Related Factors of Marijuana Crime Rate in Denver

Hongyang Zheng, Shanyu Hou, Menghan Lin, Xinman Wu

University of Colorado Denver

Abstract

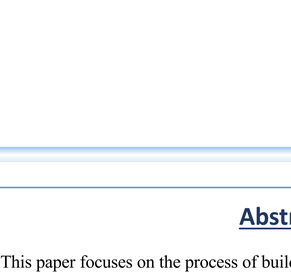
This paper focuses on the process of building a regression model revealing the significant demographic factors associating with marijuana crime in Denver Area. The exploratory, colinear checking, model selection, F-test, and diagnostic steps taken in the model building process is described in detail in this paper. There are also interpretations of the coefficients of the predictors based on the context of the dataset, explaining how this model works in predicting marijuana crime rate in Denver area.

Introduction

1. **and Background:** In Denver, both medical and retail marijuana are legal to people 21 or older. However, the use of marijuana is still forbidden in many parts of the world since physical and mental effects of marijuana may cause people to commit crimes.
2. **Interesting Points:** How marijuana crime rate is associated with different demographic features of neighborhoods in Denver.
3. **Name and Location of datasets & Purpose:** Crime Marijuana, Crime and American Community Survey_2010-2014 from the Denver Open Data Catalog. These data are gathered by Denver Police Department and US Census Bureau to measure the crime rate and demographic features.
4. **Variables:** One is related to crimes, including marijuana crime rate per capital, crime rate per capital and whether there are police stations in the neighborhood; The other indicates the demographic features in the neighborhood, such as percentage of black people, percentage of male, and percentage of people have a college degree.

Method

Summarize data



We transform crime_rate in logarithm since the plot (log(m_crime_rate) vs crime_rate) indicates there is a logarithm relationship between them.

Collinearity

The combined data contains 26 variables in total. Since the neighborhood name cannot be evaluated as numerical or categorical predictor, we removed it at first. After checking the pairwise correlations, we found that PCT_female and PCT_male, PCT_non_fam and PCT_married have perfect collinearity. Thereafter, we amputated PCT_female and PCT_married as needed.

Example of soling collinearity problem: We rebuilt a linear model with the former response and the rest of predictors. When computing the variance inflation(VIF) for the new model, we found several large VIFs, where the VIF of PCT_hispanic is the largest. In the meantime, the PCT_hispanic also has large correlations with other predictors, such as PCT_White, PCT_LeCollege, and PCT_other fam. When checking the variance decomposition proprotions, PCT_hispanic, PCT_white, and PCT_black have very large variance decomposition proportions for the second largest index. Thus PCT_hispanic is amputated. Repeating the process of building new model, we found that the problems of VIF are significantly improved.

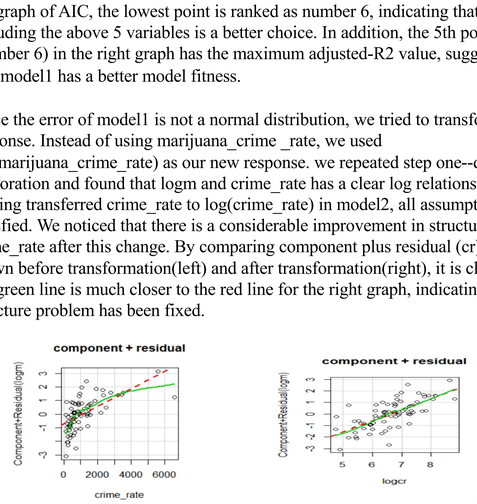
The VIF of the new dataset and the variance decomposition proportions suggest that there is still possible collinearity problem among PCT_age65Plus, PCT_LeCollege and PCT_white. Then, we amputated PCT_white, PCT_age65plus and PCT_LeCollege.

Since we were interested in whether PCT_white and PCT_non fam have association with m_crime_rate, we added those two predictors back. Although the VIFs increased after this step, both the VIF and variance decomposition proportions are still in the acceptable range. Therefore, we finally kept 19 variables in the final dataset.

Model Selection

After excluding collinearity, we used the remaining 19 variables to do model selection. Regressing marijuana_crime_rate on the rest of variables, we found that the model (model1) containing pct_male, pct_native_American, pct_Asian, pct_graduate, pct_renter, crime_rate and an intercept is a preferable model, according to the graph of Akaike information criterion(AIC) and adjusted-R2. In the graph of AIC, the lowest point is ranked as number 6, indicating that including the above 5 variables is a better choice. In addition, the 5th point (number 6) in the right graph has the maximum adjusted-R2 value, suggesting that model1 has a better model fitness.

Since the error of model1 is not a normal distribution, we tried to transform our response. Instead of using marijuana_crime _rate, we used log(marijuana_crime_rate) as our new response. we repeated step one--data exploration and found that logm and crime_rate has a clear log relationship. Having transferred crime_rate to log(crime_rate) in model2, all assumptions are satisfied. We noticed that there is a considerable improvement in structure of crime_rate after this change. By comparing component plus residual (cr) plots drawn before transformation(left) and after transformation(right), it is clear that the green line is much closer to the red line for the right graph, indicating that the structure problem has been fixed.



In the next step, we replaced crime_rate with log(crime_rate) and ran the model selection using the updated 19 variables. The best model suggested by AIC, Bayesain information Criterion(BIC), Mallows' s Cp statistic (CP) and adjusted-R2 plot is slightly different: model containing 7, 2, 17, and 8 variables is preferable respectively according to each graph and all of them pass the gvlma test. Since the general assumptions are satisfied, we firstly did F-test for these models in order to exclude unpreferable models.

Hypothesis test

F-test	(lmodn2, lmodn7)	(lmodn2, lmodn8)	(lmodn2, lmodn17)	(lmodn7, lmodn8)
p-value	0.02714	0.03294	0.2849	0.2849

lmodn2:log(marijuana_crime_rate) = PCT_POVERTY + log(crime_rate).
lmodn7: log(marijuana_crime_rate) = PCT_male + PCT_WHITE + PCT_ASIAN + PCT_Otherfamaily + MEDIAN_EARNINGS + PCT_FB + log(crime_rate).
lmodn8: log(marijuana_crime_rate) = PCT_male + PCT_WHITE + PCT_ASIAN + PCT_graduate + PCT_Otherfam + MEDIAN_EARNINGS + PCT_FB + log(crime_rate).

1. From F-test, lmodn7 and lmodn8 are better than lmodn2 but there is no evidence to conclude that lmodn17 is better.
2. Compared lmodn7 and lmodn8: the lmodn7 is more suitable than lmodn8 according to the p-value. As are result, we chosen lmodn7.
3. The coefficient for pct_male is not statistically significant in lmodn7. Since the p-value is 0.1699 which means that removing pct_male doesn't influence our model fitness much, we choose a simpler model as ideal model(model3), which regresses log(m_crime_rate) on pct_white, pct_Asian, pct_other_family, median_earnings, pct_fb and log(crime_rate).

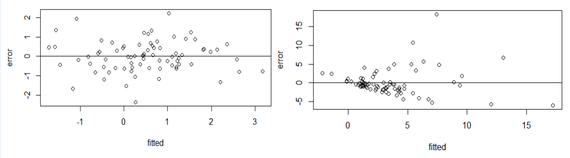
Diagnostics

--Error Assumption

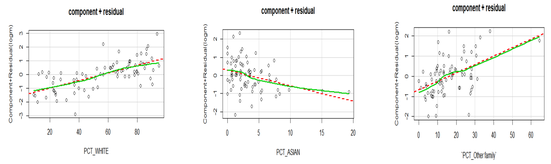
Model3 satisfies all those error assumptions well. The scatter plot of residuals against fitted value shows random pattern and constant thickness when going along horizontal axis. The residuals seem to have some cyclical pattern. However, the plot of successive pairs of residuals doesn't show obvious correlated relationship. The p-value of Durbin-Watson test is about 0.12, which convinced us the residuals of this model are not serial correlated. When checking the normality of the residuals, we found that the residuals fits the normal distribution pretty well in the qq-plot, with all the points fall in the confidence band.

--Structural Assumptions

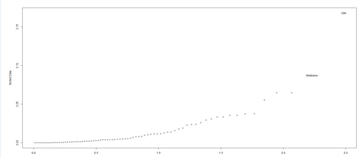
We used two methods to check the structural assumptions. Firstly, we drew fitted-value vs. residuals plot. A random scatter of points around 0 in the below graph suggests that the linearity assumption for model3 holds. Therefore, we conclude that the general structure of model3 is proper. Compared with model3, although model1 also passed link function test, the fitted-value vs. residuals graph(the second one) for model1 is not as good as model3's, since the scatter of



In the next step, we used cr plot as a tool to verify that there is no nonlinear relationship between each variable included in model3 and response, as cr plot can show the impact of a regressor on the fitted values. The cr plots for 3rd diagnostics in the appendix clearly show that the scatter plots for the five regressors are all linear. Especially, the green line and the red line are very close, indicating that it is not necessary to do further transformation. In a nutshell, the structure assumptions hold for model3 and the structure of each regressor in model3 is also proper.



-- influential observations

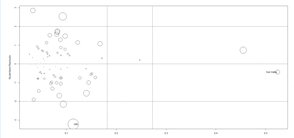


#DIA

	Est. 1	SE 1	Est. 2	SE 2
(Intercept)	-7.0113	1.2470	-7.7098	1.3074
PCT_WHITE	0.0204	0.0105	0.0277	0.0109
PCT_ASIAN	-0.0844	0.0316	-0.0863	0.0336
PCT_Otherfamily	0.0419	0.0153	0.0419	0.0163
median	-0.0285	0.0104	-0.0318	0.0110
PCT_FB	0.0192	0.0171	0.0349	0.0174
log2	1.0224	0.1078	1.0463	0.1145

#Wellshire

	Est. 1	SE 1	Est. 2	SE 2
(Intercept)	-7.7479	1.2595	-7.7098	1.3074
PCT_WHITE	0.0252	0.0106	0.0277	0.0109
PCT_ASIAN	-0.0807	0.0324	-0.0863	0.0336
PCT_Otherfamily	0.0380	0.0158	0.0419	0.0163
median	-0.0337	0.0106	-0.0318	0.0110
PCT_FB	0.0327	0.0168	0.0349	0.0174
log2	1.0901	0.1116	1.0463	0.1145



	Est. 1	SE 1	Est. 2	SE 2
(Intercept)	-7.8729	1.3687	-7.7098	1.3074
PCT_WHITE	0.0291	0.0115	0.0277	0.0109
PCT_ASIAN	-0.0810	0.0360	-0.0863	0.0336
PCT_Otherfamily	0.0480	0.0218	0.0419	0.0163
median	-0.0320	0.0111	-0.0318	0.0110
PCT_FB	0.0340	0.0177	0.0349	0.0174
log2	1.0453	0.1152	1.0463	0.1145

Results and Interpretation

Variable	Estimate	Estimatedstandard error	Test statistic	p-value
Percentage of White	0.028	0.011	2.528	0.014
Percentage of Asian	-0.086	0.034	-2.568	0.012
Percentage of Other Family	0.042	0.016	2.567	0.012
Median earning (per thousand)	-0.031	0.019	-2.896	0.005
Percentage of Foreign Born	0.035	0.017	2	0.049
log(crime rate)	1.046	0.115	9.138	0.000

- **White:** If two places are identical except one has 1% more white people, we expect that the place with more white people is associated with a multiplicative effect of 1.028 on the mean of marijuana crime rate.
- **Asian:** If two places are identical except one has 1% more Asian, we expect that the place with more Asian is associated with a multiplicative effect of 0.917 on the mean of marijuana crime rate.
- **Other family:** If two places are identical except one has 1% more other family household that a family only has mother or father, we expect that the place with more other family household is associated with a multiplicative effect of 1.0283 on the mean of marijuana crime rate.
- **Median Earning per thousand:** If two places are identical except one has 1000 dollar more median earning per year, we expect that the place with higher median wage is associated with a multiplicative effect of 0.96 on the mean of marijuana crime rate.
- **Foreign born:** If two places are identical except one has 1% more foreign born population, we expect that the place with more foreign born people is associated with a multiplicative effect of 1.0283 on the mean of marijuana crime rate.
- **Crime rate per 1000 people:** If two places are identical except one has a change in crime rate from x to 2*x is associated with a multiplicative effect of 2.0 on the mean of marijuana crime rate

Conclusion

1. **Results:** The predictors such as pct_white, pct_Asian, pct_other_family, median_earnings, pct_fb and log(crime_rate) significantly influence the log(marijuana crime rate) in each neighborhood in Denver Area.
2. **Causal Conclusion:** we cannot make a causal conclusion since the dataser is an observational data
3. **The extent of** results**:** As the data only cover the Denver area, it is unavailable to extend the result to a larger population. Different areas have different definitions of marijuana crime under different circumstances.
4. **Improving Study:** Our group only focused on normal distributed data and fitted it into linear model. However, there are many useful models other than linear and other possible distributions such as Poission. we could improve our study and the way of handling data as well as collecting data.
5. **Future Reaserch**: we could explore the related factors of marijuana crime rate in other areas in Colorado or other States and compare the significant factors in the Denver area to which in other places. Figuring out the cause of the similarities and difference is another valuable and interesting topic.

Reference

- Denver Open Data Category. (2017). Crime. Retraced from https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime.
- Denver Open Data Category. (2017). Crime Marijuana. Retraced from https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime-marijuana.
- Denver Open Data Category. (2017). American Community Survey Blk Grp (2010-2014). Retraced from https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-american-community-survey-blk-grp-2010-2014.