Analysis of opioid drug crisis

Yuanfang Zhang, Shuo Miao, Yuting Li

Harbin Huade University, Harbin, Heilongjiang, 150025

Email:306492896@qq.com

Keywords: multivariable linear regression model, Factor Analysis, Cluster Analysis

Abstract: In recent years, with the proliferation of opioid use, the country is suffering from a serious drug crisis. This article will study this. First, we used the data provided by NFLIS to use Excel software to plot bar graphs of Total Drug Reports status data for five states from 2010 to 2017. By analyzing the data, we decided to use a multiple linear regression model. We used Matlab software to calculate the relevant data, and then used Google Earth Pro software to map the trend of the total amount of opioids in five states. Then we added the instability factors of family and marriage to the multiple linear regression model established by the first question to achieve the purpose of Optimizing the first part of the problem model. Finally, we conclude that opioid abuse is associated with destabilizing factors in family marriages. In unstable families, families are less able to bind their families and make family members more susceptible to depression. In the third part, we build a complete evaluation system through principal component analysis with the multi-possible variables to make the model perfect fit data to establish a complete evaluation system by searching data and data.

1. Introduction

With the proliferation of opioid use in recent years, the country is facing a serious crisis. Opioids are widely used in the medical community to control pain and treat diseases. However, the use of illegal use will not only have a negative impact on people's mind and body, but also seriously affect the country's economic and cultural development. For example, young people regard opioids as recreational drugs, which will affect the physical and mental development of young people and affect the overall level of the country. Therefore, it is imperative to restrict the large-scale use of opioids in illegal use. Therefore, many departments and agencies in the United States have paid a lot of manpower, resources and financial resources to control the spread of illegal use of opioids.

In recent years, the illegal use of opioid drugs has become more rampant, and opioids are a very effective drug if they are used in medical treatment. The United States has also mobilized a large amount of manpower and resources to restrict the spread of opioids.

The first part of the question calls for the use of NFLAIS data from Ohio, Kentucky, Suffolk, Virginia, and Tennessee to create a mathematical model. The second part of the question asks to use the US Census socioeconomic data to determine and analyze why the current large use of opioids has reached the current level. The third part of the question calls for the identification of possible strategies for the opioid crisis in conjunction with the results of the first and second parts.

The submission should include:

(1) One-page Summary Sheet

(2) One- to Two-page memo

(3) The solution should be no more than 20 pages and a maximum of 23 pages, a summary memo.

(4) The reference list and any appendices, excluding the 23 page limit, should appear after the solution.

2. The description of the problem

Problem statement

  DOI: 10.25236/issec.2019.150

In recent years, the illegal use of opioid drugs has become more rampant, and opioids are a very effective drug if they are used in medical treatment. However, if the illegal use of opioids is a recreational drug, it will not only have a negative impact on the people's mind and body, but also affect the country's economic and cultural development. The United States has also mobilized a large amount of manpower and resources to restrict the spread of opioids.

The first part of the question calls for the use of NFLAIS data from Ohio, Kentucky, Suffolk, Virginia, and Tennessee to create a mathematical model. To look for opioids and heroin events. The characteristics of communication between these five states and their counties. This mathematical model is used to determine the possible locations where opioids begin to be used in these five states.

The second part of the question asks to use the US Census socioeconomic data to determine and analyze why the current large use of opioids has reached the current level. Who is using opioids? What has led to an increase in the number of opioid addictions? Is the trend in the use of opioids related to the US Census socioeconomic data? If so, please modify the first part of the model to include this data. In the question, you want to gather any important factor.

The third part of the question calls for the identification of possible strategies for the opioid crisis in conjunction with the results of the first and second parts. And use the model to test the effectiveness of this strategy. And determine the success or failure depends on any important parameter boundaries.

In addition to the main report, a one to two page memo will be included to give the lead administrator the DEA/NFLIS database. Summarize any reconsiderations or results you find during the modeling process. The submission should include:

(5) One-page Summary Sheet

(6) One- to Two-page memo

(7) The solution should be no more than 20 pages and a maximum of 23 pages, a summary memo.

(8) The reference list and any appendices, excluding the 23 page limit, should appear after the solution.

3. The first part of the problem

Problem hypothesis

1) Suppose the data we selected are very incomplete but not affected

2) Hypothesis model is built in ideal state

3) Assuming that the model is not disturbed by external conditions

4. Model establishment

4.1 Analysis of the basis and model of model establishment

First, based on the information given in the title, we use excel software to map the total number of drug reports in five states in the United States from 2010 to 2017. Trying to find the relationship between the total number of drug reports in five states and time. The total number of drug reports in the five states is in the histogram of 2010 to 2017, as shown in Figure 1.

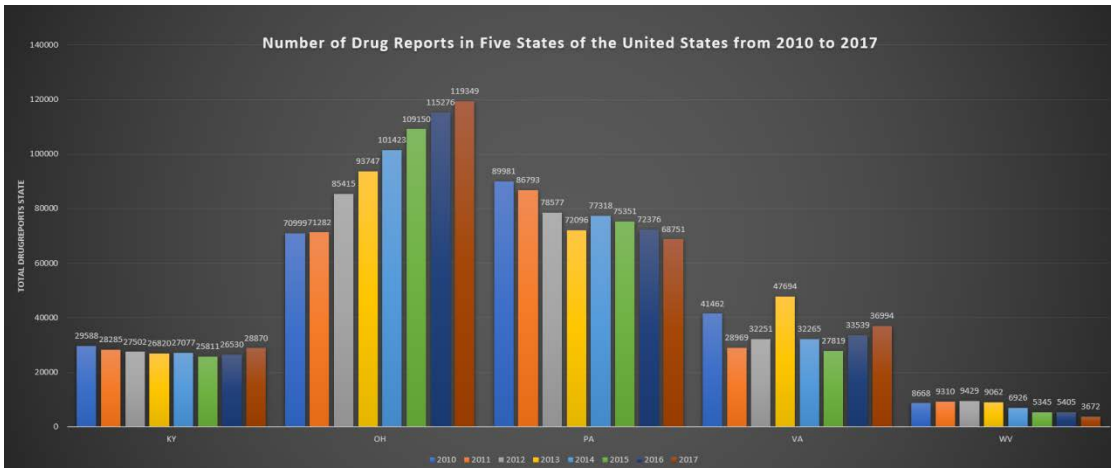

Fig.1: Number of Drug Reports in Five States of the United States from 2010 to 2017

From the above fig.1, we can easily see that the total number of drug reports in each of the five states is normally distributed. Thus we have drawn statistical charts of the five state drug reports from 2010 to 2017.

The statistical chart is shown in Fig.2.

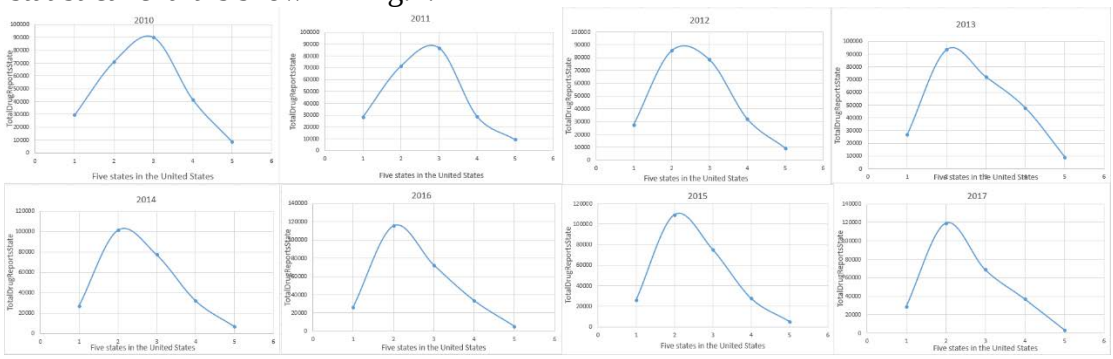

Fig2: statistical charts of the five state drug reports from 2010 to 2017

In Fig. 2, '1"2"3"4"5' in the abscissa represents KY, OH, PA, VA, WV from left to right. Through the analysis of Fig.2, we can find that the data in 2012 is relatively flat. In order to eliminate the error caused by the accident, we chose the relatively stable 2012 data as the research object. A multivariate linear regression curve model can be used to fit the normal distribution curve of the total drug reported in 2012, so that we have found opioid and heroin events in these five states and their counties by establishing mathematical models.

The specific analysis process of the multiple linear regression analysis model is as follows:

We can think of 'FIPS_Combined' as the information used to represent the state and country locations in this question.

In order to determine the possible locations where opioids begin to be used in these five states.

We use the four data sets 'FIPS_Combined' , 'DrugReports' , 'TotalDrugReportsCounty' and 'TotalDrugReportsState' in the most stable 2012 data in the MCM_NFLIS_Data.xlsx data table with 'FIPS_Combined' as the dependent variable, 'DrugReports', 'TotalDrugReportsCounty ' and 'TotalDrugReportsCounty' as the independent variable. And a multiple liner regression model is established to use 'Drugreprots', 'TotalDrugReportsCounty' and 'TotalDrugReportsState' to estimate the possible locations where opioids begin to be used in these five states.

4.2 Establishment of the multivariable linear regression model

According to the previous analysis, we can establish the multivariable linear regression model to know the relationship between ‘DrugReports’ and ‘DrugReports’,‘TotalDrugReportCounty’ and‘TotalDrugReportsState’ by using the data from MCM_NFLIS_Data.xlsx.

The multiple linear regression model we built is as follows:

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \delta^2) \end{cases} \quad (1)$$

$\beta_0, \beta_1, ..., \beta_m$ are regression coefficients which can indicate the strength of the relationship between dependent variable 'DrugReports' and Independent variables 'DrugReports', 'TotalDrugReportCounty' and 'TotalDrugReportsState'. In this question, since we only study the relationship between the three independent variables and the dependent variable, we take the above formula: m=3

And, we set the observation value of x_1, x_2, x_3, y to $a_{i1}, a_{i2}, a_{i3}, d_i$

$$X = \begin{bmatrix} 1 & a_{11} & a_{12} & a_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{2764,1} & a_{2764.2} & a_{2764.3} \end{bmatrix}, Y = \begin{bmatrix} d_1 \\ \vdots \\ d_{2764} \end{bmatrix}$$

4.3 Solution of multiple linear regression model

First, we use the mathematical method of least squares to find the predicted values of z_0, z_1, z_2, z_3. Therefore, we should choose the predicted value as $\hat{z}_j$, when $z_j = \hat{z}_j$, $j = 0,1,2,3$ its's error sum of squares is

$$Q = \sum_{i=1}^{2764} \varepsilon_i^2 = \sum_{i=1}^{2764} (d_i - \hat{d}_i)^2 = \sum_{i=1}^{2764} (d_i - z_0 - z_1 a_i - z_2 a_{i2} - z_3 a_{i3})^2$$

Where Q is the sum of the squared errors between our predicted FIPS Combined and actual values.

In order to minimize the sum of the squared errors between the predicted and actual values, we make:

$$\frac{\partial Q}{\partial c_j} = 0, \ j = 0,1,2,3$$

So we can get the normal equations:

$$\left[\hat{z}_0, \hat{z}_1, \hat{z}_2, \hat{z}_3 \right] = (X^T X)^{-1} X^T Y$$

We use the Matlab to solve the normal equations to get the values of z_0, z_1, z_2, z_3, which are predicted values.

$$\hat{z}_0 = 3929.445, \hat{z}_1 = -1.3242, \hat{z}_2 = 0.1157, \hat{z}_3 = 0.0056$$

We can get the regression equation:

$$y = 3929.445 - 1.3242 x_1 + 0.1157 x_2 + 0.0737 x_3 \quad (2)$$

The function describing the strength of the relationship between FIPS_Combined and 'DrugReports' , 'TotalDrugReportsCounty' and 'TotalDrugReportsState' is (2).

4.4 Test of multiple linear regression model

(1)F Test

We can see if there is a linear relationship between the dependent variable y and the independent variable x_1, x_2, x_3 through the test. If $|\hat{z}_j|, j = 1,2,3$ are small, then the linear relationship between

y and x_1, x_2 , x_3 is not obvious. So we can use the original assumption:

$$H_0 : z_j = 0,\ j = 1,2,3 \tag{3}$$

Make $m = 3, n = 2764$, then

$$Q = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (d_i - \hat{d}_i)^2, U = \sum_{i=1}^{n} (d_i - d)^2$$

$$\hat{d}_i = \hat{z}_0 + \hat{z}_1 a_{i1} + \cdots + \hat{z}_m a_{im} (i = 1, \cdots, n), \bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$$

When H_0 is true,

$$F = \frac{U / m}{Q / (n - m - 1)} \sim F(m, n - m - 1)$$

Under the significance level of α if:

$$F_{1-\frac{\alpha}{2}}(m, n - m - 1) < F < F_{\frac{\alpha}{2}}(m, n - m - 1)$$

Then accept H_0, otherwise refuse it.

We use the Matlab to get the statistics $F = 0.7709$,

Therefore, we deny the original hypothesis (3). The model passed the F test at this time.

(1)T Test

If H_0 in (3) is rejected, β_j is not all 0. But we do not rule out that there are several cases where β_j is equal to 0. So we have to carry out further tests, the specific process is as follows:

$$H_0^{(j)} : z_j = 0,\ j = 0,1,\cdots,m, \tag{4}$$

When $H_0^{(j)}$ is true,

$$t_g = \frac{\hat{\beta}_g / \sqrt{c_{gg}}}{\sqrt{Q / (N - M - 1)}} \sim t(n - m - 1)$$

In the formula, c_{jj} is the element (j, j) of $(X^T X)^{-1}$. We can accept $H_0^{(j)}$ if $|t_j| < t_{\frac{\alpha}{2}}(n - m - 1)$ under the given α , otherwise it is not acceptable. We use Matlab to solve the problem and get the statistics:

$$t_0 = 86.823,\ \ t_1 = -0.743, t_2 = 1.139, t_3 = 0.694$$

By testing T, we can get the $\frac{\alpha}{2}$ Quantile:

$$t_{0.2}(2760) = 0.8418$$

For the test of the formula (4.13), if $\alpha = 0.4$, $H_0^{(j)} : c_j = 0 (j = 2,3)$ is unacceptable while $H_0^{(j)} : z_j = 0, (j = 0,1)$ is acceptable.

So variable x_1 has a significant impact on the model. When building a linear model, x_1 , α must be used to be a significant level.

4.5 Conclusion of multiple linear regression model

First, based on the multiple linear regression equations and data from the first problem, we can see that the reported amount of drugs is better controlled in some states. We used the Excel to get the Scatter diagram and its trend line of opioid event reports in five states in 2010-2017 and found that the reported volume of drugs in the OH is increasing year by year. While that in the PA is declining. In general, the total number of drug reports in the five states is on the rise due to the improvement in the living standards of the population in recent years. The United States is experiencing a national crisis regarding the use of synthetic and non-synthetic opioids. To this end, government departments should strictly control the use of opioids and reduce their dependence on and demand for opioids.

Statistics show that the US population accounts for only 5% of the world's population, but opioids consumption accounts for 80% of the world's total. In this case, it is inevitable that a large number of people will become addicted to it and die. To this end, the US government should strictly control its opioids and consider that the spread of opioids is caused by profit-driven health care, inadequate regulation of the pharmaceutical market, or the drive of economic crisis. According to the scatter diagram and its trend line of opioid event reports in five states in 2010-2017

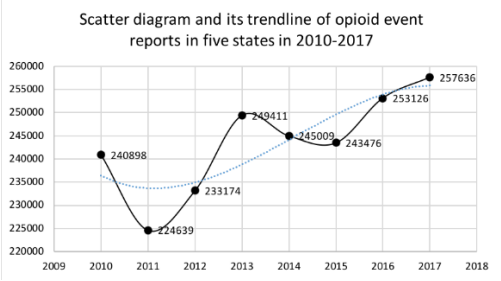

Fig.3: Scatter diagram and its trend line of opioid event reports in five states in 2010-2017

We can see that the opioid incident report in 2010-2016 is generally on the rise from fig.3. To analyze the state's share of trends more accurately, we produced a pie chart of the total number of drug reports for the five states in 2010-2017, as shown in fig.4.

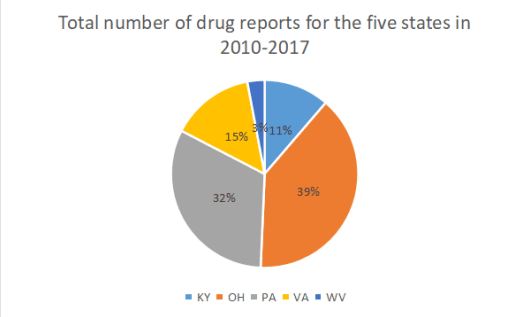

Fig.4: Total number of drug reports for the five states in 2010-2017

By analyzing fig. 4, it is found that PA and OH account for more than 70% of the five state opioid events. The threshold for drug levels was identified by analyzing the situation in both states. According to the model in the first part, the use of opioids in PA has been growing in recent years and the trend of growth has decreased with time. Through the above model, the threshold value is predicted to be about 130,000, and the fluctuation range is within 5000. It is predicted that it will occur in PA and OH in 2018-2020 with a probability of occurrence of 87%.

5. Establishment of Principal Component Analysis Model

Five data selected by the second portion of the cluster, and 10 sets of data in the table is selected by analyzing ACS_10_5YR_DP02_with_ann.csv need to implement strategies as variables by principal component analysis of the data to ensure be simplified without affecting the results. That is, the weft processing of the high latitude space variable. The scores obtained by principal component analysis are combined with the actual situation to calculate the validity.

roots decrease rapidly, so we can think that the factors obtained by principal component analysis

$$\tilde{a_{ij}} = \frac{a_{ij} - \mu_j}{s_j}, i = 1, 2, \cdots, 464; \ j = 1, 2, \cdots, 10,$$

Among them, $\mu_j = \frac{1}{n}\sum_{i=1}^{464} a_{ij}; s_j = \sqrt{\frac{1}{464-1}\sum_{i=1}^{464}\left(a_{ij} - \mu_j\right)^2}, j = 1, 2, \cdots, 10,$ are the sample mean and sample standard deviation of j indicators.

Correspondingly,

$$\tilde{x}_j = \frac{x_j - \mu_j}{s_j},\ j = 1, 2, \cdots, 10$$

Table 1 Principal component analysis result

Serial number	Contribution rate	feature	Cumulative contribution
1	72.988	7.299	72.988
2	12.369	1.237	85.358
3	6.655	0.666	92.013
4	4.147	0.415	96.160
5	2.450	0.245	98.611
6	0.720	0.072	99.331
7	0.400	0.040	99.732
8	0.188	0.019	99.920

By observing the above table, we can know that the cumulative contribution rate of the first eight eigenvalues reaches 99.9% or more. Next, we select the first 8 data for comprehensive evaluation. The first 8 feature roots are as follows

Table 2

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$	$\tilde{x}_9$	$\tilde{x}_{10}$
1	0.352	0.364	0.365	0.142	0.033	0.322	0.352	0.335	0.362	0.345
2	-0.072	-0.003	-0.042	0.623	0.758	-0.094	0.045	-0.116	-0.021	-0.032
3	-0.075	0.023	-0.031	0.730	-0.651	-0.129	-0.081	-0.114	-0.005	-0.003
4	-0.372	-0.102	-0.218	-0.003	-0.008	0.632	-0.034	-0.428	0.124	0.452
5	-0.179	0.284	0.024	-0.223	0.015	-0.390	0.551	-0.526	0.314	-0.070
6	-0.097	-0.125	-0.111	-0.060	0.024	-0.486	0.099	0.201	-0.043	0.783
7	0.622	-0.211	0.059	0.016	-0.001	-0.189	-0.414	-0.452	0.333	0.207
8	-0.224	0.663	0.385	-0.010	0.004	-0.080	-0.513	-0.171	-0.221	0.122

Thus, the eight principal components are respectively

$$y_1 = 0.352\tilde{x}_1 + 0.364\tilde{x}_2 + \cdots + 0.345\tilde{x}_{10}$$
$$y_2 = -0.072\tilde{x}_1 - 0.003\tilde{x}_2 + \cdots - 0.032\tilde{x}_{10}$$
$$y_3 = -0.075\tilde{x}_1 + 0.023\tilde{x}_2 + \cdots - 0.003\tilde{x}_{10}$$
$$y_4 = -0.372\tilde{x}_1 - 0.102\tilde{x}_2 + \cdots + 0.452\tilde{x}_{10}$$
$$y_5 = -0.179\tilde{x}_1 + 0.284\tilde{x}_2 + \cdots - 0.070\tilde{x}_{10}$$
$$y_6 = -0.097\tilde{x}_1 - 0.125\tilde{x}_2 + \cdots + 0.783\tilde{x}_{10}$$
$$y_7 = 0.622\tilde{x}_1 - 0.211\tilde{x}_2 + \cdots + 0.207\tilde{x}_{10}$$
$$y_8 = -0.224\tilde{x}_1 + 0.663\tilde{x}_2 + \cdots + 0.122\tilde{x}_{10}$$

We use the contribution rate of 8 principal components to establish a principal component comprehensive evaluation model.

$$Z = 0.73y_1 + 0.124y_2 + 0.067y_3 + 0.042y_4 + 0.025y_5 + 0.007y_6 + 0.004y_7 + 0.002y_8$$

Bringing the eight principal components of each county into the above model, we can get the county effectiveness score of 76.52%.

5.1 Principal Component Analysis Model Conclusion

By using Principal Component Analysis on the basis of the second part, the difficulty of data

processing at high latitudes is solved. By analyzing the factors affecting the effectiveness, the independent variables are roughly selected by principal component analysis to accurately select the independent variables with large influence. Accurately and efficiently calculate the score for effectiveness.

6. Evaluation of the model

6.1 Dvantages of the model

1) Making full use of data when dealing with more data, and have better implementability in dealing with big data.

2) The model is solved using professional mathematical software with high reliability.

3) Making full use of the relationship between data to select more useful data when dealing with the problem of more data,, which is convenient for the later calculation and scoring system.

6.2 Disadvantages of the model

1) It is difficult to achieve ultra-high precision in the face of accuracy data, which is relatively cumbersome in information processing and sorting.

2) The model establishment environment is an ideal environment, and there will be errors in the real environment.

3) When dealing with problems, it is necessary to consider the influence of multiple factors on the results, so it is difficult to accurately process some data.

References

[1] In Jae Myung, Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology,2003:78-86

[2] Yonghong Hu, Enhui He, Comprehensive evaluation metho[M].Beijing: Science Press, 2000:172-179.

[3] Diekmann O, Heesterbeek J A P. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation [M]. Wiley, 2000.

[4] Si Shoukui, Sun Xijing, Mathematical Modeling, Beijing, National Defense Industry Press,2011 19-27

[5] Zhuo Jinwu, Application of MATLAB in Mathematical Modeling, Beijing, Beijing University of Aeronautics and Astronautics Press,2011 56-59

[6] Miller,Factor Analysis: Statistical Methods and Application Issues,Shanghai, Gezhi Publishing House, 2007 103-104

[7] Watts D J, Strogatz S H. Collective dynamics of ‘small-world’networks [J]. nature, 1998, 393(6684): 440-442.