RMISDAT: A computer program to estimate the population correlation coefficient based on incomplete data sets

ULBE BROUWER and PIETER VIJN
*University of Amsterdam,
Amsterdam 1018 WB, The Netherlands*

Several estimates of the population correlation coefficient, based on incomplete data sets, are given in Boas (1969) and Dixon and Brown (1977). In the RMISDAT program presented here, various Bayesian estimates of the population correlation are computed. These estimates are based on likelihood functions given by Mehta and Swamy (1975). Apart from Bayesian estimates, product-moment correlations based on complete observation pairs and the so-called Pearson correction estimate (in restriction of range, Case 1) are also computed.

The assumption is that the variables X and Y have been measured at interval level and are normally distributed. Four Bayesian estimators can be defined optionally, namely:

$$\max_{\rho} \text{Prob}(\rho | k = \hat{k}, \mu_x = \bar{X}, \mu_y = \bar{Y}, \text{data}), \quad (\text{rb1}) \quad (1)$$

$$\max_{\rho} \text{Prob}(\rho | \mu_x = \bar{X}, \mu_y = \bar{Y}, \text{data}), \quad (\text{rb2}) \quad (2)$$

$$\max_{\rho} \text{Prob}(\rho | k = \hat{k}, \mu_y = \bar{Y}, \text{data}), \quad (\text{rb3}) \quad (3)$$

and

$$\max_{\rho} \text{Prob}(\rho | \mu_y = \bar{Y}, \text{data}), \quad (\text{rb4}) \quad (4)$$

where $\hat{k} = S_y^2/S_x^2$ = quotient of the sample variances of variables Y and X, respectively; $\bar{X}$ = sample mean of variable X; and $\bar{Y}$ = sample mean of variable Y. For the formulas, refer to Brouwer and Vijn (Note 1).

A priori knowledge concerning ρ can be expressed formally by specification of a priori distribution for ρ . The prior for ρ (in RMISDAT) is as follows:

$$\text{Prob}(\rho) \sim (\rho - \rho_1)^{a-1} (\rho_u - \rho)^{b-1},$$

where ρ_1 and ρ_u are the a priori possible lower and upper bounds for ρ .

The parameters a and b represent the strength of the a priori knowledge regarding ρ . For $a = b = 1$, one opts for a uniform prior for ρ , which means that every value for ρ on the interval (ρ_1, ρ_u) is equally likely. For a mildly symmetrical prior for ρ , $a = b = 3$ is sufficient.

U. Brouwer is with the Technical Center, Department of Social Sciences; P. Vijn is with the Psychological Laboratory.

The prior distribution for the parameters μ_x , μ_y , $\ln(\sigma_x^2)$, and $\ln(\sigma_y^2)$ is uniform in RMISDAT. The assumption is that the various priors are mutually independent.

Optionally, the program computes the Pearson correction estimate, R:

$$R = \frac{(S_x/s_x)r}{\sqrt{1 + r^2 [(S_x^2/s_x^2) - 1]}},$$

where S_x = the sample variance of variable X; s_x = the sample variance of variable X (with $X < X_0$); and r = the sample correlation between X and Y (for the complete pairs).

The estimator is often selected in the so-called restriction of range (Case 1) situation (see Greener & Osburn, 1979; Gullickson & Hopkins, 1976). This means that the possible range in values for variable Y is restricted to, for instance, all values Y_i for which $X_i < X_0$, where X_0 is a presumably known critical cutting score on, for instance, a test. Simulation research (Brouwer & Vijn, 1978) has shown that in random missing-data patterns, this estimator also gives better results when the population correlation is reasonably high than does the classical product-moment correlation.

The possibility exists that an estimated correlation matrix based on several cases is not positive definite. It is optionally possible in RMISDAT to "smooth" this matrix, which is written on an alternate output file. This means the matrix is converted into a positive-definite one, as follows: $B = ADAT^T$, where A = the n by n matrix having eigen vectors of the original correlation matrix (R) and D = a diagonal matrix with the eigen values of R. All negative or zero eigen values are reset to .01.

The resulting covariance matrix B can be converted into a correlation matrix. Smoothing makes it possible to execute a standard multivariate analysis.

Some Remarks. Only one prior can be chosen for ρ or for Σ (the correlation matrix). This will not present problems in the bivariate case, but it will in the multivariate case. The researcher cannot always rely on prior estimates of correlations between variables, and, besides, these prior estimates can be very dissimilar. For this reason, the uniform prior over -1 and $+1$ has been chosen as default prior for ρ . In later RMISDAT versions, a separate prior will be introduced for every correlation to be estimated.

The Bayesian estimates rb2 and rb4 are computed by means of a numerical integration method (DCADRE from the IMSL library). It may occur that DCADRE prints a warning in the output that computation of the integral is difficult. Careful inspection of the results is then advisable.

In the bivariate case, plots are given optionally of the density function and the cumulative distribution for ρ . This makes it relatively simple to give a verdict on ρ (e.g., the probability that $\rho < .75$ is .95, etc.).

The raw data are transformed by the program into values between -1 and $+1$. This is so the function values of the density function will stay within an admissible range. This transformation does not influence the finally obtained estimates.

The maximum allowed number of cases is 100. Simulation research (Brouwer & Vijn, 1979) has shown that the advantage of using Bayesian estimators, in the case of larger numbers, even when the amount of missing data is relatively high (e.g., 40%), is no longer of any practical value.

The "smoothing" procedure is still in an experimental phase. More simulation or Monte Carlo studies will be needed to prove the usefulness of this method.

Availability. The program RMISDAT is available in the form of a card deck (or magnetic tape) with the FORTRAN IV source text at the Technical Center, Roetersstraat 15, Amsterdam 1018 WB, The Netherlands. It has been tested on a Cyber CDC 73 computer.

REFERENCE NOTE

1. Brouwer, U., & Vijn, P. *Missing data and Bayesian estimators for the correlation coefficient*. Unpublished report, University of Amsterdam, Technical Centre FSW, 1980.

REFERENCES

- BOAS, J. A note on the estimation of the covariance between two random variables using extra information in the separate variables. *Statistica Neerlandica*, 1969, **21**, 120-122.
- BROUWER, U., & VIJN, P. De empirische power van 2 procedures in "restriction of range" (case I). *Tijdschrift voor Onderwijs Research*, 1978, **3**, 106-112.
- BROUWER, U., & VIJN, P. Bayesiaanse schatters voor de correlatie coefficient in "restriction of range" (geval 1). *Tijdschrift voor Onderwijs Research*, 1979, **6**, 281-290.
- DIXON, W. J., & BROWN, M. D. (Eds.). *BMDP-77 manual*. Los Angeles: University of California Press, 1977.
- GREENER, J. M., & OSBURN, H. G. An empirical study of the accuracy of corrections for restriction of range due to explicit selection. *Applied Psychological Measurement*, 1979, **3**, 31-41.
- GULLICKSON, A., & HOPKINS, K. Interval estimation of correlation coefficients corrected for restriction of range. *Educational and Psychological Measurement*, 1976, **36**, 9-25.
- MEHTA, J. S., & SWAMY, P. Bayesian analysis of a bivariate normal distribution where some observations are missing. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics*. Amsterdam: North Holland, 1975.

(Accepted for publication May 30, 1980.)