MVA

The Missing Value procedure provides descriptions of missing value patterns; estimates of means, standard deviations, covariances, and correlations (using a listwise, pairwise, EM, or regression method); and imputation of values by either EM or regression.

Notation

The following notation is used throughout this chapter unless otherwise noted:

$\mathbf{X}$	Data matrix
x_{ij}	Value of the *i*th case, *j*th variable
v	Number of variables
n	Number of cases
n_i	Number of nonmissing values of the *i*th variable
n_{ij}	Number of nonmissing value pairs of the *i*th and *j*th variables
n_c	Number of complete cases
J	Index of all variables
$J_{\#} = J(\text{condition})$	Index of variables satisfying “condition”
I	Index of all cases
$I(k_1,\ldots,k_l)$	Index of cases at which variables $(k_1,\ldots,k_l)$ are not missing
$I(J)$	Index of complete cases
$\mathbf{a}=[a_i]$	Vector whose *i*th element is a_i
$\mathbf{A}=[a_{ij}]$	Matrix whose *i*th row, *j*th column element is a_{ij}

Example to Illustrate Notation

$$\mathbf{X} = \begin{bmatrix} 43 & 76 & 34 \\ . & 45 & 72 \\ 44 & 15 & 52 \\ . & . & 65 \\ . & . & 43 \\ 54 & 12 & . \\ 43 & 67 & 34 \end{bmatrix}$$

$x_{2,3} = 72$	The 2nd row, 3rd element
$v = 3$	Number of variables
$n = 7$	Number of cases
$n_2 = 5$	Number of nonmissing values in the 2nd variable
$n_{2,3} = 4$	Number of nonmissing value pairs in the 2nd and 3rd variables
$n_c = 3$	Number of complete cases
$J = \{1,2,3\}$	Index of variables
$J(2 \text{ or more missing}) = \{1,2\}$	The 1st and 2nd variables have two or more missing values
$I = \{1,2,3,4,5,6,7\}$	Index of cases
$I(2) = \{1,2,3,6,7\}$	Index of cases at which the 2nd variable is not missing
$I(2,3) = \{1,2,3,7\}$	Index of cases at which the 2nd and 3rd variables are not missing
$I(J) = \{1,7\}$	Index of complete cases
$\bar{x}_2 = 43.0$	The 2nd element of the vector $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \bar{x}_3]$

Univariate Statistics

The index j refers to quantitative variables.

Mean

$$\bar{\mathbf{x}} = [\bar{x}_j] = \left[\sum_i x_{ij} / n_j;\ i \in I(j)\right]$$

Standard Deviation

$$\hat{\sigma} = \left[\hat{\sigma}_j\right] = \left[\left(\sum_i \left(x_{ij} - \bar{x}_j\right)^2 / \left(n_j - 1\right)\right)^{1/2}; \quad i \in I(j)\right]$$

Extreme Low

$$\text{NL} = \left[\text{nl}_j\right] = \left[\text{number of } x_{ij} \text{ values } < \text{ low\_limit}_j\right]$$

Extreme High

$$\text{NH} = \left[\text{nh}_j\right] = \left[\text{number of } x_{ij} \text{ values } > \text{ high\_limit}_j\right]$$

where

$$\text{low\_limit}_j = \begin{cases} \bar{x}_j - 2 * \hat{\sigma}_j & \text{if} \quad v * n * \log_{10}(n) > 150{,}000 \\ 25th \text{ percentile of the } j\text{th varible} & \text{if} \quad v * n * \log_{10}(n) \le 150{,}000 \end{cases}$$

and

$$\text{high\_limit}_j = \begin{cases} \bar{x}_j + 2 * \hat{\sigma}_j & \text{if} \quad v * n * \log_{10}(n) > 150{,}000 \\ 75th \text{ percentile of the } j\text{th variable} & \text{if} \quad v * n * \log_{10}(n) \le 150{,}000 \end{cases}$$

Separate Variance T Test

The index k refers to quantitative variables, and index j refers to all variables.

$$t_{jk} = \frac{\bar{x}_{jk}^{P} - \bar{x}_{k|\text{variable } j \text{ is missing}}}{\left(\frac{\hat{\sigma}_{jk}^{P}}{n_{jk}} + \frac{\hat{\sigma}_{k|\text{variable } j \text{ is missing}}}{n_{kk} - n_{jk}} \right)^{1/2}}$$

where $\bar{x}_{jk}^{P}$ and $\hat{\sigma}_{jk}^{P}$ are defined below in **Pairwise Statistics**.

$$\text{df}_{jk} = \frac{\left(\frac{\hat{\sigma}_{jk}^{P}}{n_{jk}} + \frac{\hat{\sigma}_{k|\text{variable } j \text{ is missing}}}{n_{kk} - n_{jk}} \right)^{2}}{\frac{\left(\hat{\sigma}_{jk}^{P}\right)^{2}}{n_{jk} - 1} + \frac{\left(\hat{\sigma}_{k|\text{variable } j \text{ is missing}}\right)^{2}}{n_{kk} - n_{jk} - 1}} \quad p(2\text{-tail})_{jk} = 1 - 2 * \left| 0.5 - \text{tcdf}\left(t_{jk}, \text{df}_{jk}\right) \right|$$

where “tcdf” is the t cumulative distribution function

Listwise Statistics

The indices j and k refer to quantitative variables.

Mean

$$\bar{\mathbf{x}}^{L} = \left[\bar{x}_{j}^{L}\right] = \left[\sum_{i} x_{ij} / n_{c};\ i \in I(J)\right]$$

Covariance

$$\mathbf{C}^{L} = \left[c_{jk}^{L}\right] = \left[\sum_{i} \left(x_{ij} - \bar{x}_{j}^{L}\right) * \left(x_{ik} - \bar{x}_{k}^{L}\right) / \left(n_{c} - 1\right); \quad i \in I(J)\right]$$

Correlation

$$\mathbf{R}^L = \left[r_{jk}^L\right] = \left[c_{jk}^L / \left(c_{jj}^L * c_{kk}^L\right)^{1/2}\right]$$

Pairwise Statistics

The indices *j* and *k* refer to quantitative variables, and *l* refers to all variables.

Mean

$$\overline{\mathbf{X}}^P = \left[\overline{x}_{lk}^P\right] = \left[\sum_i x_{ik} / n_{lk};\ i \in I(l,k)\right]$$

Standard Deviation

$$\hat{\boldsymbol{\sigma}}^P = \left[\hat{\sigma}_{lk}^P\right] = \left[\left(\sum_i \left(x_{ik} - \overline{x}_{lk}^P\right)^2 / \left(n_{lk} - 1\right)\right)^{1/2}; \quad i \in I(l,k)\right]$$

Covariance

$$\mathbf{C}^P = \left[c_{jk}^P\right] = \left[\sum_i \left(x_{ik} - \overline{x}_{jk}^P\right) * \left(x_{ij} - \overline{x}_{kj}^P\right) / \left(n_{jk} - 1\right); \quad i \in I(j,k)\right]$$

Correlation

$$\mathbf{R}^P = \left[r_{jk}^P\right] = \left[c_{jk}^P / \left(\hat{\sigma}_{jk}^P * \hat{\sigma}_{kj}^P\right)\right]$$

Regression Estimated Statistics

The indices *j* and *k* refer to quantitative variables, and *l* refers to predictor variables.

Estimates of Missing Values

$$x_{ij}^R = \begin{cases} x_{ij} & \text{if } x_{ij} \text{ is not missing} \\ \text{regression estimated } x_{ij} & \text{if } x_{ij} \text{ is missing} \end{cases}$$

Regression Estimated x_{ij}

$$x_{ij}^R = \beta_{0,ij} + \sum_l \beta_{l,ij} * x_{il} + \varepsilon_{ij} \qquad l \in J_1 = J\left(l\colon x_{il} \text{ not missing and } l \neq j\right)$$

where:

- $\left[\beta_{0,ij}, \beta_{l,ij}\right]$ is computed from $\text{Diag}\left(\overline{\mathbf{X}}^P\right) = \left[\bar{x}_{jj}^P\right]$

  and by pivoting on the “best” “q” of the J_1 diagonals of $\mathbf{C}^P$.

- “best” is forward stepwise selected.

- “q” is less than or equal to the user-specified maximum number of predictors; it may also be limited by the user-specified *F*-to-enter limit.

- “ε_{ij}” is the optional random error term, as specified:

  i. residual of a randomly selected complete case

  ii. random normal deviate, scaled by the standard error of estimate

  iii. random t(df) deviate, scaled by the standard error of estimate, df is specified by the user

  iv. no error term adjustment

Note that for each missing value x_{ij}, a unique set of regression coefficients $\left(\beta_{0,ij}, \beta_{l,ij}\right)$ and error terms ε_{ij} is computed.

Mean

$$\overline{\mathbf{x}}^R = \left[\overline{x}_j^R\right] = \left[\sum_i x_{ij}^R / n; \qquad i \in I\right]$$

Covariance

$$\mathbf{C}^R = \left[c_{jk}^R\right] = \left[\sum_i \left(x_{ij}^R - \overline{x}_j^R\right) * \left(x_{ik}^R - \overline{x}_k^R\right) / (n-1); \qquad i \in I\right]$$

Correlation

$$\mathbf{R}^R = \left[r_{jk}^R\right] = \left[c_{jk}^R / \left(c_{jj}^R * c_{kk}^R\right)^{1/2}\right]$$

EM Estimated Statistics

The indices *j* and *k* refer to quantitative variables, and *l* refers to predictor variables.

Estimates of Missing Values, Mean Vector, and Covariance Matrix

$$\overline{\mathbf{x}}_0 = \left[\overline{x}_j^0\right] = \mathrm{Diag}\left(\overline{\mathbf{X}}^P\right) = \left[\overline{x}_{jj}^P\right]$$
$$\mathbf{C}_0 = \left[c_{jk}^0\right] = \mathbf{C}^P = \left[c_{jk}^P\right]$$

For $m = 1$ to M, or Until Convergence Is Attained

If x_{ij} is not missing then $x_{ij}^m = x_{ij}$.

If x_{ij} is missing then it is estimated in the *m*th iteration as:

$$x_{ij}^m = \beta_{0,ij}^{m-1} + \sum_l \beta_{l,ij}^{m-1} * x_{il}; \qquad l \in J_2 = J\left(l{:}\, x_{il} \text{ is not missing and } l \neq j\right)$$

where $\left[\beta_{0,ij}^{m-1}, \beta_{l,ij}^{m-1}\right]$ is computed from $\overline{\mathbf{x}}_{m-1}$ and $\mathbf{C}_{m-1}$.

$$\overline{\mathbf{x}}_m = \left[\overline{x}_j^m\right] = \left[\sum_i w_i * x_{ij}^m \,/ \sum_i w_i; \qquad i \in I\right]$$

$$\mathbf{C}_m = \left[c_{jk}^m\right] = \left[\frac{\sum_i w_i * x_{ij}^m \left(x_{ij}^m - \overline{x}_j^m\right) * \left(x_{ik}^m - \overline{x}_k^m\right) + \sum_i \sum_s c_{j,s|J2}^{m-1}}{(n-1) * \sum_i w_i \,/\, n}; \; i \in J_2,\, s \notin J_2, \text{ and } s \neq j\right]$$

where $c_{j,s|J2}^{m-1}$ is the *j*th row, *s*th element of the J_2 pivoted $\mathbf{C}_{m-1}$.

Note that some sources (Little & Rubin, 1987, for example) simply use *n* as the denominator of the formula for $\mathbf{C}_m$, which produces full maximum likelihood (ML) estimates. The formula used by MVA produces restricted maximum likelihood (REML) estimates, which are *n*/(*n*-1) times the ML estimates.

$$w_i = \begin{cases} 1 & \text{for multivariate normal} \\ \dfrac{1-\alpha+\alpha*\lambda^{1+p/2}*\exp\left((1-\lambda)*D^2/2\right)}{1-\alpha+\alpha*\lambda^{p/2}*\exp\left((1-\lambda)*D^2/2\right)} & \text{for contaminated normal} \\ (\mathrm{df}+p)/\left(\mathrm{df}+D^2\right) & \text{for } t(\mathrm{df}) \end{cases}$$

$\alpha =$ **proportion** of contamination

$\lambda =$ **ratio** of standard deviations

$p =$ number of predictors = number of indices in J_2

$D^2 =$ Mahalanobis distance square of the current case from the mean

$$= \sum_{jk} \left(x_{ij}^m - \bar{x}_j^m\right) * \left(c_{jk}^m\right)^{-1} * \left(x_{ik}^m - \bar{x}_k^m\right)$$

where $\left(c_{jk}^m\right)^{-1}$ is the *jk*th element of $\mathbf{C}_m{}^{-1}$.

Convergence

The algorithm is declared to have converged if, for all *j*,

$$\left|c_{jj}^m - c_{jj}^{m-1}\right| / c_{jj}^m \le \text{CONVERGENCE}$$

Filled-In Data

$$\mathbf{X}_i^E = \left[x_{ij}^E\right] = \left[x_{ij}^{m'}\right]$$

where m' is the last value of m.

Mean

$$\bar{\mathbf{x}}^E = \left[\bar{x}_j^E\right] = \bar{\mathbf{x}}_{m'} = \left[\bar{x}_j^{m'}\right]$$

Covariance

$$\mathbf{C}^{E} = \left[c_{jk}^{E}\right] = \mathbf{C}_{m'} = \left[c_{jk}^{m'}\right]$$

Correlation

$$\mathbf{R}^{E} = \left[r_{jk}^{E}\right] = \left[c_{jk}^{E} / \left(c_{jj}^{E} * c_{kk}^{E}\right)^{1/2}\right]$$

Little's MCAR Test

$$\chi^2_{\text{MCAR}} = \sum\nolimits_{\text{each unique pattern}} (\text{no. of cases in pattern}) * \left(\text{Mahalanobis } D^2 \text{ of pattern mean from } \bar{\mathbf{x}}^{E}\right)$$

$$\text{DF}_{\text{MCAR}} = \sum\nolimits_{\text{each unique pattern}} (\text{no. of nonmissing variables}) - v$$

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, B*,. 39: 1–38.

Dixon, W. J., ed. 1983. *BMDP statistical software*. Berkeley: University of California Press.

Little, R. J. A. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83: 1198–1988.

Little, R. J. A. and Rubin, D. B. 1987. *Statistical analysis with missing data.* New York: John Wiley & Sons, Inc.

Louise, T. A. 1982. Finding the observed information matrix when using the *EM* algorithm. *Journal of the Royal Statistical Society, B,* 44(2): 226–233.

Orchard, T. and Woodbury, M. A. 1972. Missing information principal: Theory and applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1. Berkeley: University of California Press, 697–715.

Rubin, D. B. 1987. Multiple imputation for nonresponse data in surveys. New York: John Wiley & Sons, Inc.