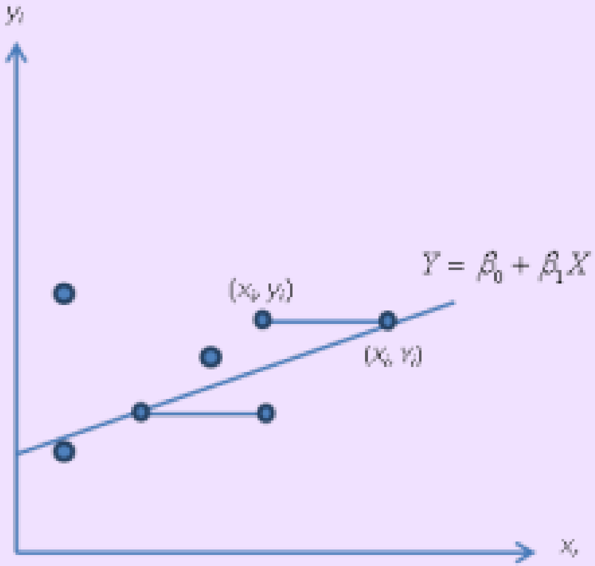
Data problems

Professor V.M. Auken, PhD

Reverse regression method

The reverse (or inverse) regression approach minimizes the sum of squares of horizontal distances between the observed data points and the line in the following scatter diagram to obtain the estimates of regression parameters.



Reverse regression method

The reverse regression has been advocated in the analysis of sex (or race) discrimination in salaries. For example, if *y* denotes salary and *x* denotes qualifications and we are interested in determining if there is a sex discrimination in salaries, we can ask:

> “Whether men and women with the same qualifications (value of *x*) are getting the same salaries (value of *y*). This question is answered by the **direct regression**.”

Alternatively, we can ask:

> “Whether men and women with the same salaries (value of *y*) have the same qualifications (value of *x*). This question is answered by the **reverse regression**, i.e., regression of *x* on *y*.”

The regression equation in case of reverse regression can be written as $x_i = \beta_0^* + \beta_1^* y_i + \delta_i \quad (i = 1, 2, ..., n)$

where δ_i 's are the associated random error components and satisfy the assumptions as in the case of usual simple linear regression model.

The reverse regression estimates $\hat{\beta}_{OR}$ of β_0^* and $\hat{\beta}_{1R}$ of β_1^* for the model are obtained by interchanging the *x* and *y* in the direct regression estimators of β_0 and β_1. The estimates are obtained as

$$\hat{\beta}_{OR} = \overline{x} - \hat{\beta}_{1R}\overline{y}$$

and

$$\hat{\beta}_{1R} = \frac{s_{xy}}{s_{yy}}$$

for β_0^* and β_1^* respectively.

The residual sum of squares in this case is

$$SS_{res}^* = s_{xx} - \frac{s_{xy}^2}{s_{yy}}.$$

Note that

$$\hat{\beta}_{1R} b_1 = \frac{s_{xy}^2}{s_{xx} s_{yy}} = r_{xy}^2$$

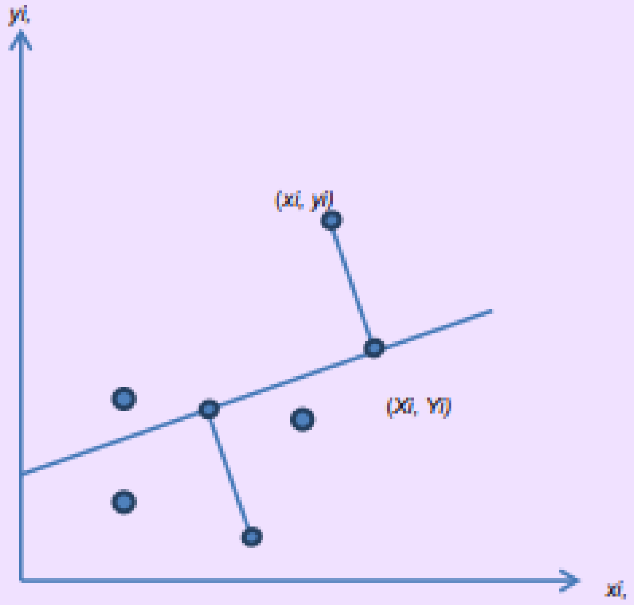
where b_1 is the direct regression estimator of slope parameter and r_{xy} is the correlation coefficient between *x* and *y*. Hence if r_{xy}^2 is close to 1, the two regression lines will be close to each other.

An important application of reverse regression method is in solving the calibration problem.

Orthogonal regression method (or major axis regression method)

The direct and reverse regression methods of estimation assume that the errors in the observations are either in *x*-direction or *y*-direction. In other words, the errors can be either in dependent variable or independent variable. There can be situations when uncertainties are involved in dependent and independent variables both. In such situations, the orthogonal regression is more appropriate. In order to take care of errors in both the directions, the least squares principle in orthogonal regression minimizes the squared perpendicular distance between the observed data points and the line in the following scatter diagram to obtain the estimates of regression coefficients. This is also known as **major axis regression method**. The estimates obtained are called as **orthogonal regression estimates** or **major axis regression estimates** of regression coefficients.

If we assume that the regression line to be fitted is $Y_i = \beta_0 + \beta_1 X_i$, then it is expected that all the observations $(x_i, y_i),\ i = 1, 2, ..., n$ lie on this line. But these points deviate from the line and in such a case, the squared perpendicular distance of observed data $(x_i, y_i)(i = 1, 2, ..., n)$ from the line is given by $d_i^2 = (X_i - x_i)^2 + (Y_i - y_i)^2$ where (X_i, Y_i) denotes the i^{th} pair of observation without any error which lie on the line.



Orthogonal or major axis regression method

The objective is to minimize the sum of squared perpendicular distances given by $\sum_{i=1}^{n} d_i^2$ to obtain the estimates of β_0 and β_1.

The observations $(x_i, y_i)(i = 1, 2, \ldots, n)$ are expected to lie on the line

$$Y_i = \beta_0 + \beta_1 X_i$$

so let

$$E_i = Y_i - \beta_0 - \beta_1 X_i = 0.$$

The regression coefficients are obtained by minimizing $\sum_{i=1}^{n} d_i^2$ under the constraints E_i's using the Lagrangian's multiplier method. The Lagrangian function is

$$L_0 = \sum_{i=1}^{n} d_i^2 - 2\sum_{i=1}^{n} \lambda_i E_i$$

where $\lambda_1, \ldots, \lambda_n$ are the Lagrangian multipliers.

The set of equations are obtained by setting

$$\frac{\partial L_0}{\partial X_i} = 0, \frac{\partial L_0}{\partial Y_i} = 0, \frac{\partial L_0}{\partial \beta_0} = 0 \text{ and } \frac{\partial L_0}{\partial \beta_1} = 0 \ (i = 1, 2, \ldots, n).$$

Thus we find

$$\frac{\partial L_0}{\partial X_i} = (X_i - x_i) + \lambda_i \beta_1 = 0$$

$$\frac{\partial L_0}{\partial Y_i} = (Y_i - y_i) - \lambda_i = 0$$

$$\frac{\partial L_0}{\partial \beta_0} = \sum_{i=1}^{n} \lambda_i = 0$$

$$\frac{\partial L_0}{\partial \beta_1} = \sum_{i=1}^{n} \lambda_i X_i = 0.$$

Since

$$X_i = x_i - \lambda_i \beta_1$$
$$Y_i = y_i + \lambda_i$$

so substituting these values in E_i, we obtain

$$E_i = (y_i + \lambda_i) - \beta_0 - \beta_1 (x_i - \lambda_i \beta_1) = 0$$
$$\Rightarrow \lambda_i = \frac{\beta_0 + \beta_1 x_i - y_i}{1 + \beta_1^2}.$$

Also using this λ_i in the equation $\sum_{i=1}^{n} \lambda_i = 0,$ we get

$$\frac{\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)}{1 + \beta_1^2} = 0$$

and using $(X_i - x_i) + \lambda_i \beta_1 = 0$ and $\sum_{i=1}^{n} \lambda_i X_i = 0,$ we get

$$\sum_{i=1}^{n} \lambda_i (x_i - \lambda_i \beta_1) = 0.$$

Substituting λ_i in this equation, we get

$$\frac{\sum_{i=1}^{n} (\beta_0 x_i + \beta_1 x_i^2 - y_i x_i)}{(1 + \beta_1^2)} - \frac{\beta_1 \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2}{(1 + \beta_1^2)^2} = 0. \qquad (1)$$

Using λ_i in the equation and using the equation $\sum_{i=1}^{n} \lambda_i = 0$, we solve

$$\frac{\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)}{1 + \beta_1^2} = 0 .$$

The solution provides an orthogonal regression estimate of β_0 as

$$\hat{\beta}_{0OR} = \overline{y} - \hat{\beta}_{1OR}\overline{x}$$

where $\hat{\beta}_{1OR}$ is an orthogonal regression estimate of β_1.

Now, substituting β_{0OR} in equation (1), we get

$$\sum_{i=1}^{n}(1+\beta_1^2)\left[\overline{y}x_i - \beta_1\overline{x}x_i + \beta_1 x_i^2 - x_i y_i\right] - \beta_1\sum_{i=1}^{n}\left(\overline{y} - \beta_1\overline{x} + \beta_1 x_i - y_i\right)^2 = 0$$

or $$(1+\beta_1^2)\sum_{i=1}^{n}x_i\left[y_i - \overline{y} - \beta_1(x_i - \overline{x})\right] + \beta_1\sum_{i=1}^{n}\left[-(y_i - \overline{y}) + \beta_1(x_i - \overline{x})\right]^2 = 0$$

or $$(1+\beta_1^2)\sum_{i=1}^{n}(u_i + \overline{x})(v_i - \beta_1 u_i) + \beta_1\sum_{i=1}^{n}(-v_i + \beta_1 u_i)^2 = 0$$

where

$$u_i = x_i - \overline{x},$$
$$v_i = y_i - \overline{y}.$$

Since $\sum_{i=1}^{n}u_i = \sum_{i=1}^{n}v_i = 0$, so

$$\sum_{i=1}^{n}\left[\beta_1^2 u_i v_i + \beta_1(u_i^2 - v_i^2) - u_i v_i\right] = 0$$

or

$$\beta_1^2 s_{xy} + \beta_1(s_{xx} - s_{yy}) - s_{xy} = 0.$$

Orthogonal regression method (or major axis regression method)

Solving this quadratic equation provides the orthogonal regression estimate of β_1 as

$$\hat{\beta}_{1OR} = \frac{(s_{yy} - s_{xx}) + sign(s_{xy})\sqrt{(s_{xx} - s_{yy})^2 + 4s_{xy}^2}}{2s_{xy}}$$

where *sign*(s_{xy}) denotes the sign of s_{xy} which can be positive or negative. So

$$sign(s_{xy}) = \begin{cases} 1 & \text{if } s_{xy} > 0 \\ -1 & \text{if } s_{xy} < 0. \end{cases}$$

Notice that this gives two solutions for $\hat{\beta}_{1OR}$. We choose the solution which minimizes $\sum_{i=1}^{n} d_i^2$.

The other solution maximizes $\sum_{i=1}^{n} d_i^2$ and is in the direction perpendicular to the optimal solution.

The optimal solution can be chosen with the sign of s_{xy}.

QUESTIONS!