CORRELATION ESTIMATOR FOR UNBALANCED DATA

Jai W. Choi, National Center for Health Statistics, J. Richard Landis,
University of Pennsylvania School of Medicine
Jai W. Choi, National Center for Health Statistics, 6525 Belcrest
Road, Hyattsville MD 20782

KEY WORDS: Intra -cluster correlation, Categorical data, Estimation, Variance, Testing.

1. INTRODUCTION

In this paper, we present the correlations for inter- and intra-levels and for intra-cluster, and show its closed forms of variance derived by Delta method. In Section 2, we define a population and three types of correlation. Section 4 introduces sample estimate of these correlations and its variance. In Section 4, asymptotic distributions are discussed. Finally in Section 5, we present a drug test example.

2. POPULATION

The population U of interest is decomposed into A clusters $U = (U_1, \ldots, U_i, \ldots, U_A)$. The cluster U_i consists of B_i final units as $U_i = (U_{i1}, \ldots, U_{ij}, \ldots, U_{iBi})$. Within the final unit, U_{ij}, the generic measurement level will be denoted by h, with a total of r levels or cells, and the cells are mutually exclusive and exhaustive. Total number of population units is $N = \Sigma_{i \epsilon A} B_i$.

Let the random variable y_{hij} be the realization of the unit U_{ij}, continuous or discrete, and define a model for y_{hij} as

$$y_{hij} = \Pi_h + \beta_{hi} + e_{hij}$$

for $h = 1, \ldots, r$, $i = 1, \ldots, A$, and $j = 1, \ldots, B_i$.

We assume the random variables β_{hi}'s and e_{hij}'s are uncorrelated and the existence of the first two moments of y_{hij}.s. The expectation and variance are expressed as $E(y_{hij}) = \Pi_h$ and $\mathrm{var}(y_{hij}) = \sigma^2_{bhh} + \sigma^2_{ehh}$.

2.1. DEFINITION OF CORRELATION

We assume that the clusters are independent. But any two units in the same cluster are correlated by the common intra cluster correlation matrix R with ρ_{hh} on the diagonal for h = h' and $\rho_{hh'}$ on the off-diagonal for h ≠ h':

$$\rho_{hh'} = \left| \begin{array}{l} \dfrac{\sigma^2_{\beta hh}}{\sigma^2_{\beta hh} + \sigma^2_{ehh}} \\[2ex] \dfrac{\sigma^2_{\beta hh'}}{\sqrt{\left(\sigma^2_{\beta hh} + \sigma^2_{ehh}\right) \times \left(\sigma^2_{\beta h'h'} + \sigma^2_{eh'h'}\right)}} \end{array} \right.$$

The correlation over all cells is defined:

$$\rho = \frac{\Sigma_h \, \sigma^2_{\beta hh}}{\Sigma_h \, (\sigma^2_{\beta hh} + \sigma^2_{ehh})}$$

The sum of the trace in nonsingular matrix is the same as the sum of its eigenvalues, that is the sum of diagonal elements. Thus, the sum of diagtonal elements is used for the estimation of overall correlation. This definition provides only positive correlation, which arises in most practical situations. Negative correlation sometimes occurs in actual data and we might set $\rho = 0$ for conservative inferences.

2.2. CATEGORICAL VARIABLE

Let the random variable $y_{hij} = 1$ if the U_{ij}-th unit falls into the h-th cell, and = 0 otherwise.

Let $y_h = \Sigma_{i=1,A} \Sigma_{j=1,B_i} y_{hij}$

be the population counts in the h-th cell, and the cells are mutually exclusive and exhaustive and

$N = \Sigma_{h=1,r}\, y_h$ be the entire units in the population. Denote the vector of population cell frequencies by $y^T = (y_1, .., y_h, .., y_r)$ and the corresponding proportions by vector $\Pi^T = (\Pi_1, .., \Pi_h, .., \Pi_r)$. We impose some restrictions on these proportions: $\Pi_h = y_h/N$, $\Pi_h > 0$ for all h, and $\Sigma_h \Pi_h = 1$.

Define pairwise probability for the members in a same cluster as

$p(y_{hij}=y_{h'i'j'}=1)=\delta_{hh'}$ if i=i' j≠j' h≠h',

$p(y_{hij}=y_{h'i'j'}=1)=\delta_{hh}$ if i=i' j≠j' h=h',

where $\delta_{hh'}$ is the probability that one member of pair falls in the cell h while the other in the cell h'. δ_{hh} is the probability that both members fall into the same cell h.

Note that the sum over all cells, $\Sigma_{hh'}\ \delta_{hh'} = 1$. If both members of pair fall in the same cell or they are completely dependent, the off-diagonal elements are zero and the sum of the diagonal elements is one.

For categorical variables, we may define the intra-level and inter-level correlation by

$$\rho_{hh'} = \left[\begin{array}{c} \dfrac{\delta_{hh} - \Pi_h^2}{\Pi_h\,(1-\Pi_h)} \\ \\ \dfrac{\delta_{hh'} - \Pi_h \Pi_{h'}}{\sqrt{\Pi_h\,(1-\Pi_h)\,\Pi_{h'}(1 - \Pi_{h'})}} \end{array}\right.$$

and the overall correlation by

$$\rho = \frac{\Sigma_{h=1,r}\ (\delta_{hh} - \Pi_h{}^2)}{\Sigma_{h=1,r}\ \Pi_h\,(1-\Pi_h)}.$$

These two definitions involve only the parameters, Π_h's, δ_{hh}, and $\delta_{hh'}$ while the definitions of continuous variables involve $\sigma_{chh}{}^2$ and $\sigma_{ehh}{}^2$.

3. SAMPLE

We now assume that a probability sample S is taken from the population U described above. The estimations are to be made on the basis of a sample $S=\{(i,j):i \in S^*; j \in S_i\}$, where S^* is a sample of "a" clusters out of A clusters, and S_i is a sample of b_i units from the B_i units in the ith cluster.

We also assume that the sample fractions are ignorable under the model we use. In fact the sampling design is not important under the model assumption. We do not specify the sampling design in both stages except that it is a probability measure on the set of all possible samples.

The sample clusters are indexed by i; $S^* = (S_1, ..., S_i, ..., S_a)$, and the final sample units in the i-th cluster are indexed by j; $S_i = (u_{i1}, ..., u_{ij}, ..., u_{ib_i})$. Note that the indexes i and j for the population unit U_{ij} are not the same as those used for the sample unit u_{ij}. The cluster sizes are assumed known, but of different size, and the total number of sample units is $n = \Sigma_i\, b_i$. The lower case y_{hij} is the realization of observation on the sample unit u_{ij}.

3.1. ESTIMATION

There are three types of correlations to estimate. The first case ρ_{hh} arises when any two members in a same cluster fall in the same category h. The second case $\rho_{hh'}$ arises when one member falls into one cell h while remaining member into a different cell h'. The third case ρ is the correlation for intra-cluster or overall levels.

Landis and Koch (1977) obtained the estimator of ρ in terms of ANOVA mean squares. We express the correlation ρ in terms of

$$d_{hij} = (y_{hij} - \overline{Y}_h)$$

for the level h to simplify the notation.

There would be four combinations

of d_{hij} for units and cells as (h=h' and j=j'), (h=h' and j≠j'), (h≠h' and j=j'), and (h≠h'and j≠ j') for each cluster. These four situations are described below and involved in the variance estimation shown in Section 3.3.

Since the clusters are independent, we find the variance and covariance of each cluster and sum them all for total.

(1) The variance of y_{hij} for h=h' and j=j' in cluster i is estimated as

$$s_{hhi} = \frac{\Sigma_j d_{hij}^2}{b_i}$$

The mean ms_{hh} of s_{hhi} is the sum of all s_{hhi}'s divided by "a", and the variance of the variance s_{hhi} is estimated by

$$V\hat{a}r(s_{hhi}) = \frac{1}{(a-1)} \Sigma_i (s_{hhi} - ms_{hh})$$

(2) Similarly the covariance between y_{hij} and $y_{hij'}$ for h = h' and j ≠ j' is estimated by

$$t_{hhi} = \frac{\Sigma_{j \neq j'} d_{hij} d_{hij'}}{b_i (b_i - 1)} .$$

The mean mt_{hh} of t_{hhi} is the sum of all t_{hhi}'s divided by "a". The variance of the covariances of t_{hhi}'s is estimated by

$$V\hat{a}r(t_{hhi}) = \frac{\Sigma_i (t_{hhi} - mt_{hh})^2}{a-1}$$

(3) The covariance of y_{hij} and $y_{h'ij}$ for h≠ h' and j =j' is estimated as

$$s_{hh'i} = \frac{\Sigma_j d_{hij} \; d_{h'ij}}{b_i} .$$

The mean $ms_{hh'}$ is the sum of all $s_{hh'i}$'s divided by "a". The variance of the covariace of $s_{hh'i}$'s is estimated by

$$V\hat{a}r(s_{hh'i}) = \frac{\Sigma_i (s_{hh'i} - ms_{hh'})^2}{a-1}$$

(4) The covariance of y_{hij} and $y_{h'ij'}$ or cross product of d_{hij} and $d_{h'ij'}$ for j≠ j' and h≠h' is

$$t_{hh'i} = \frac{\Sigma_{j \neq j'} \; d_{hij} d_{h'ij'}}{b_i (b_i - 1)}$$

The mean $mt_{hh'}$ of $t_{hh'i}$ is the sum of all $t_{hh'i}$'s divided by "a". The variance of the covariance is estimated by

$$V\hat{a}r(t_{hh'i}) = \frac{\Sigma_i (t_{hh'i} - mt_{hh'})^2}{a-1}$$

The covariance between s_{hhi} and t_{hhi} is estimated by

$$\hat{C}(s_{hhi} t_{hhi}) = \frac{\Sigma_i (s_{hhi} - ms_{hh})(t_{hhi} - mt_{hh})}{a-1}$$

Other covariances can be obtained similarly.

Define design constants for the i-th first stage unit to be
$g_i = (n - 1)/[(a - 1)b_i]$,
$c_i = [g_i \, b_i - d]/b_i$ and
$d = (n^2 - \Sigma_{i=1,a} b_i^2)/(n(a-1))$.
d is the average number of second stage units within each first stage unit and $n = \Sigma_{i=1,a} \, b_i$.

Express an estimator of correlation of $\rho_{hh'}$ by

$$\hat{\rho}_{hh'} = \frac{U_{hh'}}{\sqrt{D_{hh} \; D_{h'h'}}}$$

$U_{hh'}=(1/a)\Sigma_{i=1,a}\ [(g_i-1)s_{hh'i}+g_i\ t_{hh'i}]$,

$D_{hh}=(1/a)\Sigma_{i=1,a}\ [(c_i+d-1)s_{hhi}+c_i\ t_{hhi}]$,

$D_{h'h'}=(1/a)\Sigma_i[(c_i+d-1)s_{h'h'i}+c_i\ t_{h'h'i}]$.

It is easy to see that $E(U_{hh})=\sigma^2_{chh}$ and $E(D_{hh}) = (\sigma^2_{chh}+\sigma^2_{ehh})$ for h = h'.

For the balanced data of $b_i = b$, above constants are reduced to $d = b$, $n = ab$, $g_i = (n-1)/(n-b)$, and $c_i = (b-1)/(n-b)$.

However, the estimator $\hat\rho_{hh}$ is biased due to the dependency of numerator and denominator. But the bias is in the order of $a^{-1/2}$ because the bias is less than or equal to $CV(D_{hh})xSE(\hat\rho_{hh})$, the coefficient of variation of D_{hh} is bounded and the standard error of $\hat\rho_{hh}$ is in the order of $a^{-1/2}$.

It is not difficult to show directly the bias is in the order 1/a, so the bias becomes small for a large "a".

The estimator of overall correlation coefficient ρ is given by

$$\hat\rho = \frac{\Sigma_{h=1,r}U_{hh}}{\Sigma_{h=1,r}D_{hh}}$$

The bias of overall correlation also becomes small for a large "a".

3.2 CATEGORICAL VARIABLE

The correlation of categorical varables, defined in Section 2.2, can be estimated as following.

The two parameters π_h and δ_{hh} are unbiasedly estimated by

$$\hat\pi_h = \frac{\Sigma_{i=1,a}\Sigma_{j=1,b_i}\ y_{ijh}}{n}$$

$$\hat\delta_{hh} = \frac{\Sigma_{i=1,a}\ (y^2_{i+h} - \Sigma_{j=1,b_i}\ y^2_{ijh})}{H}$$

where $H = \sum_{i=1,a} b_i(b_i-1)$.

It can be shown that sample estimator of ρ_{hh} for h = h' for the categorical variables is

$$\tilde\rho_{hh} = \frac{\hat\delta_{hh} - \hat\pi^2_h + \dfrac{(\hat\pi_h - \hat\delta_{hh})}{n}}{\hat\pi_h(1-\hat\pi_h) - \dfrac{H(\hat\pi_h - \hat\delta_{hh})}{n^2}}$$

We can similarly estimate the inter-class estimator for h ≠ h' for categorical variables. The overall correlation is estimated by

$$\tilde\rho = \frac{\Sigma_h\ (\hat\delta_{hh} - \hat\pi_h{}^2) + \dfrac{(1-\Sigma_h\hat\delta_{hh})}{n}}{(1-\Sigma_h\ \hat\pi^2_h) - \dfrac{H(1-\Sigma_h\hat\delta_{hh})}{n^2}}.$$

The second terms in the numerator and denominator becomes small for a large "n". Although the numerator and denominator are unbiased, the correlation is biased estimator due to the correlation between the numerator and denominator, but the bias is in the order of 1/n, and becomes small for large n.

The asymptotic variance of the correlations for categorical variables are shown in another study (Choi, 1987).

These correlation estimators are consistent since the estimates of a parameter function is consistent as the same function of consistent estimators of the parameters as seen in maximum likelihood estimation.

Although two sets of formulas are developed for different variables, one for continuous and another for categorical variables, both give almost same results when applied to categorical variables, as seen from the previous study (Choi,1987). The first set may be used when total sum of squares is available, while the second set is easier to use when we have information on π_h and δ_{hh}.

3.3. VARIANCE ESTIMATION

we attempt to find the variances

of these correlations by Delta method. The variance of $\hat{\rho}_{hh}$, $\hat{\rho}_{hh'}$, and $\hat{\rho}$ is presented below:

Variance of $\hat{\rho}_{hh}$

Denote the partial derivatives of $\hat{\rho}_{hh}$ with respect to (s_{hhi}, t_{hhi}), evaluated at $(s_{hhi},t_{hhi})=(S_{hhi}, T_{hhi})$ by $d_{Shhi}=(1/D_{hh}^{2})[(g_i-1)D_{hh}-(c_i+ d-1)U_{hh}]$ and $d_{thhi} = (1/D_{hh}^{2})[\ g_i\ D_{hh} - c_i\ U_{hh}]$.

$$Var(\hat{\rho}_{hh}) = \frac{1}{A}\Sigma_i\ [\ (d^2_{shhi}Var(s_{hhi})$$

$$+d^2_{thhi}Var(t_{hhi})$$

$$+2\,d_{shhi}d_{thhi}Cov(s_{hhi}\,t_{hhi})\ ]$$

Variance of $\hat{\rho}_{hh'}$

$\hat{\rho}_{hh'}$ involves six variables:

(1) t_{hhi}, (2) $t_{h'h'i}$, (3) $t_{hh'i}$, (4) s_{hhi}, (5) $s_{h'h'i}$, and (6) $s_{hh'i}$. Let the variance of these six variables expressed by V_{11i}, V_{22i}, V_{33i}, V_{44i}, V_{55i}, and V_{66i}, where the subscripted number corresponds to the variable numbers (1) to (6), and the covariances between these six variables by V_{12i}, V_{13i}, V_{14i}, V_{15i}, V_{16i}, V_{23i}, V_{24i}, V_{25i}, V_{26i}, V_{34i}, V_{35i}, V_{36i}, V_{45i}, V_{46i}, and V_{56i}

Also denote the partial derivative of $\hat{\rho}_{hh'}$ with respect to each of these six variables by d_{1i}, d_{2i}, d_{3i}, d_{4i}, d_{5i}, and d_{6i} for

$\partial\rho_{h'h}/\partial t_{hhi} =- c_i\ U_{hh'}/[2\ (D_{h'h'})^{1/2}(D_{hhi})^{3/2}]$

$\partial\rho_{h'h}/\partial t_{h'h'i} =- c_i\ U_{hh'}/[2\ (D_{hh})^{1/2}(D_{h'h'i})^{3/2}]$

$\partial\rho_{h'h}/\partial t_{h'hi} = g_i\ /(D_{hh}\ D_{h'h'})^{1/2}$

$\partial\rho_{h'h}/\partial s_{hhi} =-(c_i\ +d-1)U_{hh'}/[2D_{hhi}^{3/2}(D_{h'h'})^{1/2}]$

$\partial\rho_{h'h}/\partial s_{h'h'i}=-(c_i+d-1)U_{hh'}/[2(D_{h'h'i})^{3/2}D_{hh}^{1/2}]$

$\partial\rho_{h'h}/\partial s_{h'hi} =(g_i-1)/(D_{hh}\ D_{h'h'})^{1/2}$

$$Var(\hat{\rho}_{hh'}) = \frac{1}{A}\Sigma_i\ [\textstyle\sum_{k=1}^{6} d^2_{ki}\ V_{kki}$$

$$+ \textstyle\sum^{6}_{k\neq k'}\ d_{ki}\,d_{k'i}\ V_{kk'i}]$$

The Var($\hat{\rho}_{hh'}$) reduces to Var($\hat{\rho}_{hh}$) when we replace the estimates with subscripts hh' and h'h' by those of hh.

Variance of $\hat{\rho}$

For Var($\hat{\rho}$), we need the partial derivatives of overall correlation with respect to s_{hhi} and t_{hhi}, evalusted at $(s_{hhi}, t_{hhi})= (S_{hhi}, T_{hhi})$. Denote them by d_{rshhi}, and d_{rthhi} for

$\partial\hat{\rho}/\partial S_{hhi}=(1/D^2)[(g_i -1)D-U(c_i+d - 1)]$,

$\partial\hat{\rho}/\partial T_{hhi}=(1/D^2)[g_i\,D - U\ c_i]$.

$$Var(\hat{\rho}) = \frac{1}{A}\Sigma_i\{\textstyle\sum_h [d^2_{rshhi}Var(s_{hhi})$$

$$+d^2_{rthhi}Var(t_{hhi})$$

$$+2\,d_{rshhi}d_{rthhi}Cov(s_{hhi}\,t_{hhi})\ ]$$

$$+\textstyle\sum_{h\neq h'}[d_{rshhi}d_{rsh'h'i}Cov(s_{hhi}\,s_{h'h'1})$$

$$+d_{rshhi}d_{rth'h'i}Cov(s_{hhi}\,t_{h'h'1})$$

$$+d_{rsh'h'i}d_{rthhi}Cov(s_{h'h'i}\,t_{hhi})$$

$$+d_{rthhi}d_{rth'h'i}Cov(t_{hhi}\,t_{h'h'i})\ ]\}$$

The estimator of above variances also can be obtained by replacing each component with its respective sample estimate based on the “a” sample clusters.

Unless we need a closed form of variance to investigate which parts do major roles in variance, we may use a resampling method. If randomly executed, resampling methods produce similar results as current method does.

We assumed that partial derivatives were possible and that non-linear terms could be negligible for delta method used in this section.

4. ASYMPTOTIC DISTRIBUTION

The asymptotic test z_{hh} shown below is based mainly on the three components: $E(\hat{\rho}_{hh})=\rho_{hh} + o_p(a^{-1})$, variance, and null hypothesis

$\rho_{hh} = \rho_{0hh}$:

$$z_{hh} = \frac{\hat{\rho}_{hh} - \rho_{0hh}}{\sqrt{var(\hat{\rho}_{hh})}} \rightarrow N(0,1)$$

for a large "a". Note that $\sqrt{a}(\hat{\rho}_{hh} - \rho_{hh}) \rightarrow N(0, D^T VD)$ as a→∞

where V is Var($\hat{\psi}_{hh}$), and D is the deritive matrix of $\hat{\rho}_{hh}$ with respect to $\hat{\psi}_{hh}$, evaluated at $\psi^T_{hh} = (S_{hh}, T_{hh})$. This is the direct result of the linear expansion of $\hat{\rho}_{hh} = \rho_{hh} + \frac{\partial \rho_{hh}}{\partial \psi}(\hat{\psi} - \psi) + o(a^{-1})$.

Similarly, the test statistic for H_o: $\rho = \rho_o$ is

$$z = \frac{\hat{\rho} - \rho_o}{\sqrt{Var(\hat{\rho})}} \rightarrow N(0,1)$$

where Var($\hat{\rho}$) is shown previously.

The asymptotic test statistics for catagorical variables can be similarly formulated as those of above.

5. DRUG TEST

Miller and Landis (1991) studied a clinical data to see if individual is improved by treatment. 107 persons are treated by drug, and 102 by placebo, and classified them as (1) worse or no change, (2) slight improvement, or (3) more improvement or cured. Nine investigators examined the patients, considered as nine clusters. Patients treated by an investor are considered as the units in cluster.

The Intra-cluster Correlation of Treatment group is 0.223 with standard deviation of 0.178 and z =1.253, while that of Placebo Group is 0.109 with standard deviation of 0.130.and z = 0.838.

Individual correlations are not significant at $\alpha = 0.05$ under the null hypothesis of no correlations.

The over-all correlations, 0.223 and 0.109, are useful to estimate the design effect for correlated members in cluster. For example, the design effect for the treatment group is

[1 + 0.223(12-1)] = 3.453,

and that of placebo group is

[1 + 0.109(11.3-1)] = 2.123.

The number 12 and 11.3 are the average number of cluster members for treatment and placebo groups, respectively.

Although above individual correlations, 0.223 and 0.109, are not significant, the design effects, 3.453 and 2.123, arising from the correlation are quite big, and should not be ignored.

If normal test or t-test statistic were calculated as if data were based on a simple random sample, it should be adjusted by dividing it with the square root of design effect.

For example, the difference between two proportions for two levels, level 1 and level 3, in the treatment group is 44/107 (=24/107 - 68/107). Normal test score is z = 0.0624 under the simple random sample assumption. Being adjusted by the design effect for intra-cluster correlation, corrected score is z' = 0.0336 = $z/(3.453)^{1/2}$ under the null hypothesis of equal proportion.

REFERENCES

J.W. Choi (1987). A Direct Estimation of Intracluster Correlation. 1987 Proceeding of American Statistical Association on Survey Research Methods. 1047-51.

J. Richard. Landis and Gary. G. Koch (1977). A one way components of variance model for categorical data. Biometrics 33, 671-79.

M.E. Miller and J.R. Landis (1991). Generalized Variance Component Models for Clustered Categorical Response Variables. Biometrics, 47, 33-44.