Study of cigarette sales in the United States

Ge Cheng[1, a],

[1]Department of pure mathematics and mathematical statistics, university of Cambridge, Cambridge, CB3 0BU, the United Kingdom

[a] gc541@cam.ac.uk

Keywords: cigarette sales, linear model, LASSO, principle component

Abstract: Nowadays more and more teenagers and adults are addicted to smoking, which makes an increasing number of people focus on this topic. Therefore, to analysis the factors that influence cigarette sales, we can try to find a strong approach to better control the smoking in the United States. Here we use several statistical models to finally find one that best describes our data.

1. Introduction

1.1 Background to cigarette sales

A cigarette is a small cylinder of finely cut tobacco leaves rolled in thin paper for smoking. It will make people addictive through the main chemical in tobacco called Nicotine. In fact, cigarettes can also carry serious health risks to people that are more widely spread than other tobacco products. According to the research, nearly fifty percent of cigarette-smokers die of tobacco-related disease and lose on average 14 years of life [1]. Therefore, cigarette sales face strict laws in different places over the world [2]. For example, The Food and Drug Administration (FDA) reissued that the sales of electronic cigarettes are banned to minors [3]; only the retailers who have the tobacco retail license can sale cigarette to public in California. These rules are helpful both to curb smoking in teenagers and to reduce sale opportunities to keep people from becoming addicted to them as easily. Nowadays, sale patterns for cigarette attract more and more researchers and scientists to study in order to find way that can limit cigarette consumptions. Our research plans to find this pattern in the United State.

1.2 Objectives and data description

In this project, we will try to fins possible factors influencing rates of cigarette sales over the years, and it is important to carefully consider potentially existing correlations between the variables.

The data we discuss here contains the sales of cigarette and some related properties that were recorded from 1963 to 1992 over 46 states in the United States. And there are exactly 8 variables concerning the topic.

State - state abbreviation for 46 states
Year - the year
Price - price per pack of cigarettes
Pop – population in different states and different years
Pop16 - population above the age of 16
Ndi - per capita disposable income
Sales - cigarette sales in packs per capita
Pimin - minimum price in adjoining states per pack of cigarettes

Clearly, out response variable is sales, and other seven items are the factors that may have a significant influence on cigarette sales. For example, the states have passed different laws on cigarette sales, which may have different impacts; people who obtain more disposable income have more opportunities to buy cigarettes than low-income smokers; also, taken population above 16 into account rather than the whole population may be meaningful as the cigarettes are banned to teenagers these years. All of these will be discussed in our model.

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

2. Exploratory analysis

In the exploratory analysis, we intend to analyze the raw materials, which normally need to be transformed or altered in the formal analyze part. Here we find that two population variables need to be transformed as the histogram of them are not properly distributed. In figure 1, both of them are extremely skewed in right tails and we should use log transformation.

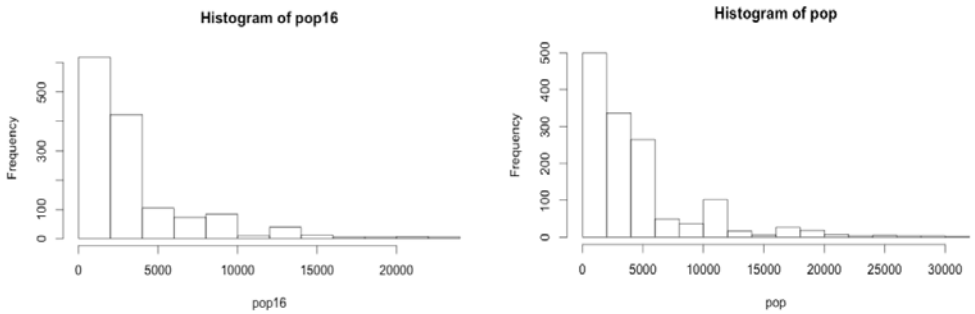

Figure 1 Histogram for improperly distributed variables

Then looking at the pair plots of these variables, in figure 2, there seems exist highly correlated relationships among variables. Indeed, the price of cigarettes would follow the trend of minimum price, as all the cigarette company were under the same environment and forced to compete with each other. Also, the welfare of the society like disposable income per person could lead to these company adjust the price of their products into a lower or higher level. As for total population and population above 16, it is obvious that they are linear related as population above 16 is definitely a subset of whole population, but the sales of cigarette might be related more with population above 16, because they are the target costumers of cigarette. We should note that, in a linear model, it is dangerous to put the highly correlated variables in the model, so we need to carefully choose the variables for our final model.

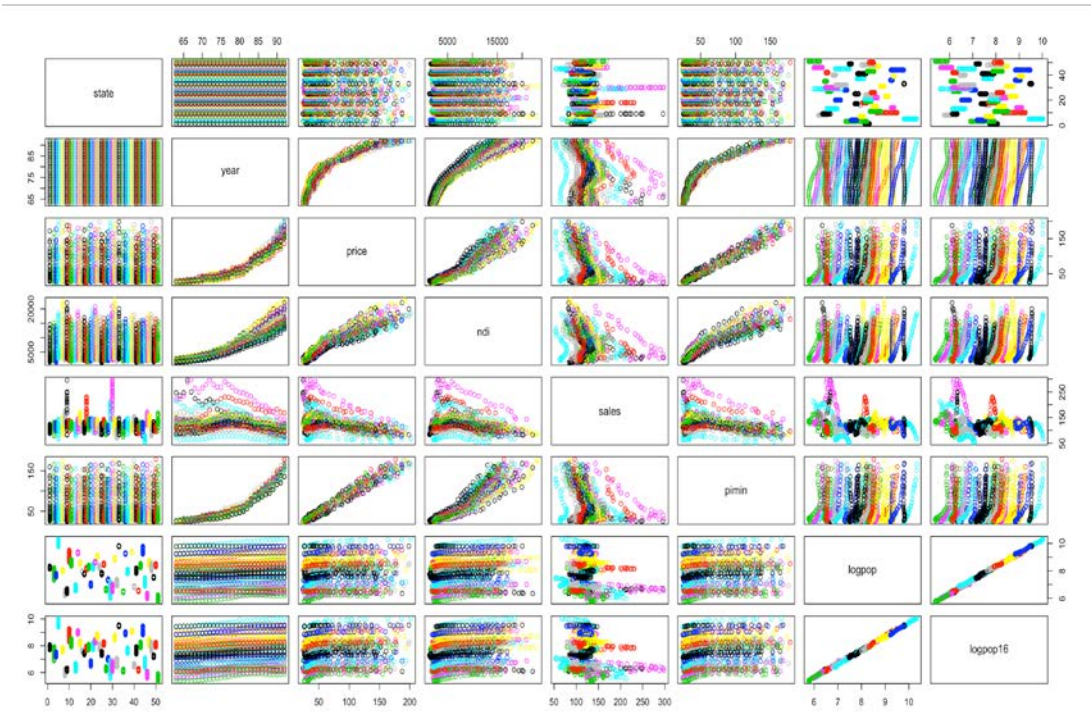

Figure 2 Pair plots for 8 variables in cigarette data

3. Modeling and results

3.1 Simple Linear models

Initially I will use the linear model and consider the correlation of 0.8 to my threshold. Thus using sales as the response variables and choosing combination of no highly related variables in the model, I will find the best model under this circumstance by using the AIC and adjust R Squares when all the models have the same number of parameters. Then the most proper simple linear model is shown below:

Sales ~ factor (state) + price + logpop16

It has an AIC value of 11360 and uses state1 as a baseline, comparing other states to it. Actually, in different states the sales are significantly different (with P-value less than 0.05 from ANOVA model, as it has so many levels that I will not put them into the outcome table.), which means the factor state has a significant influence on the sale of cigarette. In addition, the price, which is also significant in the model (P value less than 0.05), has the estimated coefficient -0.318, which means the sales of cigarette will reduce by -0.318 units with the price increased by 1 unit when other variables keep constant. Also, the coefficient for logpop16 is 37.975, which means the sales rate of cigarette will increase by 37.975 units with logarithm of the population above 16 increased by 1 units when other variables keep constant, and it is also significant in the model with p-value less than 0.05, which means they have imparts on the sales.

Table 1 Summary for simple linear model

	Coefficients	Standard deviation	P value
Price	-0.318	0.014	<0.001
Logpop16	37.975	3.573	<0.001

Finally, for the linear model, we should look at the diagnosis plot to check the model fitting. In figure 4, the plot of residuals against fitted value indicates there may still exist a little trend in the residuals, and the variance is not constant. In addition, the qq-plot, which used to check the normality assumption, shows some deviance in both tails. In other words, the normality of the model does not hold. Therefore, the linear model with two variables seems fit proper, but may be not the best one.

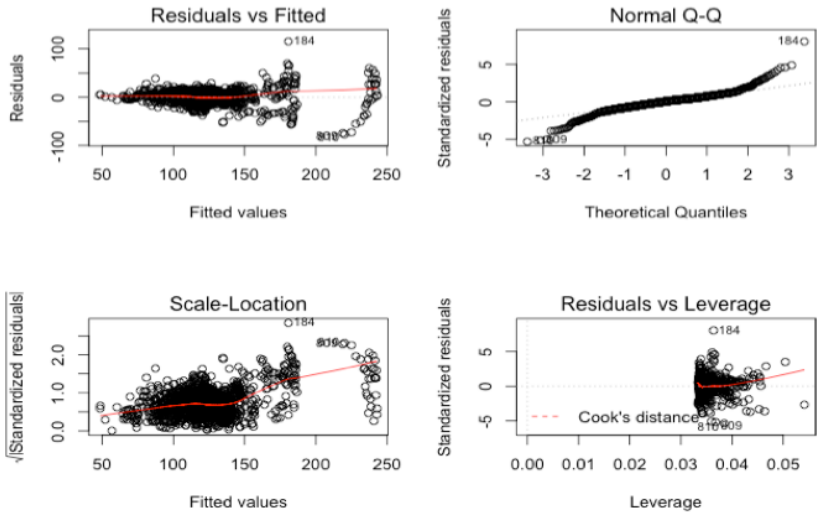

Figure 3 Diagnosis plot from the linear model

3.2 Linear model with Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO, which is similar to ridge regression, has constraint: sum of square coefficients should be less than a constant, t [4]. It tends to “shrink” the OLS coefficients towards zero, especially for

smaller values of t. Meanwhile, it can deal with highly correlated variable through forcing some coefficients to be 0.

Using the cigarette data in R with LASSO approach, I get the cp and the order of adding variables (figure 4). Then I choose p=8 with smallest cp and get the regression coefficients for original variables (Year: 0.000, price: -1.304, ndi: 0.004, pimin: 0.571, logpop: -140.245, logpop16: 135.045). From the coefficients we find that the coefficient of year has been shrunk to 0, and other coefficients has been reduced too.

At this model, one-unit increase of price per pack cigarette is companied by the 1.304 decrease in the sale rate of cigarette for overall states, while other variables remain unchanged. However, one-unit increase in minimum price of cigarette or people disposable money will lead to the sale rates increase by 0.004 or 0.571 units respectively if only changing one variables at the same time in a model. Then, the total population have a positive effect on sale rates while the population of people larger than 16 has a negative effect on it.

Finally, by calculation the predicted value, I find the AIC in this model is 15037, which need to be compared with other models.

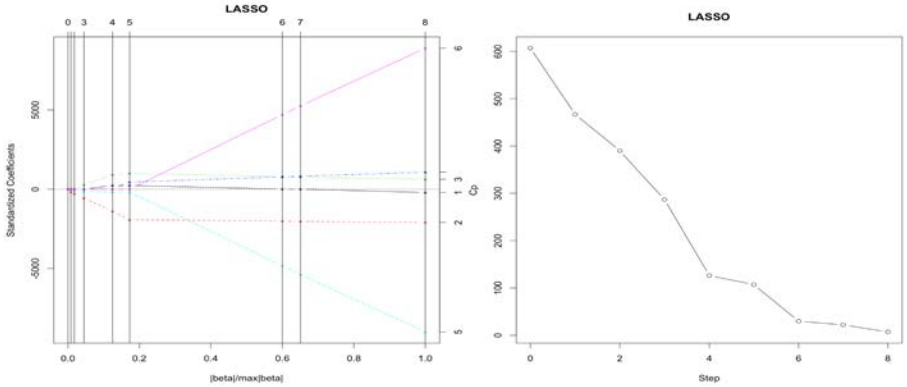

Figure 4 plot of lasso model: order of variables and the cp

3.3 Linear model with Principle components

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [5]. The original variables are highly correlated which can pose a problem for some model fitting (like regression), so I plan to construct new variables that will be uncorrelated by principle component.

Table 2 shows the importance of the components. I will use 0.9 as the threshold for proportion of variance that we want retained, and then, keep the number of eigenvalues that will give at least this proportion of variance. Finally, 2 principle components should be included in my model in terms of it.

Table 2 Importance of components

Component	1	2	3	4	5	6
Standard deviation	1.970	1.394	0.326	0.240	0.110	0.020
Proportion of variance	0.647	0.324	0.018	0.010	0.002	0.000
Cumulative proportion	0.647	0.971	0.988	0.998	0.999	1.000

According to table 2 and table 3, the first component accounts for 64.7% of total variation, and all the loadings have the same sign and only two different magnitudes, with about -0.5 and -0.1 respectively. So this component appears to be the sum of two averages measurement (the average of year, price, ndi and pimin plus the average of logpop and logpop16). Then, the second component has different signs. Hence it is a contrast between the price measurements (year, price, ndi and pimin) and population (logpop and logpop16). The expressions are showed as below:

Component 1 = -0.5*Year-0.5*Price-0.5*Ndi-0.5*Pimin-0.1*logpop-0.1*logpop16 (approx.)
Component 2 = -0.1*Year-0.1*Price-0.06*Ndi-0.08*Pimin-0.7*logpop-0.7*logpop16 (approx.)

Table 3 Loadings of two principle components

	Year	Price	Ndi	Pimin	Logpop	Logpop16
Component 1	-0.486	-0.494	-0.493	-0.493	-0.112	-0.131
Component 2	-0.089	-0.085	-0.058	-0.083	0.699	0.693

The formula for the final PC model is below, where PC denotes the 2 components together:

Sales ~ factor (state) + PC

Again, the state is significant in terms of the ANOVA function, which indicates that it definitely has an important influence on the rate of cigarette sales. Both components are all significant with p value extremely less than 0.05. Meanwhile, the estimated coefficients are positive, which means the components are positive related with the rate of cigarette sales. For this model, the AIC is 11497, which is larger than the simple linear model. And the residual plots did not change much compared with the previous one.

Finally, comparing all of this three models according to their AIC and the number of parameters, principle component method and simple linear model are better than the LASSO approach, but the AICs just show little differences between PC model and simple linear model. As statistical basis, we usually prefer a simple linear model as we tend to choose the easy model when the standard criteria is nearly same.

4. Conclusion

In conclusion, having considered these three models (the simple linear model, LASSO, Principle Component) in terms of their AIC (The smaller the AIC is, the better the model fit) and the number of parameters, the first model definitely fits better than the other three. So the variables price and logpop16 are included in the final model.

However, it is not the unique way to explain the data, because there are many correlated variables, and any combination of them will fit the model proper and thus leading to some limitations. These problems may result from the fact that the variables are similar in some way. For example, with time going by, the economy of the society will be better and people will become wealthier than before, then the disposable money per person will clearly increase due to the time passing by, all the the cigarette companies were planning to earn money as much as possible, so they would adjust the price to higher level when people had more money to spend on the cigarette. In turn, as the price of the cigarette became higher, it was obviously that the minimum price of the cigarette would increase too due to the wealthier environment. Further, pop16 is the subset of pop. When the total population increased, we could infer that the population of people larger than 16 would follow this trend too. Also, in terms of this analysis, there are just three valid variables in total, which is relatively small for a 1380 observations dataset. Therefore, in order to solve the limitation of the data, we should further find more variables that may be related with the rate of cigarette sales like prevalence and risk [6]. The bad fitting for the linear model may also due to the limited variables.

Additionally, in order to promote the accuracy of the linear model, we need to explore more useful variations. For example, the transport situation may be included, which indicates the degree that people had access to the cigarettes. Also, the spreading of information like advertisement may also attract large number of people, then promote the sales. The limited information and the highly correlated variables together in the dataset should be responsible for the deviance in the model.

Finally, the main purpose of this study is to control the cigarette sale in the whole society. And we conclude that the price has a negative impact on the rate of cigarette sale. Actually, everyone will be sensitive about the price, because we all prefer to by the products with low price and will be

h5esitated when the item is expensive. Therefore, policy makers can then improve tax on cigarettes or simply improve cigarette price to reduce the consumption of cigarettes. In addition, the population above 16 has a positive effect on cigarette sales, which partly shows that some policies have successfully inhabited the smoking in teenagers. And also, they can develop some strict rules for adult regarding smoking like giving extra bonus on these who are not smoking and totally forbidding smoking cigarette in public areas. When all the citizens follow these rules, both the cigarette sale and smoker will decrease.

5. Reference

[1]. Doll R., Peto R., Wheatley K., Gray R., Sutherland I. Mortality in relation to smoking: 40 years' observations on male British doctors. BMJ. 1994;309(6959):901–911.

[2]. Jason LA, Ji PY, Anes MD, et al. (1991) Active enforcement of cigarette control laws in the prevention of cigarette sales to minors. JAMA 266:3159–3161

[3]. Food and Drug Administration. Regulations restricting the sale and distribution of cigarettes and smokeless tobacco to protect children and adolescents; final rule. August 28, 1996; Fed Regist. 1996 [Accessed December 12, 2007.]. p. 44395-44445.

[4]. R. Tibshirani, "Regression Shrinkage and Selection via the LASSO", J. Royal Statistical Soc. B, vol. 58, no. 1, pp. 267-288, 1996.

[5]. I. T. Jolliffe, Principle Component Analysis, 1986, Springer-Verlag.

[6]. Kuri-Morales PA, Cortes-Ramirez M, Cravioto-Quintana P. [Prevalence and risk factors related to sale of cigarettes to minors in stores in Mexico City] Salud Pública de México. 2005; 47:402–412.