A REGRESSION MODEL

S. SZABÓ

Department of Civil Engineering Mathematics,
Technical University, H-1521 Budapest

Received June 20, 1987
Presented by prof. Dr. J. Reimann

Abstract

In the normal linear regression the least square estimation of the coefficients has a series of nice properties. In addition the needed numerical calculations may be done a fairly efficient way. For the case when the application of this widely used least square method is not justified some further methods are suggested. But experience about their statistical and numerical properties can hardly be found. This paper intends to help the practitioner to become familiar with a type of non-least-square regression methods.

Introduction

Besides other mathematical methods some random ones can be used in engineering sciences. The first task is to select from the stochastic models. Then using the data of the investigated phenomenon or systems we have to calibrate our model. In other words we have to estimate the values of the corresponding parameters of the model. After testing the goodness of the model we may use it for predictions provided it was acceptable. In this paper we focus our attention to describing a regression model and then estimating its parameters and testing them.

There are a lot of theoretical results about regression. Not only pure theory but also practice is well developed and widely used thanks to the symplicity of the least square method.

Sometimes the use of the least square principle is not justified. There is no lack in theoretically well founded methods which may solve these problems. They are generally more complicated from a numerical point of view than the least square one. Probably this is the reason why they are used very seldom.

This paper does not give new theoretical results but intends to give the model of the formulae and interpretations in a detailed way which may ease the choice from the models and their implementation and interpretation for the practitioner.

The intuitive background

We start with the simplest well known regression model which is the following. Two quantities, say x and y, are measured. We know y is a function of x. For the sake of simplicity we suppose that it is $y = ax$ with an unknown

coefficient a. The reader may think that y is the mass and x is the volume of a given homogeneous material. Consequently the parameter a is the density of the material. Next we have n bodies made from this material with volumes known precisely. These, together with the measured quantities of masses, are listed in the next table

$$\begin{array}{ll} y_1, & x_1 \\ y_2, & x_2 \\ \vdots & \vdots \\ y_n, & x_n \,. \end{array}$$

While the x values are quite precise y is measured with a random defect. Since $y_1, \ldots, y_n$ are contaminated by random errors they are not necessarily equal to $ax_1, \ldots, ax_n$. Thus

$$e_1 = y_1 - ax_1, \ldots, e_n = y_n - ax_n$$

are the random errors. It is reasonable to assume that they are independent random variables of the same normal distribution and so

$$M(e_1) = \ldots = M(e_n) = 0 \quad \text{and} \quad D(e_1) = \ldots = D(e_n) = \sigma$$

where σ is unknown.

According to the maximum likelihood principle the maximum place of

$$L(a) = \sum_{i=1}^{n} ln \left(\frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y_i - ax_i)^2/2\sigma} \right),$$

the so-called likelihood function, provides an estimation $\hat{a}$ for a which has a series of good properties. For example it is unbiased that is $M(\hat{a}) = a$ and $D(\hat{a})$ is minimal among other estimations and tends to zero when n tends to infinity.

It is easy to realise that $L(a)$ reaches its maximum where

$$S(a) = \sum_{i=1}^{n} (y_i - ax_i)^2$$

reaches its minimum. Thus the maximum likelihood principle is the base of the least square principle (Fig. 1).

We would like to point out that the model sketched above is widely used in a theoretically different situation. Namely, when y is not a function of x but there is an association between them. For instance consider a population and let x be the altitude and let y be the weight of a randomly chosen person.

It is clear that y is not a function of x. In spite of this the previous model is applicable and useful. Suppose we have a list about n measured persons. If we want to forecast the persons' weight we meet in the street, we can first do is to guess $M(y)$ which can be estimated by the arithmetic average of the measured weights. The uncertainty of this forecast can be characterised by

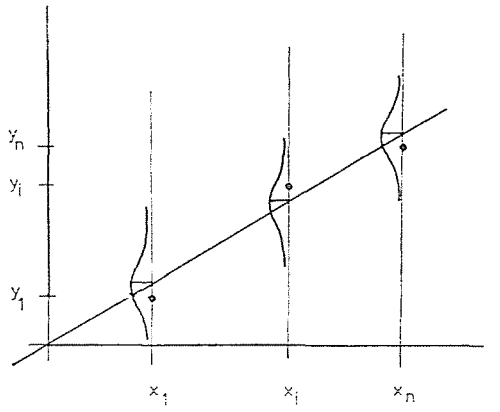

Fig. 1

$D(y)$. If we know x the persons' height then our guess may be $M(y|x)$ which can be estimated by $y = ax$ from the regression. The uncertainty of this forecast is $\sigma = D(e_i)$ which may be smaller than $D(y)$. The quantity

$$\frac{D(y) - D(e_i)}{D(y)}$$

indicates how large a part of the uncertainty of y can be explained by x in our model.

Distribution of errors

In the previous section it was assumed that the random errors follow a normal distribution with a zero expected value and an unknown standard deviation. The errors in real life sometimes follow this distribution, and sometimes they do not. We now introduce a wider collection for the error distributions which will be relatively easy to handle.

We have to admit that they do not necessarily cover all real situations. But we hope they may provide a finer tool than restricting our investigations to the single normal distribution. We will suppose that the density function of the error has the form

$$f_\gamma(x) = \alpha e^{-\beta|x|^\gamma}, \quad \gamma \geq 1$$

Since

$$\int_{-\infty}^{\infty} f_\gamma(x)dx = 1$$

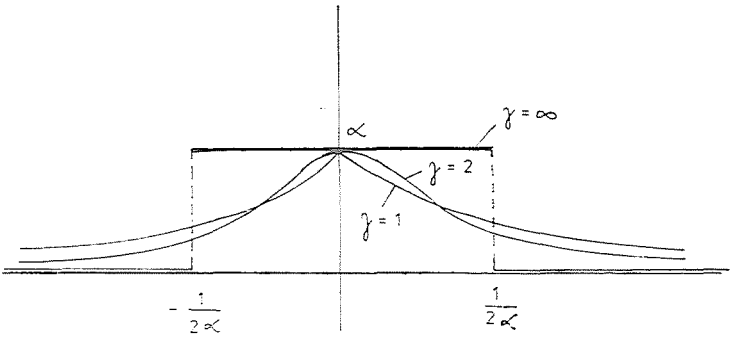

Fig. 2

α, β, γ are not independent. If $\gamma = 1$, $\gamma = 2$ or $\gamma = \infty$ then the respective density functions are the double exponential and the normal and uniform distribution functions. Figure 2 illustrates these functions.

Regression in norm

We can generalize the model of the first section in two directions at the same time. Let the random variable y be a linear function of the non-random variables $x_1, \ldots, x_s$ and the random variable e so that

$$y = a_0 + a_1 x_1 + \ldots + a_s x_s + e\,.$$

With matrix notation

$$y = \mathbf{x}^T \mathbf{a} + e,$$

where

$$\mathbf{a}^T = (a_0, a_1, \ldots, a_s) \quad \text{and} \quad \mathbf{x}^T = (1, x_1, x_2, \ldots, x_s).$$

We have n corresponding measured values, the so-called sample, which are arranged in the next table

$$\begin{array}{llllll} y_1, & 1, & x_{11}, & x_{12}, \ldots, & x_{1s} \\ y_2, & 1, & x_{21}, & x_{22}, \ldots, & x_{2s} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n, & 1, & x_{n1}, & x_{n2}, \ldots, & x_{ns} \end{array}$$

For the sake of simplicity let

$$\mathbf{x}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{is})\,.$$

We suppose that $M(e) = 0$ but the density function of the error is f_γ, that is, it may be non-normal.

The likelihood function now is

$$L(\mathbf{a}) = \sum_{i=1}^{n} ln\, \alpha e^{-\beta |y_i - \mathbf{x}_i^T \mathbf{a}|^\gamma}$$

Since β must be positive $\hat{\mathbf{a}}$ the maximum likelihood estimation of $\mathbf{a}$ coincides with the minimum place of function

$$S_\gamma(\mathbf{a}) = \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \mathbf{a}|^\gamma$$

This expression has a clear geometrical meaning since it is a distance between the first column of the sample matrix and the subspace spanned by the remaining colums. The basis of this distance is the γ-norm defined by

$$||\mathbf{z}||_\gamma = ||(z_1, \ldots, z_n)||_\gamma = |z_1|^\gamma + \ldots + |z_n|^\gamma .$$

In case $\gamma = \infty$

$$||\mathbf{z}||_\infty = \max(|z_1|, \ldots, |z_n|).$$

Suppose that the value of γ is known. For the sake of simplicity let it be an integer. The $\hat{\mathbf{a}}$ that is the minimum place of $S_\gamma(\mathbf{a})$ can be defined by the gradient method. Its algorithm is:

Choose an arbitrary initial value for $\hat{\mathbf{a}}$, say $\hat{\mathbf{a}}_0$. Set $h = S_\gamma(\hat{\mathbf{a}}_0)\, n$, where n is the number of the steps. Then apply the iteration

$$\hat{\mathbf{a}}_{i+1} = \hat{\mathbf{a}}_i - h\mathbf{u},$$

where $\mathbf{u} = \text{grad}\, S_\gamma(\hat{\mathbf{a}}_i)/(\text{grad}\, S_\gamma(\hat{\mathbf{a}}_i))^2$.

Do this while $S_\gamma(\hat{\mathbf{a}}_{i+1}) \leq S_\gamma(\hat{\mathbf{a}}_i)$.

The last $S_\gamma(\hat{\mathbf{a}}_i)$ approaches the value of the minimum with an accuracy h.

We remind that $S_\gamma(\mathbf{a})$ cannot be differentiated if γ is odd. In these cases the sign x can be viewed as the generalized derivative of $|x|$.

Although this way of estimation of $\mathbf{a}$ is easy to implement but it is rather inefficient.

If $\gamma = 2$, then the partial derivatives of $S_\gamma(\mathbf{a})$ are linear functions so its minimum place satisfies a system of linear equations. Namely $\mathbf{a}$ is the solution of

$$\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{y}$$

where $\mathbf{y}$ is the first column while $\mathbf{X}$ consists of the futher columns of the sample matrix.

From a numerical point of view its solution is the simplest among our problems.

So our policy may be the following. First we suppose $\gamma = 2$ and estimate $\mathbf{a}$ from $\mathbf{X}^T\mathbf{X}\hat{\mathbf{a}} = \mathbf{X}^T\mathbf{y}$. If the distribution of the residuals

$$e_i = y_i - (\hat{a}_0 + \hat{a}_1 x_1 + \ldots + \hat{a}_s x_s) = y_i - \mathbf{x}_i^T\hat{\mathbf{a}} \quad i = 1, \ldots, n$$

is normal then the problem is solved. Otherwise we take case $\gamma = 1$ or $\gamma = \infty$ depending on whether the tail of its distribution is too long or too short. In the next section we suggest an efficient way of estimating the parameters in these cases. But first we wish to draw the readers' attention to a technical detail. Testing the normality is possible by sketching the histogram of the residuals or by a finer statistical method. In the last case one should keep in mind that the unbiased estimation of the standard deviation of the residuals is $\left(\sum_{i=1}^{n}(e_i-\bar{e})^2\right)/(n-s)$ since $\hat{a}_0, \ldots, \hat{a}_s$ were estimated from the sample.

Linear programming for the case $\gamma = 1$ and $\gamma = \infty$

Now formulate the problem of finding the minimum of

$$S_1(\mathbf{a}) = \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T\mathbf{a}|$$

by means of linear programming.
Let $|y_i - \mathbf{x}_i^T\mathbf{a}| = u_i$. So the problem is to minimize $u_1 + u_2 + \ldots + u_n$ subject to

$$\begin{aligned} -u_1 &\leq y_1 - (a_0 + a_1x_{11} + \ldots + a_sx_{1s}) \leq u_1 \\ &\vdots \\ -u_n &\leq y_n - (a_0 + a_1x_{n1} + \ldots + a_sx_{ns}) \leq u_n . \end{aligned}$$

The simplex tableau without the objective function is

a_0	a_1	a_2	$\ldots$	a_s	u_1	u_2	$\ldots$	u_n	
1	x_{11}	x_{12}	$\ldots$	x_{1s}	-1	0	$\ldots$	0	y_1
$\vdots$									$\vdots$
1	x_{n1}	x_{n2}	$\ldots$	x_{ns}	0	0	$\ldots$	-1	y_n
-1	$-x_{11}$	$-x_{12}$	$\ldots$	$-x_{1s}$	-1	0	$\ldots$	0	$-y_1$
$\vdots$									$\vdots$
-1	$-x_{n1}$	$-x_{n2}$	$\ldots$	$-x_{ns}$	0	0	$\ldots$	-1	$-y_n$

Since $a_0, a_1, \ldots, a_s$ may be negative, we introduce b_i and c_i so that b_i and c_i are not less than zero and $b_i - c_i = a_i$ for $i = 0, 1, \ldots, s$.

The new simplex tableau with matrix notation is

$\mathbf{b}^T$	$\mathbf{c}^T$	$\mathbf{u}^T$	
$\mathbf{X}$	$-\mathbf{X}$	$-\mathbb{I}_n$	$\mathbf{y}$
$-\mathbf{X}$	$\mathbf{X}$	$-\mathbb{I}_n$	$-\mathbf{y}$

If $(\hat{\mathbf{b}}^T, \hat{\mathbf{c}}^T, \hat{\mathbf{u}}^T)$ is one of the solutions, then $\hat{\mathbf{a}} = \hat{\mathbf{b}} - \hat{\mathbf{c}}$ is a minimum place of $S_1(\hat{\mathbf{a}})$.

We would like to remark that this $\hat{\mathbf{a}}$ is a robust estimation of $\mathbf{a}$. The problem of minimizing

$$S_\infty(\mathbf{a}) = \max\left(|y_1 - \mathbf{x}_1^T\mathbf{a}|, \ldots, |y_n - \mathbf{x}_n^T\mathbf{a}|\right)$$

can also be reduced to a linear programming problem.

Let u be the maximum of $|y_1 - \mathbf{x}_1^T\mathbf{a}|, \ldots, |y_n - \mathbf{x}_n^T\mathbf{a}|$. So the problem is to minimize u subject to

$$\begin{aligned} -u &\leq y_1 - (a_0 + a_1x_{11} + \ldots + a_sx_{1s}) \leq u \\ &\vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ -u &\leq y_n - (a_0 + a_1x_{n1} + \ldots + a_sx_{ns}) \leq u \end{aligned}$$

Thus setting $\mathbf{a} = \mathbf{b} - \mathbf{c}$ where $\mathbf{b} \geq \mathbf{0}$ and $\mathbf{c} \geq \mathbf{0}$ the simplex tableau is

$\mathbf{b}^T$	$\mathbf{c}^T$	$\mathbf{u}$	
$\mathbf{X}$	$-\mathbf{X}$	$-\mathbf{e}_n$	$\mathbf{y}$
$-\mathbf{X}$	$\mathbf{X}$	$-\mathbf{e}_n$	$-\mathbf{y}$

where $\mathbf{e}_n^T = (1, \ldots, 1)$.

If $(\hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\mathbf{u}})$ is a solution of this problem then $\hat{\mathbf{a}} = \mathbf{b} - \hat{\mathbf{c}}$ is the minimum place of $S_\infty(\mathbf{a})$.

The next consideration shows that this estimation of $\hat{\mathbf{a}}$ is fairly good if the errors are distributed uniformly but as we can see it is not robust.

Consider our model when $s = 0$ that is let

$$y = a_0 + e.$$

Here $y_1, \ldots, y_n$ is the sample.

Consequently $\hat{a}_0$ the estimation of a_0 is the solution of the next linear programming problem. Minimize u subject to

$$-u \leq y_1 - a_0 \leq u$$
$$\vdots$$
$$-u \leq y_n - a_0 \leq u.$$

Its solution is $\hat{a}_0 = (\max(y_1, \ldots, y_n) + \min(y_1, \ldots, y_n))/2$ Note that $\hat{a}_0$ is an unbiased estimation of $M(y)$ and so is $\bar{y}$. But the first one is better than the second since, as is known

$$\frac{D^2(\hat{a}_0)}{D^2(\bar{y})} = \frac{6n}{(n+1)(n+2)}.$$

Testing hypothesis

The previous regression model provides a good tool for comparing the expected values of several independent (univariate) populations. Denote the expected values by $a_1, \ldots, a_s$ and by $y_{i1}, \ldots, y_{in_i}$ the sample from the i-th population, where $n_1, \ldots, n_s$ are the respective sample sizes. Our model is

$$y_{ij} = a_i + e_{ij} \qquad i = 1, \ldots, s;\ j = 1, \ldots, n_i$$

and the common density function of the independent random variables e_{ij} is f_γ.

We wish to test hypothesis

$$H:\ a_1 = a_2 = \ldots = a_s$$

We know from the maximum likelihood principle that

$$S_\gamma(\mathbf{a}) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} |y_{ij} - a_i|^\gamma$$

must be minimum independently from the hypothesis. Let Q be this minimum.

If the hypothesis is true then

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} |y_{ij} - a|^\gamma$$

must be minimum, where $a = a_1 = \ldots = a_s$. Let Q' be this minimum.

If e_{ij}'s have a normal distribution that is if $\gamma = 2$, then our problem is the case of the one-way layout version of the analysis of variance. In this case the statistics

$$T_\gamma = \frac{Q' - Q}{Q}$$

is the basis of the decision over hypothesis H.

In practice this test is used when this condition of the model is not fulfilled, that is when $\gamma \neq 2$. In the rest of this section we will show that the deviation can be controlled by an easy calculation if either $\gamma = 1$ or $\gamma = \infty$ since T_1, T_2 and T_∞ can be evaluated easily.

We need Q and Q'.

Obviously

$$\min \left(\sum_{i=1}^{s} \sum_{j=1}^{n_i} |y_{ij} - a_i|^\gamma \right) = \sum_{i=1}^{s} \min \left(\sum_{j=1}^{n_i} |y_{ij} - a_i|^\gamma \right)$$

and

$$\min \left(\max_{i,j} |y_{ij} - a_i| \right) = \max_{i} \left(\min \left(\max_{j} (|y_{ij} - a_i|) \right) \right)$$

In the cases $\gamma = 1$, $\gamma = 2$, $\gamma = \infty$ the respective minimum places are

$$\hat{a}_i = \operatorname{median} (y_{i1}, \ldots, y_{in_i})$$

$$\hat{a}_i = \operatorname{average} (y_{i1}, \ldots, y_{in_i})$$

$$\hat{a}_i = 1/2 \, (\max (y_{i1}, \ldots, y_{in_i}) + \min (y_{i1}, \ldots, y_{in_i})).$$

Consequently Q in order is

$$\sum_{i=1}^{s} \sum_{j=1}^{n_s} |y_{ij} - \operatorname*{median}_{j} y_{ij}|$$

$$\sum_{i=1}^{s} \sum_{j=1}^{n_s} (y_{ij} - \operatorname*{average}_{j} y_{ij})^2$$

$$(1/2) \max_{i} (\max_{j} y_{ij} - \min_{j} y_{ij}) = (1/2) \max_{i} \operatorname*{range}_{j} y_{ij}$$

Finally, Q' is

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} |y_{ij} - \operatorname*{median}_{i,j} y_{ij}|$$

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} (y_{ij} - \operatorname*{avevage}_{i,j} y_{ij})^2$$

$$(1/2) \operatorname*{range}_{i,j} y_{ij}$$

respectively.

References

1. Anderson, T. W.: An introduction to the multivariate analysis. Wiley, New York 1958.
2. Rao, C. R.: Linear statistical inference and its applications. Wiley, New York 1973.
3. Scheffe, H.: The analysis of variance. Wiley, New York 1959.

Dr. Sándor Szabó H-1521 Budapest