Colloquium Biometricum 39
2009, 177–185

A NOTE ON ORTHOGONAL PROJECTION METHOD

Magdalena Ćwiklińska, Zofia Hanusz

Department of Applied Mathematics and Computer Science
University of Life Sciences in Lublin
Akademicka 13, 20-950 Lublin
E-mails: magdalena.cwiklinska@.up.lublin.pl, zofia.hanusz@.up.lublin.pl

Summary

In the paper some measure of fit in the orthogonal projection method is presented. This measure is compared to the determination coefficient in the classical regression given by the least squares method on the example of linear function. Some disadvantages of the projection method are shown and illustrated by the numerical example.

Key words and phrases: regression function, estimation of parameters, measure of fit

Classification AMS 2000: 62J05, 62J02

1. Introduction

The aim of this paper is to introduce some measure of fit of the function obtained in orthogonal projection method and to show some disadvantages of the method and the measure. Estimation of the function describing the relation between two random variables X and Y is basic in statistical analysis of experimental data sets. In general, regression functions are obtained by the least squares method. Fitting of the regression function is measured by a determination coefficient. However, in the literature of this subject, some method based on the orthogonal projection on the function is also proposed (see Cramer (1946),

Przybysz (1980), Krysicki et al. (1999), Hanusz et al. (2006)). In the paper some measure of fit for the projection method is proposed. This measure is compared with the determination coefficient in the regression method. As the result of the comparison of methods some critics of the projection method is presented. In Section 2 formulas for slopes and intercepts in the regression and projection methods are presented. The measure of fit in the orthogonal projection method is proposed in Section 3. In Section 4 some concluding remarks are enclosed.

2. Estimation of parameters in linear relation of two variables

The regression method describing the relation between two dependent random variable is the most applicable for analyzing data sets in agricultural experiments. A case where there is not known in advance which variable is independent and has influence for the other one is considered. In such the case both regression function $Y=f(X)$ or $X=g(Y)$ could be evaluated. Moreover, it seems that the best function describing the relation between X and Y is that which is based on orthogonal projection on the function. Estimation of the parameters of such kind of function was presented in Hanusz et al. (2006). In the literature, the same approach can be found, for example, in Cramer (1946), Krysiski et al. (1999), Przybysz (1980) and others. The measure of fit was proposed in Hanusz et al. (2006). However, this measure gives too high fitting level and does not rich null value even in the case of no correlated variables. In this paper the other measure of fit of the experimental point to estimated function is proposed. Our attention is restricted to linear function of dependency.

Let us consider n random bivariate variables denoted by $(X_1,Y_1),...,(X_n,Y_n)$. Moreover, let us assume that the relation between X and Y can be described by linear function of the form $Y=\beta X+\alpha$, where β and α are unknown slope and intercept, respectively. When the least squares method is used, the estimator of β and α have very known forms, namely,

$$\hat{\beta}=\frac{S_{xy}}{S_x^2}, \qquad \hat{\alpha}=\overline{Y}-\hat{\beta}\overline{X},$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$,

$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

In the case where function $X=g(Y)$ is estimated, the least squares method gives function $X = \beta_0 Y + \alpha_0$ with estimators $\hat{\beta}_0 = \frac{S_{xy}}{S_y^2}$ and $\hat{\alpha}_0 = \bar{X} - \hat{\beta}_0 \bar{Y}$, where

$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

When method of orthogonal projections points (X_i, Y_i) $(i = 1, \cdots, n)$ on a function is used, estimators of a slope and an intercept have more complicated forms. Let us denote the intercept and the slope in the projection method by A and B, respectively. Estimators of A and B have the following forms: $\hat{\mathrm{A}} = \bar{Y} - \hat{\mathrm{B}}\,\bar{X}$ and $\hat{\mathrm{B}}$ is the solution of the quadratic equation $\mathrm{B}^2 + \frac{S_x^2 - S_y^2}{S_{xy}}\mathrm{B} - 1 = 0$, minimizing function $h(\mathrm{A},\mathrm{B}) = \sum_{i=1}^{n}\frac{(\mathrm{B}X_i - Y_i + \mathrm{A})^2}{\mathrm{B}^2+1}$. Explicit form for $\hat{\mathrm{B}}$ can be found in Krysicki et al. (1999) and Hanusz et al. (2006).

To show mutual locations of strait lines obtained in the discussed methods, some simulation studies were done for n=100 bivariate variables generated from normal distribution with fixed expectations, standard deviations and correlations. The results are shown in Fig. 1-4. It can be noticed that the lines corresponding to the method of orthogonal projection is always located between the lines in the least square methods. In the case where variance for variable Y is smaller than for the variable X (see Fig.1), then the orthogonal projection line is close to the line representing the regression $X=g(Y)$. For the opposite relation between variances, the opposite behavior as presented in Fig. 2 is obtained. For the same variances of both variables (see Fig.3, 4), the orthogonal projection line is situated in the center between the regression lines.

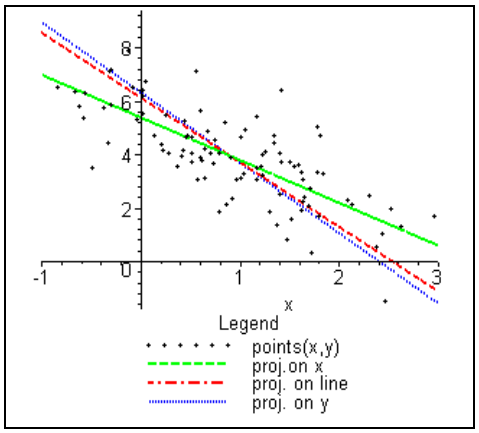

Fig. 1. Exposure the results of three methods for ρ=-0.8, μ_x=1, μ_y=4, σ_x=1, σ_y=2

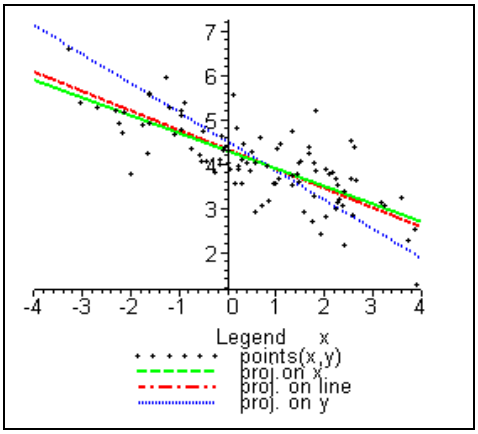

Fig. 2. Exposure the results of three methods for ρ=-0.8, μ_x=1, μ_y=4, σ_x=2, σ_y=1

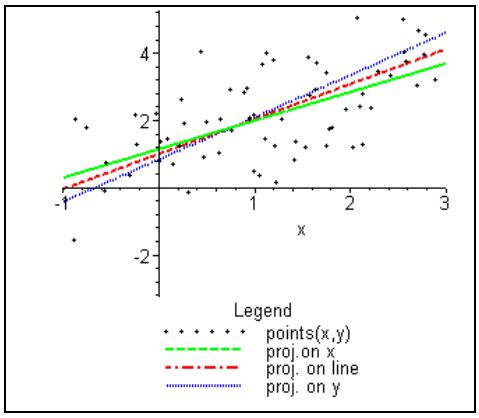

Fig. 3. Exposure of the results of three methods for ρ=0.85, μ_x=1, μ_y=2, σ_x=2, σ_y=2

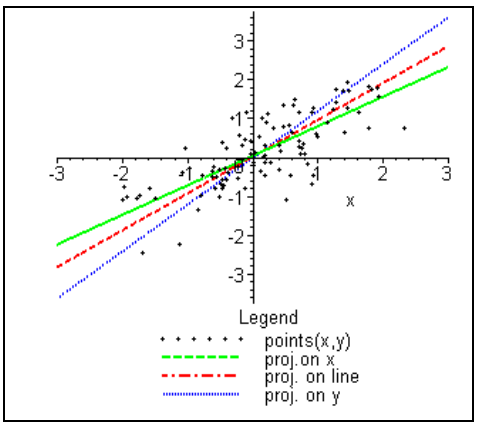

Fig. 4. Exposure of the results of three methods for ρ=0.8, μ_x=0, μ_y=0, σ_x=1, σ_y=1

3. Measure of fit in the orthogonal projection method

In Hanusz et al. (2006) some measure of fit was proposed. The main disadvantage of that measure was that it has taken too high values. It has never reached the value close to null, even in the case of independent variables. In this paper the other measure of fit, which gets values between 0 and 1, is proposed. To define

this measure some relations showed in Hanusz et al. (2006) are remembered. Namely, total sums of variability for both variables can be divide as follows:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\tilde{Y}_i)^2+\sum_{i=1}^{n}(\tilde{Y}_i-\bar{Y})^2,\quad \sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}(X_i-\tilde{X}_i)^2+\sum_{i=1}^{n}(\tilde{X}_i-\bar{X})^2,$$

where $\tilde{X}_i, \tilde{Y}_i$ (i=1,...,n) denote coordinates of orthogonal projection points (X_i, Y_i) on the line $Y = \mathrm{B}X + \mathrm{A}$, respectively. The measure of fit is defined in the following way:

$$\tilde{R}^2=\frac{\sum_{i=1}^{n}(\tilde{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\cdot\frac{\sum_{i=1}^{n}(\tilde{X}_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

To check goodness of this measure, the simulation studies were done for the n=100 generated data sets from bivariate normal distribution with some fixed expectations and standard deviations. Moreover, the correlation coefficient between variables were changed from -1 to 1 with the step 0.05. Calculated measures of fit were compared with the determination coefficient in the linear regression method. The results of simulation studies for different correlations, expectations and standard deviations of X and Y, together with the fitting quadratic functions getting by the least squares method are presented in Fig. 5–8.

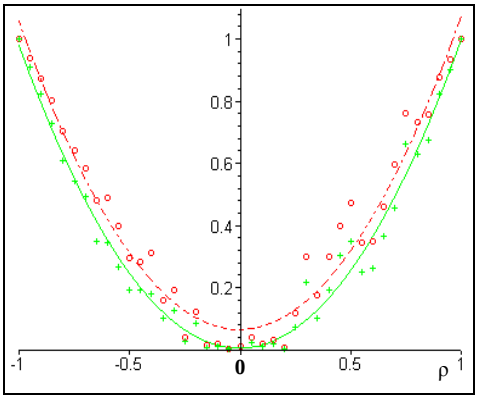

Fig. 5. Measure of fit for μ_x=1, μ_y=4, σ_x=1, σ_y=2

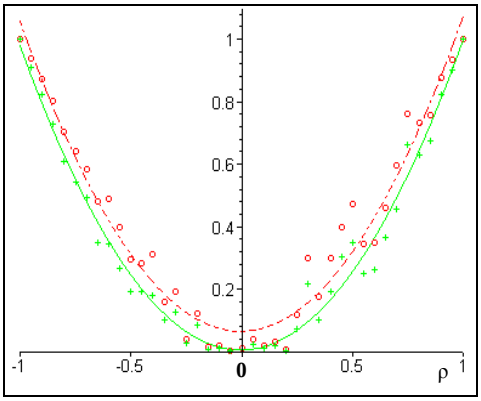

Fig. 6. Measure of fit for μ_x=1, μ_y=4, σ_x=2, σ_y=1

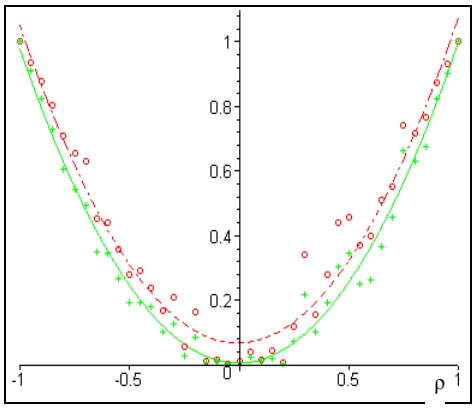

Fig. 7. Measure of fit for $\mu_x=1$, $\mu_y=2$, $\sigma_x=1$, $\sigma_y=0.5$

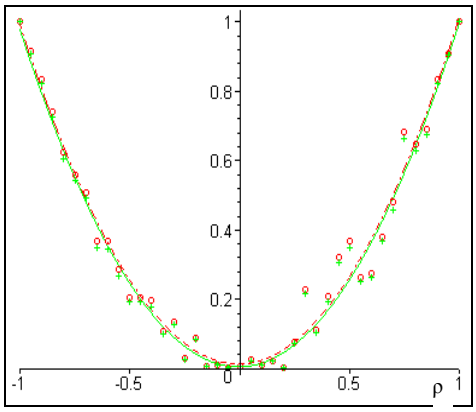

Fig. 8. Measure of fit for $\mu_x=1$, $\mu_y=20$, $\sigma_x=1$, $\sigma_y=5$

In Fig. 5–8, points and the regression line in the projection method on *x* axe are denoted by cross and a solid line, while in the orthogonal projection method on line by circle and dashed line, respectively. It can be noticed that the fitting function for ρ=0 in orthogonal projection method (circle) could take a measure greater then null (see Fig. 5–7) but there exist points getting null value of coefficient. Moreover, Fig. 8 shows that measures $\tilde{R}^2$ are quite different from that in Fig. 7 although the only variable *Y* was multiply by 10 while the determination coefficient in the regression method left the same. This fact states that the orthogonal projection method on line gives different functions depending on the units for the variables *X* an *Y*. This fact has also influence on the measure of fit. Fig. 8 shows that in the case where variable *Y* was multiply by 10, $\tilde{R}^2$ goes to R^2 in the regression method $Y=f(X)$.

4. Concluding remarks on the orthogonal projection method

In the previous sections the orthogonal projection method, which can be used to describe dependency between two random variables, was compared with classical regression method. In the case, where experimenter knows which variable is dependent and which one is independent, then the classical regres-

sion method is the best one. However, in the case where one variable is forecasted for some fix value of the second one, and there is not known which of them is estimated, then the method of orthogonal projection on function seems to be better than the least squares method. Nevertheless, the orthogonal projection method is more complicated even in the case of linear relation between variables. In spite of this, the main disadvantage consists in a fact that estimated parameters of the function depend on the measurable units. It is easy to prove this fact. Namely, let us take the transformation $Z=aY+b$. In the projection method $\hat{\mathrm{A}} = a\overline{Y} + b - \hat{\mathrm{B}}\,\overline{X}$ and $\hat{\mathrm{B}}$ are now the solution of the quadratic equation $\mathrm{B}^2 + \frac{S_x^2 - a^2 S_y^2}{aS_{xy}}\mathrm{B} - 1 = 0$, which differs from that given in Section 2. Similarly, the measure of fit has a form:

$$\tilde{\tilde{R}}^2 = \frac{\sum_{i=1}^{n}\left(\tilde{Z}_i - a\overline{Y} - b\right)^2}{a^2\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} \cdot \frac{\sum_{i=1}^{n}\left(\tilde{\tilde{X}}_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \left(\neq \tilde{R}^2\right),$$

and there is not the same as the determination coefficient for X and Y, where $\tilde{\tilde{X}}_i$ and $\tilde{Z}_i$ (i=1,...,n) are coordinates of orthogonal projection on line $Z = \mathrm{B}\,X + \mathrm{A}$. This fact is illustrated by the following numerical example.

Example. Let us consider six points for which the linear functions describing the relation between variables are obtained by the projection method. Two different cases were considered. In the first one, the linear function between variables Y and X and at the second one the function between Z=10Y and X were estimated. In both cases parameters of functions and the corresponding measures of fit were calculated. Simultaneously, forecasts of Y in both cases were calculated. The results are enclosed in Table 1.

Using the projection method, we have got the linear function $Y = 0.984046X + 0.039173$ for the data enclosed in two first columns of Table 1 and $Z = 9.88156X + 0.247846$ for data enclosed in the first and the third column. It can be noticed that the ratio of slopes is not 10 as we would get using the least squares method. Moreover, the prognosis $\tilde{z}_i$ is not the same as for $10\tilde{y}_i$ (i=1,...,n).

Table 1. Values of variables and estimates in the projection method

	x_i	y_i	$z_i = 10y_i$	$\tilde{x}_i$	$\tilde{y}_i$	$\tilde{\tilde{x}}_i$	$\tilde{z}_i$
	1	1.2	12	1.088379	1.110188	1.187382	11.98104
	2	1.8	18	1.896381	1.905299	1.798555	18.02039
	3	3.1	31	3.054338	3.044781	3.110936	30.98877
	4	3.9	39	3.962327	3.938284	3.922455	39.00785
	5	4.8	48	4.920309	4.880983	4.834146	48.01678
	6	6.1	61	6.078266	6.020465	6.146527	60.98517
Sample mean	3.5	3.48(3)	34.8(3)	3.4999	3.48(3)	3.5	34.8(3)

To evaluate the measures of fit, the sums of squares for X and Y were calculated:

$$\sum_{i=1}^{6}(y_i-\bar{y})^2 = 16.9483, \quad \sum_{i=1}^{6}(\tilde{y}_i-\bar{y})^2 = 16.9118, \quad \sum_{i=1}^{6}(y_i-\tilde{y}_i)^2 = 0.0366,$$

$$\sum_{i=1}^{6}(x_i-\bar{x})^2 = 17.5, \quad \sum_{i=1}^{6}(\tilde{x}_i-\bar{x})^2 = 17.4646, \quad \sum_{i=1}^{6}(x_i-\tilde{x}_i)^2 = 0.0354.$$

Using the above results, the measure of fit gets value $\tilde{R}^2 = \frac{16.9118}{16.9483}\cdot\frac{17.4646}{17.5} \cong 0.996$.

Similar results for the variables X and $Z=10Y$ are the following:

$$\sum_{i=1}^{6}(z_i-\bar{z})^2 = 1694.8333, \quad \sum_{i=1}^{6}(\tilde{z}_i-\bar{z})^2 = 1694.8319, \quad \sum_{i=1}^{6}(z_i-\tilde{z}_i)^2 = 0.0014,$$

$$\sum_{i=1}^{6}(x_i-\bar{x})^2 = 17.5, \quad \sum_{i=1}^{6}(\tilde{\tilde{x}}_i-\bar{x})^2 = 17.3, \quad \sum_{i=1}^{6}(x_i-\tilde{\tilde{x}}_i)^2 = 0.0354.$$

In the second case, the measure of fit is equal to, $\tilde{\tilde{R}}^2 = \frac{1694.8319}{1694.8333}\cdot\frac{17.357}{17.5} \cong 0.992,$

and is not the same as in the first case. This fact testifies against using the orthogonal projection method.

References

Cramer H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Hanusz Z., Ćwiklińska M., Siarkowski Z. (2006). Dependency function of two variables based on orthogonal projection. *Colloquium Biometricum* **36**, 245–256 (in Polish).

Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M. (1999). *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN.

Przybysz T. (1980). Dopasowanie prostej metodą najmniejszych kwadratów odległości ortogonalnych. *Rocznik nauk rolniczych* T. C-74-4.

PEWNA UWAGA O METODZIE RZUTU ORTOGONALNEGO

W pracy została przedstawiona pewna miara dopasowania w metodzie rzutu ortogonalnego na wykres funkcji. Miara ta, na przykładzie funkcji liniowej, została porównana ze współczynnikiem determinacji w metodzie najmniejszych kwadratów dla obserwacji generowanych z rozkładu normalnego z zadanymi parametrami. Pewne wady metody rzutu ortogonalnego na funkcję oraz zaproponowanej miary dopasowania zostały w pracy wskazane oraz zilustrowane przykładem liczbowym.

Słowa kluczowe: funkcja regresji, estymacja parametrów, miara dopasowania

Klasyfikacja AMS 2000: 62J05, 62J02