PROCEEDINGS OF THE YEREVAN STATE UNIVERSITY

Physical and Mathematical Sciences 2010, № 1, p. 27–31

Mathematics

REGRESSION MODELS GENERATED BY DISTRIBUTIONS OF MODERATE GROWTH

Kh. L. VARDANYAN*

Chair of Probability Theory and Mathematical Statistics, YSU

Regression model generated by two-parametric distribution of moderate growth and arising in bioinformatics is considered. Consistency (in a weak sense) of least square estimates for parameters is proved. Distributions of least square estimates for parameters and of Gaussian noise variation estimate are obtained. Results may be used for statistical hypothesis testing with regard to parameters of model.

Keywords: consistent least square estimate, regression model, distribution of moderate growth.

§ 1. Introduction. Two-parametric distribution of *moderate growth* is introduced in [1]. It takes the following form

$$\begin{cases} p_n(\alpha) = p_0(\alpha)\dfrac{\theta^n}{\psi_n}\displaystyle\prod_{m=0}^{n-1}\left(1+\frac{c-1}{\psi_m}\right), \quad n=1,2,..., \\ p_0(\alpha) = \left\{1+\displaystyle\sum_{n\ge1}\frac{\theta^n}{\psi_n}\prod_{m=0}^{n-1}\left(1+\frac{c-1}{\psi_m}\right)\right\}^{-1} \end{cases} \tag{1.1}$$

with parametric set $A=\{\alpha=(c,\theta):\, 0<c<+\infty,\, 0<\theta\le 1\}$. The moderate growth of distribution $\{p_n(\alpha)\}_0^\infty$. (1.1) is defined by conditions on sequence $\{\psi_n\}_1^\infty$:

$$\psi_0=1,\ \{\psi_n\}_1^\infty \text{ is non-decreasing, } \lim_{n\to+\infty}\psi_n=+\infty,\ \lim_{n\to+\infty}\left(\psi_n/\psi_{n-1}\right)=1$$

under the constraint

$$S_\psi=\sum_{n\ge1}\left(1/\psi_n\right)<+\infty. \tag{1.2}$$

Below $N(a,\sigma^2)$ denotes the Gaussian distribution function with *mean a* and *variance* σ^2.

In [2] the following regression model generated by model (1.1)–(1.2) is considered. We take logarithm from both sides of (1.1) and replace $\ln\left(1+\dfrac{c-1}{\psi_m}\right)$ by $(c-1)/\psi_m$ (so called *linearization*). Then, denoting $f_\alpha(n)=\eta\cdot n+(c-1)S_\psi(n)$,

* E-mail: khcho65@mail.ru

$\eta = \ln\theta$, $S_\psi(n) = \sum_{m-0}^{n-1} 1/\psi_m$ and $y_n = \ln\left(p_n(\alpha)/p_0(\alpha)\right) + \ln\psi_n$, we build the *regression model* of the form

$$y_n = f_\alpha(n) + \varepsilon_n, \quad n = 1, 2, ..., N. \tag{1.3}$$

Given the integer $N \geq 1$ and the *observed* sequence $\{y_n\}_1^N$ one needs to find estimates for *unknown* parameters c and η under, for instance, the assumption: $\varepsilon_n \sim N(0, \sigma^2)$, $n = \overline{1, N}$, *are independent Gaussian noises with unknown variation* σ^2. So, the estimate of σ^2 is also needed.

Denote $r_N = c_N \cdot S_N - \left(\sum n \cdot S_\psi(n)\right)^2$, $c_N = \sum n^2$, $S_N = \sum\left(S_\psi(n)\right)^2$.

Here and everywhere below in sums limits on n (from 1 to N) are omitted for simplicity. In [2] for *unbiased* least square estimates (LSEs) $\hat{c}_N$ and $\hat{\eta}_N$ of parameters c and $\eta\left(= \ln\theta\right)$ in the model (1.3) the following formulas are found:

$$\begin{cases} \hat{c}_N - 1 = r_N^{-1}\left\{c_N\left(\sum y_n \cdot S_\psi(n)\right) - \left(\sum n \cdot y_n\right)\left(\sum n \cdot S_\psi(n)\right)\right\}, \\ \hat{\eta}_N = r_N^{-1}\left\{S_N\left(\sum n \cdot y_n\right) - \left(\sum n \cdot S_\psi(n)\right)\left(\sum y_n \cdot S_\psi(n)\right)\right\}. \end{cases} \tag{1.4}$$

In this paper, first of all, based on (1.4) the *normality* and the *consistency* in a weak sense of LSEs $\hat{c}_N$ and $\hat{\eta}_N$ are established. Next, we consider the estimate $\hat{\sigma}_N^2 = \dfrac{1}{N-2}\sum e_n^2$ for variation σ^2, where $e_n = y_n - \hat{f}_\alpha(n)$ are so called *residuals* (*remainders*) of regression (1.3), and $\hat{f}_\alpha(n) = \hat{\eta}_N n + (\hat{c}_N - 1)S_\psi(n)$, $n = 1, 2, ..., N$, obviously, are LSEs for $f_\alpha(n)$ (the predicted value of $f_\alpha(n)$).

We prove some "good" properties of LSEs $\hat{\sigma}_N^2, \hat{c}_N$ and $\hat{\eta}_N$.

§ 2. Normality and Consistency.

Theorem 1. The LSEs $\hat{\eta}_N$ and $\hat{c}_N$ for η and c have distribution functions $N\left(\eta, \dfrac{\sigma^2}{r_N} S_N\right)$ and $N\left(c, \dfrac{\sigma^2}{r_N} c_N\right)$ respectively. They are consistent in a weak sense, i.e. $\hat{\eta}_N \xrightarrow{p} \eta$, $\hat{c}_N \xrightarrow{p} \eta$ as $N \to +\infty$. Here the sign „ $\xrightarrow{p}$ '' denotes the convergence in probability.

Proof. Denote $y = (y_1, ..., y_N)`$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)`$, $\beta = (\eta, c-1)`$, where « ` » is the symbol of allocation, and present the *regression model* (1.3) in the following form $y = X\beta + \varepsilon$,

$$X = \begin{pmatrix} 1 & S_\psi(1) \\ \vdots & \vdots \\ n & S_\psi(n) \end{pmatrix}. \tag{2.1}$$

It is well-known (see, for instance [3]), that LSE $\hat{\beta}$, minimizing the regression remainders squares sum for (2.1), due to $e`e = \sum e_n^2$, $e = y - X\hat{\beta}$, takes the form

$$\hat{\beta} = (X`X)^{-1} X`y` . \tag{2.2}$$

Easily seen that the matrix $X`X$ takes the form

$$X`X = \begin{pmatrix} c_N & x`S_\psi \\ \vdots & \vdots \\ x`S_\psi & S_N \end{pmatrix}$$

with $x = (1,2,...,N)$, $S_\psi = (S_\psi(1), S_\psi(2),..., S_\psi(N))$, where we assume that $\det(X`X) = r_N \neq 0$. Then, evaluations lead to the form (2.2) of estimate $\hat{\beta}$.

From Gauss–Markov theorem [3] it follows that the estimate $\hat{\beta}$ is *optimal* (in the sense of minimums of variations $D\hat{\eta}_N$ and $D\hat{c}_N$) in the class of linear with respect to y, unbiased estimates for parameter β. Taking into account that y_n, $n = \overline{1,N}$, has distribution function $N(f_\alpha(n), \sigma^2)$, the estimate (2,2) is linear with respect to y, and X is not random, we conclude that LSEs $\hat{c}_N$ and $\hat{\eta}_N$ are Gaussian. Further, from the representation (2.2)

$$\hat{\beta} = (X`X)^{-1} X`(X\beta + \varepsilon) = \beta + (X`X)^{-1} X`$$

it follows $E\hat{\beta} = \beta + (X`X)^{-1} X`E\varepsilon = \beta,$ where E denotes the sign of mathematical expectation.

So, $\hat{\beta}$ is the *unbiased* estimate for parameter β. Let us evaluate the variances $D\hat{\eta}_N$ and $D\hat{c}_N$. For this purpose we need in covariance matrix $V\hat{\beta}$ of estimate $\hat{\beta}$: $V\hat{\beta} = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})` = EA\varepsilon(A\varepsilon)` = A(E\varepsilon\varepsilon`)A` = \sigma^2 AA`$, where $A = (X`X)^{-1} X`$. Since $AA` = (X`X)^{-1}$, therefore,

$$V\hat{\beta} = \sigma^2 (X`X)^{-1}. \tag{2.3}$$

On the other hand, because of the form

$$(X`X)^{-1} = r_N^{-1} \begin{pmatrix} S_N & -(x`S_\psi) \\ \vdots & \vdots \\ -(x`S_\psi) & c_N \end{pmatrix},$$

due to (2.3), we obtain $D\hat{\eta}_N = \dfrac{\sigma^2}{r_N} S_N$, $D\hat{c}_N = \dfrac{\sigma^2}{r_N} c_N$.

Let us pass to the proof of consistency of estimates $\hat{c}_N$ and $\hat{\eta}_N$. It is enough to show that $D\hat{\eta}_N \to 0$ and $D\hat{c}_N \to 0$ as $N \to +\infty$ (see [4]).

Due to Couchy–Shwartz inequality, we have $0 \le 1 - \dfrac{\left(\sum n \cdot S_\psi(n)\right)^2}{\sum n^2 \sum \left(S_\psi(n)\right)^2} < 1$.

That is why

$$\lim_{n \to +\infty} \frac{r_N}{\sum \left(S_\psi(n)\right)^2} = +\infty, \tag{2.4}$$

which follows from the representation

$$\frac{r_N}{S_N}=\sum n^2-\frac{\sum n\cdot S_\psi(n)}{\sum\left(S_\psi(n)\right)^2}=\sum n^2\left(1-\frac{\left(\sum n\cdot S_\psi(n)\right)^2}{\left(\sum n^2\right)\left(\sum\left(S_\psi(n)^2\right)^2\right)}\right).$$

The limit relation (2.4) says that $D\hat{\eta}_N\to 0$ as $N\to+\infty$.

Similarly one may prove that $D\hat{c}_N\to 0$ as $N\to+\infty$. Theorem 1 is proved.

§ 3. Properties of LSE $\hat{\sigma}^2_N$. Let the constraint (1.2) holds.

Theorem 2. The statistics (1.5) is unbiased estimate for $\hat{\sigma}^2$, and the statistics $\chi^2_{N-2}=(N-2)\frac{\hat{\sigma}^2_N}{\sigma^2}$ has χ^2-distribution $H^2(N-2)$ with $(N-2)$ degrees of freedom.

Proof. Let us present the predicted (by regression) value $\hat{y}=X\hat{\beta}$ in the form $\hat{y}=(X`X)^{-1}X`y=My$, and the vector of regression remainders $e=y-\hat{y}$ in the form $e=y-X\hat{\beta}=(I_N-M)y=By=B(X\beta+\varepsilon)=B\varepsilon$, because $Bx=0$. Here I_N is a usual unit matrix of order N.

The matrixes H and B satisfy conditions: $H`=H, H^2=H,\ B`=B, B^2=B.$

Write down the following chain of equalities:

$$E(e`e)=E(\sum e_n^2)=\mathrm{Etr}(ee`)=\mathrm{Etr}(B\varepsilon\varepsilon`B)=\mathrm{tr}(BE(\varepsilon\varepsilon`)B`)=\sigma^2\cdot\mathrm{tr}(BB`)=\sigma^2\cdot\mathrm{tr}B,$$

where "tr" denotes the *trace* of matrix B. On the other hand, $\mathrm{tr}B=\mathrm{tr}(I_N)-\mathrm{tr}(M)$ and $\mathrm{tr}(M)=\mathrm{tr}(X(X`X)^{-1}X`)=\mathrm{tr}(X`X(X`X)^{-1})=\mathrm{tr}I_2=2$. That is why, finally, we obtain $E(ee`)=\sigma^2(N-2)$, i.e. $E\hat{\sigma}^2_n=\sigma^2$.

Next we have $\chi^2_{N-2}=\frac{1}{\sigma^2}(ee`)B(\varepsilon/\sigma)$, where $(\varepsilon/\sigma)$ is an $N$-dimensional standard Gaussian vector (with zero means and unit variations). Since $B`=B$, $B^2=B$, therefore, (see [3]) the statistics χ^2_{N-2} has χ^2-distribution with $k=\mathrm{rank}(B)$ degrees of freedom. But in this case of B we have $\mathrm{rank}(B)=\mathrm{tr}B=N-2$. Theorem 2 is proved.

Remark. It is of interest that for a given N:

The estimates $\hat{\sigma}^2_N$, $(\hat{c}_N,\hat{\eta}_N)$ are independent. (3.1)

Indeed, taking into account that $\hat{\sigma}^2_n=\frac{1}{N-2}(ee`)$, it is enough to prove that LSEs $\hat{\beta}=\hat{\beta}_N=(\hat{\eta}_N,\hat{c}_N)$ and remainders of regression vector E are non-correlated, because they have Gaussian distribution.

Since $Ee=0$, therefore, $\mathrm{cov}(\hat{\beta}_N,e)=E(\hat{\beta}_N-\beta)e`=EA\varepsilon(\varepsilon B)`=AE(\varepsilon\varepsilon`)B`=$ $=\sigma^2(AB)=0$, where the equalities $AB=(X`X)^{-1}X`(I_N-X(X`X)^{-1}X`)=0$ were used.

§ 4. Properties of LSEs $\hat{c}_N$ and $\hat{\eta}_N$. Let the constraint (1.2) holds.

Theorem 3. The statistics

$$t_{N-2}^{(1)} = \frac{(\hat{\eta}_N - \eta)}{\hat{\sigma}_N}\left(\frac{r_N}{S_N}\right)^{1/2} \quad \text{and} \quad t_{N-2}^{(2)} = \frac{(\hat{c}_N - c)}{\hat{\sigma}_N}\cdot\left(\frac{r_N}{S_N}\right)^{1/2}$$

have Student's distribution $T(N-2)$ with $N-2$ degrees of freedom.

Proof. Due to Theorem 1, $\hat{\eta}_N - \eta$ has Gaussian distribution $N\left(0, \sigma_{\hat{\eta}_N}^2\right)$, where $\sigma_{\hat{\eta}_N}^2 = D\hat{\eta}_N = \dfrac{\sigma^2}{r_N} S_N$.

Let us take as an estimate for $\sigma_{\hat{\eta}_N}^2$ the statistics $\dfrac{\sigma_N^2}{r_N} S_N$, i.e. $S_{\hat{\eta}_N}^2 = \hat{\sigma}_{\hat{\eta}_N}^2 = (\hat{\sigma}_N^2 S_N)/r_N$.

Due to Theorem 2, the statistics $\chi_{N-2}^2 = (N-2)\dfrac{\hat{\sigma}_N^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum e_n^2$ has distribution $H^2(N-2)$. Now let us consider the following statistics:

$$t_{N-2}^{(1)} = \frac{(\hat{\eta}_N - \eta)/\sigma_{\hat{\eta}_N}}{S_{\hat{\eta}_N}/\sigma_{\hat{\eta}_N}} = \frac{\xi_0}{\sqrt{\dfrac{1}{N-2}\chi_{N-2}^2}},$$

where $\xi_0 = (\hat{\eta}_N - \eta)/\sigma_{\hat{\eta}_N}$ has distribution $N(0,1)$ and

$$\frac{S_{\hat{\eta}_N}}{\sigma_{\hat{\eta}_N}} = \frac{\hat{\sigma}_N\left(S_N/r_n\right)^{1/2}}{\sigma\left(S_N/r_n\right)^{1/2}} = \frac{\hat{\sigma}_N}{\sigma} = \sqrt{\frac{1}{N-2}\chi_{N-2}^2}.$$

According to the Remark, $\hat{\eta}_N$ and e are independent. That is why random variables ξ_0 and χ_{N-2}^2 are independent too. It implies, due to definition of random variable, which has Student's distribution, that the statistics $t_{N-2}^{(1)}$ has Student's distribution $T(N-2)$ with $N-2$ degrees of freedom.

Theorem 3 is proved.

Received 17.12.2009

REFERENCES

1. **Astola F., Danielian E.** Regularly Varying Skewed Distributions Generated by Birth-Death Process. Tampere, TICSP, Series 27, 2004, 98 p.
2. **Vardanian Kh.L., Hovhannisyan H.G.** Vestnic Gosudarstvennogo Ingenernogo Universiteta Armenii, 2009, issue 12, v. 2, p. 46–52 (in Russian).
3. **Magnus Ya., Khatishev P., Peresedski A.** Economics. Initial Course. M.: Delo, 2004 (in Russian).
4. **Borovkov A.A.** Mathematical Statistics. M.: Nauka, 1984, 472 c. (in Russian).

Proc. of the Yerevan State Univ. Phys. and Mathem. Sci., 2010, № 1, p. 27–31.

Խ. Լ. Վարդանյան

Չափավոր աճի բաշխումներով ծնված ռեգրեսիոն մոդելներ

Դիտարկվում է կենսաինֆորմատիկայում առաջ եկած չափավոր աճի երկպարամետրական բաշխման միջոցով ծնված ռեգրեսիոն մոդել: Ապացուցվում է մոդելի պարամետրերի նվազագույն քառակուսիների գնահատականների ունակությունը թույլ իմաստով: Ստացված են պարամետրերի նվազագույն քառակուսիների գնահատականների և գաուսյան աղմուկի դիսպերսիայի գնահատականների բաշխումները, որոնք կարող են օգտագործվել մոդելի պարամետրերին վերաբերող վարկածների ստուգման համար:

Х. Л. Варданян.

Регрессионные модели, порожденные распределением умеренного роста

Рассматривается регрессионная модель, порожденная двухпараметрическим распределением умеренного роста, которое возникает в биоинформатике. Доказывается состоятельность в слабом смысле оценок наименьших квадратов параметров модели. Получены распределения оценок наименьших квадратов параметров и оценки дисперсии гауссовского шума, которые могут быть использованы для проверки гипотез относительно параметров модели.