I came to this conclusion by comparing r scores for the models which measures how much variability it was able to explain. Larceny had an r score of .8165 whereas total crime was .7282. Thus larceny prediction will be more accurate in general.

2) My second question was about the influence population groups had on a models ability to predict crime based on unemployment and Labor force. The variables used in my second model are Index_Count, Population, Unemployment, and Labor. Index_Count represents the amount of crimes recorded in a year for a particular county. For my second model after reading the advice from the previous assignment I decided I would go with a glm model. After some research I first went with a poisson regression model. I included the model below for reference. When I saw the results I realized that a poisson model was not the best method because of the amount of over dispersion present in my data. So after more research I settled on a negative binomial regression because they are able to fit data better when the variance is higher than the mean which was definitely the case. The difference was staggering between the models. The second model did the trick and most residual deviance were only off my 20 or so from the degrees of freedom. Compared to the poisson where there was a difference of tens of thousands. In conclusion as the I went through the population groups the more people in a population group increased the AIC. Which is the score of a models goodness of fit and simplicity. Thus population does affect the ability to accurate predict crime based on unemployment and labor force because it affects the how well the model fits to the data and thus affecting prediction accuracy.

3) For my third I question I'm trying to figure out if a models is able to predict crime more accurately if the data is aggregated rather than if I train based on crime in each county. In the previous assignment I attempted to answer whether unemployment and labor pct could predict certain types of crime better than others. But it was too similar to what I did with my first question so I decided to do something different. I decided to go with a negative binominal regression and used the New York crime data set to answer my question. After using these models I came to the conclusion that aggregating crime data provided more accurate results for the state rather than a county by county basis was more accurate. I came to this conclusion by comparing the AIC scores of the two regressions. Aggregate AIC was 841.17 and non aggregate AIC was 37436. This shows clearly that aggregating data is a better method for predicting crime with unemployment and labor force pct.

#1st Project Question ####Preprocessing Data for PCA

Maryland PCA

#2nd Project Question ####NewYork Preprocessing

pop group prep

poisson model 2

negative binomial

#3rd Project question