Model Construction

The multivariable linear regression model(MLR)

1 Construction of MLR

We develop a multivariable linear regression model to analyze the differences between the number of drug reports among counties. According to the data provided from 2010 to 2016, we make an elementary assumption that there is a linear relationship between independent variables and the number of drug reports. Since there are many possible *socio-economic factors* relative to the results, the multivariable linear regression model can explain the complexity of the system and explore the most important factors that contribute to the results more comprehensively and accurately.

此处需要在之前先引入变量，表示社会经济因素序号为HC01\_VCi，县编号为lk的数据，一列小x（i, lk）合成向量xi。p（lk）表示县编号为lk的drug reports 数量，合成列向量P表示目标向量。

To reflect the hidden relationship between the severity of drug abuse and various factors, we include as many possible factors as we can to modify our multivariable linear regression model.

令bi表示影响因素xi的回归系数。e为假设的随机偏差值，满足E（e）=0，COV(e，e)=sigma方In，sigma方为未知参数，D（et）=sigma方，即对e1,…en做无偏、等方差假定（变量表）

We introduce a basic multivariable linear equation to the model：

P=b0+b1x1+…bnxn+e

The equation set is showed below:

P=Xb+e

E（e）=0

COV(e,e)=sigma方In

We abbreviate the equation set to (P,Xb,COV)

And the regression function is

p一弯=b0+b1x1+…+bnxn

2. Results

Since the system involves quite a lot of underlying factors and there is complex multiple linearity between several elements, we used stepwise regression to better calculate the data.

Below are some indicators related to the regression equation

Intercept(Constant b0) 0.00314727

RESE (Root mean squared error) 0.0152012

R-square(Coefficient of determination) 0.949601

Adj R-sq (Degree-of-freedom adjusted coefficient of determination) 0.945062

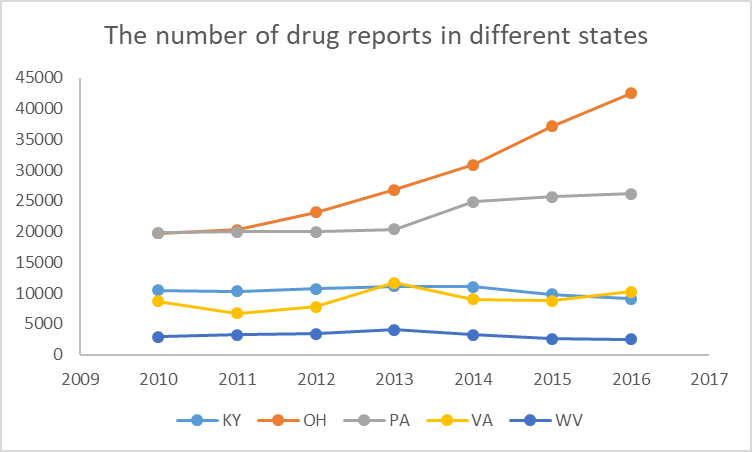
According to the data listed above, the regression is able to fit the actual data well.We finally select 38 factors with the most significant impact through the stepwise regression, some of which are listed below.

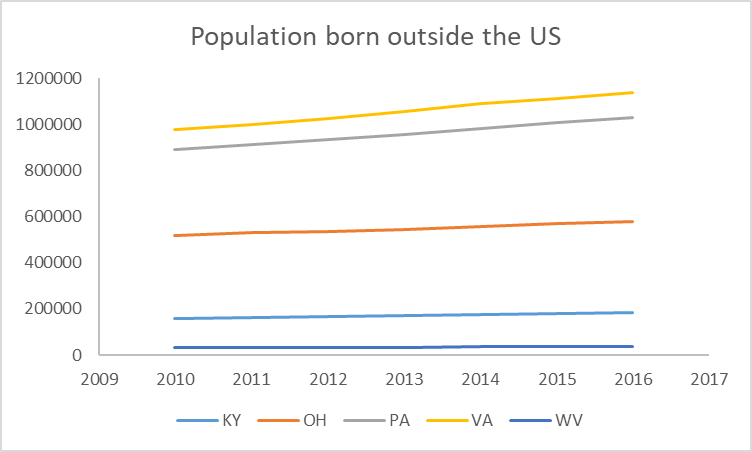
表，列了一些影响因素，要稍微解释一下

|  |  |  |  |
| --- | --- | --- | --- |
| **HC01\_VC11** | **3.0505** | 3.0505 | HOUSEHOLDS BY TYPE - Family households (families) |
| **HC01\_VC144** | **2.6515** | 2.6515 | Population born outside the United States |
| **HC01\_VC12** | **2.6013** | -2.6013 | Estimate; HOUSEHOLDS BY TYPE - Family households (families) - Female householder, no husband present, family - With own children under 18 years |
| **HC01\_VC151** | **2.5784** | -2.5784 | Estimate; YEAR OF ENTRY - Entered 2010 or later |
| **HC01\_VC28** | **1.572** | 1.572 | Estimate; RELATIONSHIP - Child |
| **HC01\_VC14** | **1.3058** | 1.3058 | Estimate; HOUSEHOLDS BY TYPE - Nonfamily households - Householder living alone |
| **HC01\_VC177** | **1.2417** | -1.2417 | Estimate; LANGUAGE SPOKEN AT HOME - Language other than English - Other languages |
| **HC01\_VC91** | **1.0572** | 1.0572 | Estimate; EDUCATIONAL ATTAINMENT - Graduate or professional degree |

’-‘ REPRESENTS A NEGATIVE IMPACT

The table shows us a lot of information. ‘Family households(families)’, ‘Population born outside the United States’ have the greatest promoting influence on the number of drug reports. While ‘Family households (families) -Female householder -With own children under 18 years’ and ‘YEAR OF ENTRY - Entered 2010 or later’ restrain the acceleration of drug abuse.





聚类分析

Cluster analysis

To analyze the data more efficiently and minimize , we then categorize these factors based on cluster analysis. The classifications including ……, and the specific category is listed in the table below.