Scientific Software International

Types of variables

PRELIS can deal with three types of variables: continuous, ordinal and censored.

Continuous variable

Observations are assumed to come from an interval or a ratio scale and to have metric properties. Means, variances, and higher moments of these variables will be computed in the usual way.

Ordinal variable

Observations are assumed to represent responses to a set of ordered categories, such as a five-category Likert scale. Here, it is only assumed that a person who responds in one category has more of a characteristic than a person who responds in a lower category. For each ordinal variable x, it is assumed that there is a latent continuous variable ξ that is normally distributed with mean zero and unit variance. The assumption of normality is not testable given only x; but for each pair of variables where x is involved, PRELIS attempts a test of the assumption of bivariate normality.

Assuming that there are k categories on x, we write $x = i$ to mean that x belongs to category i. The actual score values in the data may be arbitrary and are irrelevant as long as the ordinal information is retained. That is, low scores correspond to low-order categories of x that are associated with smaller values of ξ, and high scores correspond to high-order categories that are associated with larger values of ξ.

The connection between x and ξ is that $x = \xi$ is equivalent to $\alpha_{i-1} < \xi \le \alpha_i$, where $\alpha_0 = -\infty, \alpha_1 < \alpha_2 < \ldots < \alpha_{k-1}$ and $\alpha_k = +\infty$ are parameters called threshold values. If there are k categories, there are $k - 1$ unknown thresholds.

Censored variable

Variable x represents a latent variable ξ observed on an interval scale above a threshold value A. Below A, the value $x = A$ is observed:

$$
\begin{aligned}
x &= \xi \quad if \quad \xi < A \\
x &= A \quad if \quad \xi \le A.
\end{aligned}
$$

The value A is known and is equal to the smallest observed value of x. The latent variable ξ is assumed to be normally distributed with unknown mean μ and standard deviation σ, which are estimated by the maximum likelihood method.

The censored variable just defined will be said to be *censored below*. PRELIS can also deal with variables that are *censored above*:

$$x = \xi \quad if \quad \xi < B$$
$$x = B \quad if \quad \xi \geq B.$$

Variables that are censored *both above and below* are handled by PRELIS.

Censored variables have a high concentration of cases at the lower or upper end of the distribution. The classical example of this is in Tobit analysis where, for example, x = the price of an automobile purchased in the last year, with $x = 0$ if no car was purchased. Here ξ may represent a propensity to consume capital goods. Other examples may be x = number of crimes committed or x = number of days unemployed. Test scores that have a "floor" or a "ceiling" (a large proportion of cases with no items or with all items correct) are censored variables. Attitude questions where a large fraction of the population is expected to have the lowest or highest score or category may also be considered censored variables.

A key concept in the way PRELIS treats ordinal and censored variables is the use of normal scores.

For an ordinal variable, let n_j be the number of cases in the j-th category. The threshold values are estimated from the (marginal) distribution of each variable as

$$\hat{\alpha}_i = \Phi^{-1}\left(\sum_{j=1}^{i} n_j / N\right) \quad i = 1, 2, \ldots, k-1$$

where Φ^{-1} is the inverse standard normal distribution function, and N is the total number of real observations on the ordinal variable.

The *normal score* z_i corresponding to $x = i$ is the mean of ξ in the interval $\alpha_{i-1} < \xi \leq \alpha_i$, which is (see Johnson & Kotz, 1970, pp.81-82)

$$z_i = \frac{\phi(\alpha_{i-1}) - \phi(\alpha_i)}{\Phi(\alpha_i) - \Phi(\alpha_{i-1})}$$

where ϕ and Φ are the standard normal density and distribution function, respectively. This normal score can be estimated as:

$$\hat{z}_i = (N / n_i)[\phi\left(\hat{\alpha}_{i-1}\right) - \phi\left(\hat{\alpha}_i\right)]$$

As can be readily verified, the weighted mean of the normal scores is 0.

For a variable censored below A, PRELIS uses the normal score associated with the interval $\xi \leq A$, which is

$$\hat{z}_A = \hat{\mu} - \frac{\phi[(A - \hat{\mu}) / \hat{\sigma}]}{\Phi[(A - \hat{\mu}) / \hat{\sigma}]} \hat{\sigma}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimates of μ and σ.

For a variable censored above *B*, the normal score associated with the interval $\xi \geq B$ is:

$$\hat{z}_A = \hat{\mu} + \frac{\phi[(B - \hat{\mu}) / \hat{\sigma}]}{\Phi[(B - \hat{\mu}) / \hat{\sigma}]} \hat{\sigma}$$

References

Johnson, N.I. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, New York: John Wiley & Sons.