Analysis of the crime in America

Group memeber: Ye Rougang, Tan Chunxi, Han Ruijian
April 10, 2017

1 INTRODUCTION

The dataset we choose is about crime rates in 59 US cities, which was assembled by Steven Levitt for the paper "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime." The observations dates back to 1969-1992 for 59 large U.S. cities. Our analysis covers the period of 1971-1990. Since there are a few missing data, i.e. less than 10% of whole data, we discard them.

Though Steven Levitt had collected many different types of variables, we pick up some most related variables as covariates to analyze their relationship with the crime rates. We divide different crimes to two categories to mainly explore changes of similar crimes:

- Violent Crimes: Murder, Rape, Robbery, Aggravated assault
- Property Crimes: Burglary, Larceny, Motor vehicle theft

The dataset after preprocessing is in the attachment.

We mainly focus on using linear models here, i.e. regression with l^2 regularization for handling collinearity among features.

2 LINEAR REGRESSION USED BY STEVEN LEVITT

First we try to use the linear regression proposed by Steven Levitt. The assumption is that there is no difference among different cities. So one can put all of data of different cities together to model.

	coef	std err	t	P >|t|
rincpc	0.0008	0.002	0.459	0.646
econgrow	-0.0156	0.078	-0.201	0.841
unmep	-0.0142	0.106	-0.135	0.893
citybla	-0.0030	0.012	-0.249	0.803
sta_educ	-1.779e-05	2.09e-05	-0.852	0.394
sta_welf	1.183e-05	2.58e-05	0.458	0.647
price	-0.0056	0.008	-0.692	0.489
ln(police)	-0.0106	0.007	-1.595	0.111

Table 2.1: Results of the linear model by Steven Levitt (Violent Crime)

Then the model is defined as follows:

$$\Delta \ln(Crime_{c,t}) = \beta_1 \ln(Police_{c,t-1}) + X_{c,t}^T \Theta + \mu_{c,t}$$

The dependent variable is $\Delta \ln(Crime_{c,t})$, change of $\ln(Crime_{c,t}$ in year t of city c. Here crime refers to the crime rate per capita of either violent crime, property crime, or an specific crime category. With the consideration of the hysteresis effect of police, we choose covariate of effect of the police as $ln(Police_{c,t-1})$, which is the police per capita of city c in previous years $t-1$. Other covariates we choose are as follows:

- "rincpc"– real income per capita in state
- "econgrow" – percent change in rincpc
- "unemp" – state unemployment rates
- "citybla" – percentage of city population that is black
- "sta_educ" – Combined state and local spending per capita on education in 1992 dollars.
- "sta_welf" – Combined state and local spending per capita on public welfare in 1992 dollars.
- "price" – Consumer Price Index

Steven Levitt use this model to estimate the coefficients by ordinary least squares and Two-Stage Least Squares. However, there are correlations within the covariates. So we use this model to estimate the coefficients by l^2 regularization -ridge regression.

We did the ridge regression using python package *statsmodels*. First, we choose the violent crime category as the dependent variable. The result is presented in table 2.1

We then separate data into train data and test data by ratio 3:1 and choose the coefficient of penalty λ which has the minimal MSE error in the cross validation. λ is 0.01 for this model. Yet we can see from the Table 2.1, under the significance level $\alpha = 0.05$, none of the coefficients are significant that means no covariates play significant roles for the regression here. So this model is not as good as we expected.

	coef	std err	t	P > |t|	95%C.I.
rincpc	0.0169	0.004	3.811	0.000	(0.008,0.026)
econgrow	-0.0450	0.197	-0.228	0.820	(-0.433,0.343)
unmep	-0.0076	0.268	-0.028	0.978	(-0.533,0.518)
citybla	0.1248	0.031	4.083	0.000	(0.065,0.185)
sta_educ	5.402e-05	5.3e-05	1.020	0.308	(-5e-05,0.000)
sta_welf	-5.277e-05	6.56e-05	-0.805	0.421	(-0.000,7.6e-05)
price	-0.0332	0.020	-1.622	0.105	(-0.073,0.007)
ln(police)	-0.3194	0.017	-18.848	0.000	(-0.353,-0.286)

Table 3.1: Violent Crime($\lambda \approx 0.01$)

3 Improvement of Used Model

3.1 Violent Crime

The linear model proposed by Steven Levitt does not fit very well due to the insiginficant coefficients. We make some small changes on the original linear models with replacing $\Delta \ln(Crime_{c,t})$ to $Crime_{c,t}/Crime_{c,t-1}$. So our linear model is defined as:

$$Crime_{c,t}/Crime_{c,t-1} = \beta \ln(Police_{c,t-1}) + X_{c,t}^T \Theta + \mu_{c,t}$$

Maybe due to the scaling, the regression model here performs well for both of violent crimes and property crimes. The result for violent crime is showed in Table 3.1 where the λ of minimal MSE in the cross validation is nearly 0.01. And we also show the the relationship between lambda and coefficients in Figure 3.1. From the table 3.1, we can see "rincpc", "citybla" and "ln(police)" is significant while other features are insignificant under the significance level $\alpha = 0.05$, . So we can conclude that violent crimes has strong relation with "rincpc", "citybla" and "ln(police)". The police has negative effect on crime rates while "rincpc" and "citybla" which represent economic conditions are positively correlated with violent crime rates. These interpretations concur with our intuition.

3.2 Property Crime

We also use the same model to fit the data of property crimes. The results for property crime is in Table 3.2 where the λ of minimal MSE in the cross validation is nearly 0.005 now. And we show changes of coefficients with penalty parameter λ in Figure 3.2. Different from the case of violent crimes, under the significance level $\alpha = 0.05$, there are more significant coefficients here including "rincpc", "citybla", "sta_welf", "price" and "ln(police)". The new factors "sta_welf" and "price" have negative correlations with the property crime rates. This makes sense for more money is spent on the society welfare, property crime should decrease more. Yet here the unemployment rates keeps showing insignificant in both regression models which does not meet one's expectation.

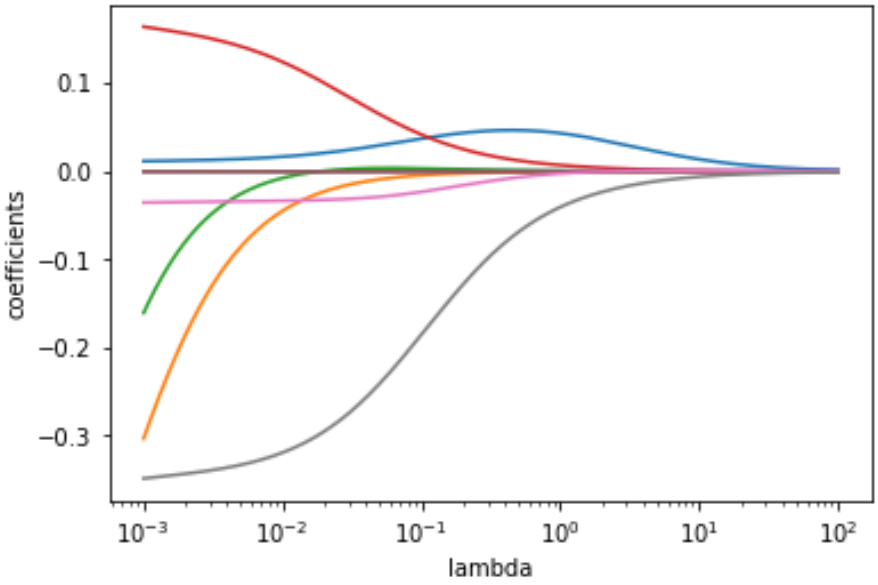

Figure 3.1: Violent Crime: the relationship between lambda and coefficients

	coef	std err	t	P > |t|	95%C.I.
rincpc	0.0180	0.003	5.325	0.000	(0.011,0.025)
econgrow	-0.1513	0.151	-1.004	0.316	(-0.447,0.145)
unmep	0.0118	0.204	0.058	0.954	(-0.389,0.413)
citybla	0.1567	0.023	6.713	0.000	(0.111,0.202)
sta_educ	6.693e-05	4.04e-05	1.656	0.098	(-1.24e-05,0.000)
sta_welf	-0.0002	5.01e-05	-3.769	0.000	(-0.000,-9.04e-05)
price	-0.0486	0.016	-3.112	0.002	(-0.079,0.018)
ln(police)	-0.3147	0.013	-24.328	0.000	(-0.340,-0.289)

Table 3.2: Property Crime($\lambda \approx 0.005$)

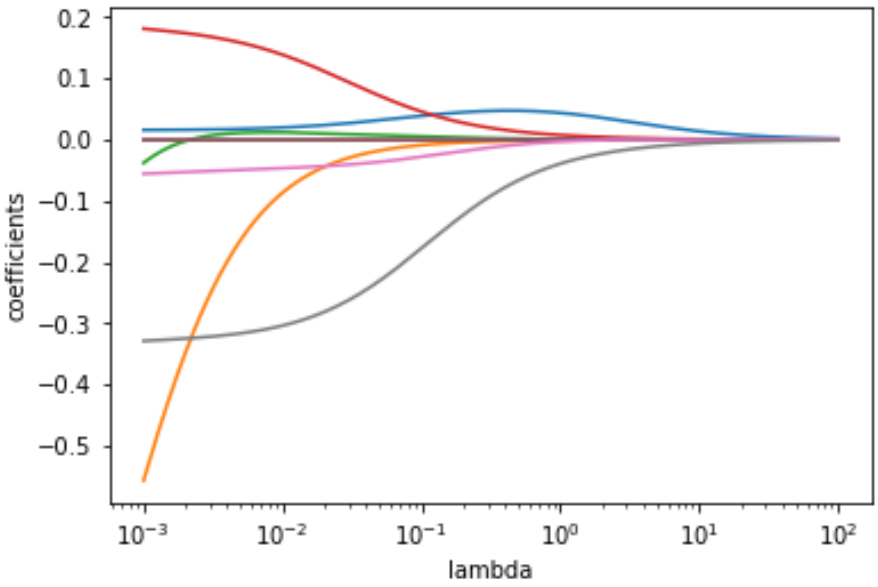

Figure 3.2: Property Crime: the relationship between lambda and coefficients

4 Summary

Actually there remains much exploration on this dataset, we chose the most directed method, regression model to figure out the relationship between crimes and a list of related variables. During the process, we found it is flexible to exert transformation on the data of different features. The transformation is usually necessary due to the magnitude of real data. Yet varied transformation might bring up different statistics results such as hypothesis test. Thus when analyzing data, we need to put enough attention on the processing of data with concerning the meaning and interpretation of this changing.

5 Remark of Contribution

Han Ruijian preprocessed the dataset and devised the regression model after reading the original article.

Tan Chunxi wrote python codes for regression model and generated the figures.

Ye Rougang summarized results of different models and wrote this report.