# The Analysis and Intervention Model of Strokes' Environmental Factors

Jun-er Ma[1], Meng-yuan Li[2], Dong-ming Li[3]

*[1]Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China*
*[2]Hangzhou Tian Hang Experimental School, Hangzhou 310004, China*
*[3]Hangzhou Xia Yan Middle School, Hangzhou 310017, China*
*maje@zjweu.edu.cn, limengyuan1989@126.com, ldm_630910@hotmail.com*

ABSTRACT: Through large amounts of data analysis and processing, combined with multiple regression analysis theory, this paper obtains 48 groups of sample points counted by the 07-10 years' environmental factors (pressure, temperature and relative humidity) and morbidity per month. This model first utilizes the five variables' linear regression models with MATLAB software for solving and analysis. By improving the five variables' pure quadratic polynomial regression model: $y = b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{5} b_{ii} x_i^2$ to obtain a more suitable model formula. Finally, enter the value of software's interactive graphics window argument $x_0$ to give the corresponding fitted value $y_0$, forecasted the numbers of disease.

KEYWORD: Stroke, Pathogenesis, Multiple regression equation, Prevention

## 1 THE RESEARCH BACKGROUND

Stroke is a serious disease that currently threatens human life, its occurrence is a long process. Once contract the disease, it's difficult to reverse. Causes of the disease have been linked to environmental factors, including a close relationship between the air temperature and humidity. Analysis the environmental factor on the incidence of stroke, its purpose is to conduct a risk assessment of the disease. At the same time, through the establishment of a data model to master the regularity in disease incidence, it has practical significance for health authorities and medical institutions to allocate the medical strength more reasonable, improving diagnosis and treatment environment, deploying the beds and medical drugs.
Data (see Appendix-C1) comes from some city hospitals in China, includes the stroke patients' information from January 2007 to December 2010 and the corresponding period of local daily meteorological data (see Appendix-C2).

## 2 ASSUMPTIONS OF THE MODEL

1) Assuming that main factor for the incidence of stroke is decided by the temperature, pressure and average humidity.

2) Assuming monthly average maximum and minimum temperature and barometric pressure and relative humidity are normal distribution.

3) Assuming that whether the incidence among patients is independent

4) All the missing data and erroneous data won't do analysis

5) Assuming that the total population of the city remained unchanged in four years, a total population of $M_0$.

## 3 EXPLANATION OF SYMBOLS

$X_1$: Monthly average maximum gas pressure;
$X_2$: Monthly average minimum pressure;
$X_3$: Monthly average maximum temperature;
$X_4$: Monthly average minimum temperature;
$X_5$: Monthly average humidity;
y: The number of monthly incidence;
$b_i : (i = 0,1,\cdots,5)$ Regression coefficients.

## 4 MODEL'S ESTABLISHMENT AND SOLUTION

This paper establishes a mathematical model to research the relationship between the incidence of stroke, air temperature, barometric pressure and relative humidity. Then solving the mathematical model.
Method 1.

### 4.1 *Establishment of multiple linear regression equation*

In practical problems, a random variable (such as count number of the incidence of stroke Y) is often associated with multiple variables $X_1, X_2, \cdots, X_m$ $(m \geq 2)$ with correlation. This model analyzes problems with multiple linear regression, set up, $X_1$, $X_2$, $\cdots$, $X_m$ follow a normal distribution, conduct independent experiments, experimental data were obtained as follows: $\mathrm{x}_{1\mathrm{k}}$, $\mathrm{x}_{2\mathrm{k}}$, $\cdots$, $\mathrm{x}_{m\mathrm{k}}$ and $\mathrm{y}_{\mathrm{k}}$ (k = 1,2, $\cdots$, n), respectively $X_1$, $X_2, \cdots$, $X_m$ and Y values in the k th observation test.

If there is a linear correlation between the random variable Y and variable $X_1, X_2, \cdots, X_m$, then we can use the follow multiple linear equations to describe.

$$\mathrm{y}=\mathrm{b}_0+\mathrm{b}_1 x_1+\mathrm{b}_2 x_2+\cdots+\mathrm{b}_\mathrm{m} x_m \qquad (1)$$

Using least square method to determine the unknown paramet $\mathrm{b}_0, \mathrm{b}_1, \mathrm{b}_2, \cdots, \mathrm{b}_\mathrm{m}$ (Denoted $b_0 = a$), considering sum of deviations squares.

$$S_1=\sum\nolimits_{k=1}^{n}(y_k-a-b_1x_{1k}-b_2x_{2k}-\cdots-b_mx_{mk})^2 \qquad (2)$$

In order to make $S_1$ the minimum values, seeking the partial derivatives of $S_1$ to $a$, $\mathrm{b}_1, \mathrm{b}_2, \cdots, \mathrm{b}_\mathrm{m}$ and let them equal to zero, organizing get the following equation:

$$\begin{cases} \mathrm{na}+\left(\sum\limits_{k=1}^{n} x_{1k}\right)b_1+\cdots+\left(\sum\limits_{k=1}^{n} x_{mk}\right)b_m=\sum\limits_{k=1}^{n} y_k \\ \left(\sum\limits_{k=1}^{n} x_{1k}\right)a+\left(\sum\limits_{k=1}^{n} x_{1k}^{2}\right)b_1+\cdots+\left(\sum\limits_{k=1}^{n} x_{1k}x_{mk}\right)b_m \\ =\sum\limits_{k=1}^{n} x_{1k}y_k \\ \cdots\cdots \\ \left(\sum\limits_{k=1}^{n} x_{mk}\right)a+\left(\sum\limits_{k=1}^{n} x_{mk}x_{1k}\right)b_1+\cdots+\left(\sum\limits_{k=1}^{n} x_{mk}^{2}\right)b_m \\ =\sum\limits_{k=1}^{n} x_{mk}y_k \end{cases} \qquad (3)$$

Set $\overline{x_i}=\dfrac{1}{n}\sum\nolimits_{k=1}^{n}x_{ik}, \overline{y}=\dfrac{1}{n}\sum\nolimits_{k=1}^{n}y_k,$

Where i = 1,2, $\cdots$, m

$$I_{\mathrm{ij}}=I_{ji}=\sum\nolimits_{k=1}^{n}(x_{ik}-\overline{x_i})\;(x_{jk}-\overline{x_j})=\sum\nolimits_{k=1}^{n}x_{ik}x_{jk}-n\overline{x_i}\,\overline{x_j}$$

Where i = 1,2, $\cdots$, m; j = 1,2, $\cdots$, m;

In particular, when i = j,

ther $I_{\mathrm{ij}}=\sum\nolimits_{k=1}^{n}(x_{ik}-\overline{x_i})^2=nS_i^2$, where $S_i^2$ represents the sample variance of $x_{i1}$, $x_{i2}$, $\cdots$, $x_{in}$ which indicate the observations of $\xi_i$;

$$I_{\mathrm{iy}}=\sum\nolimits_{k=1}^{n}(x_{ik}-\overline{x_i})\;(\mathrm{y}_\mathrm{k}-\overline{\mathrm{y}})=\sum\nolimits_{k=1}^{n}x_{ik}\mathrm{y}_\mathrm{k}-n\overline{x_i}\,\overline{y}$$

where i = 1,2, $\cdots$, m

Using the elimination method, the equations (3) is easy to be reduced to the following equations:

$$\begin{cases} a+\overline{x_1}b_1+\overline{x_2}b_2+\cdots+\overline{x_m}b_m=\overline{y} \\ I_{11}b_1+I_{12}b_2+\cdots+I_{1m}b_m=I_{1y} \\ I_{21}b_1+I_{22}b_2+\cdots+I_{2m}b_m=I_{2y} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ I_{m1}b_1+I_{m2}b_2+\cdots+I_{mm}b_m=I_{my} \end{cases} \qquad (4)$$

So, we can start with the last m equations to get the solution $\hat{b}_1$, $\hat{b}_2, \cdots, \hat{b}_m$; then substituted into the first equation, that was, $\hat{a}=\hat{y}-\hat{b}_1\overline{x}_1-\cdots-\hat{b}_m\overline{x}_m$.

Substituted the solutions ($\hat{a}$, $\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_m$) into the equation (1), which is obtained from equations (4), to get the multiple linear regression equation:

$$\hat{y}=\hat{a}+\hat{b}_1x_1+\hat{b}_2x_2+\cdots+\hat{b}_mx_m$$

### 4.2 *Statistical Analysis*

Let Y's fitting is $\hat{Y}$, to fit the error e = $Y-\hat{Y}$ called residuals. While $Q=\sum\nolimits_{i=1}^{n}e_i^2=\sum\nolimits_{i=1}^{n}(y_i-\hat{y}_i)^2$ is the sum of residual squares (or residual sum of squares), i.e Q ($\hat{\beta}$), reflecting the effect of random errors on y.

Now decomposing the sample variance of Y,

$$S_2=\sum\nolimits_{i=1}^{n}(y_i-\hat{y}_i)^2, S_2=Q+U,$$

where $U=\sum\nolimits_{i=1}^{n}(\hat{y}_i-\overline{y})^2$ called regression sum of squares, reflecting the independent variable's influence on Y.

## 5 HYPOTHESIS TESTING OF THE REGRESSION MODEL

Whether there exist a model between dependent variable Y and Independent variable $x_1$, $x_2$, $\cdots$, $x_m$.

$$y=b_0+b_1x_1+\cdots+b_mx_m(m\geq 2)$$

The linear relationship shown above need to be inspected, so make null hypothesis as

$H_0 : b_j = 0$ (j=1, 2, $\cdots$, m)

Test methods is similar as variance analysis, when $H_0$ is set up, previously defined U, Q satisfy

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

There is $1-\alpha$ quantile $F_{1-\alpha}(m, n-m-1)$ in the significance level $\alpha$, if $F < F_{1-\alpha}(m, n-m-1)$, then $H_0$ is acceptable; Otherwise, reject.

## 6 SOLVING MULTIPLE LINEAR REGRESSION EQUATION

Note: the monthly average maximum pressure is $x_1$, monthly average minimum pressure is $x_2$, monthly average maximum temperature is $x_3$, monthly average minimum temperature is $x_4$, monthly average relative humidity is $x_5$, the number of monthly incidence of disaster is y, set $y = b_0 + \sum_{i=1}^{5} b_i x_i$.

Now disaggregate the data statistical by month, the data is omitted.

By MATLAB programming, seeking

$b_0 = -9936.4$, confidence interval $[-72335, 52462]$;

$b_1 = 89.2$, confidence interval $[-150, 328]$;

$b_2 = -75.7$, confidence interval $[-283, 131]$;

$b_3 = -129.1$, confidence interval $[-285, 27]$;

$b_4 = 162.4$, confidence interval $[-18, 342]$;

$b_5 = -31.4$, confidence interval $[-63, 0]$;

$R^2 = 0.0000$, $F = 0.0000$, $p = 0.0000$.

As for $R^2 = 0.0000$, and the confidence intervals for each regression coefficient contains zero, therefore reject the hypothesis of linear regression above to improve the model's performance.

Method 2.

### 6.1 *Establishment of multiple binomial regression equations*

Set up pure quadratic regression equation as follow:

$$y = b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{5} b_{ii} x_i^{\,2} \qquad (5)$$

By MATLAB programming, get figure 1.

Derivation of the regression coefficients:

$b_0 = -545336$, $b_1 = -10897.4$,

$b_2 = 12031.02$, $b_3 = -115.8216$,

$b_4 = 4.2813$, $b_5 = -399.0992$,

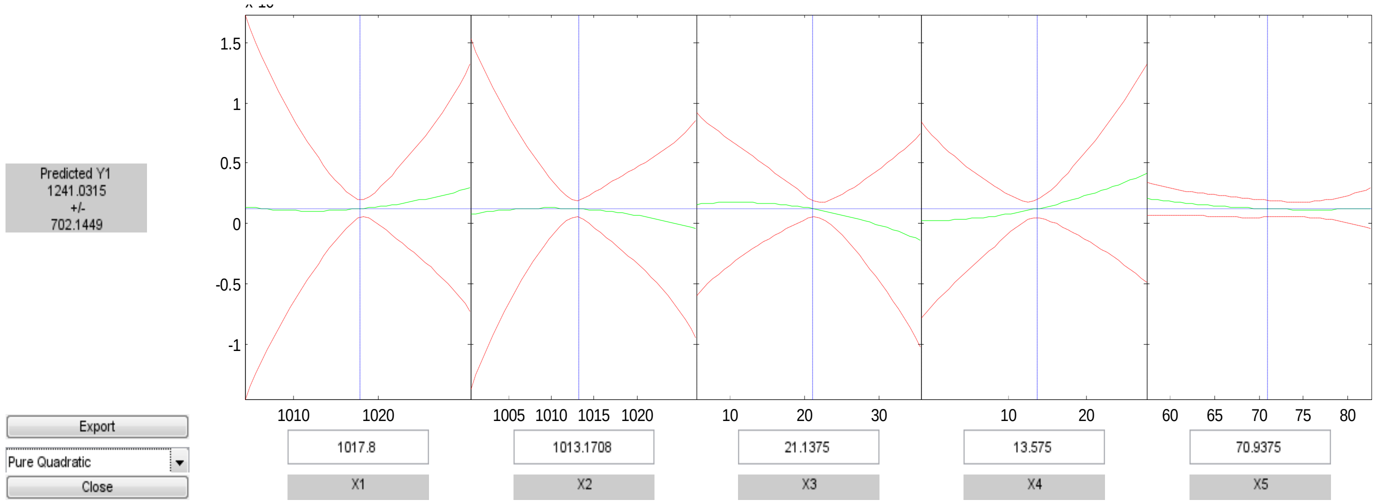
$b_{11} = 5.3836$, $b_{22} = -5.9565$,

$b_{33} = -5.2225$, $b_{44} = 4.9394$,

$b_{55} = 2.6226$

Residual standard deviation: $rmse = 456.6278$

As long as the above-mentioned factor substitution (5), that was binomial regression equation.



Predicted Y1
1241.0315
+/-
702.1449

Pure Quadratic

1017.8 | 1013.1708 | 21.1375 | 13.575 | 70.9375

Figure 1 Interactive drawing

### 6.2 *Using binomial regression model to predict*

Using the regression model

$$y = b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{5} b_{ii} x_i^{\,2}$$

runs in the MATLAB interface, it will be able to enter a given $x_0 = (x_{01}, x_{02}, \cdots, x_{05})$ to predict $y_0$, such as:

When $x_{01} = 1017.8$, $x_{02} = 1013.1708$, $x_{03} = 21.1375$, $x_{04} = 13.575$, $x_{05} = 70.9375$.

Get $y_0 = 1241.0315 \pm 702.1449$.

By forecast, it will be helpful to arrange the number of doctors and hospital beds reasonable.

## REFERENCES

[1] Shen Hengfan. Probability and Mathematical Statistics Course. Bei Jing: Higher Education Press. 1995.05

[2] Xiao Shutie. Mathematics Experiment. Bei Jing: Higher Education Press. 1999.07

[3] Wang Moran. MATLAB6.0 and Scientific Computing. Bei Jing: Electronic Industry Press. 2001.09