Math. Maced.
Vol. 1 (2003)
9-13

PARTIAL ERROR-FREE POLYNOMIAL REGRESSION

TIBOR K. POGÁNY

Abstract. New parametric regression method is proposed (called *partial error-free regression*) in the polynomial regression case, in which the sum of errors in sample nodes vanish for all subsamples. Special attention is given to the linear partial error-free regression. In this case convergence results are presented, and the connections to some kind of numerical integration quadrature formulæ are exposed.

1. Introduction to error-free regression

Interested in the polynomial regression problem for the two-dimensional sample $\mathbb{U} := \{(x_j, y_j) : 1 \le j \le n\}$ we introduce the following approach. Assume that regression polynomial is of the form

$$y = P_{m-1}(x) = p_0 + \cdots + p_{m-1}x^{m-1}, \qquad m \le n. \tag{1}$$

Here we are looking for the unknown parameter array $\mathbf{p} := (p_0, \cdots, p_{m-1})$. In this goal denote $\epsilon_j := P_{m-1}(x_j) - y_j, j = \overline{1,n}$ the elementwise regression error in the model (1) with respect to the sample $\mathbb{U}$. To find $\mathbf{p}$ we consider the case

$$\sum_{j=1}^{n} \epsilon_j = \sum_{j=1}^{n} (P_{m-1}(x_j) - y_j) = 0. \tag{2}$$

Now, the term *error-free* is selfexplanatory. Since (2) possesses unique solution only when $m = 1$ we will assume that $m > 1$ and we decompose $\{1, 2, \cdots, n\}$ into an union of nonempty, disjoint subsets $J_1, \cdots, J_m$. The resulting index-sets J_l generate subsamples U_l of $\mathbb{U}$ in the manner that $(x_j, y_j) \in U_s$ when $j \in J_s$, $1 \le s \le m$.

Repeating the procedure (2) for all decompositions $\mathbb{J} = \{J_1, \cdots, J_m\}$, say, we get the linear algebraic system in $\mathbf{p}$:

$$\sum_{k=0}^{m-1} \overline{\mathbf{x}_l^k} p_k = \overline{\mathbf{y}_l} \qquad 1 \le l \le m, \tag{3}$$

where $\overline{\mathbf{x}_l^k} := \frac{1}{\#J_l} \sum_{j \in J_l} x_j^k$, $\overline{\mathbf{y}_l} := \frac{1}{\#J_l} \sum_{j \in J_l} y_j$. The decomposition $\mathbb{J}$ is *admissible* when the system (3) has unique solution. Endly, averaging all p_j's by the number of all admissible $\mathbb{J}$ we get the optimal error-free coefficient array $\mathbf{p}^\star = (p_0^\star, \cdots, p_{m-1}^\star)$.

2000 Mathematics Subject Classification. 62-07, 62 F 10, 62 J 02, 62 P 30.

Key words and phrases. Admissible sample, error-free regression, heteroscedastic linear regression model, Monte Carlo method, polynomial regression, quadrature formulæ, sample decomposition.

Concerning the admissibility of $\mathbb{J}$ the following simple geometrical criterion could be given. Namely, $\mathbb{J}$ *is admissible* **iff** *at least two first coordinates of centroids of* J_k *differ* $k = \overline{1,m}$, see [5, Theorem 2.1, Theorem 2.2]. Also it could be mentioned that the error-free regression polynomial

$$y = \sum_{k=0}^{m-1} p_k^\star x^k$$

contains the centroid of the plane polygon $\mathbb{U}$. Additional comparations with the Minimal Least Squares regression polynomial are given in [6].

The main lack of the error-free regression method is that we have to restrict ourselves to the set of admissible decompositions of $\{1, \cdots, n\}$. To avoid this inconvenience, we introduce the so-called *partial error-free* approach.

2. Partial error-free regression

Assume that we know already the regression polynomial degree $m-1$, say. Then let $U_l^{(m)}$ one of the admissible subsamples of $\mathbb{U}$, which size equals m. Denote $\mathfrak{N}_{\mathbb{U}}$ the cardinality of the set of all m-sized admissible subsamples of $\mathbb{U}$. Obviously, $\mathfrak{N}_{\mathbb{U}} \le \binom{n}{m}$.

Consider the system in $\mathbf{p}$:

$$\sum_{k=0}^{m-1} p_k x_j^k = y_j, \qquad \forall (x_j, y_j) \in U_l^{(m)}. \tag{4}$$

Since for all different x's in $\mathbb{U}$ the system determinant in (4) is of Vandermonde type, there exists an unique solution $\mathbf{p}_l := (p_0^{(l)}, \cdots, p_{m-1}^{(l)})$, say. In the same time we deduce that any m-sized subsample $U_l^{(m)}$ of the sample $\mathbb{U}$ is admissible, iff all first coordinates of its nodes are different. Consequently, if $\mathbb{U}$ is admissible, then

$$\mathfrak{N}_{\mathbb{U}} = \binom{n}{m}.$$

Denote

$$\widehat{\mathbf{p}} = (\widehat{p}_0, \cdots, \widehat{p}_{m-1}) := \left(\frac{1}{\mathfrak{N}_{\mathbb{U}}} \sum_{l=1}^{\mathfrak{N}_{\mathbb{U}}} p_0^{(l)}, \cdots, \frac{1}{\mathfrak{N}_{\mathbb{U}}} \sum_{l=1}^{\mathfrak{N}_{\mathbb{U}}} p_{m-1}^{(l)} \right). \tag{5}$$

All this results in

$$y = f(x) \approx \widehat{P}_{m-1}(x) = \widehat{p}_0 + \cdots + \widehat{p}_{m-1} x^{m-1},$$

where the approximation $\approx$ is in the *partial error-free sense* used, having on mind the constraint (4).

3. Linear partial error-free regression

The very important linear case of regression models, in the Minimal Least Squares manner used, is well-covered in literature. Therefore, here we will discuss the linear regression model taking partial error-free method instead of the MLS one.

Consider an admissible sample $\mathbb{U}$ which index set $\{1, \cdots, n\}$ possesses a decomposition $\mathbb{J}$. It is clear that we can restrict ourselves to samples with $x_j \in [0,1]$ without any loss of generality. Indeed, taking $\min_{1\le j\le n} x_j = a$ while $\max_{1\le j\le n} x_j = b$ instead of the initial sample $\mathbb{U}$, we deal with $(b-a)^{-1}\mathbb{U} := \{(x_j/(b-a), y_j) : 1 \le j \le n\}$. So, let $(x_j, y_j), (x_k, y_k) \in U_l^{(2)}$ for some fixed $l \in \{1, \cdots, \binom{n}{2}\}$. We are looking for the values of the coefficients $p_0^{(jk)}, p_1^{(jk)}$ in this restricted variant of the system (4). We get

$$p_1^{(jk)} = \frac{y_j - y_k}{x_j - x_k}, \qquad p_0^{(jk)} = y_j - \frac{y_j - y_k}{x_j - x_k} x_j. \tag{6}$$

According to (5) it will be

$$\widehat{p}_0 = \binom{n}{2}^{-1} \sum_{1\le j<k\le n} p_0^{(jk)}, \qquad \widehat{p}_1 = \binom{n}{2}^{-1} \sum_{1\le j<k\le n} p_1^{(jk)}, \tag{7}$$

and the regression line is

$$y = \widehat{P}_1(x) = \widehat{p}_0 + \widehat{p}_1 x. \tag{8}$$

Let us transform (8) by using concrete values of the coefficients for some fixed $\xi \in [0,1]$. We get

$$\begin{aligned}
\widehat{P}_1(\xi) &= \binom{n}{2}^{-1} \sum_{l=1}^{n} (n-l) y_l + \binom{n}{2}^{-1} \sum_{1\le j<k\le n} \frac{y_j - y_k}{x_j - x_k} (\xi - x_j) \\
&= \underbrace{\frac{2n}{n-1} \sum_{l=1}^{n} \left(1 - \frac{l}{n}\right) y_l \Delta x_{n,l}}_{\mathbf{I}_s} + \underbrace{\frac{2n}{n-1} \sum_{1\le j<k\le n} \frac{y_j - y_k}{x_j - x_k} (\xi - x_j) \Delta x_{n,j} \Delta x_{n,k}}_{\mathbf{I}_d},
\end{aligned} \tag{9}$$

where

$$\Delta x_{n,l} := x_{n,l} - x_{n,l-1} \equiv \tfrac{l}{n} - \tfrac{l-1}{n}.$$

When the sample $\mathbb{U}$ is generated by a bounded function $y = f(x)$, i.e. $f \in M[0,1]$, say, we recognize $\mathbf{I}_s$ as the Riemannian simple-integral sum of the function $(1-t)f(t)$ on the integration domain $t \in [0,1]$. Similarly $\mathbf{I}_d$ is the Riemannian double-integral sum of function $(f(t) - f(s))(\xi - t)(t - s)^{-1}$ for the integration domain $(t,s) \in [0,1] \times (t,1]$. Letting $n \to \infty$ in (9), we deduce the following result:

$$\lim_{n\to\infty} \widehat{P}_1(\xi) = 2\int_0^1 (1-t) f(t)dt + 2\int_0^1 \int_t^1 \frac{f(t) - f(s)}{t - s} (\xi - t) dtds. \tag{10}$$

Theorem 1. *Let $f \in M[0,1]$ generates the sample $\mathbb{U}$ with the admissible decomposition $\mathbb{J}$. Then we have*

$$\lim_{\substack{n\to\infty \\ \max \Delta x_{n,j} \to 0}} \widehat{P}_1(\xi) = 2\int_0^1 \int_t^1 \frac{f(s)(t-\xi) - f(t)(s-\xi)}{t - s} dtds. \tag{11}$$

Proof. Elementary transformations of (10) lead to (11). □

As the straightforward consequence of the Theorem 1 it follows the following quadrature formula:

$$\int_0^1\int_t^1 \frac{f(s)(t-\xi)-f(t)(s-\xi)}{t-s}dtds$$

$$\approx \frac{1}{n(n-1)}\left(\sum_{l=1}^{n}(n-l)y_l + \sum_{1\le j<k\le n}\frac{y_j-y_k}{x_j-x_k}(\xi-x_j)\right) \tag{12}$$

such that is valid for all $f \in M\,[0,1]$.

One of our next goals is to find the distance between $\widehat{P}_1(x)$ and the point

$$\mathsf{C}(\overline{x}_n \equiv \tfrac{1}{n}\sum_{j=1}^{n} x_j, \overline{y}_n \equiv \tfrac{1}{n}\sum_{j=1}^{n} y_j),$$

where C is the centroid of the sample $\mathbb{U}$. In this purpose balancing with suitable ξ in (11) one gets the desirable result taking $\xi = \overline{x}_n$. By this we have $\xi \to \frac{1}{2}$ as $x_{n,j} \sim \frac{j}{n}$ when n is growing. On the other side, deducing by the Monte Carlo method, we obtain the approximation

$$\overline{y}_n = \frac{1}{n}\sum_{j=1}^{n} y_j \xrightarrow[n\to\infty]{} \int_0^1 f(t)dt.$$

Now, by (11) we deduce:

$$\widehat{P}_1(\overline{x}_n) - \overline{y}_n = \int_0^1\int_t^1\left(\frac{f(s)(2t-1)-f(t)(2s-1)}{t-s} - \frac{f(t)}{1-t}\right)dtds. \tag{13}$$

Taking vanishing integrand in (13), the problem: *"When is $\widehat{P}_1(x)$ containing the centroid C of the sample $\mathbb{U}$?"* will be reduced to the functional equation

$$\frac{f(s)(2t-1)-f(t)(2s-1)}{t-s} = \frac{f(t)}{1-t}, \qquad (t,s) \in [0,1]\times(t,1]. \tag{14}$$

It is not hard to see that the unique solution of (14) is $f(x) = \gamma(1-x)$, for all $x \in [0,1]$ with some real constant γ. Hence, the following result is proved.

Corollary 1.1. *The centroid of an admissible sample $\mathbb{U}$ belongs to the partial error-free regression line $y = \widehat{P}_1(x)$* **iff** *the sample $\mathbb{U}$ takes the form*

$$\mathbb{U}_\gamma = \{(x_j, \gamma(1-x_j)) : \; x_j \in [0,1],\; 1 \le j \le n\}.$$

Remark 1. *Of course, the admissibility is not a prerequisite to applying the partial error-free quadrature formula* (12). *So, using* linear partial error-free approach *instead of the* error-free *one, we are loosing the desirable property that the centroid C belongs to the regression line. At this point we can only conjecture that the suitably weigthed centroid $\mathsf{C}_w(\sum_{j=1}^n w_j x_j, \sum_{j=1}^n w_j y_j)$, where $\sum_{j=1}^n w_j = 1$, $w_j \ge 0$, could avoid this anomaly.*

To find the coefficients of the regression polynomial $\widehat{P}_{m-1}(x)$ we can accept mathematically more sophisticated approaches. One of them is the estimation of solutions of system of linear algebraic equations by the theorem on extremal value of nonnegative bounded functions, compare [3, Chapter 2]. The same item is treated in [2, Chapter 17], but there the Monte Carlo method is the main mathematical tool. The interested reader can consult [4] too.

Finally, let us given the linear statistical model

$$Y_j = p_0 + p_1 X_j + \varepsilon_j, \tag{15}$$

where the statistical sample $\mathbf{X} = (X_1, \cdots, X_n)$ is defined on given probability space $(\Omega, \mathfrak{F}, \mathsf{P})$ and the random variables X_j are i.i.d. Let $Y = f(X)$, f bounded Borel. It would be of interest to observe the homoscedastic case in (15), i.e. the i.i.d.r.v. case $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$. But, having on mind the partial error-free approach (4) with $m = 2$, this is not a realistic assumption (compare (2) as well). Therefore the linear partial error-free polynomial regression has to be a heteroscedastic procedure with the noise i.r.v's distributed like

$$\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \qquad j = \overline{1, n}.$$

The detailed treatment of the heteroscedastic Minimal Least Square regressions can be found in [1, Chapter 7, §4], [7].

References

[1] Demidenko, E.Z., *Optimization and Regression*, Nauka, Moscow, 1989. (Russian)

[2] Demidovich, B.P. & Maron, I.A., *Computational Mathematics*, Mir Publishers, Moscow, 1987.

[3] Girko, V.L., *Multivariate Statistical Analysis*, Vishcha Shkola, Kiev, 1988. (Russian)

[4] Levin, B.R., *Theoretical Foundations for Statistical Radio Engineering*, 3rd. ed. rev. and compl., Radio i Svyaz', Moscow, 1989. (Russian)

[5] Pogány, T. & Tudor, M., On the error-free polynomial regression model **I.**, *Proceedings of the* **KOI'93**, Rovinj, 1993, (L. Neralić *et al. eds.*), Hrvatsko društvo za operacijska istraživanja, Zagreb, 1993, 207-216.

[6] Pogány, T., On the error-free polynomial regression model **II.** (Least squares estimation under error-free condition), *Proceedings of the* **KOI'93**, Rovinj, 1993, (L. Neralić *et al. eds.*), Hrvatsko društvo za operacijska istraživanja, Zagreb, 1993, 201-206.

[7] Taylor, W., The heteroscedastic linear model: exact finite sample results, *Econometrica* **43**, No. 3, 1978, 663-676.

Tibor K. Pogány, Department of Sciences, Faculty of Maritime Studies, University of Rijeka, 51000 Rijeka, Studentska 2, Croatia

E-mail address: poganj@pfri.hr