MEASURING ASSOCIATION AND MODELLING RELATIONSHIPS BETWEEN INTERVAL AND ORDINAL VARIABLES

John Schollenberger, Alan Agresti and D. D. Wackerly, University of Florida

0. ABSTRACT

Measurement of association between an interval variable X and an ordinal variable Y, is usually accomplished by considering both variables as either interval or ordinal. However, measures analogous to Kendall's tau-b (τ_b) and Goodman and Kruskal's gamma (γ) can be defined which utilize the interval nature of the variable X. Consider a random pair of observations (X_i,Y_i), (X_j,Y_j) on (X,Y). An analog of τ_b is defined as the correlation between (X_i-X_j) and $S(Y_i-Y_j)$ where S is the sign function. We define an analog of γ by weighting each concordant or discordant pair by (X_i-X_j).

An alternative approach that also leads to a measure of interval-ordinal association is the use of a linear logistic model to represent $P(Y_i>Y_j)$ as a monotonic function of (X_i-X_j).

1. INTRODUCTION

The usual solution to measuring the association between an interval variable and an ordinal variable is to consider both variables as either interval or ordinal. If both variables are treated as ordinal, well known ordinal measures of association can be used but then not all of the available information is being utilized. If a metric is imposed upon the levels of the ordinal variable, more sophisticated interval-level techniques can be used but the validity of the results is questionable.

In a related problem Mayer (1973) introduced the monotone coefficient as an estimator of the correlation ratio when one interval variable is directly observed but only a monotone transformation of the other (interval) variable is observed. The properties of the monotone coefficient are derived and discussed in Mayer and Robinson (1978). Their measures correspond naturally to monotone regression functions for modelling an interval scale dependent variable in terms of one or more ordinal scale variables.

This article considers alternative approaches to the problems of measuring interval-ordinal association and modelling interval-ordinal relationships. Our measures and models are defined in terms of scores for pairs of observations -- sign scores for the ordinal variable and distance scores for the interval variable. We also use a linear logistic model for describing the dependence of ordinal sign scores on interval distance scores. Hence, unlike Mayer and Robinson, our emphasis in model-building is on the case in which the ordinal variable is the dependent variable, a case commonly encountered in social science research.

We assume throughout this article that we have n observations (x_i,y_i) on a bivariate random variable (X,Y) with X a discrete interval variable and Y a discrete ordinal variable. Unless otherwise noted, we also assume full multinomial sampling wherein the sample of n individuals or objects is cross-classified according to the categories of X and Y and the observed cell frequencies can be described as in the (r x c) contingency table below with $y_{(1)}<\ldots<y_{(r)}$ and $x_{(1)}<\ldots<x_{(c)}$.

Table 1.1

$y_{(1)}$	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
$y_{(2)}$	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
$y_{(r)}$	n_{r1}	n_{r2}		n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n
	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(c)}$	

The same system of notation is used to describe the population cell proportions $\{p_{ab}\}$ noting $p_{a.}=\sum_b p_{ab}$, $p_{.b}=\sum_a p_{ab}$ and $\sum\sum_{ab} p_{ab}=1$. Finally, note that a pair of observations say (x_i,y_i) and (x_j,y_j) is a concordant pair if $(x_{(i)}-x_{(j)})(y_{(i)}-y_{(j)})>0$ and discordant if $(x_{(i)}-x_{(j)})(y_{(i)}-y_{(j)})<0$.

2. MEASURES OF ASSOCIATION

We define interval-ordinal measures of association analogous to the ordinal measures Kendall's τ_b and Goodman and Kruskal's γ.

Kendall (1970) defines a generalized correlation coefficient Γ which includes his own τ_b and other correlation coefficients as special cases which arise when particular methods of scoring are adopted. Given n observations (x_i,y_i), to each pair of observations assign an x-score, denoted by a_{ij} and subject only to the condition $a_{ij}=-a_{ji}$. Similarly assign y-scores denoted by b_{ij}, with $b_{ij}=-b_{ji}$. Then Kendall's generalized correlation coefficient is defined by

$$\Gamma = \frac{\Sigma a_{ij}b_{ij}}{\{\Sigma a_{ij}^2 \Sigma b_{ij}^2\}^{\frac{1}{2}}} . \qquad (2.1)$$

If, for example, a_{ij} and b_{ij} are defined by $a_{ij}=S(x_j-x_i)$ and $b_{ij}=S(y_j-y_i)$ where

$$S(z) = \begin{Bmatrix} +1 & z>0 \\ 0 & z=0 \\ -1 & z<0 \end{Bmatrix} \text{ is the sign function,}$$

then the resulting measure is Kendall's τ_b.

2.1 An interval-ordinal analog of tau-b

In the situation where X is an interval variable and Y is an ordinal variable, it seems intuitive to allow the x-score, a_{ij}, to be (x_j-x_i) vice $S(x_j-x_i)$. Thus $\hat{\tau}'_b$, the sample version of our interval-ordinal measure of association, is given by

$$\hat{\tau}'_b = \frac{\Sigma(x_j-x_i)S(y_j-y_i)}{\{\Sigma(x_j-x_i)^2\Sigma S(y_j-y_i)^2\}^{\frac{1}{2}}} \qquad (2.2)$$

and it is the usual sample correlation between appropriately defined x-scores and y-scores. The population value of τ'_b is defined by

$$\tau'_b = \frac{E[(X_j-X_i)S(Y_j-Y_i)]}{\{V(X_j-X_i)V(S(Y_j-Y_i))\}^{\frac{1}{2}}} \qquad (2.3)$$

for a randomly selected pair of observations.

If the n observations are cross-classified into an (r x c) table (Table 1.1), then following the notation of Goodman and Kruskal (1972), $\hat{\tau}'_b$ can be written as

$$\hat{\tau}'_b = \frac{C_w - D_w}{\{\sum_{b'>b}\Sigma n_{.b}n_{.b'}(x_{(b')}-x_{(b)})^2\}^{\frac{1}{2}}\{\frac{1}{2}(n^2-\sum_a n_{a.}^2)\}^{\frac{1}{2}}} \qquad (2.4)$$

where $C_w = \sum_{a=1}^{r}\sum_{b=1}^{c} n_{ab} \sum_{\substack{a'>a\\b'>b}}\Sigma\, n_{a'b'}(x_{(b')}-x_{(b)})$ and

$$D_w = \sum_{a=1}^{r}\sum_{b=1}^{c} n_{ab} \sum_{\substack{a'<a\\b'>b}}\Sigma\, n_{a'b'}(x_{(b')}-x_{(b)}).$$

Notice that C_w is the weighted number of concordant pairs of observations with weights equal to $(x_{(b')}-x_{(b)})$. Similarly D_w is the weighted number of discordant pairs of observations. We call C_w the sample "weight of concordance" and D_w the sample "weight of discordance".

2.2 An interval-ordinal analog to gamma

Formula 2.4 suggests the definition of an analog of Goodman and Kruskal's gamma (γ). We define the sample version of our interval-ordinal analog of γ by

$$\hat{\gamma}' = \frac{C_w - D_w}{C_w + D_w} \qquad (2.5)$$

The population value, γ', can be defined by replacing the observed cell frequencies $\{n_{ab}\}$ in C_w and D_w by the population cell proportions $\{p_{ab}\}$.

3. PROPERTIES OF TAU-b' AND GAMMA'

The relationship between Kendall's $\hat{\tau}_b$ and $\hat{\tau}'_b$ is obvious. They are both special cases of Γ and if (x_j-x_i) is replaced by $S(x_j-x_i)$, $\hat{\tau}'_b$ becomes $\hat{\tau}_b$. Similarly, when this same substitution of ordinal for interval information is made for $\hat{\gamma}'$, $\hat{\gamma}'$ becomes $\hat{\gamma}$. Thus τ'_b and γ' are natural generalizations of τ_b and γ to the situation of an interval and an ordinal variable.

Some additional properties of $\hat{\tau}'_b$ and $\hat{\gamma}'$ follow.

Result 3.1: $-1 \le \hat{\tau}'_b \le 1$, $-1 \le \hat{\gamma} \le 1$.

Result 3.2: $\hat{\tau}'_b = \hat{\gamma}' = 0$ when $C_w = D_w$.

Result 3.3: For data categorized into a contingency table with arbitrary fixed marginal distributions, the maximum values of $\hat{\tau}'_b$ and $\hat{\gamma}'$ occur when there are no discordant pairs of observations. The maximum value of $\hat{\tau}'_b$ is a function of the values of X and the maximum value of $\hat{\gamma}'$ is one.

Result 3.4: The asymptotic variances of $\hat{\tau}'_b$ and $\hat{\gamma}'$ are easily obtained using the procedures of Goodman and Kruskal (1972). If the population value of the measure, defined in terms of the population cell proportions, is denoted by ν/δ and we define

$$\nu'_{ab} = \frac{\partial\nu}{\partial p_{ab}}, \quad \delta'_{ab} = \frac{\partial\delta}{\partial p_{ab}},$$

$$\phi_{ab} = \nu\delta'_{ab} - \delta\nu'_{ab}, \quad \bar{\phi} = \sum_{ab}\Sigma\, p_{ab}\phi_{ab};$$

then the asymptotic variance of the sample value of ν/δ (under full multinomial sampling) is σ^2/n where

$$\sigma^2 = \frac{1}{\delta^4}\sum_{ab}\Sigma\, p_{ab}(\phi_{ab}-\bar{\phi})^2 = \frac{1}{\delta^4}(\sum_{ab}\Sigma\, p_{ab}\phi_{ab}^2-\bar{\phi}^2) \qquad (3.1)$$

Now if τ'_b is defined by replacing the n_{ab}'s in $\hat{\tau}'_b$ by p_{ab}'s then

$$\nu'_{ab} = \sum_{\substack{a'b'\\a'>a}}\Sigma\, p_{a'b'}(x_{(b')}-x_{(b)}) - \sum_{\substack{a'b'\\a'<a}}\Sigma\, p_{a'b'}(x_{(b')}-x_{(b)}) \qquad (3.2)$$

and

$$\delta'_{ab} = \tfrac{1}{2}\delta\left\{\frac{\sum_{b'\neq b}\Sigma p_{.b'}(x_{(b')}-x_{(b)})^2}{\sum_{b'<b''}\Sigma\, p_{.b'}p_{.b''}(x_{(b')}-x_{(b'')})^2} - \frac{p_{a.}}{\frac{1}{2}(1-\sum_a p_{a.}^2)}\right\} \qquad (3.3)$$

For γ', ν'_{ab} is the same as for τ'_b since they have the same numerator while

$$\delta'_{ab} = \sum_{\substack{a'b'\\a'\neq a}}\Sigma\, p_{a'b'}|x_{(b')}-x_{(b)}|. \qquad (3.4)$$

Also, for γ', $\bar{\phi}$ can be shown to equal zero. Now the asymptotic variances for both $\hat{\gamma}$ and $\hat{\tau}'_b$ are obtained by substituting (3.2)-(3.4) into (3.1). The asymptotic variances under product multinomial sampling can be derived similarly.

4. MODEL BUILDING

Recall we have n observations (x_i,y_i) on (X,Y) where X is a discrete interval variable and Y is a discrete ordinal variable. If a positive association exists between X and Y, then it seems reasonable that the probability that a randomly selected pair of observations is concordant, given their respective values on X, should increase as the difference between the X values increases.

We denote the probability that the ij^th pair of observations is concordant, given it is untied on Y, by $P[C(x_i,x_j)]$. Formally, we define

$$P[C(x_i,x_j)]=P[(X_i-X_j)(Y_i-Y_j)>0|X_i=x_i,X_j=x_j,Y_i\neq Y_j] \quad (4.1)$$

for $x_i \neq x_j$. For most bivariate distributions encountered in practice it seems reasonable that as $|x_i-x_j|\to 0$, $P[C(x_i,x_j)]\to\frac{1}{2}$. Therefore we define $P[C(x_i,x_i)]=\frac{1}{2}$. We also assume that $P[C(x_i,x_j)]$ is a monotonic function of (x_j-x_i).

A linear logistic model may be defined by

$$\Lambda(x_i,x_j)=\ln\left\{\frac{P[C(x_i,x_j)]}{1-P[C(x_i,x_j)]}\right\}=\beta(x_j-x_i) \quad (4.2)$$

for all x_i,x_j. A no intercept model is used since $\Lambda(x_i,x_i)=0$ by the definition of $P[C(x_i,x_i)]$.

Recall the observations can be cross-classified into an (r x c) contingency table where $x_{(1)}<\ldots<x_{(c)}$ and $y_{(1)}<\ldots<y_{(r)}$ denote the distinct values of X and Y respectively. Thus, since we assume that $P[C(x_i,x_j)]$ is constant given x_i and x_j, we have

$$P[C(x_i,x_j)]=P[C(x_{(b)},x_{(b')})] \quad (4.3)$$

for all $x_i=x_{(b)}$ and $x_j=x_{(b')}$ $(b,b'=1,\ldots,c)$, and we can base our analysis on these $\frac{1}{2}c(c-1)$ values of $P[C(x_{(b)},x_{(b')})]$ for $b<b'$.

First note that

$$P[C(x_{(b)},x_{(b')})]=\frac{\sum\limits_{a'>a} p_{ab}p_{a'b'}}{\sum\limits_{a'<a} p_{ab}p_{a'b'}+\sum\limits_{a'>a} p_{ab}p_{a'b'}}. \quad (4.4)$$

Thus our model becomes

$$\Lambda(x_{(b)},x_{(b')})=\ln\left\{\frac{\sum\limits_{a'>a} p_{ab}p_{a'b'}}{\sum\limits_{a'<a} p_{ab}p_{a'b'}}\right\}=\beta(x_{(b')}-x_{(b)}) \quad (4.5)$$

for $1 \le b < b' \le c$. Using vector notation our model becomes $\underline{\Lambda}=\beta\underline{d}$ where

$$\underline{\Lambda} = (\Lambda(x_{(1)},x_{(2)}),\ldots,\Lambda(x_{(c-1)},x_{(c)}))' \quad (4.6)$$

and $\underline{d}$ is the known vector

$$\underline{d} = (x_{(2)}-x_{(1)},\ldots,x_{(c)}-x_{(c-1)})' \quad (4.7)$$

It is well known that $\underline{p}_n=\frac{1}{n}(n_{11},\ldots,n_{rc})$ is the maximum likelihood estimator of $\underline{p}=(p_{11},\ldots,p_{rc})$ and has a multivariate normal asymptotic distribution. Notice $\underline{\Lambda}$ is simply a vector function of $\underline{p}$ and if $\underline{\lambda}$ is defined as $\underline{\Lambda}$ evaluated at $\underline{p}=\underline{p}_n$, then the method of generalized (or weighted) least squares can be applied to $\underline{\lambda}$ using an empirically estimated covariance matrix to produce an estimate of β. This procedure is a generalization of a method used in Cox (1970) and relies heavily on the asymptotic theory of functions of maximum likelihood estimators as given in Rao (1973). It is anticipated that under suitable regularity conditions this estimate of β will have a normal asymptotic distribution.

If the necessary regularity conditions hold, then the extension to higher order models is immediate -- $\underline{d}$ become a known matrix D. The multiple regression situation (three-way and higher order tables) is more complicated.

To provide some motivation for the multiple regression case , suppose we have two discrete interval variables $X^{(1)}$ and $X^{(2)}$ and let $\underline{X}_i$ denote $(X_i^{(1)},X_i^{(2)})$. We consider probabilities of the form

$$P[C(\underline{x}_i,\underline{x}_j)]=P[Y_i-Y_j>0|\underline{X}_i=\underline{x}_i,\underline{X}_j=\underline{x}_j,Y_i\neq Y_j], \quad (4.8)$$

ordering the pairs of observations such that $x_i^{(1)}\le x_j^{(1)}$ and if $x_i^{(1)}=x_j^{(1)}$ such that $x_i^{(2)}\le x_j^{(2)}$. Notice this probability is actually a generalization of $P[C(x_i,x_j)]$ since in the single regression case we ordered the pairs of observations so that $x_i<x_j$ $(x_j-x_i>0)$. We consider models of the form

$$\Lambda(\underline{x}_i,\underline{x}_j)= \ln\left\{\frac{P[C(\underline{x}_i,\underline{x}_j)]}{1-P[C(\underline{x}_i,\underline{x}_j)]}\right\}$$

$$= \beta_1(x_j^{(1)}-x_i^{(1)}) + \beta_2(x_j^{(2)}-x_i^{(2)}). \quad (4.9)$$

Now the method outlined for the single regression situation can be used to obtain an estimate for $\underline{\beta}$. Results are anticipated in this area in the near future.

5. CONCLUDING REMARKS

The statistics $\hat{\tau}_b'$ and $\hat{\gamma}'$ are reasonable, easily understood measures of interval-ordinal association. Indeed it seems very intuitive that the larger the difference in the x-values of a pair of observations, the more weight that pair should have in determining the weight of concordance or the weight of discordance. The model building section uses the familiar framework of linear models to model the logit of the probability of concordance as a function of the difference in x-values.

REFERENCES

Cox, D.R. The analysis of Binary Data. London: Chapman and Hall, 30-32, 1970.

Goodman, L.A. and Kruskal, W.H. "Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances", JASA, 67, 415-421, 1972.

Kendall, M.G. Rank Correlation Methods. London: Griffen, 19-22, 1970.

Mayer, L.S. "Using Monotone Regression to Estimate a Correlation Coefficient", Sociological Methodology, ed. H.L. Costner. San Francisco: Jossey Bass, 200-212, 1973.

Mayer, L.S. and Robinson, J.A. "Measures of Association for Multiple Regression Models with Ordinal Predictor Variables", Sociological Methodology, ed. K.F. Schuesster, San Francisco: Jossey Bass, 141-163, 1978.

Rao, C.R. Linear Statistical Inference and its Applications. New York: John Wiley and Sons, 385-389, 1973.