When Do Plots of Regressions of *X* on *Y* and of *Y* on *X* Coincide?

Czesław Stępniak* and Katarzyna Wąsik

Institute of Mathematics, University of Rzeszów, Rejtana 16 A, 35-959 Rzeszów, Poland

Abstract: It is well known that the plots of the regressions of a random variable *X* on *Y* and of *Y* on *X*, in general, does not coincide. We study conditions for such coincidence.

Keywords: Linear regression, overall regression, regression of *X* on *Y*, regression of *Y* on *X*, coincidence.

1. INTRODUCTION

To find the regression function of a random variable *X* on *Y*, one of our students used the equation $y = f(x)$, of the regression of *Y* on *X*, and solved it with respect to *x*. The aim of this note is to answer the question, when such a procedure is appropriate or, in the other words, when do plots of these two regression equations coincide.

Let *X* and *Y* be random variables such that *X* is observable, while *Y* is of our interest. Then we are seeking for a (possibly best) predictor, say $y = f(x)$ of *Y* on *X*. The classical measure of the error of such predictor is the expected Mean Squared Error, i.e. $E(Y - f(X))^2$. Thus we are looking for a function minimizing the MSE in a class, say F, of respective functions. It is well known (cf., e.g., [1], Sec. 3.2, or [2], Sec. 3.7) that, among all possible functions, this minimum is attained when $f(x)$ coincides with the mean of the conditional distribution of *Y* on $X = x$, say $f_1(x)$. We shall refer to such a function f_1 as to the first kind (or overall) regression of *Y* on *X*.

If we restrict ourselves to the linear functions of type $f(x) = ax + b$, then the solution is respectively modified. We shall refer to this solution, say f_2, as to the second kind (or linear) regression of *Y* on *X*.

By changing the role of the variables *X* and *Y* we reach to the regressions $x = g_1(y)$ and $x = g_2(y)$, of the first and the second type, respectively, of *X* on *Y* . It is well known that the plots of the functions f_1 and g_1 (or f_2 and g_2), in general, do not coincide. The aim of this note is to answer the question, in terms of the joint distribution of *X* and *Y*, when they do.

2. LINEAR REGRESSION CASE

Let *X* and *Y* be random variables with finite but nonzero variances σ_X^2 and σ_Y^2 and with correlation coefficient $\rho = \rho_{XY} = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$. Then the linear regression $y = f_2(x)$, of *Y* on *X* is defined by the equation

$$y - EY = \rho \frac{\sigma_Y}{\sigma_X}(x - EX) \tag{2.1}$$

(cf. [3], Sec. 3.3) while the linear regression $x = g_2(y)$, of *X* on *Y* , is defined by

$$x - EX = \rho \frac{\sigma_X}{\sigma_Y}(y - EY). \tag{2.2}$$

Thus a necessary and sufficient condition for the coincidence of the plots of the equations (2.1) and (2.2) is

$$\rho_{XY}^2 = 1. \tag{2.3}$$

We shall present this condition in a more readable form

$$P(Y = aX + b) = 1 \text{ for some deterministic } a \neq 0 \text{ and } b. \tag{2.4}$$

One can verify immediately that (2.4) implies (2.3). For the reverse implication we shall use the following inequality.

Chebyshev's inequality: (cf. [4], p. 58 and 93). For any random variable *X* with finite variance σ^2

$$P(|X - EX| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \text{ for every } \varepsilon > 0.$$

In consequence of this inequality we get the following corollary.

Corollary 1. *For any random variable X its variance is equal to zero, if and only if* $P(X = c) = 1$ *for some deterministic c*.

Now we are ready to show the implication (2.3) $\Rightarrow$ (2.4). By the property $\rho_{aX+b,cY+d} = sgn(ab)\rho_{XY}$ one can assume, without loss of generality, that $EX = EY = 0$ and $\sigma_X = \sigma_Y = 1$. Then $\rho_{X,Y} = \sigma_{XY}$.

Suppose $\rho^2 = 1$. Then either $\rho = 1$, or $\rho = -1$.

If $\rho = 1$ then,

$$\sigma_{X-Y}^2 = 2(1 - \sigma_{XY}) = 2(1 - \rho) = 0,$$

Thus, by Corollary 1, we get the desired condition (2.4).

Otherwise, if $\rho = -1$,

$$\sigma_{X+Y}^2 = 2(1 + \sigma_{XY}) = 2(1 + \rho) = 0,$$

and again we get (2.4). In this way we have shown that the conditions (2.3) and (2.4) are equivalent, and each of them is

*Address correspondence to this author at the Institute of Mathematics, University of Rzeszów, Rejtana 16 A, 35-959 Rzeszów, Poland; E-mail: cees@univ.rzeszow.pl

1876-5270/09

necessary and sufficient for the coincidence of the plots of the linear regression functions.

Now we shall present an application of this result in descriptive statistics.

Let $(x_1, y_1),\ldots,(x_n, y_n)$ be a sample with the sample statistics

$$\bar{x} = \frac{1}{n}\sum_i x_i, \quad s_x^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 > 0,$$
$$\bar{y} = \frac{1}{n}\sum_i y_i, \quad s_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2 > 0, \text{ and}$$
$$s_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}).$$

If the values $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$ are different then such a sample may be identified with a distribution of some random variables X and Y, taking values $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$, respectively, with probabilities $P(X = x_i, Y = y_i) = \frac{1}{n}$.

Let us recall that the empirical (i.e. the Least Squares) regressions, of y on x, and of x on y, are defined by

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \tag{2.5}$$

and

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \tag{2.6}$$

(cf. [5]. Sec.1.7) and they coincide with theoretical linear regressions of Y on X and of X on Y, respectively. Therefore, the empirical regressions (2.5) and (2.6) coincide, if and only if, all the points $(x_1\ y_1),\ldots,(x_n, y_n)$ lie on a straight line.

3. OVERALL REGRESSION CASE

In this section we shall restrict ourselves to the case of discrete random variables.

Let X and Y be discrete random variables taking values in some discrete sets $\mathsf{X} = \{x_1,\ldots,x_m\}$ and $\mathsf{Y} = \{y_1,\ldots,y_k\}$, respectively. For convenience, we shall identify the joint distribution of these variables with a matrix $P = (p_{ij})$, where $p_{ij} = P(X = x_i, Y = y_j)$.

Introduce also the symbols $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$ for $i = 1,\ldots,m$ and $j = 1,\ldots,k$. Without loss of generality we may (and shall) assume that $p_{i.}$ and $p_{.j}$ are positive for all i and j. In this context the overall (i.e. the first type) regression of Y on X is the function $y = f_1(x)$ from X onto Y, defined by

$$f_1(x_i) = E[Y / X = x_i] = \sum_j \frac{p_{ij}}{p_{i.}} y_j, \tag{3.1}$$

and the overall regression of X on Y is the function $x = g_1(y_j)$ from Y onto X, defined by

$$g_1(y_j) = E[X / Y = y_j] = \sum_i \frac{p_{ij}}{p_{.j}} x_i, \tag{3.2}$$

(cf. Goldberger [3], Sec. 3, or Fisz [2], Sec. 3.7).

It can be easily verified that, if $k = m$ and the joint distribution P satisfies the condition:

(C) Each row and each column in the matrix P has exactly one non-zero entry: then the plots of the regression functions (3.1) and (3.2) coincide. So one can ask, whether the condition (C) is also necessary.

Assuming $k = m$ let us arrange the possible values of X in the increasing order $x_1 < x_2 < \ldots < x_m$ and let

$$y_i = f_1(x_i) \text{ for } i = 1,\ldots,m.$$

We shall start from the following assumptions.

Assumption 1. The sequence $y_1,\ldots,y_m$ is strictly monotone.

Assumption 2. The integer $m \leq 3$.

We will show that under any of these assumptions the condition (C) is necessary and sufficient for the equivalence of the equations (3.1) and (3.2).

Under Assumption 1, the necessity of (C) may be proved by induction with respect to m. We will show one step of this induction. Suppose $y_1 < y_2 < \ldots < y_m$.

Then

$$E[Y / X = x_i] = \sum_j \frac{p_{1j}}{p_{1.}} y_j$$

and it equals y_1, if and only if,

$$p_{1j} = \begin{cases} p_{1.} & \text{if } j = 1 \\ 0, & \text{if } j \neq 1. \end{cases}.$$

Similarly we get

$$p_{i1} = \begin{cases} p_{.1}, & \text{if } i = 1 \\ 0, & \text{if } i \neq 1. \end{cases}$$

and the problem reduces to $m - 1$ values $x_2,\ldots,x_m$ and $y_2,\ldots,y_m$.

Now let us go to the Assumption 2. For $m = 2$ the necessity of (C) may be verified directly, while for $m = 3$ it only remains to consider the case $E[Y/X = x_1] = y_1$, where y_1 lies between y_2 and y_3. Since $E[X/Y = y_1] = x_1$ we get

$$p_{i1} = \begin{cases} p_{.1}, & \text{if } i = 1 \\ 0, & \text{if } i \neq 1 \end{cases}$$

and, similarly,

$$p_{i3} = \begin{cases} p_{.3}, & \text{if } i = 3 \\ 0, & \text{if } i \neq 3 \end{cases}.$$

Now, searching the conditions $E[Y/X = x_1] = y_1$ and $E[Y/X = x_3] = y_3$ we get the desired result (C).

At this moment one can suspect that the condition (C) is also necessary in general. We shall demonstrate, by example, that it is not this case.

Let us set $x_i = y_i = i$ for $i = 1, 2, 3, 4$ and

$$P = \begin{bmatrix} 0.2 & 0.1 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0.1 & 0.2 \end{bmatrix}.$$

One can verify that $E[Y/X = 1] = 2$, $E[Y/X = 2] = 1$, $E[Y/X = 3] = 4$, $E[Y/X = 4] = 3$, $E[X/Y = 2] = 1$, $E[X/Y = 1] = 2$, $E[X/Y = 4] = 3$ and $E[X/Y = 3] = 4$. Thus the plots of the regression functions f_1 and g_1 coincide, while the condition (C) does not hold.

ACKNOWLEDGMENT

Our thanks go to a couple of reviewers for their helpful comments.

REFERENCES

[1] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II (2nd Ed.), New York: Wiley, 1971.
[2] M. Fisz, *Probability Theory and Mathematical Statistics*, (3rd Ed.), New York: Wiley, 1963.
[3] A.S. Goldberger, *Econometric Theory*, New York: Wiley, 1964.
[4] P. Brémaud, *An Introduction to Probabilistic Models, Corrected 2nd Printing*, New York: Springer, 1994.
[5] H. Theil, *Principles of Econometrics*, New York: Wiley, 1971.

Received: December 26, 2008 Revised: March 06, 2009 Accepted: March 31, 2009