MAKING USE OF INCOMPLETE OBSERVATIONS FOR REGRESSION IN BIVARIATE NORMAL MODEL

Joanna Tarasińska, Lublin

(Received November 15, 2000, in revised version February 6, 2001)

Abstract. Two estimates of the regression coefficient in bivariate normal distribution are considered: the usual one based on a sample and a new one making use of additional observations of one of the variables. They are compared with respect to variance. The same is done for two regression lines. The conclusion is that the additional observations are worth using only when the sample is very small.

Keywords: bivariate normal distribution, regression coefficient

MSC 2000: 62F11

1. Introduction

Let random variables (y,z) have joined binormal distribution with expectation vector $[\mu_1,\mu_2]'$ and variance-covariance matrix $\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_y^2 & \sigma_{yz}\\ \sigma_{yz} & \sigma_z^2\end{bmatrix}$. Let vectors $Y_0=[y_1,\ldots y_m]'$ and $Z_0=[z_1,\ldots z_m]'$ make a sample. The usual estimate of the regression coefficient $\beta=\sigma_{yz}/\sigma_y^2$ is $\tilde\beta=\sum\limits_{i=1}^{m}(y_i-\overline{Y}_0)(z_i-\overline{Z}_0)\Big/\sum\limits_{i=1}^{m}(y_i-\overline{Y}_0)^2$, where $\overline{Y}_0$ and $\overline{Z}_0$ are means of Y_0 and Z_0, respectively. Thus, $\tilde\beta$ is an unbiased estimate of β with variance equal to $\sigma_z^2(1-\varrho^2)/\sigma_y^2(m-3)$ [1].

Let us assume that the vector Y_0 can be enlarged by taking $k=n-m$ additional observations of the variable y. Now let us consider the estimate

$$\hat\beta=\frac{\sum\limits_{i=1}^{m}(y_i-\overline{Y}_0)(z_i-\overline{Z}_0)}{\sum\limits_{i=1}^{n}(y_i-\overline{Y})^2},\qquad \text{where } \overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i.$$

In the paper we will compare the unbiased estimates $\beta^* = \frac{n-1}{m-1}\hat{\beta}$ and $\tilde{\beta}$. We will also compare the predictors based on two regression lines $z = \overline{Z}_0 + \beta^*(y - \overline{Y})$ and $z = \overline{Z}_0 + \tilde{\beta}(y - \overline{Y}_0)$.

By $Y = \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}$ we denote the vector made of Y_0 and Y_1, where Y_1 contains the additional $n - m$ observations of the variable y. We will assume $m > 3$ throughout the paper.

2. Comparison of β^* and $\tilde{\beta}$

We see that $\hat{\beta}$ is a linear form of the quantities z_i and for a given Y it has a normal distribution with mean $E(\hat{\beta} \mid Y) = \beta \sum_{i=1}^{m} (y_i - \overline{Y}_0)^2 \Big/ \sum_{i=1}^{n} (y_i - \overline{Y})^2$ and variance $\mathrm{Var}(\hat{\beta} \mid Y) = \sigma_z^2(1 - \varrho^2) \sum_{i=1}^{m} (y_i - \overline{Y}_0)^2 \Big/ \big[\sum_{i=1}^{n} (y_i - \overline{Y})^2\big]^2$. To compute the mean and the variance of $\hat{\beta}$ we will use the formulas

$$E(\hat{\beta}) = E[E(\hat{\beta} \mid Y)], \quad \mathrm{Var}(\hat{\beta}) = E[\mathrm{Var}(\hat{\beta} \mid Y)] + E[E(\hat{\beta} \mid Y)^2] - [E(\hat{\beta})]^2.$$

Random variables $\sigma_y^{-2} \sum_{i=1}^{m} (y_i - \overline{Y}_0)^2$, $\sigma_y^{-2} \sum_{i=m+1}^{n} (y_i - \overline{Y}_1)^2$ and $\sigma_y^{-2} m(n-m)n^{-1} \times (\overline{Y}_0 - \overline{Y}_1)^2$ are independent χ^2 variables with, respectively, $m - 1$, $n - m - 1$ and 1 degrees of freedom. This implies that $b = \sum_{i=1}^{m} (y_i - \overline{Y}_0)^2 \Big/ \sum_{i=1}^{n} (y_i - \overline{Y})^2$ is the beta variable with parameters $\frac{1}{2}(m-1)$ and $\frac{1}{2}(n-m)$. Variables b and $\sum_{i=1}^{n} (y_i - \overline{Y})^2$ are independent ([2] p. 38). Thus

$$\begin{aligned} E(\hat{\beta}) &= \beta \frac{m-1}{n-1}, \\ E[E(\hat{\beta} \mid Y)^2] &= \varrho^2 \frac{\sigma_z^2}{\sigma_y^2} \cdot \frac{m^2-1}{n^2-1}, \\ E[\mathrm{Var}(\hat{\beta} \mid Y)] &= \frac{\sigma_z^2}{\sigma_y^2} \cdot \frac{m-1}{(n-1)(n-3)} \cdot (1 - \varrho^2), \end{aligned}$$

and finally we have

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma_z^2}{\sigma_y^2} \cdot \frac{m-1}{n-1} \cdot \Big(\frac{1-\varrho^2}{n-3} + \frac{2(n-m)}{n^2-1} \varrho^2 \Big).$$

Of course, $\hat{\beta}$ underestimates the parameter β so we will take into consideration the unbiased estimate $\beta^* = \frac{n-1}{m-1}\hat{\beta}$ with variance

$$\mathrm{Var}(\beta^*) = \frac{\sigma_z^2}{\sigma_y^2(m-1)} \cdot \Big[\frac{n-1}{n-3}(1 - \varrho^2) + \frac{2(n-m)}{n+1} \varrho^2 \Big].$$

Theorem 2.1. *The unbiased estimate β^* has smaller variance than $\tilde{\beta}$ if and only if*

$$\varrho^2 < \frac{n+1}{n+1+(n-3)(m-3)}. \tag{1}$$

Proof. It is sufficient to compare $\mathrm{Var}(\beta^*)$ with $\mathrm{Var}(\tilde{\beta}) = \frac{\sigma_z^2(1-\varrho^2)}{\sigma_y^2(m-3)}$. □

Now, for fixed values of m and ϱ^2 and for $n = m + k$ let us consider $\mathrm{Var}(\beta^*)$ as a function of k. To make further considerations easier let us assume that k is a continuous variable. The derivarive of $\mathrm{Var}(\beta^*)$ with respect to k is

$$[\mathrm{Var}(\beta^*)]' = \frac{2\sigma_z^2}{\sigma_y^2(m-1)(m+k-3)^2(m+k+1)^2} \cdot [ak^2 + bk + c], \tag{2}$$

where $a = \varrho^2(m+2) - 1$, $b = 2(m+1)[\varrho^2(m-2) - 1]$, $c = (m+1)[(m^2 - 5m + 10)\varrho^2 - (m+1)]$.

Corollary 2.1.

(a) *If $\varrho^2 \leqslant \frac{1}{m+2}$ then β^* is better than $\tilde{\beta}$ for each $k > 0$. The bigger k we take the smaller the variance of β^* is. We have*

$$\frac{m-3}{m-1} \leqslant \frac{\mathrm{Var}(\beta^*)}{\mathrm{Var}(\tilde{\beta})} \leqslant \frac{m-3}{m-1} \cdot \Big[\frac{m+k-1}{m+k-3} + \frac{2k}{(m+1)(m+k+1)}\Big].$$

(b) *If $\frac{1}{m+2} < \varrho^2 < \frac{1}{m-2}$ then β^* is better than $\tilde{\beta}$ for each $k > 0$. To minimize $\mathrm{Var}(\beta^*)$ the value $k = m - 1$ for $m > 4$ and $k = 2$ for $m = 4$ is recommended and then we have*

$$\frac{m-3}{m-2} + \frac{m-3}{m(m+1)} \leqslant \frac{\mathrm{Var}(\beta^*)}{\mathrm{Var}(\tilde{\beta})} \leqslant \frac{m-3}{m-2} + \frac{1}{m}.$$

(c) *If $\varrho^2 > \frac{m+2}{m^2-4m+8}$ then $\tilde{\beta}$ is better than β^* for each $k > 0$.*

Proof. (a) For $\varrho^2 < \frac{1}{m+2}$ the values a, b, c in (2) are less than 0 so the variance of β^* decreases for each $k > 0$. We get the inequalities for the quotient $\frac{\mathrm{Var}(\beta^*)}{\mathrm{Var}(\tilde{\beta})}$ by putting $\varrho^2 = \frac{1}{m+2}$ for the right inequality and $\varrho^2 = 0$, $k = \infty$ for the left one.

(b) When ϱ^2 is between $\frac{1}{m+2}$ and $\frac{1}{m-2}$ we have $a > 0$, $b \leqslant 0$, $c < 0$ in formula (2). So $\mathrm{Var}(\beta^*)$ decreases for each k less than a certain value k_0. It increases when $k > k_0$ but it stays less than $\mathrm{Var}(\tilde{\beta})$ because $\lim_{k\to\infty} \mathrm{Var}(\beta^*) = \frac{\sigma_z^2(1+\varrho^2)}{\sigma_y^2(m-1)} < \frac{\sigma_z^2(1-\varrho^2)}{\sigma_y^2(m-3)}$. The minimum value of $\mathrm{Var}(\beta^*)$ is achieved when we take k as an integer closest to k_0.

Of course k_0 depends on the unknown ϱ^2. The closer ϱ^2 to $\frac{1}{m+2}$ is the greater k_0 is. When ϱ^2 tends to $\frac{1}{m-2}$, the optimal value k_0 tends to be in the interval $(m-2, m-1)$.

(c) Now the question is: how large must ϱ^2 be to be sure that $\tilde{\beta}$ is better than β^* for all $k > 0$? Transforming the inequality (1) with respect to k we get $k < \frac{m+1-\varrho^2(m^2-5m+10)}{\varrho^2(m-2)-1}$. □

Unfortunately, $\frac{m+2}{m^2-4m+8}$ converges to zero rather quickly with $m \to \infty$. Thus the improvement of $\tilde{\beta}$ is possible only for small m and ϱ^2. In Table 1 we have the optimal values of k_0 (for which $\mathrm{Var}(\beta^*)$ achieves the minimum) and the corresponding values of $\frac{\mathrm{Var}(\beta^*)}{\mathrm{Var}(\tilde{\beta})}$ for different m and ϱ^2. The calculations were made with Maple V Release program.

k_0 $\frac{\mathrm{Var}(\beta^*)}{\mathrm{Var}(\tilde{\beta})}$	$\varrho^2 = 0.1$	$\varrho^2 = 0.2$	$\varrho^2 = 0.3$	$\varrho^2 = 0.4$	$\varrho^2 = 0.5$
$m = 5$	∞ 0.611	15 0.737	4 0.838	2 0.917	1 0.976
$m = 6$	∞ 0.733	8 0.869	2 0.954		
$m = 8$	∞ 0.873	0.982			
$m = 10$	∞ 0.951				
$m = 12$	10 0.983				

Table 1. The optimal values of k_0 and the corresponding values of $\frac{\mathrm{Var}(\beta^*)}{\mathrm{Var}(\tilde{\beta})}$.

3. Comparison between two prediction equations

Let us consider the predicted value of z for a given y of the form $z^* = \overline{Z}_0 + \beta^*(y - \overline{Y})$ and let us compare it with the predicted value based only on complete pairs of observations, i.e. $\tilde{z} = \overline{Z}_0 + \tilde{\beta}(y - \overline{Y}_0)$. The variance of the predictor z^* for a given value y $(y \neq y_i,\ \ i = 1, \ldots, n)$ is

$$\mathrm{Var}(z^*) = \frac{\sigma_z^2}{\sigma_y^2}[a_1(y - \mu_1)^2 + d_1], \tag{3}$$

where $a_1 = \frac{1}{m-1}\left[\frac{n-1}{n-3}(1-\varrho^2) + 2\frac{n-m}{n+1}\varrho^2\right]$, $d_1 = \frac{\sigma_y^2}{n}\left[a_1 + \frac{n-m\varrho^2}{m}\right]$. The variance of the predictor $\tilde{z}$ is (it is enough to put $n = m$ in formula (3))

$$\mathrm{Var}(\tilde{z}) = \frac{\sigma_z^2}{\sigma_y^2}[a_2(y - \mu_1)^2 + d_2],$$

where $a_2 = \frac{1-\varrho^2}{m-3}$, $d_2 = \sigma_y^2 a_2 \frac{m-2}{m}$. The question is: when $\mathrm{Var}(z^*) < \mathrm{Var}(\tilde{z})$? Let us denote

$$\begin{aligned} a &= a_1 - a_2, \\ d &= d_1 - d_2, \\ \varrho_1^2 &= \frac{n+1}{n+1+(n-3)(m-3)}, \\ \varrho_2^2 &= \frac{n+1}{n+1+\left(1+4\frac{m-1}{nm+3-n-m}\right)(n-3)(m-3)}. \end{aligned}$$

Theorem 3.1.

(a) *If $\varrho^2 < \varrho_2^2$ then for each value of y the predictor z^* is better than $\tilde{z}$.*

(b) *If $\varrho^2 > \varrho_1^2$ then for each value of y the predictor $\tilde{z}$ is better than z^*.*

(c) *If $\varrho_2^2 < \varrho^2 < \varrho_1^2$ then $\tilde{z}$ is better for values of y such that $|y - \mu_1| < \sqrt{-d/a}$ and z^* is better for the other y's.*

Proof. $\mathrm{Var}(z^*) < \mathrm{Var}(\tilde{z}) \Leftrightarrow a(y-\mu_1)^2 + d < 0$. Let us notice that

$$\begin{aligned} a < 0 \Leftrightarrow \varrho^2 < \varrho_1^2 &= \frac{n+1}{n+1+(n-3)(m-3)}, \\ d < 0 \Leftrightarrow \varrho^2 < \varrho_2^2 &= \frac{n+1}{n+1+\left(1+4\frac{m-1}{nm+3-n-m}\right)(n-3)(m-3)}. \end{aligned}$$

Because $\varrho_2^2 < \varrho_1^2$ so three cases are possible: $(a < 0,\ d < 0)$, $(a < 0,\ d > 0)$ and $(a > 0,\ d > 0)$. □

Corollary 3.1.

(a) *If $\varrho^2 < \frac{1}{m-2}$ then, for each k (where $k = n - m$) and for each y, z^* is better than $\tilde{z}$.*

(b) *If $\varrho^2 > \frac{m+2}{m^2-4m+8}$ then, for each k and for each y, $\tilde{z}$ is better than z^*.*

Proof. It is enough to note that ϱ_2^2 and ϱ_1^2 are decreasing functions of k, $\lim_{k\to\infty} \varrho_2^2 = \frac{1}{m-2}$ and $\varrho_1^2 = \frac{m+2}{m^2-4m+8}$ for $k = 1$. □

Thus only for small values of m the predictor z^* which utilizes all available observations can be better than $\tilde{z}$.

Example: $m = 6$.

a) If $\varrho^2 < 0.25$ then, for each k and for each y, z^* is better than $\tilde{z}$.

b) If $\varrho^2 > 0.4$ then, for each k and for each y, $\tilde{z}$ is better than z^*.

c) If $\varrho^2 = 0.3$ then Table 2 shows some k's and the corresponding values of y for which z^* is better than $\tilde{z}$.

k	1	2	3	4	5
$\lvert y-\mu_1\rvert$	$> 0.22\sigma_y^2$	$> 0.37\sigma_y^2$	$> 0.49\sigma_y^2$	$> 0.60\sigma_y^2$	$> 0.72\sigma_y^2$

Table 2. The values of y for which z^* is better than $\tilde{z}$ for different k, $m = 6$, $\varrho^2 = 0.3$.

References

[1] *H. Cramer*: Mathematical Methods of Statistics. Princ. Univ. Press, Princeton, 1966.

[2] *N. L. Johnson, S. Kotz*: Continuous Univariate Distributions–2. Houghton Mifflin Company, Boston, 1970.

Author's address: *J. Tarasińska*, Department of Applied Mathematics, University of Agriculture, Akademicka 13, 20-934 Lublin, Poland, email: johata@ursus.ar.lublin.pl.