Article

Linear regression as great joke of great scientists

L.V. Nedorezov
Research Center for Interdisciplinary Environmental Cooperation RAS, nab. Kutuzova 14, Saint-Petersburg, 191187, Russia
E-mail: l.v.nedorezov@gmail.com

Received 27 May 2017; Accepted 21 June 2017; Published 1 September 2017

Abstract
Analysis of problems of simplest variant of linear regression between two variables is presented. It was demonstrated that in classic situation problem of estimation of regression line parameters hasn’t a correct solution. It was also obtained that “paradox of two regressions” (Szekely, 1986) cannot be solved as a presentation of two intervals of (possible) changing of parameters of real regression line. Numerical examples allowed demonstrating that real parameters of regression line can be out of intervals defined by parameters of two regressions.

Keywords linear regression model; paradox of two regressions; ecological models.

Proceedings of the International Academy of Ecology and Environmental Sciences
ISSN 2220-8860
URL: http://www.iaees.org/publications/journals/piaees/online-version.asp
RSS: http://www.iaees.org/publications/journals/piaees/rss.xml
E-mail: piaees@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Statistics with its methods, objects, methodology and so on became a science many years ago when abacus played the role of strongest apparatus for mathematical calculations. But modern students (may be, all of them) may never heard about “abacus-computer” and how it is possible to use this “computer”. What is the reason to have useless information about abacus if we have modern computer? Moreover, at every time we can find in Internet useful and detailed instructions for calculations we want to realize. Sometimes in Internet we can find on-line services which allow providing required calculations without problems.

In most cases we try to follow instructions and recommendations described in books, textbooks etc., and at that moments we don’t think about “genetics of statistics” which exists in modern statistics from ancient “abacus’ years”. Every science has its own genetics, and statistics isn’t an exception from this rule. This “own genetics” can be observed in scientific books, textbooks, software… And we analyze datasets, we make qualitative and quantitative conclusions on the base of these calculations, and we don’t think about correctness of providing calculations.

Everybody can ask – is it bad condition for calculations if we use methods with “abacus’ genetics”? Honest answer is following – yes, in various situations use of these methods can lead to appearance of incorrect or false results.

2 Traditional Linear Regression and Comfortless Questions

Let $\{(x_k, y_k)\}$, $k = 1,...,N$, be an initial sample. N is a number of provided observations (sample size). For example, x_k can be a diameter of tree, and y_k is a height of the same tree. Respectively, k is a number of this tree. These amounts can also be heights and weights of separated individuals, lengths and weights of crocodiles and so on. Additionally, k can be (discrete) time, thus existing sample is changing in time of any parameters of one and the same object. At the beginning let's consider a case when k is a number of object.

2.1 First variant

Measurements cannot be provided without any errors; in various situations it is assumed that these stochastic errors have Normal distribution with zero average (Lakin, 1990; Bard, 1974; Draper and Smith, 1981; McCallum, 2000; Nedorezov, 2012).

Assumption about Normality of errors is a first serious problem of modern biometrics. First of all, Normal distribution is unbounded, and with positive probability we can get very big number with plus or minus. It is a good condition to ask – what is real condition of researcher if he or she can write a phrase "weight of larva is equal to minus one kilogram"? If we assume that errors have a Normal distribution with positive probability we can get error in several tons for weight of larva. From that point of view assumption about Normality does not look normal.

One more question is following: what is a reason to assume that errors have Normal distribution? One of typical answers is next: in nature various amounts have Normal distribution, we have Central Limit Theorem (Borovkov, 1984; Sevastianov, 1982) and so on. But this theorem doesn't allow concluding that all amounts in the nature have Normal distribution…

Let's assume that for selected intervals of changing of variables we can observe linear dependence:

$$y = ax + b, \qquad (1)$$

This assumption is our hypothesis, and we have to check it using existing sample. In other words, if we know errors of measurements (ε_k for y_k, and δ_k for x_k) we can determine/calculate real values of variables which (according to our hypothesis) belong to straight line. It is possible to point out parameters a and b, and following equation is truthful:

$$y_k - \varepsilon_k = a(x_k - \delta_k) + b.$$

It can be presented in other form:

$$y_k - ax_k - b = \varepsilon_k - a\delta_k. \qquad (2)$$

In left hand part of equation (2) there are two unknown amounts (parameters a and b), in right hand side there are three unknown amounts (it is true for every fixed value of k). Problem is following: find/estimate values of parameters a and b when distance between straight line and sample's points (x_k, y_k) has its minimum. But it looks like a disaster – on the plane (x, y) distance between two points doesn't determine. We cannot use Euclidean distance because in such a situation we'll need to summarize meters with kilograms. But where is an exit out of this blind alley? Many years ago following solution was recommended: to summarize squared expressions in left-hand side of equation (2) and to minimize obtained sum using respective values of parameters a and b:

$$Q(a,b) = \sum_{k=1}^{N} \left(y_k - ax_k - b\right)^2. \qquad (3)$$

But before finding a minimum of expression (3) we have to find acceptable explanation for next unobvious points: what is a reason to use squared expressions? What is real relation of expression (3) to considering biological problem? And what is a real sense of expression $y_k - ax_k - b$? Answer on two first questions is obvious: expression (3) has no relation to considering biological problem, and there are no objective and obvious reasons to use squared expressions. May be, sum of squared differences is most comfortable expression which allows (after primitive mathematical calculations) finding of minimum of functional form (2). But if we want to use other degrees in (3) (in most cases it will lead to use absolute values for differences $y_k - ax_k - b$) it can lead to serious problem in finding of minimum of functional form (3).

Answer on last question is rather obvious too: taking into account that $ax_k + b$ is value of linear function calculated in point x_k, then difference $y_k - ax_k - b$ is equal to distance between straight line and point (x_k, y_k) calculated along y ordinate line. In other words, it means that we postulate that values y_k of sample were obtained with (Normal) errors and values x_k were determined without errors. But it is not true because we know that both variables were estimated with errors.

Straight line (1) with parameters which gives minimum for (3) has name “regression of y on x” (Lakin, 1990). As it was pointed out above this straight line has no relation to real regression line we want to find because it is based on non-correct assumption that x_k were determined without errors. We can rename our variables and find one more straight line which is called “regression of x on y” (Lakin, 1990). This second line is based on assumptions that all values y_k were determined without errors, and x_k were obtained with (Normal) errors.

Finally, we have two straight (regression) lines which are based on non-correct assumptions. Existence of two regression lines got a name “paradox of two regressions” (Szekely, 1986). One of possible solutions of this paradox is following: we must point out intervals for parameters:

$$\min(a_1, a_2) \le a \le \max(a_1, a_2),$$

$$\min(b_1, b_2) \le b \le \max(b_1, b_2).$$

It was supposed that it will give us best solution of paradox. But it is good condition to ask these inequalities: who can give guarantees that pointed out inequalities will be truthful for all possible samples? And one more question: if both regression lines have no relation to biological problem who can give guarantees that pointed out inequalities will give us something which has any relation to problem?

Let’s consider one more important side of considering problem. Initially it was postulated that errors for both variables (2) have Normal distribution. It looks natural and obvious that after determination of values of parameters we must check respective hypotheses. We must check correspondence of error’s distribution to Normal, equivalence of average to zero, and independence of observed errors. If one of these hypotheses cannot be accepted than hypothesis about existence of linear dependence between variables must be rejected. The question is: can we do it or not (Fig. 1)?

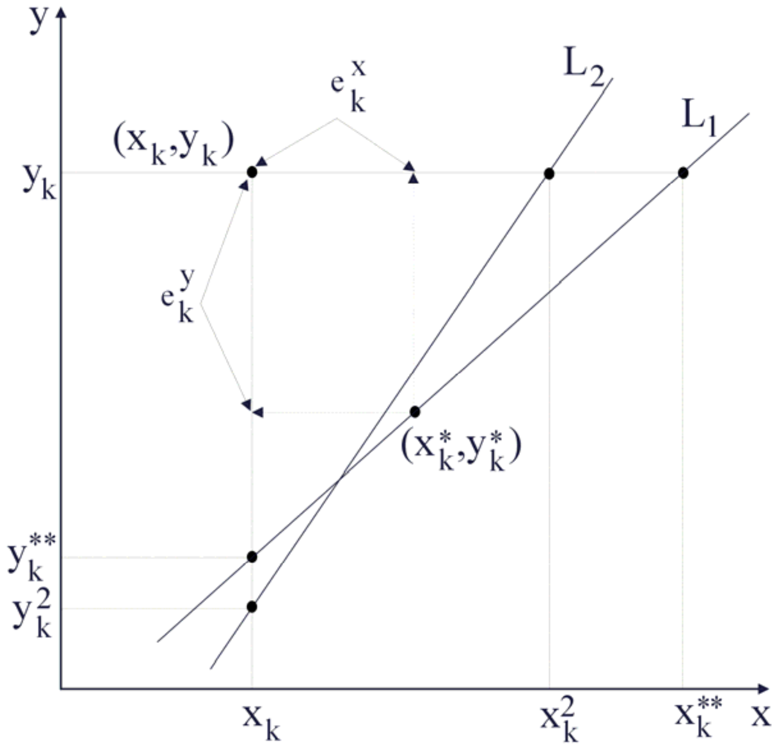

Fig. 1 Notations are presented in text.

Let’s assume that we know values of real regression line (L_1, fig. 1), and we know also real values (x_k^*, y_k^*) which belong straight line L_1. As it was assumed above these values were estimated with errors:

$y_k = y_k^* + e_k^y,\ x_k = x_k^* + e_k^x.$

Thus Normal distribution must be observed for linear combination (Fig. 1)

$\varepsilon_k = ae_k^x - e_k^y.$

On the other hand, for real initial sample we have no idea about points (x_k^*, y_k^*); if we know occupation of line L_1 we can calculate deviations $x_k^{**} - x_k$ and $y_k^{**} - y_k$ (fig. 1), but in general case these deviations have no correspondence with errors ε_k. Thus it looks rather strange if we decide to check normality of these deviations. Moreover, a’priori we have no ideas about occupations of points (x_k^*, y_k^*) and straight line L_1 too. We can determine, for example, occupation of line L_2 (regression of y on x) with parameters which give minimum for functional (3). Respectively, we can obtain two sets of deviations $x_k^2 - x_k$ and $y_k^2 - y_k$ (fig.1). But these deviations have no relations to postulates too. In a result of provided activities described above we have absurd situation: we postulate properties of concrete set of stochastic variables, but we have no possibilities to check postulated properties and have to check properties for deviations which have no relation to postulates…

Our simple objections allow us concluding following statement: problem of linear regression in the form described above haven’t correct solution. In this sense existing solution of problem of linear regression looks like a great joke of great scientists.

2.2 Second variant

Let’s consider other variant when k corresponds to time moments of providing measurements of variables x and y. In this case we have two correlated time series (for example, corresponding to changing of weight and height in time for one and the same organism, diameter and height of one and the same tree and so on). For every time series we have to formulate its own hypothesis. For example,

$$x(t) = a_x f(t) + b_x + \varepsilon^x(t). \qquad (4)$$

In (4) $f(t)$ can be a non-linear function (for example, if we are talking about biomass, or about diameter and height of tree this function can be a monotonic increasing S-type function). Amounts $\varepsilon^x(t)$ are independent stochastic variables with Normal distribution and zero average. Taking into account that time t are measured without errors then after estimation of values of parameters a_x and b_x (may be with using of functional (3)-type) we have good possibilities to check properties of following deviations:

$$\varepsilon_k^x = x_k - a_x f(t_k) - b_x.$$

Note, that properties of these deviations were postulated above.

If in a result of calculations we have to demonstrate that there is a linear dependence between observed variables, we have to check the following hypothesis for variable y:

$$y(t) = a_y f(t) + b_y + \varepsilon^y(t). \qquad (5)$$

If both sets of deviations have required properties, then we have a good background for conclusion about existence of linear dependence between variables x and y.

Finally, after consideration of two different variants we have to conclude that in the second case we can get correct solution of problem. But for finding a solution we have to have two additional hypotheses about changing of variables in time. And we have to check these hypotheses.

In first case correct solution cannot be obtain in principle. It is important to note that various biological problems can get correct solutions because in various situations together with basic variables researches determine age classes and/or time moments.

Remark. If it is assumed that between variables x and y we have allometric relation (Terskov and Terskova, 1980; Kofman, 1986):

$$y = ax^b,$$

then we have to check two following hypotheses:

$$x(t) = a_x f(t) + \varepsilon^x(t),$$

$$y(t) = a_y f^b(t) + \varepsilon^y(t).$$

Thus, for checking of any hypothesis (linear or non-linear) we have to have two or more hypotheses of *other level*. In other words, correct solution of regression problems requires existence of respective theory or at least existence of group of combined hypotheses. If we haven't good theory we haven't correct solution. Various statistical models without serious biological interpretations of parameters cannot save a situation.

2.3 Correct testing of hypotheses

Let's assume that our basic hypothesis is following: we have a linear dependence between variables x and y. Additionally we'll assume that values of these variables were estimated at time moments t_k with errors, $x_k = x(t_k)$, $y_k = y(t_k)$, $k = 1,...,N$. As it was noted above, before checking of basic hypothesis we have to check two other hypotheses (4) and (5). Checking of two pointed out hypotheses must contain following steps. In spaces of parameters (a_x, b_x) and (a_y, b_y) (let's note that these spaces of parameters are different) we have to find stochastic points and for every points we have to check properties of deviations $\{e_k^x = x_k - a_x f(t_k) - b_x\}$ and $\{e_k^y = y_k - a_y f(t_k) - b_y\}$. In 4-dimensional space point belongs to feasible set Ω^* if and only if following properties are truthful (for a'priori determined significance level, for example, for 5%):

1. Distribution densities of deviations are symmetric with respect to ordinate line. It can be provided with tests of homogeneity of two samples, for example, for $\{e_k^{x+}\}$ and $\{-e_k^{x-}\}$ where $\{e_k^{x+}\}$ are positive deviations and $\{-e_k^{x-}\}$ are negative deviations with sign minus. For checking of homogeneity of two samples Kolmogorov – Smirnov test, Lehmann – Rosenblatt test, Munn – Whitney test and othe criterions can be used (Kobzar, 2006; Likes and Laga, 1985; Bolshev and Smirnov, 1983).

Symmetry of distribution density means that errors of measurements to both sides must be observed with equal probabilities.

2. Branches of density functions for negative arguments (when $x < 0$ and $y < 0$) are monotonic increasing functions, and branches are monotonic decreasing functions for positive arguments. For checking of such behavior of branches of density functions Spearmen rank correlation coefficient can be used (Bolshev and Smirnov, 1983; Lakin, 1990; Nedorezov, 2015, 2016 a,b,c,d).

Let's consider a particular case for positive deviations $\{e_k^{x+}\}$ only, and let $\{e_{k,up}^{x+}\}$ be an ordered set of these deviations: $e_{1,up}^{x+} < e_{2,up}^{x+} < ... < e_{m,up}^{x+}$. m is a size of this sample. Monotonic decreasing of branch of density function for positive arguments means that thickness of points on a straight line decreases with increase of argument. In other words, in ideal case length of first interval $[0, e_{1,up}^{x+}]$ is less than length of second interval $[e_{1,up}^{x+}, e_{2,up}^{x+}]$ and so on. We can compare ideal variant with situation which is determined by existing sample. And we have to reject Null hypothesis $\rho = 0.5$ where ρ is Spearmen rank correlation coefficient, for fixed significance level and alternative hypothesis $\rho > 0.5$.

Monotonic decreasing of branch of density function for positive arguments means that bigger errors can be observed with smaller probabilities.

3. We have to have a background for conclusion that $\{e_k^x\}$ and $\{e_k^y\}$ are values of independent stochastic variables. If it is not true we can say that our model (in considering situation it is $af(t) + b$) is not suitable for fitting of time series. In such a situation we have to modify model, or construct a new one.

For checking of respective hypotheses we can use, for example, Swed – Eizenhart test, test of "jumps up – jumps down" (Bard, 1974; Draper and Smith, 1981; Likes and Laga, 1985; Hettmansperger, 1987; Hollander and Wolfe, 1973).

Remark. It is very important to note that feasible set Ω^* can be considered as *confidence domain*:

For every element of this set deviations satisfy to all selected statistical criterions, and we have no reasons for saying that model isn't suitable for fitting of time series.

3 Numerical Example

Let's consider (artificial) numerical example for checking of described above solution of "paradox of two regressions". Let $a = 1$ and $b = 0$. Initial sample (without additive stochastic deviations) is following: $x_1 = 1.0$, $x_2 = 1.05$,..., $x_{20} = 1.95$, $x_{21} = 2.0$. In table 1 there are initial samples with Normal errors (with zero average and various values of σ). "Errors of measurements" were modeled with standard Excel random generator.

3.1 Case 1

For first column (table 1) we have following results: $\bar{x} = 0.00876$, standard error is equal to 0.0222; for Shapiro – Wilk test we have $p - value = 0.4126$, for Anderson – Darling test we have $p - value = 0.3779$, for Cramer – von Mises test $p - value = 0.3872$, for Lilliefors test $p - value = 0.4159$, for chi-squared test $p - value = 0.1991$, and for Shapiro – Francia test $p - value = 0.4538$ (Shapiro, Wilk, 1965; Anderson, Darling, 1952, 1954; Lilliefors, 1967, 1969; Thode, 2002; Vasiliev, Melnikova, 2009; Shapiro, Francia, 1972). Thus, for first column we have no reasons for rejecting Null hypothesis about Normality of deviations.

For second column we have: $\bar{y} = 0.010905$, standard error is equal to 0.0183; for Shapiro – Wilk test $p - value = 0.9978$, for Anderson – Darling test $p - value = 0.9723$, for Cramer – von Mises test $p - value = 0.9523$, for Lilliefors test $p - value = 0.9488$, for chi-squared test $p - value = 0.9554$, and for Shapiro – Francia test $p - value = 0.9871$. Thus, in this situation we have to accept Null hypothesis because amounts of $p - value$ are very big.

Table 1 Datasets for testing example.

$\sigma = 0.1$		$\sigma = 0.5$		$\sigma = 1$	
x	y	x	y	x	y
1.021	1.034	0.064	1.113	1.086	1.852
1.157	0.96	0.249	0.274	0.421	0.632
1.059	1.113	1.633	1.07	0.908	1.649
0.995	0.984	1.482	1.259	0.475	1.39
1.139	1.169	1.195	1.678	1.848	0.922
1.16	1.25	0.591	1.414	1.033	0.658
1.307	1.294	0.836	0.866	2.108	1.435
1.248	1.277	1.089	1.469	1.441	0.926
1.304	1.482	0.763	0.962	0.926	2.135
1.385	1.431	1.168	1.927	1.367	1.649
1.689	1.506	1.427	1.023	0.779	1.081
1.65	1.65	2.233	1.589	1.257	1.33
1.681	1.5	1.915	1.787	2.13	1.064
1.702	1.71	2.017	0.859	1.638	2.42
1.694	1.836	0.966	1.518	1.386	1.475
1.75	1.801	1.15	1.971	1.577	0.844
1.902	1.901	2.896	1.392	2.245	1.347
2.069	1.775	1.435	1.509	1.497	1.49
1.989	1.937	1.864	2.453	1.562	1.446
1.885	2.124	2.049	2.936	1.121	1.541
1.898	1.995	3.519	1.775	2.276	0.921

Linear regression of y on x is determined by the equation:

$$y = 0.9464x + 0.083,\ R^2 = 0.8718. \qquad (6)$$

Average of deviations from this straight line (calculated along ordinate line y) is equal to $5.66 \cdot 10^{-16}$. Standard error is equal to 0.02775. For Shapiro – Wilk test for these deviations $p - value = 0.2851$, for Anderson – Darling test $p - value = 0.08987$, for Cramer – von Mises test $p - value = 0.0524$, for Lilliefors test $p - value = 0.06144$, for chi-squared test $p - value = 0.09158$, and for Shapiro – Francia test $p - value = 0.2035$. Thus, for 5% significance level hypothesis about Normality of deviations from regression line (6) cannot be rejected by all used tests. But for several criterions amount of $p - value$ is very close to threshold value. For three criterions Null hypothesis must be rejected with 10% significance level.

Comparison of known deviations (for second column of table 1) with deviations from regression line (6) shows that for Munn – Whitney test $p - value = 0.9603$, for Kolmogorov – Smirnov test $p - value = 0.8531$. We have amazing picture: we haven't a background for conclusion that two sets of deviations have different distribution functions. On the other hand, distribution of known deviations is close to Normal, and we have no reasons to say the same about deviations from line (6).

Linear regression of x on y is determined by the equation:

$$y = 1.0855x - 0.1269. \qquad (7)$$

Amount of R^2 is rather big, $R^2 = 0.8718$. Two linear equations (6) and (7) allow us presenting following inequalities:

$0.9464 < a < 1.0855$, $-0.1269 < b < 0.083$.

These inequalities are truthful for real values of regression line.

Average of deviations from this straight line (calculated along ordinate line x) is equal to 3.54·10^{-16}. Standard error is equal to 0.0274. For Shapiro – Wilk test for these deviations $p-value = 0.1221$, for Anderson – Darling test $p-value = 0.07642$, for Cramer – von Mises test $p-value = 0.05116$, for Lilliefors test $p-value = 0.1046$, for chi-squared test $p-value = 0.06999$, and for Shapiro – Francia test $p-value = 0.09731$. Thus, with 5% significance level hypothesis about Normality for these deviations from regression line (7) cannot be rejected. But for several criterions amount of $p-value$ is very close to threshold value. For four criterions Null hypothesis must be rejected with 10% significance level.

Comparison of known deviations (for first column of table 1) with deviations from regression line (7) shows that for Munn – Whitney test $p-value = 0.7059$, for Kolmogorov – Smirnov test $p-value = 0.9829$. We can observe amazing picture again: we haven't a background for conclusion that two sets of deviations have different distribution functions. On the other hand, distribution of known deviations is close to Normal, and we have no reasons to say the same about deviations from line (7).

3.2 Case 2

For deviations of third column (table 1) we have $\overline{x} = -0.0457$. Standard error is equal to 0.1411. For Shapiro – Wilk test for these deviations $p-value = 0.2827$, for Anderson – Darling test $p-value = 0.4308$, for Cramer – von Mises test $p-value = 0.5223$, for Lilliefors test $p-value = 0.7711$, for chi-squared test $p-value = 0.406$, and for Shapiro – Francia test $p-value = 0.2747$.

For deviations of fourth column (table 1) we have $\overline{y} = -0.03124$. Standard error is equal to 0.09841. For Shapiro – Wilk test for these deviations $p-value = 0.7549$, for Anderson – Darling test $p-value = 0.6691$, for Cramer – von Mises test $p-value = 0.6433$, for Lilliefors test $p-value = 0.8604$, for chi-squared test $p-value = 0.06999$, and for Shapiro – Francia test $p-value = 0.6855$. Thus, we have no reasons for rejecting Null hypotheses for both columns with 5% significance level.

Linear regression of y on x (for fourth and third columns of table 1) is determined by following equation:

$$y = 0.2912x + 1.0452,\ R^2 = 0.1701. \qquad (8)$$

Average of deviations from this straight line is equal to 4.1237·10^{-16}. Standard error is equal to 0.1159. For Shapiro – Wilk test for these deviations $p-value = 0.595$, for Anderson – Darling test $p-value = 0.598$, for Cramer – von Mises test $p-value = 0.6162$, for Lilliefors test $p-value = 0.6309$, for chi-squared test $p-value = 0.3232$, and for Shapiro – Francia test $p-value = 0.4795$. Thus, even with 32% significance level Null hypothesis for deviations from line (8) cannot be rejected (it is observed for all used tests).

Comparison of known deviations (for fourth column of table 1) with deviations from regression line (8) shows that for Munn – Whitney test $p-value = 0.9405$, for Kolmogorov – Smirnov test $p-value = 1$. We haven't a background for conclusion that two sets of deviations have different distribution functions. Moreover, we have to accept Null hypothesis.

Linear regression of x on y is determined by following equation:

$$y = 1.71164x - 1.02053. \qquad (9)$$

From two linear equations (8) and (9) we get following inequalities for parameters of real regression line (1):
$0.2912 < a < 1.71164$, $-1.02053 < b < 1.0452$.
Note, these inequalities are truthful.

Average for deviations from line (9) is equal to $1.269 \cdot 10^{-16}$. Standard error is equal to 0.1641. For Shapiro – Wilk test for these deviations $p-value = 0.08182$, for Anderson – Darling test $p-value = 0.05871$, for Cramer – von Mises test $p-value = 0.05521$, for Lilliefors test $p-value = 0.1708$, for chi-squared test $p-value = 0.1546$, and for Shapiro – Francia test $p-value = 0.06298$. Thus, with 5% significance level all used criterions don't allow rejecting Null hypothesis about Normality of deviations from regression line (9). But for four criterions Null hypothesis must be rejected with 10% significance level.

Comparison of known deviations (for third column of table 1) with deviations from regression line (9) shows that for Munn – Whitney test $p-value = 0.901$, for Kolmogorov – Smirnov test $p-value = 0.8531$. We can observe amazing picture again: we haven't a background for conclusion that two sets of deviations have different distribution functions. On the other hand, distribution of known deviations is close to Normal, and we have no reasons to say the same about deviations from line (9).

3.3 Case 3

For deviations of fifth column (table 1) we have $\bar{x} = -0.11519$. Standard error is equal to 0.09991. For Shapiro – Wilk test for these deviations $p-value = 0.5641$, for Anderson – Darling test $p-value = 0.5847$, for Cramer – von Mises test $p-value = 0.5607$, for Lilliefors test $p-value = 0.5753$, for chi-squared test $p-value = 0.3232$, and for Shapiro – Francia test $p-value = 0.6638$.

For deviations of sixth column (table 1) we have $\bar{y} = -0.15681$. Standard error is equal to 0.11748; For Shapiro – Wilk test for these deviations $p-value = 0.0722$, for Anderson – Darling test $p-value = 0.02441$, for Cramer – von Mises test $p-value = 0.01498$, for Lilliefors test $p-value = 0.0135$, for chi-squared test $p-value = 0.02306$, and for Shapiro – Francia test $p-value = 0.08726$. For deviations of fifth column we have no background for rejection of Null hypothesis even with 32% significance level. For deviations of sixth column with 10% significance level we have to reject Null hypothesis for all used criterions. Four criterions allow rejecting Null hypothesis with 5% significance level.

Linear regression of y on x (for sixth and fifth columns) is determined by the equation:

$$y = -0.0336x + 1.3897, \; R^2 = 0.0016. \qquad (10)$$

Average for deviations from this regression line (10) is equal to $-6.1855 \cdot 10^{-16}$. Standard error is equal to 0.10031. For Shapiro – Wilk test for these deviations $p-value = 0.5716$, for Anderson – Darling test $p-value = 0.5084$, for Cramer – von Mises test $p-value = 0.4644$, for Lilliefors test $p-value = 0.7223$, for chi-squared test $p-value = 0.5037$, and for Shapiro – Francia test $p-value = 0.4739$. Thus, even with 46% significance level Null hypothesis cannot be rejected.

Comparison of known deviations (for sixth column of table 1) with deviations from regression line (10) shows that for Munn – Whitney test $p-value = 0.2606$, for Kolmogorov – Smirnov test $p-value = 0.365$. Thus, we haven't a background for conclusion that two sets of deviations have different distribution functions.

Linear regression of x on y is determined by the equation:

$$y = -21.6013x + 31.25689. \qquad (11)$$

Two regression lines (10) and (11) allow us to point out next inequalities:
$-0.0336, -21.6013 < a$, $b < 31.25589, 1.3897$.

Consequently, we can conclude that “paradox of two regressions” (Szekely, 1986) cannot be solved as a presentation of two intervals of (possible) changing of parameters of real regression line. As we can see both parameters are out of intervals determined by parameters of regression lines (10) and (11).

Average for deviations from this regression line (11) is equal to $1.0574 \cdot 10^{-17}$. Standard error is equal to 0.11779. For Shapiro – Wilk test for these deviations $p-value=0.6784$, for Anderson – Darling test $p-value=0.8613$, for Cramer – von Mises test $p-value=0.9311$, for Lilliefors test $p-value=0.8268$, for chi-squared test $p-value=0.7358$, and for Shapiro – Francia test $p-value=0.8335$. Thus, even with 73% significance level Null hypothesis cannot be rejected.

Comparison of known deviations (for fifth column of Table 1) with deviations from regression line (11) shows that for Munn – Whitney test $p-value=0.3963$, for Kolmogorov – Smirnov test $p-value=0.6028$. Thus, we haven’t a background for conclusion that two sets of deviations have different distribution functions.

4 Conclusion

Now we can summarize results described above. In a result of analysis of paired observations researches obtain two straight lines which are based on unreal assumptions. Respectively, these straight lines have no relation to considering problem, and obtained estimations of parameters of these lines have no relation to problem too. But researches use these results (or results for one of regression lines) and present quantitative and qualitative conclusions about dependence of variables. May be, it is a good time to reconsider results of provided investigations and re-analyze existing datasets.

Analysis of numerical examples shows that “paradox of two regressions” (Szekely, 1986) cannot be solved as a presentation of two intervals of (possible) changing of parameters of real regression line. In Example 1 we got that real parameters of regression line can be out of intervals defined by parameters of two regressions.

References

Anderson TW, Darling DA. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics, 23: 193-212

Anderson TW, Darling DA. 1954. A test of goodness-of-fit. Journal of American Statistical Association, 49: 765-769

Bard Y. 1974. Nonlinear Parameter Estimation. Academic Press Inc, New York, USA

Bolshev LN, Smirnov NV. 1983. Tables of Mathematical Statistics. Nauka, Moscow, Russia

Borovkov AA. 1984. Mathematical Statistics. Nauka, Moscow, USSR

Draper NR, Smith H. 1981. Applied Regression Analysis. Wiley and Sons Inc, New York, USA

Hettmansperger T. 1987. Statistical outputs based on ranks. Finance and Statistics, Moscow, USSR

Hollander MDA, Wolfe DA. 1973. Nonparametric Statistical Methods. John Wiley and Sons Inc, USA

Kobzar AI. 2006. Applied Mathematical Statistics. For engineers and scientists. Fizmatlit, Moscow, Russia

Kofman GB. 1986. Growth and Form of Trees. Nauka, Novosibirsk, USSR

Lakin GF. 1990. Biometrics. Visshaya Shkola, Moscow, USSR

Likes J, Laga J. 1985. Basic Tables of Mathematical Statistics. Finance and Statistics, Moscow, Russia

Lilliefors H. 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. Journal of American Statistical Association, 62(318): 399-402

Lilliefors H. 1969. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of American Statistical Association, 64(325): 387-389

McCallum H. 2000. Population parameters estimation for ecological models. Blackwell Sciences Ltd., Brisbane, Australia

Nedorezov LV. 2012. Chaos and Order in Population Dynamics: Modeling, Analysis, Forecast. LAP Lambert Academic Publishing, Saarbrucken, USA

Nedorezov LV. 2016a. Shine and poverty of ordinary least squares. Samarskaya Luka: Problems of Regional and Global Ecology, 25(1): 13-17

Nedorezov LV. 2016b. Dynamics of a “Lynx-hare” system: An application of the Lotka–Volterra model. Biophysics, 61(1): 149-154

Nedorezov LV. 2016c. Application of nonlinear model of population dynamics with phase structure to analysis of pine looper moth time series. Proceedings of the International Academy of Ecology and Environmental Sciences 6(1): 1-12

Nedorezov LV. 2016d. Non-traditional approach to fitting of time series of larch bud moth dynamics: Application of Moran - Ricker model with time lags. Selforganizology, 3(1): 25-40

Nedorezov LV. 2015. Application of generalized discrete logistic model for fitting of pine looper moth time series: Feasible sets and estimations of model parameters. Proceedings of the International Academy of Ecology and Environmental Sciences 5(1): 38-48

Sevastianov BA. 1982. Course of Probability Theory and Mathematical Statistics. Nauka, Moscow, USSR

Shapiro SS, Francia RS. 1972. An approximate analysis of variance test for normality. Journal of American Statistical Association, 67: 215–216

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality. Biometrika, 52(3): 591-611.

Szekely GJ. 1986. Paradoxes in Probability Theory and Mathematical Statistics. Akademiai Kiado, Budapest, Hungary

Terskov IA, Terskova MI. 1980. Growth of Regular High Forests. Nauka, Novosibirsk, USSR

Thode JrHC, 2002. Testing for Normality. Marcel Dekker, New York, USA

Vasiliev AV, Melnikova IN. 2009. Methods of Applied Analysis of Field Measurements in Environment. Baltic State Technical University, Saint-Petersburg, Russia