On Randomizing Estimators in Linear Regression Models

S.Ermakov and R.Schwabe

Abstract: In this work we consider a special kind of randomization in the analysis of linear regression. This randomization is connected to the Δ^2-distribution which was first introduced by S.M. Ermakov and V.G. Zolotukhin (1960) for decreasing the variance in the Monte Carlo calculation of integrals. The resulting resampling procedure allows for separating the systematic and the random components of the variance. Further, some problems are discussed in designing regression experiments.

Keywords and phrases: randomization, least squares estimator, linear regression, bootstrap, experimental design.

In the article by Ermakov and Zolotukhin(1960) a special distribution was introduced for the nodes in interpolation-cubature formulae which will be called the Δ^2-distribution. This distribution appeared to be efficient in the theory of Monte Carlo methods. Subsequently, Ermakov(1963, 1989) studied interpolation procedures where the nodes were chosen at random according to the Δ^2-distribution. That allowed for extending the application to the field of experimental design and to prove a number of existence theorems for interpolation-cubature formulae [Ermakov(1993)]. Apparently, there is a close connection between those results and the least squares estimators in the analysis of linear models though this question has not been studied, yet, properly. In general there were only common considerations discussed.

Here we will show this connection in detail and obtain some new results concerning resampling procedures as well as connections between D-optimal designs and the average of random interpolation nodes.

[1]This work was supported by the RFBR grant 96-01-00015G

1 Some properties of the Δ^2 distribution

First, we will make use of the following known results [Ermakov(1975)]:

Let $\mathcal{X}$ be a set, μ a probability measure on $\mathcal{X}$ and $f \in L^2(\mu)$ a function of interest on $\mathcal{X}$ which is square integrable with respect to μ.

Instead of f a function ξ, $\xi(x_i) = \xi_i$, $i = 1, 2, ..., m$ may be calculated (observed) at m points $x_1, ..., x_m$ from $\mathcal{X}$. Moreover, we assume that ξ is a random perturbation of f:

$$\xi_i = f(x_i) + \varepsilon_i, \qquad \mathbf{E}\varepsilon_i = 0, \qquad \mathbf{E}(\varepsilon_i\varepsilon_j) = \begin{cases} \sigma^2 & , \quad i = j \\ 0 & , \quad i \neq j \end{cases}$$

$i, j = 1, ..., m$.

A system of m functions $\{\phi_l\}_{l=1}^m$ on $\mathcal{X}$ is selected which is orthonormal with respect to μ. Now the problem is considered of approximating the function f by means of a linear combination $\sum_{l=1}^m c_l\phi_l$ of $\phi_1, ..., \phi_l$.

We will choose the points $x_1, ..., x_m$ at random according to the density

$$p(x_1, ..., x_m) = (m!)^{-1} \left(\det \|\phi_i(x_j)\|_{i,j=1}^m\right)^2$$

(Δ^2-distribution).

Then we solve the problem of interpolation for each set $x_1^{(k)}, ..., x_m^{(k)}$ with observations $\xi_1^{(k)}, ..., \xi_m^{(k)}$ where k is the replication number, i. e. we determine the coefficients $c_l^{(k)}$ from the system of equations

$$\sum_{l=1}^m c_l^{(k)}\phi_l(x_j^{(k)}) = \xi_j^{(k)}. \tag{1.1}$$

Theorem 1.1 *The expectation of $c_l^{(k)}$ is equal to the lth Fourier coefficient of the function f with respect to the system $\{\phi_l\}_{l=1}^m$:*

$$\mathbf{E}c_l^{(k)} = \int_{\mathcal{X}} f(x)\phi_l(x)\mu(dx). \tag{1.2}$$

Further, we introduce the definition of D-regularity of a system of functions which weakens standard regularity conditions [Ermakov and Kurotschka (1995)].

Definition. A system of functions is D-regular if

$$\begin{aligned}&\mu^m(\{(x_1, ..., x_m) : \det \|\phi_i(x_j)\|_{i,j=1}^m = 0 \text{ and} \\ &\qquad \det \|\phi_1(x_j), ..., \phi_{l-1}(x_j), f(x_j), \phi_{l+1}(x_j), ..., \phi_m(x_j)\|_{j=1}^m \neq 0\}) = 0\end{aligned}$$

for $l = 1, ..., m$ and every function f on $\mathcal{X}$.

This definition is, in particular, suitable for discrete measures μ as it requires regularity only on the support of μ.

Theorem 1.2 *If $\phi_1, ..., \phi_m$ is a D-regular system, then*

$$Cov(c_l^{(k)}, c_r^{(k)}) = \begin{cases} \int \left(f(x) - \sum_{i=1}^m \langle f, \phi_i\rangle \phi_i(x)\right)^2 \mu(dx) + \sigma^2, & l = r \\ 0, & l \neq r \end{cases} \tag{1.3}$$

where $\langle f, \phi_i\rangle = \int f(x)\phi_i(x)\mu(dx)$.

---

2 Least squares estimators and their randomization

In this section we consider the problem of estimating the coefficients in linear regression in the simplest formulation corresponding to the previous section. Only a finite set of points $x_1, ..., x_N$ is of interest and we use the uniform measure μ which is concentrated on the points $x_1, ..., x_N$ with equal weights $\frac{1}{N}$. It is supposed that data $\xi_1, ..., \xi_N$ can be made available, $\xi_i = f(x_i) + \varepsilon_i$, with $\mathbf{E}\varepsilon_i = 0$, $Var(\varepsilon_i) = \sigma^2$, and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

If we approximate f by a linear combination $\sum_{l=1}^m c_l\phi_l$ where the ϕ_l are linearly independent functions on the set $x_1, ..., x_N$ then the least squares estimators $\hat{c}_l$ of the coefficients c_l are expressed by

$$\hat{c}_l = \frac{\det \|\langle \phi_1, \phi_k\rangle, ..., \langle \phi_{l-1}, \phi_k\rangle, \langle \xi, \phi_k\rangle, \langle \phi_{l+1}, \phi_k\rangle, ..., \langle \phi_m, \phi_k\rangle\|_{k=1}^m}{\det \|\langle \phi_k, \phi_r\rangle\|_{k,r=1}^m} \tag{2.1}$$

where, here, $\langle \phi_k, \phi_r\rangle = \frac{1}{N}\sum_{j=1}^N \phi_k(x_j)\phi_r(x_j)$ and $\langle \xi, \phi_k\rangle = \frac{1}{N}\sum_{j=1}^N \xi_j\phi_k(x_j)$. If the functions ϕ_l are orthonormal on the set of points $x_1, ..., x_N$ then the estimators of the coefficients simplify to

$$\hat{c}_l = \frac{1}{N}\sum_{j=1}^N \xi_j\phi_l(x_j) \tag{2.2}$$

Note that if we use the Cauchy–Binet formula we have

$$\hat{c}_l = \frac{\sum_{1\le i_1<...<i_m\le N} \Delta_l(\xi, i_1, ..., i_m)\Delta(i_1, ..., i_m)}{\sum_{1\le i<...<i_m\le N} \Delta^2(i_1, ..., i_m)} \tag{2.3}$$

where

$$\Delta(i_1, ..., i_m) = \det \|\phi_k(x_{i_j})\|_{k,j=1}^m \qquad \text{and}$$

$$\Delta_l(\xi, i_1, ..., i_m) = \det \|\phi(x_{i_j}), ..., \phi_{l-1}(x_{i_j}), \xi(x_{i_j}), \phi_{l+1}(x_{i_j}), ..., \phi_m(x_{i_j})\|_{j=1}^m.$$

Considering the equality of the corresponding products of determinants when indices are interchanged we have

$$\hat{c}_l = \sum_{i_1,...,i_m} \Delta_l(\xi, i_1, ..., i_m)\Delta(i_1, ..., i_m) \Big/ \sum_{i_1,...,i_m} \Delta^2(i_1, ..., i_m)\ . \tag{2.4}$$

If we introduce a probability distribution of sets of points $x_{i_1}, ..., x_{i_m}$ by the density

$$p(i_1, ..., i_m) = \Delta^2(i_1, ..., i_m) \Big/ \sum_{i_1,...,i_m} \Delta^2(i_1, ..., i_m) \tag{2.5}$$

then the random variable χ_l defined by

$$\chi_l(i_1, ..., i_m) = \Delta_l(\xi, i_1, ..., i_m)/\Delta(i_1, ..., i_m) \tag{2.6}$$

has expectation (conditionally on ξ fixed) equal to the value of the least squares estimator $\hat{c}_l$ determined by formula (2.1).

At the same time according to Cramer's rule the values $\chi_l(i_1, ..., i_m)$ are the solutions of the system of equations

$$\sum_{l=1}^m \chi_l(i_1, ..., i_m)\phi_l(x_{i_j}) = \xi_{i_j}, \quad j = i, ..., m. \tag{2.7}$$

The corresponding determinant differs from 0 with probability 1.

Thus we see that the distribution (2.5) is a Δ^2-distribution with respect to the uniform measure on $x_1, ..., x_N$. The correctness of the formula (1.3) for the case of D-regular $\{\phi_l\}$ arises from here. In the present notation this formulae can be rewritten in the following way:

$$Cov(\chi_l, \chi_r) = \begin{cases} \frac{1}{N}\sum_{j=1}^N \left(f(x_j) - \sum_{i=1}^m \langle f, \phi_i\rangle\phi_i(x_j)\right)^2 + \sigma^2, & l = r \\ 0, & l \neq r \end{cases} \tag{2.8}$$

where $\langle f, \phi_l\rangle = \frac{1}{N}\sum_{j=1}^N f(x_j)\phi_l(x_j)$.

Next we investigate two possibilities of sampling. Each of them leads to a different value of the final variance.

1. Randomization associated to an active experiment (independent replications are possible): A replication $x_{i_1}, ..., x_{i_m}$ is chosen at random according to the distribution (2.5) and independent values ξ_{i_j} are measured (observed) at each point. If we choose a new replication $x_{i'_1}, ..., x_{i'_m}$ then the observed values $\xi'_{i'_j}$ do not depend on the ξ_{i_j} obtained earlier.

2. Randomization is carried out after that we have got all N values of the function f (bootstrap, passive experiment).

In the first case, when we choose N_1 independent replications $x_{i_1}^{(k)}, ..., x_{i_m}^{(k)}$, $k = 1, ..., N_1$, we obtain for the average $\frac{1}{N_1}\sum_{k=1}^{N_1} \chi_l^{(k)}$ of estimators for the lth regression coefficient c_l

$$Var\left(\frac{1}{N_1}\sum_{k=1}^{N_1} \chi_l^{(k)}\right) = \frac{1}{N_1}(S^2 + \sigma^2),$$

where

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}\left(f(x_i) - \sum_{l=1}^{m}\langle f, \phi_l\rangle \phi_l(x_i)\right)^2$$

is the systematic error in the approximation of f by $\phi_1, ..., \phi_m$.

If, moreover, N_2 repeated independent measurements are possible for each replication $x_{i_1}^{(k)}, ..., x_{i_m}^{(k)}$ then we have

$$Var\left(\frac{1}{N_1 N_2}\sum_{k=1}^{N_1 N_2} \chi_l^{(k)}\right) = \frac{1}{N_1}\left(S^2 + \frac{\sigma^2}{N_2}\right). \tag{2.9}$$

In this case it is possible to separate the systematic component S^2 and the random σ^2 component of the variance when we repeat the experiment in the same m points. This procedure was described earlier by Ermakov (1975). As we can see much more experiments may be needed.

In the second case it is assumed that the experiment is carried out in N points (these are supposed to be different). The only aim of our procedure is to separate the systematic and the random components of the variance (bootstrap procedures).

As the result of the experiment we have got the observations $\xi_1, ..., \xi_N$ at $x_1, ..., x_N$. $\{\phi_l\}$ is assumed to be a D-regular orthonormal system of functions and $(i_1, ..., i_m)$ and $(i'_1, ..., i'_m)$ denote two independent samples according to the distribution (2.5).

Lemma 2.1 *For the estimators χ_l and χ_l' of the lth regression coefficients based on the replications $(i_1, ..., i_m)$ and $(i_1', ..., i_m')$, respectively, which are defined by (2.6) we have*

$$Cov(\chi_l, \chi_l') = \frac{\sigma^2}{N}\,. \tag{2.10}$$

Proof. Note that the overall covariance can be decomposed into the covariance of the conditional expectation and the expectation of the conditional covariance according to

$$Cov(\chi_l, \chi_l') = Cov(\mathbf{E}(\chi_l|\varepsilon_1, ..., \varepsilon_N), \mathbf{E}(\chi_l'|\varepsilon_1, ..., \varepsilon_N)) + \mathbf{E}Cov(\chi_l, \chi_l'|\varepsilon_1, ..., \varepsilon_N).$$

Moreover, $\mathbf{E}(\chi_l|\varepsilon_1, ..., \varepsilon_N) = \mathbf{E}(\chi_l'|\varepsilon_1, ..., \varepsilon_N) = \frac{1}{N}\sum_{i=1}^N \xi_j$ and the estimators χ_l and χ_l' are conditionally independent under fixed $\varepsilon_1, ..., \varepsilon_N$. Thus the second term in the sum equals 0. For the first we have $Var(\frac{1}{N}\sum_{i=1}^N \xi_i) = \sigma^2/N$ which completes the proof.

Now, let the repeated sampling of N_1 independent replications be carried out under the fixed observations $\xi_1, ..., \xi_N$. Then

$$Var\left(\frac{1}{N_1}\sum_{k=1}^{N_1} \chi_l^{(k)}\right) = \frac{1}{N_1} Var(\chi_l) + \left(1 - \frac{1}{N_1}\right) Cov(\chi_l, \chi_l') \tag{2.11}$$

Here, the estimator $\chi_l^{(k)}$ corresponds to the kth independent replication.

The results can be summarized in the following theorem.

Theorem 2.1 *The variance of average of the l-th regression coefficient estimators (2.6) that correspond to N_l independent replics in a passive experiment is expressed by*

$$Var\left(\frac{1}{N_1}\sum_{k=1}^{N_1} \chi_l^{(k)}\right) = \frac{1}{N_1} S^2 + \frac{N + N_1 - 1}{NN_1}\sigma^2\,.$$

Proof. According to (2.11) we have

$$Var\left(\frac{1}{N_1}\sum_{k=1}^{N_1} \chi_l^{(k)}\right) = \frac{1}{N_1}(S^2 + \sigma^2) + \left(1 - \frac{1}{N_1}\right)\frac{\sigma^2}{N}$$

as was to be proved.

These results enable us to use the randomization rather flexibly as we separate the errors of different kinds in the least squares method.

Note. As was shown in Ermakov(1975) a randomized procedure for $N > M > m$ can be constructed. Namely, let as before $x_1, ..., x_N$ be points where the experiment is carried out. Let $x_{i_1}^{(k)}, ..., x_{i_M}^{(k)}$ be the kth replication distributed according to the density

$$p(i_1, ..., i_M) = \frac{\det \|[\phi_i, \phi_k]_M\|(i_1, ..., i_M)}{\sum_{j_1=1}^{N} \cdots \sum_{j_M=1}^{N} \det \|[\phi_i, \phi_k]_M\|(j_1, ..., j_M)} \tag{2.12}$$

where $[\phi_i, \phi_k]_M = [\phi_i, \phi_k]_M(i_1, ..., i_M) = \frac{1}{M}\sum_{r=1}^{M} \phi_i(x_{i_r})\phi_k(x_{i_r})$ and the list of indices in brackets denotes the replication that we use to calculate the least squares estimators

$$\chi_l^{(M)} = \frac{\det \|[\phi_1, \phi_k]_M, ..., [\phi_{l-1}, \phi_k]_M, [\xi, \phi_k]_M, ..., [\phi_m, \phi_k]_M\|(i_1, ..., i_M)}{\det \|[\phi_i, \phi_k]_M\|(i_1, ..., i_M)}.$$

Then $\mathbf{E}(\chi_l^{(M)}|\xi) = \chi_l^{(N)}$ where $\chi_l^{(N)}$ is the least squares estimator of the coefficient c_l based on all N points. Unfortunately, we did not succeed, yet, in getting an explicit expression similar to (2.8), as we cannot get rid of the denominator in this expression.

The only setting where we we are able to calculate $Var\chi_l^{(M)}$ is the case of piecewise constant functions ϕ_l (wavelets, histograms). However, the system $\{\phi_l\}$ is not D-regular in this case [Ermakov(1989)].

---

3 Δ^2-distribution in experimental design

As an example let us consider the situation of polynomial regression on the unit interval [0,1]

$$\phi_l(x) = x^{l-1}, \qquad l = 1, ..., m. \tag{3.1}$$

With respect to the Lebesgue measure the density of the Δ^2-distribution is

$$p(x_1, ..., x_m) = \mathbf{c} \prod_{i<j} (x_i - x_j)^2 \qquad (3.2),$$

i.e. up to a constant $\mathbf{c}$ it is equal to the squared Vandermonde determinant. The maximum of this determinant is known [Schur (1918)] and this allows for using the rejection method for simulating the density. By sampling realizations $x_1^{(k)}, ..., x_m^{(k)}$ and putting the received numbers in increasing order $x_{(1)}^{(k)}, ..., x_{(m)}^{(k)}$, where $x_{(1)}^{(k)} \leq x_{(2)}^{(k)} \leq ... \leq x_{(m)}^{(k)}$ and $\{x_{(1)}^{(k)}, ..., x_{(m)}^{(k)}\} = \{x_1^{(k)}, ..., x_m^{(k)}\}$, we can notice that these numbers are classified near some average values. It is interesting to study the behavior of these means $\overline{x_{(i)}} = \frac{1}{N}\sum_{k=1}^{N} x_{(i)}^{(k)}$ and, in particular, to compare these averages to the nodes of optimal experimental designs.

It is known that one of the widely used optimality criteria, the D-criterion, leads to experimental designs where the global maximum in (3.2) is attained. Consequently, the nodes of this design are the roots of the polynomial $x(1-x)P'_{m-1}(x)$, where P'_n is the derivative of the Legendre polynomial P_m of degree m.

On the other side, the so-called unbiased optimal designs (designs that minimize systematic error) are located at the roots of the Legendre polynomial [Ermakov and Sedunov(1979)].

Finally, we consider some results for the simulation ofthe Δ^2-distribution [Schwabe(1992)]. In some cases exact calculations are possible for the expectations of the ordered samples $x_{(1)}, ..., x_{(m)}$.

In order to get an impression of the shape of the randomization measure associated with the Δ^2-distribution for fitting polynomials (3.1) we generated random numbers by using the von Neumann rejection method. For $m = 2$, i.e. linear regression, we generated $N = 10,000$ samples (x_1, x_2), while $N = 1,000$ were chosen for $m = 3$ and 5, and $N = 200$ for $m = 10$. The means $\overline{x_{(1)}}, ..., \overline{x_{(m)}}$ have been calculated from the ordered samples to get an estimate for their corresponding expectations $\mathbf{E}x_{(1)}, ..., \mathbf{E}x_{(m)}$. For lower degrees up to $m = 5$ exact results were obtained for the expectations.

Despite the apparent closeness in the simulation we can notice that the means are a little bit more scattered than the roots of P_n and that they are shifted towards the corresponding nodes of the D-optimal design, which are also listed in TABLE I. This observation led us to investigate the distribution of the smallest component in the sample more closely.

Lemma 3.1 *For the Δ^2-distribution defined in* (3.2) *the smalles component* $x_{(1)} = \min(x_1, ..., x_m)$ *has distribution function* $F(x) = 1-(1-x)^{m^2}$, *density* $f(x) = m^2(1-x)^{m(m-1)}$ *and expectation* $\mathbf{E}x_{(1)} = \frac{1}{m^2+1}$.

The expectations $\mathbf{E}x_{(1)}$ listed in TABLEs I and II are by a multiplicative factor smaller than the corresponding smallest roots $z_1^{(m)}$ of the Legendre polynomials P_m. In particular, for large degree m we have $z_1^{(m)} \sim \frac{1}{4} j_1^2 m^{-2} \approx 1.45 m^{-2}$ where j_1 is the smallest positive root of the Bessel function of the first kind of order zero [Szegö(1959)] and, thus, $\mathbf{E}x_{(1)} \sim m^{-2} \approx 0.69 z_1^{(m)}$, i.e. $\mathbf{E}x_{(1)}$ decreases with the same speed as $z_1^{(m)}$ to zero, which is the corresponding smallest node of the D-optimal design.

More detailed investigation of the one- and two-dimensional marginal distribution of the order statistics $(x_{(1)}, ..., x_{(n)})$ can be found in Schwabe(1992).

The results obtained in the present section give rise to the impression that the randomized design associated with the Δ^2-distribution proposed by Ermakov(1975) is close to both the roots of P_m and to the nodes of the D-optimal design which are known to be of great use in approximation theory and experimental design, respectively.

TABLE I

$m=2$				$m=3$			
P_2	$\overline{x_{(l)}}$	$\mathbf{E}x_{(l)}$	D-opt.	P_3	$\overline{x_{(l)}}$	$\mathbf{E}x_{(l)}$	D-opt.
0.2113	0.2005	0.2000	0.0000	0.1127	0.0988	0.1000	0.0000
0.7887	0.7980	0.8000	1.0000	0.5000	0.4986	0.5000	0.5000
				0.8873	0.9008	0.9000	1.0000

$m=5$				$m=10$		
P_5	$\overline{x_{(l)}}$	$\mathbf{E}x_{(l)}$	D-opt.	P_{10}	$\overline{x_{(l)}}$	D-opt.
0.0469	0.0376	0.0385	0.0000	0.0130	0.0107	0.0000
0.2308	0.2116	0.2178	0.1727	0.0675	0.0621	0.0402
0.5000	0.4965	0.5000	0.5000	0.1603	0.1536	0.1306
0.7692	0.7816	0.7822	0.8273	0.2833	0.2828	0.2610
0.9531	0.9615	0.9615	1.0000	0.4256	0.4231	0.4174
				0.5744	0.5838	0.5826
				0.7167	0.7274	0.7390
				0.8397	0.8512	0.8694
				0.9325	0.9431	0.9598
				0.9870	0.9896	1.0000

TABLE II

m	6	7	8	9	10
$z_1^{(m)}$	0.0338	0.0254	0.0199	0.0159	0.0130
$\mathbf{E}x_{(1)}$	0.0270	0.0200	0.0154	0.0122	0.0099

References

Ermakov, S.M. (1963). Interpolation on Random Points, *J. of Comput. Math. and Math. Physics*, **3**, N 1, 550–554. (in Russian).

Ermakov, S.M. (1975). *Die Monte Carlo Methode und verwandte Fragen.* Berlin: Deutscher Verlag der Wissenschaften; and München: Oldenbourg.

Ermakov, S.M. (1989). Random Interpolation in the theory of experimental design. *Comput. Statist. Data Anal.*, **8**, 75–80.

Ermakov, S.M. (1993). On metric theorems in the interpolation theory. *Vestnik St Petersburg State Univ., ser. math., mech., astron.*, **1**, 26.-29 (in Russian).

Ermakov, S.M. and Kurotschka, V.G. (1995). Zufällige Replikationen von Regressionsexperimenten in der Modellbildung. *Preprint Nr. A-31/95.* Freie Universität, Berlin, Fachbereich Mathematik.

Ermakov, S.M. and Sedunov, E.V. (1979). On unbiased designs for regression experiments in a finite-dimensional function space. *Vestnik Leningrad State Univ., ser. math., mech., astron.*, **7**, 11–21 (in Russian).

Ermakov, S.M. and Zolotukhin, V.G. (1960). Polynomial approximations and Monte Carlo method. *J. Probability Theory Applic.*, **5**, N 4, 473–476 (in Russian).

Schur, J. (1918). Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Math. Z.*, **1**, 373–385.

Schwabe, R. (1992). On randomized designs for regression. In: *Model-Oriented Data Analysis. A Survey of Recent Methods* (V. Fedorov, W.G. Müller and I.N. Vuchkov, eds.) Heidelberg: Physica, 53–66.

Szegö, G. (1959). *Orthogonal polynomials*, revised ed. Am. Math. Soc. Coll. Pub., **23**.