

SOME GEOMETRICAL CONCEPTS IN LINEAR REGRESSION -
WITH 'MEASUREMENT' VARIABILITY - SITUATIONS

C. PHILIP COX
PROFESSOR EMERITUS

Preprint Number 96-16

June 1996

Department of Statistics
Iowa State University
Ames, IA 50011-1210

SOME GEOMETRICAL CONCEPTS IN LINEAR REGRESSION -
WITH 'MEASUREMENT' VARIABILITY - SITUATIONS

by

C. Philip Cox
Professor Emeritus
Iowa State University

ABSTRACT

1. If the rectangular axes for n points (x_i, y_i) are rotated through an angle $(\theta - \pi/2)$ the coordinates relative to the new axes are

$$x_{\theta i} = x_i \sin \theta - y_i \cos \theta, \quad y_{\theta i} = x_i \cos \theta + y_i \sin \theta .$$

An analysis is given to obtain the regression coefficient and the partitioning of the sum of squares $\sum_{i=1}^n y_{\theta i}^2$ for the regression of $y_{\theta i}$ -on- $x_{\theta i}$.

2. For linear regression when 'measurement' variability is present in both variables and the ratio δ ($0 < \delta < \infty$) of the 'measurement' variances is known, it is shown that the estimates of the parameters of the line can be obtained by minimizing the sum of the squares of oblique deviations - those making an estimated angle θ ($0 < \theta < \pi/2$) with the x-axis. Additionally:
- 3) It is noted that the point (\bar{x}, \bar{y}) lies on the fitted line.
- 4) In the standard analysis for this situation (Fuller, 1987) the estimator of the slope β_1 is obtained as the root of a quadratic equation. It is shown here that the estimator of the tangent of the angle θ is the other root of the quadratic equation.
- 5) Simple estimation procedures are given for the coordinates $\{\mu(x_i), \mu(y_i)\}$, on the line, from which the observations (x_i, y_i) deviate and it is shown that the two residuals, $x_i - \tilde{\mu}(x_i)$ and $y_i - \tilde{\mu}(y_i)$, are functionally related.
- 6) The common, measurement-variability-in-one-variable only, y-on-x and x-on-y procedures appear as special cases at the extreme angles $\theta = \pi/2$ and $\theta = -\pi$.
- 7) It is suggested that, notwithstanding the convenience, the assumption that measurement variability is present in only one variable is often fallacious; wider use of the more general procedure is advocated.

1. Introduction

When variability attends only one of two linearly related variables, x and y , the other being 'fixed', estimates of the y -on- x and x -on- y regression coefficients are obtained by minimizing the sums of squares of, respectively, the vertical and horizontal deviations of the observed points from the regression lines. It is an accordingly reasonable conjecture that minimizing sums of squares of oblique deviations should give results relevant to the much more common (it is suggested) cases when 'measurement' variability attends both the x and y variates. The conjecture is here confirmed and some implications of it are examined.

2. Estimation

Suppose that the coordinates (x_i, y_i) of n points are plotted and lines making an angle θ with the abscissa axis are drawn through the points to intercept the line $\eta = \beta_0 + \beta_1 \xi$ in the points $(\tilde{\xi}_i, \tilde{\eta}_i = \beta_0 + \beta_1 \tilde{\xi}_i)$. Let $\cos\theta = \ell$, $\sin\theta = m$ and d_i be the Euclidean distance between the points (x_i, y_i) and $(\tilde{\xi}_i, \tilde{\eta}_i)$. The d_i and $(y_i - \beta_0 - \beta_1 x_i)$, the vertical deviation from the line, are then related so that

$$(y_i - \beta_0 - \beta_1 x_i) = d_i(m - \ell\beta_1) . \quad (1)$$

The estimates b_0 and b_1 , of β_0 and β_1 , which minimize $\sum d_i^2$ are then

$$b_0 = \bar{y} - b_1 \bar{x} \quad (2)$$

so that the line passes through the point (\bar{x}, \bar{y}) , and

$$b_{1\theta} = (\ell \Sigma'yy - m \Sigma'xy) / (\ell \Sigma'xy - m \Sigma'xx) \quad (3)$$

where $\Sigma'xy = \Sigma(x_i - \bar{x})(y_i - \bar{y})$.

The situation is illustrated in Figure 1 where P is the 'observed point' (x_i, y_i) , T is the point $(\tilde{\xi}_i, \tilde{\eta}_i)$, the line PT makes an angle θ with the x -axis,

PT is the distance d_1 , M is the point (\bar{x}, \bar{y}) , the lines LM and OK are perpendicular to PT, N and K are the intersections of PT with LM and OK respectively, and OA is the intercept b_0 made on the y-axis by the fitted line AMT of slope $b_{1\theta}$. It is then apparent, for example by rotation of the coordinate

axes through the angle $-\left(\frac{\pi}{2} - \theta\right)$, that minimization of $\sum d_i^2$ involves the simple linear regression of the quantities $y_{\theta i} = KP$ on the quantities $x_{\theta i} = OK$ where

$$x_{\theta i} = mx_i - ly_i, \quad y_{\theta i} = lx_i + my_i. \quad (4)$$

The equation of the fitted line relative to the new axes is

$$y_{\theta} = \bar{y}_{\theta} + a_{1\theta}(x_{\theta} - \bar{x}_{\theta}) \quad (5)$$

where $a_{1\theta}$, the tangent of the angle \hat{TMM} , is given by

$$\begin{aligned} a_{1\theta} &= \Sigma' x_{\theta i} y_{\theta i} / \Sigma' x_{\theta i} x_{\theta i} \\ &= (\ell m \Sigma' xx - (\ell^2 - m^2) \Sigma' xy - \ell m \Sigma' yy) / (m^2 \Sigma' xx - 2\ell m \Sigma' xy + \ell^2 \Sigma' yy). \end{aligned} \quad (6)$$

The slope, $b_{1\theta}$ which the line (5) makes with the original x-axis can be obtained using the inverse of the transformation (4) or, from Figure 1, as

$$b_{1\theta} = \tan\left\{\left(\tan^{-1} a_{1\theta}\right) - \left(\frac{\pi}{2} - \theta\right)\right\}$$

which reduces to the expression given for $b_{1\theta}$ in (3).

Further, Table 1 gives the partition of the total sum of squares corresponding to the prescribed direction θ .

Table 1

The partition of the total sum of squares

source	degrees of freedom (df)	sum of squares (ss)
total	n	$\Sigma y_{i\theta}^2$
mean	1	$(\Sigma y_{i\theta})^2/n$
deviations from mean	n - 1	$\Sigma' y_{i\theta} y_{i\theta}$
slope ($a_{1\theta}$)	1	$a_{1\theta} \Sigma' x_{i\theta} y_{i\theta}$
residuals	n - 2	by subtraction = Σd_i^2

wherein

$$\Sigma' y_{i\theta} y_{i\theta} = \ell^2 \Sigma' xx + 2\ell m \Sigma' xy + m^2 \Sigma' yy$$

$$\Sigma d_i^2 = (\Sigma' xx \Sigma' yy - (\Sigma' xy)^2) / (m^2 \Sigma' xx - 2\ell m \Sigma' xy + \ell^2 \Sigma' yy)$$

and from which assessment quantities may be calculated as:

the fraction of the total sum of squares explained by the mean and slope structural components -

$$r_{S\theta}^2 = 1 - (\Sigma d_i^2) / \Sigma y_{i\theta}^2 \quad (7)$$

the fraction (adjusted for df) of the variance from the mean explained by the slope -

$$r_{A\theta}^2 = 1 - ((n-1)/(n-2)) (\Sigma d_i^2 / \Sigma' y_{i\theta} y_{i\theta}) \quad (8)$$

For the special case $\theta = \pi/2$, so that $\ell = 0$ and vertical deviations are being minimized, the slope estimate from (3) is the usual one, $b_1 = \Sigma' xy / \Sigma' xx$, for 'fixed-x' situations. Similarly $m = 0$ gives the regular estimate of the x-on-y regression coefficient. Corresponding reductions obtain for the other quantities defined above which may accordingly be regarded as a general treatment of which

the regular y-on-x and x-on-y analyses are the $\theta = \pi/2$ and $\theta = 0$ extreme special cases.

A more interesting special case is that for which $\tan\theta = m/l = -1/b_{1\theta}$ so that the

deviation vectors \vec{d}_i are perpendicular to the line. In this case (3) gives the quadratic equation

$$b_{1\theta}^2 \Sigma'xy + b_{1\theta} (\Sigma'xx - \Sigma'yy) - \Sigma'xy = 0 \quad (9)$$

of which the two solutions are at right-angles; that giving the finite and minimum sum of squares has the sign of $\Sigma'xy$.

3. $\theta?$

Introducing distributed quantities to extend the previous, purely computational, results it will now be additionally supposed that

$$\eta_i = \beta_0 + \beta_1 \xi_i, \quad x_i = \xi_i + \delta_i, \quad y_i = \eta_i + \epsilon_i,$$

$$\delta_i \sim G^* I(\sigma, \sigma_1^2), \quad \epsilon_i \sim GI(0, \sigma_2^2), \quad CV(\delta_i, \epsilon_i) = \sigma_{12} = \rho \sigma_1 \sigma_2, \quad i = 1, \dots, n$$

in which case the Fisherian log likelihood, (ℓn) is given by

$$-\ell n = (\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)^{-1} \Sigma (\sigma_2^2 (x_i - \xi_i)^2 - 2\sigma_{12} (x_i - \xi_i)(y_i - \beta_0 - \beta_1 \xi_i) + \sigma_1^2 (y_i - \beta_0 - \beta_1 \xi_i)^2) + (n/2) \ln 2\pi (\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)^{-1/2}$$

With $\tilde{\xi}_i$, b_0 and b_1 as estimates, the equation $\delta(\ell n)/\delta\tilde{\xi}_i = 0$ then gives

$$-\sigma_2^2 (x_i - \tilde{\xi}_i) - \sigma_{12} (-b_1 (x_i - \tilde{\xi}_i) - (y_i - b_0 - b_1 \tilde{\xi}_i)) - b_1 \sigma_1^2 (y_i - b_0 - b_1 \tilde{\xi}_i) = 0$$

that is

$$(x_i - \tilde{\xi}_i)(\sigma_2^2 - b_1 \sigma_{12}) + (y_i - b_0 - b_1 \tilde{\xi}_i)(b_1 \sigma_1^2 - \sigma_{12}) = 0 \quad (10)$$

Now, since $y_i - b_0 - b_1 \tilde{\xi}_i = y_i - \tilde{\eta}_i$, this indicates that the deviations of the

(x_i, y_i) , from the $(\tilde{\xi}_i, \tilde{\eta}_i)$, points on the estimate-line all make the angle

*G for Gauss(ian)

$$\bar{\theta} = \tan^{-1}(b_1\sigma_{12}-\sigma_2^2)/(b_1\sigma_1^2-\sigma_{12}) \quad (11)$$

with the abscissa axis.

The estimate b_1 can therefore be directly obtained by substituting $l = (b_1\sigma_1^2-\sigma_{12})$, $m = (b_1\sigma_{12}-\sigma_2^2)$ in (3). This gives, after simplification,

$$b_1^2(\sigma_1^2\Sigma'xy-\sigma_{12}\Sigma'xx) + b_1(\sigma_2^2\Sigma'xx-\sigma_1^2\Sigma'yy) + \sigma_{12}\Sigma'yy-\sigma_2^2\Sigma'xy = 0 \quad (12)$$

Further, it can be shown that the coefficients in the quadratic equation for $\tan\bar{\theta}$, obtained by substituting for b_1 from (11), are exactly those in (12) itself. The two roots of (12) are therefore b_1 , the slope of the line of fit, and $\tan\bar{\theta}$ where $\bar{\theta}$ is the angle (the slope of the line PTN in Figure 1) which the oblique deviations make with the abscissa axis. The value b_1 for the slope, agrees with the result given for this situation in the text Fuller (1987) and it follows that the estimates $(\hat{\beta}_0, \hat{\beta}_1)$ there obtained by minimizing statistical distances can be alternatively defined as those which minimize the sum of the squares of deviations all of which make the same angle, $\bar{\theta}$ in (11), with the x-axis.

In obtaining (12) it was assumed that σ_1^2 , σ_{12} and σ_2^2 were known quantities. From (12) itself, however, it is seen that only two of their ratios need be known and in particular that, provided σ_{12} is zero, only the ratio σ_2^2/σ_1^2 is required. It can also be seen that for the special case $\sigma_{12} = 0$, $\sigma_1^2 = \sigma_2^2$ (12) reduces to (9) and gives the regular estimator if either of σ_1^2 and σ_2^2 is zero.

4. Estimation of the 'true' values

Estimates $\bar{\xi}_i, \bar{\eta}_i$, of the 'true' values generating the observed values x_i, y_i , can be obtained by either of the two following procedures, both simple:

i) With b_1 from (12) and $b_0 = \bar{y} - b_1\bar{x}$ from (2) the estimates of ξ_i and η_i are now easily calculated as

$$\tilde{\xi}_i = x_i - \bar{d}_i \cos \bar{\theta}, \quad \tilde{\eta}_i = y_i - \bar{d}_i \sin \bar{\theta} \quad (13)$$

where, using (1),

$$\bar{d}_i \cos \bar{\theta} = (y_i - b_0 - b_1 x_i) / (\tan \bar{\theta} - b_1) \quad \text{and} \quad \bar{d}_i \sin \bar{\theta} = (\bar{d}_i \cos \bar{\theta}) \tan \bar{\theta} . \quad (14)$$

ii) The estimates required are the coordinates of the point T in Figure 1.

With reference to the rotated axes the coordinates are $x_{\theta i} = m x_i - l y_i$

from (4) and $\bar{y}_{\theta i} = \bar{y}_{\theta} + a_{1\theta} (x_{\theta i} - \bar{x}_{\theta})$ from (5). The slope $a_{1\theta}$ can be

calculated from the roots of (12) as $a_{1\theta} = (1 + b_1 \tan \theta) / (\tan \theta - b_1)$ or from

(6) with $\tan \theta = l/m$. The coordinates $(\tilde{\xi}_i, \tilde{\eta}_i)$ are then obtained using the inverse of the transformation (4).

In practice the x_i and y_i measurements will often be made, for example, on different occasions or by different observers, so that it may be reasonable to assume that $\sigma_{12} = 0$. Taking $\delta = \sigma_2^2 / \sigma_1^2$ as the known ratio of the two variances (12) then reduces to

$$b_1^2 \Sigma' xy + b_1 (\delta \Sigma' xx - \Sigma' yy) - \delta \Sigma' xy = 0 \quad (15)$$

which, since $\Sigma' xx$, $\Sigma' xy$ and $\Sigma' yy$ are respectively proportional to the sample covariances m_{xx} , m_{xy} and m_{yy} , is exactly equation (1.3.6) obtained for this case in Fuller (1987) by the method of moments.

5. Example

In the present notation the data pairs in Table 1.3.1 of Fuller (1987) give:

x_i = the number of hen pheasants sighted in Iowa by trained observers in Spring,

y_i = the similarly defined number sighted in August
for the years $i = 1$ (1976) to $i = 15$ (1962).

The model in Section 3 here was used with $\sigma_{12} = 0$ and, on the basis of other analyses, the value $\delta = \sigma_2^2/\sigma_1^2 = 1/6$ was taken for the assumed-to-be-known ratio of the variances. Summary statistics given for the data are $(\bar{x}, \bar{y}) = (10.0467, 8.6667)$ and

$$(m_{xx}, m_{xy}, m_{yy}) = (3.62124, 2.35167, 1.84952)$$

where $m_{xy} = \Sigma'xy/14$.

After substitution of numerical values, the two solutions of (15 are -0.2217 and 0.7516 and, as in Fuller (1987), the root having the same sign as m_{xy} is the estimate $b_1 = 0.7516$ of the slope while the intercept is $b_0 = \bar{y} - b_1\bar{x} = 1.1158$. With $\sigma_2^2/\sigma_1^2 = 1/6$ the slope, $b_1 = 0.75$ is, as expected, closer to that, $b_{xy} = 0.79$, of the ordinary x-on-y regression line than to that, $b_{yx} = 0.65$ of the y-on-x line. Correspondingly also, the present analysis shows that the estimates are obtained by minimizing the sum of squares of deviations making an angle $\hat{\theta} = \tan^{-1}(-0.2217) = -12^\circ 30'$ with the x-axis. Calculations of the estimates $(\tilde{\xi}_i, \tilde{\eta}_i)$ of the true values are now especially simple. Thus for 1974 with $(x_3, y_3) = (12.3, 9.8)$ $y_3 - b_0 - b_1x_3 = -0.5605$ and hence, from (14 and (13 with $\tan\theta = -0.2217$,

$$\tilde{\xi}_3 = 12.3 - .576 = 11.7, \quad \tilde{\eta}_3 = 9.8 + .128 = 9.9 .$$

Alternatively, since from (13 $(y_i - \tilde{\eta}_i) = (x_i - \tilde{\xi}_i)\tan\theta$, $\tilde{\eta}_3$ can be obtained as $9.8 + (.576)(.2217)$.

Quantities relating to the rotated axes can also be exhibited. Thus, with $\tan\theta = -0.2217$, $l = \cos\theta = 0.9763$, $m = \sin\theta = -0.2164$,

$$\bar{x}_\theta = m\bar{x} - l\bar{y} = -10.6354, \quad \bar{y}_\theta = l\bar{x} + m\bar{y} = 7.9331$$

and the slope $a_{1\theta} = (1+b_1 \tan\theta)/(\tan\theta-b_1) = -0.8564$.

Relative to the rotated axes, the equation (5 of the regression line gives

$$\tilde{y}_{\theta i} = -1.1751 - 0.8564x_{\theta i} .$$

Thus, for the year (1974), $x_{\theta 3} = mx_3 - ly_3 = -12.2295$ giving $\tilde{y}_{\theta 3} = 9.2983$ and hence $\tilde{\xi}_3 = mx_{\theta 3} + l\tilde{y}_{\theta 3} = 11.7$ and $\tilde{\eta}_3 = -lx_{\theta 3} + m\tilde{y}_{\theta 3} = 9.9$ as before.

Values of the quantities in Table 1 are given in Table 2 from which $r_A^2 = 0.83$ which compares with the value $r_A^2 = 0.81$ for the y-on-x and x-on-y regressions.

Table 2

Partition of the total sum of squares: Pheasant data

source	df	ss
total	15	979.6395
mean	1	944.0111
deviations from mean	14	35.6284
slope	1	30.0424
residuals	13	5.5860

In passing it may be noted that the Gaussian distributions assumed in Section 3 may be incompletely veridical for the pheasant data. The model inherently specifies the probability (0.5) that the observed number of pheasants will exceed(!) the true number - as suggested by the preceding calculation where $\tilde{\xi}_3 < 12.3$, the observed value. It seems improbable that repeated countings of the same pheasants could completely account for such contingencies.

5. Discussion

The linear regression situation examined here is comprehensively treated in Fuller (1987) where it is shown that, if $\sigma_{12} = 0$, the method of moments estimator of β_1 is the same as the least squares estimator. The latter is obtained by minimizing the sum of the squares of the general statistical distances between the observed points and the estimates of their 'true' values on the line of fit. And since with $\sigma_{12} = \rho\sigma_1\sigma_2$, none of which is here being estimated, the previous expression for $-\ln$ is proportional to the sum of these distances plus a constant, the general least squares estimator is also exactly that obtained from (6 above.

Mutatis mutandis, the preceding results can also be demonstrated using oblique axes and, in fact, if the Cartesian coordinates x_i and y_i are replaced by coordinates x_i/σ_1 and y_i/σ_2 referred to axes inclined at the angle $\cos^{-1}\rho$, minimization of the sum of the squared statistical distances between the paired points (x_i, y_i) and (ξ_i, η_i) is simply minimization of the sum of the squares of the actual distances between their correspondents $(x_i/\sigma_1, y_i/\sigma_2)$ and $(\xi_i/\sigma_1, \eta_i/\sigma_2)$.

It has been seen that the analogy developed between statistical distances and 'angled' Euclidean distances also provides very simple calculations of the estimates $\tilde{\xi}_i$ and $\tilde{\eta}_i$. Another outcome is that, whereas the basic model here specifies that the δ_i and ϵ_i are independent, their 'estimators', the residuals $(x_i - \tilde{\xi}_i)$ and $(y_i - b_0 - b_1\tilde{\xi}_i) = (x_i - \tilde{\xi}_i)\tan\tilde{\theta}$ are perfectly correlated. This, with the particular that the signs of the paired residuals have a fixed relationship, follows from (11 or (13 and whether or not $\sigma_{12} = 0$. This, somewhat counter-intuitive property, is simply exemplified by the values in

Fuller (1987) Tables 1.3.2 and 1.3.3, and invites further study.

Interestingly also, the fact that the estimators of the line parameters, for situations such as that for the data in Table 1.3.2, can be simply obtained via (15, the reduction of (12, shows that the point estimates are obtained without using the information that not merely their ratio but the actual values of σ_1^2 and σ_2^2 are known.

Finally here it is suggested that the convenience of the 'simple linear regression' procedure has encouraged its widespread use when a more veridical model would specify that both the x and the y observations deviate from their means on the structural relation to be estimated. One category of such examples includes the use of arbitrarily imposed, for example yearly, so-called 'Fixed', time values to index a related explanatory variable for, for example, agricultural outputs. In such cases the true explanatory value may approximate but be displaced from the assigned time-value by what may be better interpreted as a random variability element. Nor need the deviations be exclusively due to deficient measurement procedures for example technically accurate observations taken on independent individuals to relate two physiological attributes may deviate co-relatedly from their population means because of individual organic differences.

Accordingly it is this author's view that simple linear regression should preferably be exposted as a subordinate, and perhaps relatively rare, special case of the model in Section 3 which might then be appositely termed the compound linear regression model.

The review Sprent (1990) is pertinent in this whole context.

References

- Fuller, W. A. (1987). Measurement Error Models. John Wiley, New York.
- Sprent, P. (1990). Some History of Functional and Structural Relationships. In Brown, P. J. and Fuller, W. A. (Ed). Statistical Analysis of Measurement Error Models and Applications. pp. 1-15. Contemporary Mathematics, 112. Amer. Math. Soc., Providence, R.I.

Figure 1

Simple linear regression with oblique deviations

