About bi-lines of regression

Grzegorz SITEK

Abstract. We discuss the problem of estimation of the bi-linear regression parameters using the least squares method for an implicit interdependence. In general, values of parameter estimators are evaluated by means of an appropriate numerical approximation method. In a particular case it was possible to derive exact expressions for the parameter estimators.

Keywords: bi-lines, bi-linear regression function, least squares method for an implicit interdependence.

2010 Mathematics Subject Classification: 60E05, 62E99.

1. Introduction

The classic simple regression does not represent a two-dimensional population, if it is a bimodal population. Antoniewicz in [1] proposed to approximate probability distribution of a one-dimensional random variable by means of two points. Generalizing this result we decided to approximate a two-dimensional distribution by means of two lines, which we call **bi-lines**. Parameters of the bi-lines are estimated on the basis of the least squares method for an implicit interdependence introduced by Antoniewicz in [1]. Moreover, the bi-line estimator can be applied to estimation of classical bi-linear regression treated as conditional expected value of mixture of two two-dimensional probability distributions.

G. Sitek
Department of Statistics, Econometrics and Mathematics Management Faculty, University of Economics in Katowice, e-mail: grzegorz.sitek@ue.katowice.pl

R. Wituła, B. Bajorska-Harapińska, E. Hetmaniok, D. Słota, T. Trawiński (eds.), *Selected Problems on Experimental Mathematics*. Wydawnictwo Politechniki Śląskiej, Gliwice 2017, pp. 289–297.

2. The least squares method for an implicit interdependence

Two lines, none of which is parallel to the axis of the coordinate system, are described by the equation

$$(y - ax - b)(y - cx - d) = 0. \tag{1}$$

Basing on the available data, we wish to estimate the parameters a, b, c and d. This is equivalent to finding the straight lines that gives the best fit (representation) of the points in the scatter plot of the response versus the predictor variable. We estimate the parameters using the popular least squares method, which gives the lines that minimizes the sum of squares of the vertical distances from each point to the lines. The vertical distances represent the errors in the response variable.

The sum of squares of these distances can then be written as

$$S(a, b, c, d) = \sum_{i=1}^{n} [(y_i - a \cdot x_i - b)(y_i - c \cdot x_i - d)\,]^2. \tag{2}$$

The values of $\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{d}$ that minimize $S(a, b, c, d)$ are solutions to the system of nonlinear equations (see [1])

$$\begin{cases} ac^2m_{40} + ad^2m_{20} + bc^2m_{30} + bd^2m_{10} + 2bcdm_{20} + 2acdm_{30} - c^2m_{31} \\ -d^2m_{11} - 2acm_{31} - 2bcm_{21} - 2cdm_{21} - 2bdm_{11} - 2adm_{21} \\ +am_{22} + 2cm_{22} + 2dm_{12} + bm_{12} - m_{12} = 0, \\ bc^2m_{20} + bd^2 + ac^2m_{30} + ad^2m_{10} + 2acdm_{20} + 2bcdm_{10} - 2adm_{11} \\ -2bcm_{11} - 2cdm_{11} - 2bdm_{01} - d^2m_{01} - c^2m_{21} \\ -2acm_{21} + bm_{02} + 2dm_{02} + 2cm_{12} + am_{12} - m_{13} = 0, \\ a^2cm_{40} + cb^2m_{20} + da^2m_{30} + db^2m_{10} + 2abdm_{20} + 2abcm_{30} - 2acm_{31} \\ -d^2m_{11} - 2acm_{31} - 2abm_{21} - 2adm_{21} - b^2m_{11} - 2bdm_{11} - a^2m_{31} \\ -2bcm_{21} + 2am_{22} + cm_{22} + 2bm_{12} + dm_{13} - m_{13} = 0, \\ da^2m_{40} + b^2d + a^2cm_{30} + cb^2m_{10} + 2acbm_{20} + 2badm_{10} - 2acm_{21} \\ -2abm_{11} - 2adm_{11} - 2bcm_{11} - 2bdm_{01} - a^2m_{21} - b^2m_{01} + 2am_{12} \\ +2bm_{02} + cm_{12} + dm_{02} - m_{03} = 0. \end{cases} \tag{3}$$

where $m_{uv} = \sum_{i=1}^{n} x_i^u y_i^v$ and $c_{uv} = \sum_{i=1}^{n} (x_i - m_{10})^u (x_i - m_{01})^v$, $u = 0, 1, 2, \ldots$, $v = 0, 1, 2, \ldots$. The problem of parameters estimation was presented by the Author in [3]. This system can be solved only numerically. However, if $c = d = 0$, then (3) has an algebraic solution. Classic simple regression does not represent a population of two-dimensional, if it is a bimodal population. Then we approximate a two-dimensional distribution by means of two lines, which we called bi-linear regression. In this paper we will called Antoniewicz's model simply as bi-lines function, because in general it leads to approximation two-dimensional data spread by means of two lines. We use the least squares method for an implicit interdependence applying the function `optim` in R.

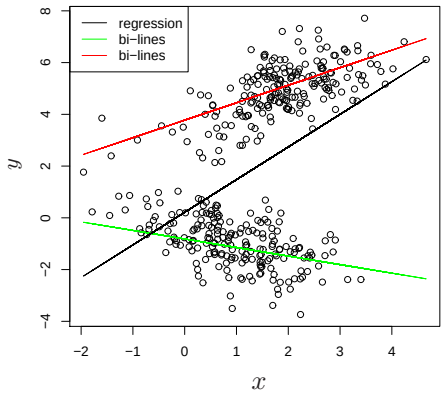

Fig. 1. Regression and bi-lines of bimodal population

We shall consider a special case, namely we assume that the two lines in question contain the origin. Then formulas (1)-(3) simplify to the following ones:

$$(y - ax)(y - bx) = 0. \tag{4}$$

$$S(a, b) = \sum_{i=1}^{n} [(y_i - ax_i)(y_i - bx_i)]^2. \tag{5}$$

$$\begin{cases} m_{13} - 2bm_{22} + b^2 m_{31} - am_{22} + 2abm_{31} - ab^2 m_{40} = 0, \\ m_{13} - 2am_{22} + a^2 m_{31} - bm_{22} + 2abm_{31} - a^2 bm_{40} = 0. \end{cases} \tag{6}$$

Under the assumption that $a \neq b$, after appropriate transformations from (6) we have:

$$-(b - a)m_{22} + (b^2 - a^2)m_{31} - ab(b - a)m_{40} = 0/(b - a),$$

$$a = \frac{m_{22} - bm_{31}}{m_{31} - bm_{40}}. \tag{7}$$

Substituting (7) to the first equation in (6) yields the following quadratic equation:

$$b^2(m_{40}m_{22} - m_{31}^2) + b(m_{31}m_{22} - m_{13}m_{40}) + m_{31}m_{13} - m_{22}^2 = 0. \tag{8}$$

Next, we have

$$\Delta = (m_{31}m_{22} - m_{13}m_{40})^2 - 4(m_{40}m_{22} - m_{31}^2)(m_{31}m_{13} - m_{22}^2),$$

$$A = m_{40}m_{22} - m_{31}^2,$$

$$B = m_{31}m_{22} - m_{13}m_{40},$$

$$C = m_{31}m_{13} - m_{22}^2.$$

If $\Delta > 0$ then there are two distinct roots

$$\hat{b}_1 = \frac{-B - \sqrt{\Delta}}{2A}, \tag{9}$$

$$\hat{b}_2 = \frac{-B + \sqrt{\Delta}}{2A}. \tag{10}$$

Substituting the results to the equation (7) gives

$$\hat{a}_1 = \frac{m_{22} - \hat{b}_1 m_{31}}{m_{31} - \hat{b}_1 m_{40}}.$$

It can be shown that $\hat{a}_1 = \hat{b}_2$. Indeed, if we assume

$$\frac{m_{22} - \hat{b}_1 m_{31}}{m_{31} - \hat{b}_1 m_{40}} = \hat{b}_2,$$

then

$$m_{22} - \hat{b}_1 m_{31} = \hat{b}_2 \cdot (m_{31} - \hat{b}_1 m_{40}),$$
$$m_{22} - \hat{b}_1 m_{31} = \hat{b}_2 \cdot m_{31} - \hat{b}_2 \hat{b}_1 m_{40},$$
$$m_{22} + \hat{b}_2 \hat{b}_1 m_{40} = (\hat{b}_1 + \hat{b}_2) \cdot m_{31}.$$

and from Viete's formulas we have:

$$\hat{b}_1 + \hat{b}_2 = \frac{-B}{A},$$

$$\hat{b}_1 \cdot \hat{b}_2 = \frac{C}{A},$$

$$m_{22} - \frac{C}{A} m_{40} = \frac{-B}{A} \cdot m_{31},$$

$$m_{22} A - C m_{40} = -B \cdot m_{31},$$

$$m_{40} m_{22}^2 - m_{22} m_{31}^2 + m_{40} \cdot m_{31} m_{13} - m_{40} m_{22}^2 = m_{40} \cdot m_{31} m_{13} - m_{22} m_{31}^2,$$

$$m_{40} \cdot m_{31} m_{13} - m_{22} m_{31}^2 = m_{40} \cdot m_{31} m_{13} - m_{22} m_{31}^2.$$

which is always true, so we have proved that $\hat{a}_1 = \hat{b}_2$.

3. Examples

The above results will be considered in the particular case of two-dimensional random variables. Moreover, distributions are defined as mixtures of two-dimension normal distributions with appropriate parameters. The considered cases are as follows:

Example 3.1. We consider the bivariate normal distribution: $N(0,0,1,1,r)$. We set needed the following moments

$$m_{40}=3,\quad m_{31}=m_{13}=3r,\quad m_{22}=1+2r^2,$$

$$\Delta=12(1-r^2)^3,$$

$$A=3(1-r^2),\quad B=-6r(1-r^2),$$

$$b_1=\frac{-B-\sqrt{\Delta}}{2A}=r-\frac{\sqrt{12(1-r^2)^3}}{6(1-r^2)}=r-\frac{\sqrt{3(1-r^2)}}{3},$$

$$b_2=\frac{-B+\sqrt{\Delta}}{2A}=r+\frac{\sqrt{3(1-r^2)}}{3}.$$

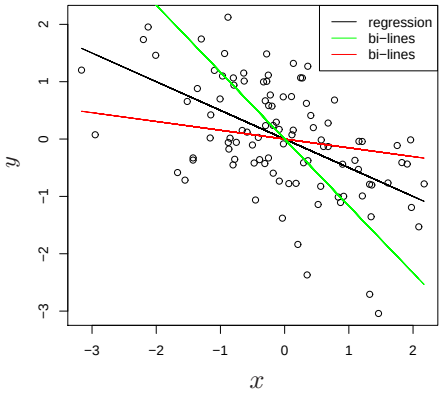

Fig. 2. Regression and bi-lines bivariate normal distribution, $r=-0.5$

Example 3.2. Bivariate normal distribution $N(0,0,\sigma_1,\sigma_2,r)$. We set needed moments

$$m_{40}=3\sigma_1^4,\quad m_{31}=3r\sigma_1^3\sigma_2,\quad m_{13}=3r\sigma_2^3\sigma_1,\quad m_{22}=\sigma_1^2\sigma_2^2(1+2r^2),$$

$$\Delta=12(1-r^2)^3\sigma_1^{10}\sigma_2^6,$$

$$A=3(1-r^2)\sigma_1^6\sigma_2^2,\quad B=-6r(1-r^2)\sigma_1^5\sigma_2^3,$$

$$b_1=\frac{-B-\sqrt{\Delta}}{2A}=\frac{\sigma_2}{\sigma_1}r-\frac{\sigma_1^5\sigma_2^3\sqrt{12(1-r^2)^3}}{6(1-r^2)\sigma_1^6\sigma_2^2}=\frac{\sigma_2}{\sigma_1}r-\frac{\sigma_2\sqrt{3(1-r^2)}}{3\sigma_1},$$

$$b_2=\frac{-B+\sqrt{\Delta}}{2A}=\frac{\sigma_2}{\sigma_1}r+\frac{\sigma_2\sqrt{3(1-r^2)}}{3\sigma_1}.$$

Finally, we get the following slope coefficients of the lines:

$$b_i=\frac{\sigma_2}{\sigma_1}r\mp\sigma_2\sqrt{1-r^2}\sqrt{\frac{\sigma_1^2}{3\sigma_1^3}}=\frac{\sigma_2}{\sigma_1}r\mp\sigma_{Y|X}\sqrt{\frac{m_2}{m_4}}.$$

Example 3.3. Bivariate t-Student distribution $[\{1,r\},\{r,1\},v]$.
We set needed moments:

$$m_{40}=\frac{3v^2}{4\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},\quad m_{31}=m_{13}=\frac{3rv^2}{4\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},$$

$$m_{22}=\frac{\left(16r^2+8\right)v^2}{32\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},\ v>4,$$

$$\Delta=\frac{12v^8(1-r^2)^3}{(v-4)^4(v-2)^4},$$

$$A=-\frac{3\left(r^2-1\right)v^4}{(v-4)^2(v-2)^2},\quad B=\frac{6r\left(r^2-1\right)v^4}{(v-4)^2(v-2)^2},$$

$$b_1=\frac{-B-\sqrt{\Delta}}{2A}=r-\frac{\sqrt{12(1-r^2)^3}}{6(1-r^2)}=r-\frac{\sqrt{3(1-r^2)}}{3},$$

$$b_2=\frac{-B+\sqrt{\Delta}}{2A}=r+\frac{\sqrt{3(1-r^2)}}{3}.$$

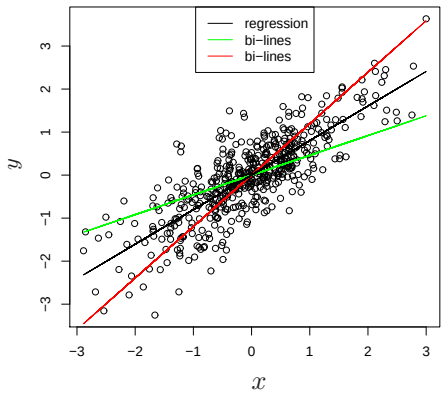

Fig. 3. Regression and bi-lines bivariate t-Student distribution, $r=0.8$

Example 3.4. Bivariate t-Student distribution $[\{\sigma_1,r\},\{r,\sigma_2\},v]$.
We set needed moments

$$m_{40}=\frac{3\sigma_1^2v^2}{4\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},\ m_{31}=\frac{3r\sigma_1v^2}{4\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},\ m_{13}=\frac{3r\sigma_2v^2}{4\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},$$

$$m_{22}=\frac{v^2\left(16r^2+8\sigma_1\sigma_2\right)}{32\left(1-\frac{v}{2}\right)\left(2-\frac{v}{2}\right)},\quad v>4,$$

$$b_i=\frac{r}{\sigma_1}\mp\frac{1}{3}\sqrt{3}\sqrt{\frac{\sigma_1\sigma_2-r^2}{\sigma_1^2}},\quad i=1,2.$$

Example 3.5. Mixture of bivariate distribution. Let the distribution be defined by the following mixture of bivariate normal distribution:

$$pN(0,0,1,1,r_1) + (1-p)N(0,0,1,1,r_2).$$

In this case the moments are as follows:

$$m_{40} = 3,\ m_{31} = m_{13} = 3pr_1 + 3(1-p)r_2,\ \ m_{22} = p(1+2r_1^2) + (1-p)(1+2r_2^2),$$

$$\begin{aligned}\Delta =& 36(pr_1 + (1-p)r_2)^2(1 - pr_1^2 + (p-1)r_2^2)^2 \\ & - 4\left(3 + 6pr_1^2 + (1-p)r_2^2 - 9(pr_1 + (1-p)r_2)^2\right) \times \\ & \times \left(9(pr_1 + (1-p)r_2)^2 - (1 + 2pr_1^2 + 2(1-p)r_2^2)^2\right),\end{aligned}$$

$$A = 3 + 6pr_1^2 + (1-p)r_2^2 - 9(pr_1 + (1-p)r_2)^2,$$

$$B = 6(pr_1 + (1-p)r_2)(pr_1^2 + (1-p)r_2^2 - 1),$$

$$b_1 = \frac{-B - \sqrt{\Delta}}{2A}, b_2 = \frac{-B + \sqrt{\Delta}}{2A}.$$

Let us consider the following mixture of normal distribution: $0.5N(0,0,1,1,0.8) + 0.5N(0,0,1,1,-0.8)$.

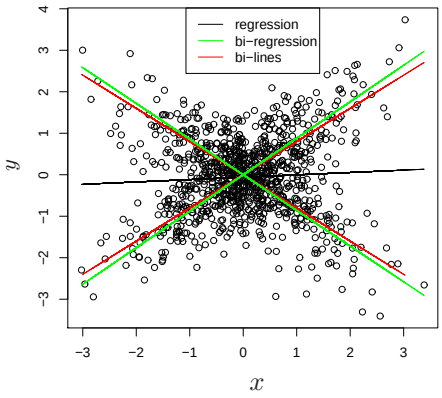

Fig. 4. Bi-lines and bi-regression of mixture normal distribution

The results are as follows:
$y_1 = 0.8x$, $y_2 = -0.8x$ is bi-regression of mixture normal distribution;
$y_1 = 0.86x$, $y_2 = -0.87x$ is bi-lines regression of mixture normal distribution.

The estimators of the bi-lines function parameters are biased estimators of parameters r_1 and r_2 of the regressions of the distribution mixture. The bias decreases, when module of the difference between the regression slopes (correlation) coefficients increases.

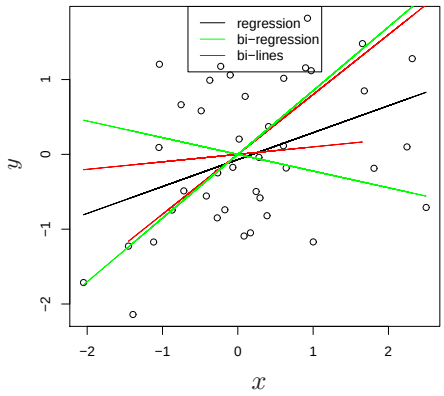

Fig. 5. Bi-lines and bi-regression of mixture of normal distribution, $r_1 = 0.1$, $r_2 = 0.8$, $n = 40$

Example 3.6. We randomly generated data (100 points around two straight lines) in R with the following parameters: $y_1 = b_1 x + \varepsilon$, $y_2 = b_2 x + \varepsilon$, $varepsilon \cong N(0, \sigma)$, $\sigma = 1$, $p = 0.5$, $b_1 = 2$, $b_2 = 4$,p-mixing proportions. The values of x was generated from the uniform distribution $U(0, 20)$.

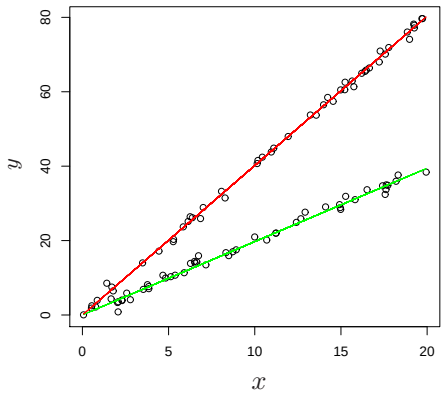

Fig. 6. Estimators of bi-regression

On the basis of appropriate expressions (9) and (10) we have the following values of estimators: $\hat{b}_1 = 1.99$, $\hat{b}_2 = 4.01$.

Example 3.7. We approximate a two-dimensional normal distribution by means of three lines. We estimate parameters of lines using numerical approximation method, applying the function optim in R. We can see that one of these lines overlaps with the regression line.

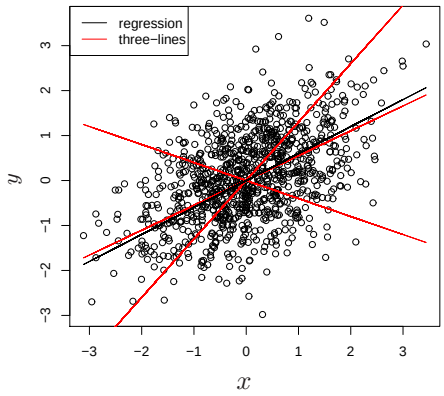

Fig. 7. Regression and three-lines bivariate normal distribution, $r = 0.5$

4. Conclusions

In this article we considered the problem of estimation of bi-lines regression parameters by means of the implicit least square method. Additionally this method demands some numerical iteration methods because in a general case it is not possible to get exact results. But, in the particular case we considered we derived exact expressions for parameter estimators. The relationship between the definitions of the bi-lines and the bi-regression functions of mixture of two-dimensional probability distribution was considered.

It turned out that the estimators of bi-lines parameters can be useful in bi-linear regression in the case when the slopes of bi-lines differ significantly.

A generalization of this approach to three or more lines will be the subject of further research in the future.

Bibliography

1. Antoniewicz R.: *The least squares method for an implicit interdependence and its application in economy*. Wyd. Akademii Ekonomicznej we Wrocławiu 1988, 134–135 (in Polish).
2. Antoniewicz R.: *Bi-linear regressions*. Zeszyty Nauk. Uniwersytetu Szczecińskiego (2001), 13–14 (in Polish).
3. Sitek G.: *Estimation of regression parameters of two dimensional probability distribution mixtures*. Studia Ekonomiczne. Informatyka i Ekonometria **7** (2016).