Calculating the Variance of the Linear Regression Coefficient

Gurami Tsitsiashvili

•

Russia, 690041, Vladivostok, Radio street, 7
Institute for Applied Mathematics
Far Eastern Branch of Russian Academy Sciences
guram@iam.dvo.ru

Abstract

In this paper, we choose such a particular formulation of the problem of calculating linear regression coefficient, when the moments of observation form an arithmetic progression. It is proved that the variance of the trend estimation in this case decreases proportionally to the third degree of the length of the series of observations. If the estimation of a linear trend is based on several independent samples, the integral estimation of the trend is constructed and its variance is determined by special optimization procedure. This procedure is based on simple geometric consideration.

Keywords: linear regression coefficient, variance calculating, independent samples of observations.

1 Introduction

The problem of studying the variance of the linear trend estimation and its dependence on the length of the time series on which this estimate is based is of both theoretical and practical interest. This problem is closely related to the problem of small samples in mathematical statistics. In reliability theory, this problem arises when using linear regression analysis to predict the safety margin of a technical system (see, for example, [1], [2]). This task can be extended to the case when there are several time series, in particular for small-scale production. In this paper we choose such a particular formulation of this problem when the moments of observation form an arithmetic progression. It is proved that the variance of the trend estimation in this case decreases proportionally to the third degree of the length of the series of observations. This makes it possible to use short series of observations to estimate the linear trend. If the estimation of a linear trend is based on several independent samples, the integral estimation of the trend is constructed and its variance is determined.

2 The variance of the estimate of linear trend for a single series of observations

Consider the following linear regression model $x(t) = y(t) + \varepsilon(t),\ y(t) = at + b$. Assume that at times $t_1, \dots, t_n,\ 0 \le t_1 < t_2 < \cdots < t_N$, measured values are $y(t_1), \dots, y(t_N)$ with random errors $\varepsilon_1, \dots, \varepsilon_N$. The random variables $\varepsilon_1, \dots, \varepsilon_N$ are assumed to be independent, equally distributed with zero mean and variance σ^2.

To solve this problem, replace the variable $\tilde{t} = t - T_N,\ T_N = \frac{\sum_{k=1}^{N} t_k}{N}$, and define a linear function

$$\tilde{y}(t) = y(t+T_n) = at + b + aT_N = at + \tilde{b},\quad \sum_{k=1}^{N} \tilde{t}_k = 0,\quad \tilde{b} = b + aT_N.$$

To do this, we compute $\tilde{t}_k,\ k = 1, \dots, N$, and construct the least squares [3], [4] estimates of the coefficients $a,\ \tilde{b}$ of the linear regression function $\tilde{y}(t) = at + \tilde{b}$ from observations

$$x_1 = \tilde{y}(\tilde{t}_1) + \varepsilon_1, \dots, x_N = \tilde{y}(\tilde{t}_N) + \varepsilon_N.$$

The solution to this problem is a random vector consisting of estimates

$$\hat{a}_N = \frac{\sum_{k=1}^{N} x_k \tilde{t}_k}{\sum_{k=1}^{N} \tilde{t}_k^2},\ \hat{b}_N = \frac{\sum_{k=1}^{N} x_k}{N}$$

of coefficients $a,\ \tilde{b}$ of linear function $\tilde{y}(t)$. The components of this vector have the following averages, variances, and covariance coefficient:

$$M\hat{a}_N = a,\ M\hat{b}_N = \tilde{b},\ D\hat{a}_N = \frac{\sigma^2}{\sum_{k=1}^{N} \tilde{t}_k^2},\ D\hat{b}_N = \frac{\sigma^2}{N},\ cov(\hat{a}_N, \hat{b}_N) = 0. \tag{1}$$

Of greatest interest to us is the denominator $S(N) = \sum_{K=1}^{N} \tilde{t}_k^2$ in Formula (1) in the definition of the variance $D\hat{a}_N$. To simplify the calculations, assume that $\tilde{t}_{k+1} - \tilde{t}_k = 1, \dots N-1$. By induction at $n = 1,2, \dots$, it is easy to obtain equalities

$$S(2n+1) = \frac{2n^3}{3} + n^2 + \frac{n}{3},\ S(2n) = \frac{2n^3}{3} - \frac{n}{6}. \tag{2}$$

Indeed, the definition implies that $S(2n+1) = 2R(n),\ R(n) = \sum_{k=1}^{n} k^2$. Looking for $R(n)$ in the form $R(n) = a_0 + a_1 n + a_2 n^2 + a_3 n^3$. Then we have the equality $R(n+1) = R_n + (n+1)^2$ and so obtain the relation

$$a_0 + a_1 n + a_2 n^2 + a_3 n^3 + (n+1)^2 = a_0 + a_1(n+1) + a_2(n+1)^2 + a_3(n+1)^3.$$

Removing the parentheses and leading like that, we get the following equalities:

$$a_0 + 1 = a_0 + a_1 + a_2 + a_3,\ a_1 + 2 = a_1 + 2a_2 + 3a_3,\ a_2 + 1 = a_2 + 3a_3.$$

The solution of this system of linear algebraic equations is $a_1 = \frac{1}{6},\ a_2 = \frac{1}{2},\ a_3 = \frac{1}{3}$. Since $R(1) = 1$, then $a_0 = 0$. The first equality in Formula (2) is proved.

Now calculate $S(2n) = \frac{2Q(n)}{2^2}$, $Q(n) = 1^2 + 3^2 + \dots + (2(n-1)+1)^2$. Looking for $Q(n)$ in the form of $Q(n) = b_0 + b_1 n + b_2 n^2 + b_3 n^3$. Then the equalities $Q(n+1) = Q(n) + (2n+1)^2$ are true and the following equalities are fulfilled

$$b_0 + b_1 n + b_2 n^2 + b_3 n^3 + (2n+1)^2 = b_0 + b_1(n+1) + b_2(n+1)^2 + b_3(n+1)^3.$$

Removing the parentheses and leading like that, we get the following equality:

$$b_0 + 1 = b_0 + b_1 + b_2 + b_3,\ b_1 + 4 = b_1 + 2b_2 + 3b_3,\ b_2 + 4 = b_2 + 3b_3.$$

The solution to this system of linear algebraic equations is $b_1 = -\frac{1}{3},\ b_2 = 0,\ b_3 = \frac{4}{3}$. Because $Q(1) = 1$, then $b_0 = 0$ and means

$$b_0 = 0,\ b_1 = -\frac{1}{3},\ b_2 = 0,\ b_3 = \frac{4}{3}.$$

The second equality in the formula (2) is also proved.

Formula (2) leads to asymptotic relations:

$$S(2n) \sim S(2n+1) \sim \frac{2n^3}{3},\ n \to \infty. \tag{3}$$

However, in the applied plan the values $S(2n),\ S(2n+1)$ are of great interest for small values n.

We now give the results of numerical calculations in the following tables.

N	2	3	4	5	6	7	8	9	10	11	12	13
$S(N)$	0.5	2	5	10	17.5	28	42	60	82.5	110	143	182

N	14	15	16	17	18	19	20	21	22	23
$S(N)$	227.5	280	340	408	484.5	570	665	770	885.5	1012

Tables of $S(N)$

4 Estimation of linear regression coefficient on several independent samples

Assume that there are m independent samples $i = 1, \dots, m$. The sample i has volume N_i. The estimate $\hat{a}^{(i)}$ of linear regression coefficient a satisfies the equalities

$$\hat{a}^{(i)} = a,\ D\hat{a}^{(i)} = d_i = \frac{\sigma^2}{S(N_i)},\ i = 1, \dots, m. \tag{4}$$

We will look for an estimate of $\hat{a}$ of the linear regression coefficient a in the form

$$\hat{a} = \sum_{i=1}^{m} c_i \hat{a}^{(i)},\ \sum_{i=1}^{m} c_i = 1,\ D\hat{a} = \sum_{i=1}^{m} c_i^2 d_i. \tag{5}$$

Choice of coefficients $c_i,\ i = 1, \dots, M$, are produced from the minimum condition

$$\min\left(\sum_{i=1}^{m} c_i^2 d_i:\ \sum_{i=1}^{m} c_i = 1\right). \tag{6}$$

Make the change of variables $f_i = c_i\sqrt{d_i},\ i = 1, \dots, m$ and we can rewrite the optimization problem (6) in the form

$$\min\left(\sum_{i=1}^{m} f_i^2 = F:\ \sum_{i=1}^{m} \frac{f_i}{\sqrt{d_i}} = 1\right). \tag{7}$$

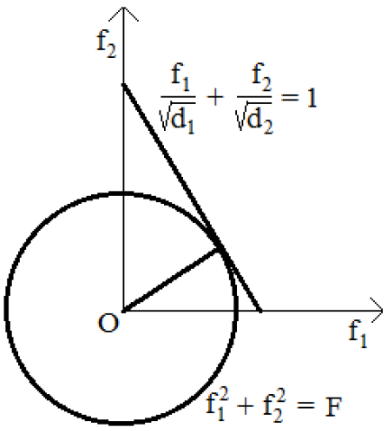

Fig 1. Geometric interpretation of the optimization problem.

From simple geometric considerations we obtain the following solution to the optimization problem (7): for $i = 1, \dots, m$

$$f_i = d_i^{-1/2}(\sum_{k=1}^m d_k^{-1})^{-1},\ c_i = d_i^{-1}(\sum_{k=1}^m d_k^{-1})^{-1},\ \sum_{i=1}^m c_i^2 d_i = (\sum_{k=1}^m d_k^{-1})^{-1}. \quad (8)$$

Thus, from Formulas (5), (8) we finally obtain:

$$\hat{a} = \sum_{i=1}^m d_i^{-1}(\sum_{k=1}^m d_k^{-1})^{-1}\hat{a}^{(i)} = \frac{\sum_{i=1}^m S(N_i)\hat{a}^{(i)}}{\sum_{k=1}^m S(N_k)},\ D\hat{a} = (\sum_{k=1}^m d_k^{-1})^{-1} = \frac{\sigma^2}{\sum_{k=1}^m S(N_k)}. \quad (9)$$

4 Conclusion

The results show that the coefficient of linear regression has a variance significantly lower than the variance of the free term. This makes it possible to raise the question of the evaluation of this coefficient separately from the evaluation of the free member. The resulting estimate can be used to predict relatively short series of observations.

The results of the estimation of the linear regression coefficient for several independent series of observations suggest that it is possible to estimate the linear trend coefficient fairly economically and accurately for a small group of series of observations. This result is obtained by simple geometric considerations that are based on the basic properties of the variance of a random variable.

5 Acknowledgements

The research of Gurami Tsitsiashvili is partially supported by Russian Foundation for Basic Research, project 17-07-00177 and by Presidium of FEB RAS, program "Priority research for the integrated development of the far Eastern branch of RAS", project 18-5-044.

References

[1] Abramov, O. V., Tsitsiashvili, G. Sh. (2018). Prediction of failure of the controlled technical system. Informatics and control systems, 3: 42–49.

[2] Abramov, O. V., Tsitsiashvili, G. Sh. (2019). Interval estimation for the task of predicting failures of complex engineering systems. (In Russian). Proceedings of the International Symposium "Reliability and quality", 1, 113–114. (In Russian).

[3] Rykov V. V. Mathematical statistics and experiment planning. MAKS Press, Moscow 2010. (In Russian).

[4] Borovkov A. A. Mathematical statistics and experiment planning. Additional chapters. Nauka, Moscow 1984. (In Russian).