Test of Association between Two Ordinal Variables while Adjusting for Covariates

Chun Li* and Bryan E. Shepherd

Department of Biostatistics, Vanderbilt University

Nashville, Tennessee 37232

**email:* chun.li@vanderbilt.edu

Running title: Conditional Ordinal-by-ordinal Association Test

Author's Footnote

Chun Li is Assistant Professor, Department of Biostatistics, Vanderbilt University, Nashville, TN 37232 (e-mail: chun.li@vanderbilt.edu); and Bryan E. Shepherd is Assistant Professor, Department of Biostatistics, Vanderbilt University, Nashville, TN 37232 (e-mail: bryan.shepherd@vanderbilt.edu). This work was partially supported by the National Institutes of Health grants R01HG004517 (C.L.) and P30AI54999 (B.E.S.). The authors thank Dr. Sean Donahue for providing data on anisometropic amblyopia and Dr. Vikrant Sahasrabuddhe for providing data on cervical neoplastic stage and condom use.

ABSTRACT

We propose a new set of test statistics to examine the association between two ordinal categorical variables X and Y after adjusting for continuous and/or categorical covariates $\boldsymbol{Z}$. Our approach first fits multinomial (e.g., proportional odds) models of X and Y, separately, on $\boldsymbol{Z}$. For each subject, we then compute the conditional distributions of X and Y given $\boldsymbol{Z}$. If there is no relationship between X and Y after adjusting for $\boldsymbol{Z}$, then these conditional distributions will be independent, and the observed value of (X, Y) for a subject is expected to follow the product distribution of these conditional distributions. We consider two simple ways of testing the null of conditional independence, both of which treat X and Y equally, in the sense that they do not require specifying an outcome and a predictor variable. The first approach adds these product distributions across all subjects to obtain the expected distribution of (X, Y) under the null and then contrasts it with the observed unconditional distribution of (X, Y). Our second approach computes "residuals" from the two multinomial models and then tests for correlation between these residuals; we define a new individual-level residual for models with ordinal outcomes. We present methods for computing p-values using either the empirical or asymptotic distributions of our test statistics. Through simulations, we demonstrate that our test statistics perform well in terms of power and type I error rate when compared to proportional odds models which treat X as either a continuous or categorical predictor. We apply our methods to data from a study of visual impairment in children and to a study of cervical abnormalities in HIV-infected women. Supplemental materials for the article are available online.

KEY WORDS: regression for ordinal outcome, residual for ordinal outcome, concondance-discordance, categorical data analysis

1. INTRODUCTION

Consider the situation where there are two ordinal categorical random variables, X and Y, and we want to examine the association between X and Y after adjusting for continuous and/or categorical covariates $\boldsymbol{Z}$. This is a common scenario in medical research and social sciences. For example, a recent study collected data on stages of cervical squamous intraepithelial lesions and condom use among HIV-infected women in Zambia (Parham et al. 2006). Cervical dysplastic stage is an ordered categorical variable with four levels according to the commonly used revised Bethesda system: normal, low-grade squamous intraepithelial lesions, high-grade squamous intraepithelial lesions, and squamous cell carcinoma. Condom use was categorized as never, rarely, almost always, and always. Researchers are interested in testing the association between cervix health and condom use after controlling for other variables such as age and CD4 count.

Proportional odds models (also known as ordinal logistic regression) (McCullagh 1980), other cumulative link models (Aitchison and Silvey 1957; Farewell 1982), and continuation ratio models (Läärä and Matthews 1985) are commonly used to examine the association between an ordinal response variable and continuous or categorical predictors. These regression models are useful because they account for the natural ordering of the outcome but do not treat the outcome as a continuous variable. However, when one of the predictors is ordered categorical, all traditional regression approaches including these for ordered categorical outcomes have to treat the ordinal predictor as either numerical or categorical. The former enforces a linearity assumption and the latter ignores the order information.

When the ordinal predictor is treated as continuous, we assume the effect of moving from level 1 to level 2 is the same as that from level 2 to level 3. Often this assumption is unreasonable. In the above example, there is little reason to suppose that the effect difference on cervical neoplastic stage between no condom use and rare condom use is the same as that between rare and almost always use. Alternatively, one could assign numbers to the categories or transform the predictor in some manner so that the assigned values reflect a linear relationship with the appropriately transformed expected outcome. The problem with this approach is that such a transformation of the predictor is difficult to choose and may lead to data dredging, ignoring uncertainty in model selection. Splines, a special type of transformation (e.g., Ramsay 1988), have other drawbacks: uncertainty in number and locations of knots, dependence of results on how the categories are coded, non-monotonic results when a monotonic relationship is expected, and difficulty when there are only three categories.

When the ordinal predictor is treated as discrete, the order information is ignored. When there are many categories, this approach may have low power due to high degrees of freedom. In addition, the effect estimates may be non-monotonic. Isotonic regression (Barlow et al. 1972) addresses the latter problem by grouping adjacent categories if their relative effect is in reverse of the general trend. However, the categorization is data driven and the results need appropriate adjustment for this source of model selection variability. In addition, the degrees of freedom may still be high, particularly if the original categories already manifest a monotonic relationship.

Another, possibly Bayesian, option could be to perform an analysis using latent variables; the ordinal variables X and Y can be thought of as some categorization of continuous latent variables, say U and V. A test of the association between X and Y conditional on $\boldsymbol{Z}$ could then be some test of the correlation between U and V conditional on $\boldsymbol{Z}$. The primary problem we see with latent variable approaches of this nature is that they require assuming a distribution for the latent variable which essentially forces one to assign a scale to the ordinal categorical variables.

Some approaches model X and Y jointly conditional on $\boldsymbol{Z}$. Examples include loglinear, especially linear-by-linear association, models (Goodman 1979; Agresti 2002) and generalized Cochran-Mantel-Haenszel test (Mantel 1963). However, these methods require assigning numeric values to the ordinal categories, and their implementation also requires grouping continuous or multivariable $\boldsymbol{Z}$ into strata, generating arbitrary cutoffs on $\boldsymbol{Z}$ and losing information.

Other approaches for measuring conditional associations of ordinal categorical data have built on classic two-variable statistics such as Kendall's concordance-discordance statistic tau, Goodman and Kruskal's gamma (Goodman and Kruskal 1954), and Spearman's rank-based correlation (Agresti 2002). Kendall's partial tau (Kendall 1948) and an extension of Kendall's partial tau to multivariable $\boldsymbol{Z}$ (Hawkes 1971) are a few examples, although the usefulness of these approaches is questionable given that the expectation of these statistics is generally not zero under the null hypothesis that X and Y are independent conditional on $\boldsymbol{Z}$ (Goodman 1959; Agresti 1977; Schemper 1991). Other proposed methods involve stratifying data according to $\boldsymbol{Z}$ and computing weighted averages of stratum-specific measures of association between X and Y (Torgerson 1956; Davis 1967; Agresti 1977). Because these techniques need multiple observations per stratum, these approaches again require collapsing continuous or multivariable $\boldsymbol{Z}$ into strata.

Our method is motivated by the following observation. If X and Y are continuous variables, we could carry out linear regression $Y = \beta_0 + \beta_1 X + \gamma_1 Z_1 + \cdots + \gamma_k Z_k + e$, and test for significance of the coefficient β_1. Alternatively, we could carry out the following procedure: (a) fit a linear regression of Y on the covariates $\boldsymbol{Z}$ and obtain the residual Y_{res} for each subject, (b) fit a linear regression of X on $\boldsymbol{Z}$, and obtain the residual X_{res} for each subject, and (c) perform a simple linear regression $Y_{res} = \alpha_0 + \alpha_1 X_{res} + e$, and test for significance of the coefficient α_1. It is well known that the coefficient estimates for β_1 and α_1 are the same (Rao 1973; Mosteller and Tukey 1977). Their corresponding significance levels are similar if the number of subjects is much larger than the number of covariates. Note that for each subject, given the covariate values $\boldsymbol{Z} = \boldsymbol{z}$, the linear regression in (a) would effectively yield a distribution of possible realizations of $Y|\boldsymbol{z}$ for the subject and the regression in (b) would yield a distribution of possible realizations of $X|\boldsymbol{z}$ for the subject. These two distributions are expected to be independent if there is no association between Y and X after adjusting for $\boldsymbol{Z}$.

Similar to the linear regression setting, our approach is to fit Y and X on $\boldsymbol{Z}$ separately using multinomial models (e.g., proportional odds models), obtain the conditional distributions $Y|\boldsymbol{Z}$ and $X|\boldsymbol{Z}$ for each subject, and use this information to construct test statistics. In sections 2-3 we describe our method. In section 4 we investigate the performance of our method through simulations. In section 5 we apply our method to two data sets: one on the association between anisometropia and amblyopia and the other on the association between cervical neoplastic stage and condom use. We discuss our results in section 6.

2. METHOD

Let Y and X be two ordinal variables with s and t categories, respectively. The categories are denoted as $1^Y, 2^Y, \cdots, s^Y$, and $1^X, 2^X, \cdots, t^X$. Note that the numbers (e.g., 1^Y) are simply symbols for categories and should not be interpreted as quantities. Without loss of generality, suppose the order of categories are $1^Y < 2^Y < \cdots < s^Y$ and $1^X < 2^X < \cdots < t^X$. We will omit the superscript when it is clear from the context which variable it belongs to. Our goal is to examine the relationship between Y and X after adjusting for k covariates, $\boldsymbol{Z} = (Z_1, \cdots, Z_k)$. We will test the null hypothesis

$$H_0 : P(Y, X|\boldsymbol{Z}) = P(Y|\boldsymbol{Z})P(X|\boldsymbol{Z}),$$

that is, conditional on $\boldsymbol{Z}$, Y and X are independent.

As described in the Introduction, we first carry out a multinomial regression analysis of Y on $\boldsymbol{Z}$, and a multinomial regression analysis of X on $\boldsymbol{Z}$. A commonly used multinomial regression model for ordinal outcomes is the proportional odds model (McCullagh 1980). Other possible models include cumulative link models (Aitchison and Silvey 1957; Farewell 1982) and continuation ratio models for ordinal outcomes (Läärä and Matthews 1985), and multinomial logit models for categorical outcomes (Mantel 1966). One can choose different multinomial models for the two ordinal variables.

For each subject, the regression analyses of Y and X on $\boldsymbol{Z}$ will provide conditional distributions of Y and X given $\boldsymbol{Z}$. If there is no relationship between Y and X after adjusting for the covariates (i.e., under the null), these two conditional distributions will be independent, and the observed value of (Y, X) for the subject is expected to follow the product distribution of these two conditional distributions. We consider two ways of testing the null. The first approach adds these product distributions across all subjects to obtain the expected marginal distribution of (Y, X) under the null of conditional independence and then contrasts this distribution with the observed marginal distribution for (Y, X) (i.e., not conditional on $\boldsymbol{Z}$). Our second approach computes "residuals" from the two multinomial regression models and then tests for correlation between these residuals. We define a new individual-level residual for models with ordinal outcomes. We will describe these two approaches separately.

2.1. Observed versus Expected Distributions

In this approach, we compare the observed joint distribution between Y and X with their expected distribution under the null. In general, the joint distribution between Y and X, $P = P(Y, X) = \{\pi_{jl}\}$, can be written as

$$P(Y, X) = \int_z P(Y, X|\boldsymbol{Z}) dP(\boldsymbol{Z}). \tag{1}$$

Under the null of conditional independence, the joint distribution between Y and X can also be written as

$$P_0(Y, X) = \int_z P(Y|\boldsymbol{Z})P(X|\boldsymbol{Z})dP(\boldsymbol{Z}). \tag{2}$$

Let $P_0 = P_0(Y, X) = \{\pi^0_{jl}\}$. Then under the null, $P = P_0$.

Assume $(X_i, Y_i, \boldsymbol{Z}_i)$, $i = 1, \cdots, n$, are i.i.d. copies from the random vector $(X, Y, \boldsymbol{Z})$. (Note that although we assume $(X_i, Y_i, \boldsymbol{Z}_i)$ are i.i.d., our methods hold if $\boldsymbol{Z}_i$ is set to specific levels.) For subject i ($i = 1, \cdots, n$), let $p^j_i = P(Y_i = j|\boldsymbol{Z}_i = \boldsymbol{z}_i)$ for $j = 1, \cdots, s$, and $q^l_i = P(X_i = l|\boldsymbol{Z}_i = \boldsymbol{z}_i)$ for $l = 1, \cdots, t$. Let $\gamma^j_i = P(Y_i \leq j|\boldsymbol{Z} = \boldsymbol{z}_i)$; for convenience, let $\gamma^0_i = 0$ and $\gamma^s_i = 1$. Then $p^j_i = \gamma^j_i - \gamma^{j-1}_i$. The probabilities q^l_i can be similarly written as differences between cumulative probabilities.

To estimate P_0, we fit separate multinomial models for $P(Y|\boldsymbol{Z})$ and $P(X|\boldsymbol{Z})$ to estimate the probabilities p^j_i and q^l_i, denoted as $\hat{p}^j_i$ and $\hat{q}^l_i$. The distribution of $\boldsymbol{Z}$ is estimated by its empirical distribution. Plugging these estimates into (2), we obtain the estimate $\hat{P}_0 = \{\hat{\pi}^0_{jl}\}$, where $\hat{\pi}^0_{jl} = \frac{1}{n}\sum_i \hat{p}^j_i \hat{q}^l_i$. Without assuming the null hypothesis, P can be estimated empirically as $\hat{P} = \{\hat{\pi}_{jl}\}$, where $\hat{\pi}_{jl} = n_{jl}/n$ and n_{jl} is the number of subjects with $Y = j$ and $X = l$.

We then summarize the observed and expected distributions separately. This can be achieved by calculating Goodman and Kruskal's gamma (Goodman and Kruskal 1954), which for a two-way probability distribution $P = \{\pi_{jl}\}$ is $\Gamma(P) = \frac{C-D}{C+D}$, where $C = \sum_{j_1<j_2, l_1<l_2} \pi_{j_1 l_1}\pi_{j_2 l_2}$ and $D = \sum_{j_1<j_2, l_1>l_2} \pi_{j_1 l_1}\pi_{j_2 l_2}$. Let $\Gamma_1 = \Gamma(\hat{P})$ and $\Gamma_0 = \Gamma(\hat{P}_0)$ be the gamma for the observed and expected distributions for (Y, X). Our test statistic will be $T_1 = \Gamma_1 - \Gamma_0$. When the multinomial models are correct, under the null, $\hat{P}_0 \to P$ and $\hat{P} \to P$ as $n \to \infty$, and thus $T_1 = \Gamma(\hat{P}) - \Gamma(\hat{P}_0) \to 0$.

Note that our test statistic accounts for the order information in Y and X, whereas a direct goodness-of-fit approach comparing the observed and expected counts using a statistic in the form of $\sum_{j,l}(\text{Observed} - \text{Expected})^2/\text{Expected}$, ignores the order information in Y and X.

2.2. Residual-based Test Statistics

Another way of constructing test statistics is to mimic the residual-based linear regression analysis described in the Introduction. We wish to calculate "residuals" for multinomial models of ordinal outcomes and test if they correlate. For ordered categorical outcome variables, however, we are not aware of a standard way of calculating individual-level residuals. We first define individual-level residuals for ordinal outcome variables and then present test statistics that are based on these residuals.

In linear regression, to calculate the residual for a subject with outcome $Y = y$ and input $\boldsymbol{Z} = \boldsymbol{z}$, we first obtain the fitted value of the outcome variable given $\boldsymbol{z}$, $\hat{y} = \mathrm{E}(Y|\boldsymbol{z})$, and then calculate the residual as $y - \hat{y}$. However, for an ordered categorical outcome variable, we cannot calculate its "fitted" value and need to rethink the derivation of residuals. It should be noted that in addition to providing the fitted value described above, a linear regression model also gives an estimated distribution of possible outcome values given $\boldsymbol{z}$, say $Y_{fit} \sim Y|\boldsymbol{z}$. The residual can be written as $y - \hat{y} = y - \mathrm{E}(Y_{fit}) = \mathrm{E}(y - Y_{fit})$. In other words, we may think of a random variable $y - Y_{fit}$, which is the difference between the observed outcome value y and a random outcome value under the model given $\boldsymbol{z}$; the residual is the expectation of this random variable. This motivates us to define residuals for ordered categorical outcome variables.

Assume a multinomial model for $P(Y|\boldsymbol{Z})$ with model parameters $\boldsymbol{\theta}^Y$. Let $Y_i = y_i$ be the observed outcome level for subject i. The corresponding distribution of possible outcome levels $Y_{i,fit} \sim Y_i|\boldsymbol{z}_i$ given covariate $\boldsymbol{z}_i$ is multinomial with probability distribution $\{p_i^j\}$. Since the outcome variable is ordered categorical, we cannot calculate the difference between y_i and $Y_{i,fit}$, but we can compare them with respect to whether y_i is at a higher or lower level than $Y_{i,fit}$. The probability for y_i to be higher than $Y_{i,fit}$ is $p_{i,high} = P(y_i > Y_{i,fit}) = \gamma_i^{y_i-1}$, where $\gamma_i^j = P(Y_i \leq j|\boldsymbol{Z} = \boldsymbol{z}_i)$ as defined earlier; similarly, the probability for y_i to be lower than $Y_{i,fit}$ is $p_{i,low} = P(y_i < Y_{i,fit}) = 1 - \gamma_i^{y_i}$; the probability for y_i to tie with $Y_{i,fit}$ is $P(y_i = Y_{i,fit}) = p_i^{y_i}$. We then assign scores to these three types of events: 1, -1, and 0, corresponding to higher, lower, and tie, respectively. The expected score $p_{i,high} - p_{i,low}$ is denoted as $Y_{i,res}$, which is a function of data $(Y_i, \boldsymbol{Z}_i)$ and model parameters $\boldsymbol{\theta}^Y$. For a fitted model with parameter estimates $\hat{\boldsymbol{\theta}}^Y$, the residual for subject i is defined as $y_{i,res} = Y_{i,res|\hat{\boldsymbol{\theta}}^Y}$. For multinomial models for $P(X|\boldsymbol{Z})$, the probabilities $q_{i,high}$ and $q_{i,low}$ can be similarly defined and so are $X_{i,res} = q_{i,high} - q_{i,low}$ and $x_{i,res}$.

When the models for $P(Y|\boldsymbol{Z})$ and $P(X|\boldsymbol{Z})$ are correct, we have $\mathrm{E}(p_{i,high}|\boldsymbol{Z}_i) = \mathrm{E}(\gamma_i^{Y_i-1}|\boldsymbol{Z}_i) = \sum_{j=1}^{s} p_i^j \gamma_i^{j-1} = \sum_{j_1<j_2} p_i^{j_1} p_i^{j_2}$, and similarly $\mathrm{E}(p_{i,low}|\boldsymbol{Z}_i) = \sum_{j_1<j_2} p_i^{j_1} p_i^{j_2}$. Therefore, $\mathrm{E}(Y_{i,res}|\boldsymbol{Z}_i) = \mathrm{E}(p_{i,high} - p_{i,low}|\boldsymbol{Z}_i) = 0$ and $\mathrm{E}(Y_{i,res}) = \mathrm{E}[\mathrm{E}(Y_{i,res}|\boldsymbol{Z}_i)] = 0$. Similarly, $\mathrm{E}(X_{i,res}|\boldsymbol{Z}_i) = 0$ and $\mathrm{E}(X_{i,res}) = 0$. Under the null, since

$$\mathrm{E}(Y_{i,res}X_{i,res}) = \mathrm{E}\left[\mathrm{E}(Y_{i,res}X_{i,res}|\boldsymbol{Z}_i)\right] = \mathrm{E}\left[\mathrm{E}(Y_{i,res}|\boldsymbol{Z}_i)\mathrm{E}(X_{i,res}|\boldsymbol{Z}_i)\right] = 0,$$

we have $\mathrm{cov}(Y_{i,res}, X_{i,res}) = \mathrm{E}(Y_{i,res}X_{i,res}) - \mathrm{E}(Y_{i,res})\mathrm{E}(X_{i,res}) = 0$, and thus the correlation between $Y_{i,res}$ and $X_{i,res}$ is zero.

Once the residuals for models for $P(Y|\boldsymbol{Z})$ and $P(X|\boldsymbol{Z})$ have been calculated, the sample correlation coefficent between $Y_{i,res}$ and $X_{i,res}$ across all subjects, T_2, can be used as a test statistic. Under the null, T_2 converges to zero as $n \rightarrow \infty$.

A variation of the above approach is to compare the observed value of (Y_i, X_i) for subject i with the distribution of possible values of (Y, X) given covariate $\boldsymbol{z}_i$. Under the null, subject i's observed value (Y_i, X_i) should follow the product distribution $\{p_i^j q_i^l\}$. Consider drawing a random value from the subject's product distribution, (Y_i', X_i'), and comparing it with the observed value (Y_i, X_i). If $Y_i' > Y_i$ and $X_i' > X_i$ (or $Y_i' < Y_i$ and $X_i' < X_i$), that is, both variables are in the same direction, then we record "concordance"; if the two variables are in opposite directions, then we record "discordance"; otherwise we record a tie. Under the null, since both (Y_i', X_i') and (Y_i, X_i) follow the same product distribution, the probability of concordance is equal to the probability of discordance. In fact, there is no need to draw (Y_i', X_i'), as under the null, we can derive the probability of concordance as $C_i = p_{i,high}q_{i,high} + p_{i,low}q_{i,low}$, and the probability of discordance as $D_i = p_{i,high}q_{i,low} + p_{i,low}q_{i,high}$. Since $C_i - D_i = (p_{i,high} - p_{i,low})(q_{i,high} - q_{i,low}) = Y_{i,res}X_{i,res}$, we have $\mathrm{E}(C_i - D_i) = \mathrm{E}(Y_{i,res}X_{i,res}) = 0$. Our third test statistic will therefore be the average difference of these probabilities across all subjects, $T_3 = \frac{1}{n}\sum_i(\hat{C}_i - \hat{D}_i) = \frac{1}{n}\sum_i y_{i,res}x_{i,res}$.

3. DISTRIBUTION OF TEST STATISTICS UNDER THE NULL

We present two approaches to obtaining p-values for our test statistics. One is based on empirical distributions generated under the null, the other is based on asymptotic distributions derived from estimating equations.

3.1. Empirical Distribution

Let T be one of the three test statistics described in the last section. To generate an empirical distribution of T, we simulate replicate data sets under the null. To simulate a replicate data set, we randomly generate one observation from the product distribution $\{\hat{p}_i^j \hat{q}_i^l\}_{j,l}$; this is done for $i = 1, \cdots, n$. Then we carry out the entire estimating procedure for the replicate data set to obtain the corresponding statistic, denoted as T^* (i.e., fit separate multinomial models, obtain predicted probabilities, and calculate test statistic). This is then repeated many, say N^{emp}, times, to get an empirical distribution of T under the null. The two-sided p-value is then computed as either

$$\#(|T^*| \geq |T|)/N^{emp}$$

or

$$2 \times \min\{\#(T^* \geq T), \#(T^* \leq T)\}/N^{emp}.$$

From our simulations the results are almost the same for these two p-values, so we will present only the first. This procedure is essentially a parametric bootstrap procedure (Efron and Tibshirani 1993).

3.2. Asymptotic Distribution

An alternative approach to computing the p-value is to use the asymptotic distributions of our test statistics under the null hypothesis. In general, we define a vector of parameters $\boldsymbol{\theta}$ of length p, whose estimate $\hat{\boldsymbol{\theta}}$ can be obtained by solving the equation $\sum_{i=1}^{n} \Psi_i(\boldsymbol{\theta}) = \mathbf{0}$, where $\Psi_i(\boldsymbol{\theta}) = \Psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ is a p-variate function that does not depend on i or n and satisfies $\mathrm{E}_{\boldsymbol{\theta}}[\Psi_i(\boldsymbol{\theta})] = \mathbf{0}$. From M-estimation theory (Stefanski and Boos 2002), if Ψ is suitably smooth, then as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to_d N(\mathbf{0}, V(\boldsymbol{\theta})),$$

where $V(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) [A(\boldsymbol{\theta})^{-1}]'$, $A(\boldsymbol{\theta}) = \mathrm{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} \Psi_i(\boldsymbol{\theta})\right]$, and $B(\boldsymbol{\theta}) = \mathrm{E}\left[\Psi_i(\boldsymbol{\theta}) \Psi_i(\boldsymbol{\theta})'\right]$. If a test statistic T is a smooth function of $\hat{\boldsymbol{\theta}}$, $T = g(\hat{\boldsymbol{\theta}})$, then from the delta method,

$$\sqrt{n}[g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})] \to_d N(0, \sigma^2),$$

where $\sigma^2 = \left[\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta})\right] V(\boldsymbol{\theta}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta})\right]'$. To estimate σ^2, we estimate $A(\boldsymbol{\theta})$ as $\frac{1}{n}\sum_i [-\frac{\partial}{\partial \boldsymbol{\theta}} \Psi_i(\hat{\boldsymbol{\theta}})]$, $B(\boldsymbol{\theta})$ as $\frac{1}{n}\sum_i \Psi_i(\hat{\boldsymbol{\theta}}) \Psi_i(\hat{\boldsymbol{\theta}})'$, and $\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta})$ as $\frac{\partial}{\partial \boldsymbol{\theta}} g(\hat{\boldsymbol{\theta}})$. If $g(\boldsymbol{\theta}) = 0$ under the null, then the p-value can be computed approximately as $2\Phi\left(\frac{-|T|}{\hat{\sigma}/\sqrt{n}}\right)$, where Φ is the cumulative distribution function of the standard normal distribution.

We now define $\boldsymbol{\theta}$, $\Psi_i(\boldsymbol{\theta})$, and $g(\boldsymbol{\theta})$ for the three test statistics introduced in Section 2. For all three statistics, the parameter vector will have the form $\boldsymbol{\theta} = (\boldsymbol{\theta}^Y, \boldsymbol{\theta}^X, \boldsymbol{\theta}^T)$, where $\boldsymbol{\theta}^T$ is different for each statistic. The corresponding estimating function $\Psi_i(\boldsymbol{\theta})$ will have the form

$$\Psi_i(\boldsymbol{\theta}) = \begin{cases} \frac{\partial}{\partial \boldsymbol{\theta}^Y} l_Y(Y_i, \boldsymbol{Z}_i; \boldsymbol{\theta}^Y) \\ \frac{\partial}{\partial \boldsymbol{\theta}^X} l_X(X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}^X) \\ \psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}), \end{cases}$$

where l_Y and l_X are the log-likelihood functions of the multinomial models that are used to model $P(Y|\boldsymbol{Z})$ and $P(X|\boldsymbol{Z})$, with parameters $\boldsymbol{\theta}^Y$ and $\boldsymbol{\theta}^X$, respectively. They are score functions and thus $\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^Y} l_Y(Y_i, \boldsymbol{Z}_i; \boldsymbol{\theta}^Y)\right] = \mathbf{0}$ and $\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^X} l_X(X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}^X)\right] = \mathbf{0}$. The function $\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta})$ will be different for each statistic.

For $T_1 = \Gamma(\hat{P}) - \Gamma(\hat{P}_0)$, we define $\boldsymbol{\theta}^T = (\pi_{11}, \cdots, \pi_{1t}, \pi_{21}, \cdots, \pi_{2t}, \cdots, \pi_{s1}, \cdots, \pi_{s,t-1})$, where $\pi_{jl} = P(Y = j, X = l)$ is not conditional on any covariates. Note that $\boldsymbol{\theta}^T$ does not contain π_{st}, which is not an independent parameter because $\sum_{j,l} \pi_{jl} = 1$. The corresponding function ψ only depends on $\boldsymbol{\theta}^T$:

$$\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}^T) = \begin{cases} I_{\{Y_i=1, X_i=1\}} - \pi_{11} \\ \vdots \\ I_{\{Y_i=s, X_i=t-1\}} - \pi_{s,t-1}, \end{cases}$$

where I_a is the indicator function of event a. By definition, $\mathrm{E}[\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}^T)] = \mathbf{0}$. Let $g(\boldsymbol{\theta}) = \Gamma(P) - \Gamma(P_0)$. Then $g(\boldsymbol{\theta}) = 0$ under the null, and $T_1 = g(\hat{\boldsymbol{\theta}})$.

For T_2, the sample correlation between residuals, $\boldsymbol{\theta}^T = (w_1, w_2, w_3, w_4, w_5)$, where $w_1 = \mathrm{E}(Y_{i,res})$, $w_2 = \mathrm{E}(X_{i,res})$, $w_3 = \mathrm{E}(Y_{i,res} X_{i,res})$, $w_4 = \mathrm{E}(Y_{i,res}^2)$, and $w_5 = \mathrm{E}(X_{i,res}^2)$. The corresponding function ψ is

$$\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \begin{cases} Y_{i,res} - w_1 \\ X_{i,res} - w_2 \\ Y_{i,res} X_{i,res} - w_3 \\ Y_{i,res}^2 - w_4 \\ X_{i,res}^2 - w_5. \end{cases}$$

By definition, $\mathrm{E}[\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta})] = \mathbf{0}$. Solving the equation $\sum_i \Psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = \mathbf{0}$, we have $\hat{w}_1 = \frac{1}{n} \sum_i y_{i,res}$, $\hat{w}_2 = \frac{1}{n} \sum_i x_{i,res}$, $\hat{w}_3 = \frac{1}{n} \sum_i y_{i,res} x_{i,res}$, $\hat{w}_4 = \frac{1}{n} \sum_i y_{i,res}^2$, and $\hat{w}_5 = \frac{1}{n} \sum_i x_{i,res}^2$. Let $g(\boldsymbol{\theta}) = (w_3 - w_1 w_2) / \sqrt{(w_4 - w_1^2)(w_5 - w_2^2)} = \mathrm{cor}(Y_{i,res}, X_{i,res})$. Then $g(\boldsymbol{\theta}) = 0$ under the null, and $T_2 = g(\hat{\boldsymbol{\theta}})$.

For $T_3 = \frac{1}{n}\sum_i(\hat{C}_i - \hat{D}_i)$, $\boldsymbol{\theta}^T$ is a single parameter $\theta^T = \mathrm{E}(C_i - D_i)$, which is zero under the null. The corresponding function ψ is

$$\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) = (C_i - D_i) - \theta^T.$$

By definition, $\mathrm{E}[\psi(Y_i, X_i, \boldsymbol{Z}_i; \boldsymbol{\theta})] = 0$. Let $g(\boldsymbol{\theta}) = \theta^T$. Then $g(\boldsymbol{\theta}) = 0$ under the null, and $T_3 = g(\hat{\boldsymbol{\theta}})$. In fact, the delta method is not needed here as $\hat{\sigma}^2$ is simply the last element of the diagonal of $\hat{V}(\boldsymbol{\theta})$.

4. SIMULATIONS

We carried out simulations to investigate the performance of our method and to compare it with four other approaches: (a) proportional odds model with X coded as a continuous variable,

$$P(Y \le j|\boldsymbol{Z}) = \left[1 + \exp(-(\alpha_j^Y + \boldsymbol{Z}\boldsymbol{\beta}^Y + \eta X))\right]^{-1},$$

and testing if $\eta = 0$; (b) proportional odds model with X coded as a categorical variable (using dummy variables),

$$P(Y \le j|\boldsymbol{Z}) = \left[1 + \exp(-(\alpha_j^Y + \boldsymbol{Z}\boldsymbol{\beta}^Y + \eta_2 I_{\{X=2\}} + \cdots + \eta_t I_{\{X=t\}}))\right]^{-1},$$

and testing if $\eta_2 = \cdots = \eta_t = 0$; (c) isotonic proportional odds model, in which X was treated as a categorical variable and adjacent categories were combined to enforce monotonicity if necessary; and (d) proportional odds model with X transformed using restricted cubic splines with three pre-selected knots. For our method, we computed the p-value using both the empirical and asymptotic distribution approaches.

We investigated the performance of these approaches under multiple simulation scenarios. The first scenario was under the null to investigate type I error rate. Next, we constructed various alternatives in manners so that different modeling assumptions would be favored. In our first alternative scenario, we generated data such that the effect of the ordinal categories $(1^X, \cdots, t^X)$ was linear, in the sense that had we simply done an analysis treating X as a continuous variable we would have gotten the correct answer. For the second alternative scenario, we generated data such that the effect of the levels of X was monotonic in a non-linear fashion. Finally, we considered a simulation scenario with a non-monotonic relationship between Y and X, a scenario that favors modeling X as categorical.

The specifics of our four data generating scenarios are as follows: We first generated a covariate Z using the standard normal distribution. Then we generated X with five categories using the proportional odds model

$$P(X \leq l|\boldsymbol{Z}) = \left[1 + \exp(-(\alpha_l^X + \beta^X Z))\right]^{-1}$$

with $\boldsymbol{\alpha}^X = (\alpha_1^X, \cdots, \alpha_4^X) = (-1, 0, 1, 2)$ and $\beta^X = 1$. The outcome variable Y was generated with four levels using the proportional odds model

$$P(Y \leq j|Z) = \left[1 + \exp(-(\alpha_j^Y + \beta^Y Z + \eta_1 I_{\{X=1\}} + \cdots + \eta_5 I_{\{X=5\}}))\right]^{-1}$$

with $\boldsymbol{\alpha}^Y = (\alpha_1^Y, \alpha_2^Y, \alpha_3^Y) = (-1, 0, 1)$, $\beta^Y = -0.5$, and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_5)$ specified as

1. $\boldsymbol{\eta} = (0, 0, 0, 0, 0)$ (the null),

2. $\boldsymbol{\eta} = (-0.4, -0.2, 0, 0.2, 0.4)$ (linear effect),

3. $\boldsymbol{\eta} = (-0.30, 0.18, 0.20, 0.22, 0.24)$ (monotonic non-linear effect), and

4. $\boldsymbol{\eta} = (-0.2, 0, 0.2, 0, -0.2)$ (non-monotonic effect).

For each simulation scenario, we generated 10,000 data sets, each consisting of 500 subjects. To obtain p-values using the empirical distribution of our test statistics, for each data set we generated 1000 replicates. For all simulation scenarios, the null was rejected if the two-sided p-value was less than 0.05.

Simulation results are summarized in Table 1. Under the null hypothesis, the type I error rate was at the nominal 5% level for all analysis methods except isotonic regression; in order to achieve the nominal level using isotonic regression we would need to account for model-selection using some sort of re-sampling procedure. Under the linear scenario, the highest power was obtained under the properly specified model with X treated as a continuous variable. Also as expected, power was lower when X was treated as a continuous variable using splines, or as a categorical variable with or without isotonic regression; this lower power was expected because of the additional degrees of freedom employed by each of these methods. Somewhat surprising was the minimal loss in power seen using any of our new test statistics, ranging around 85-86%, just below the 87% power seen in the properly specified model.

Under the non-linear monotonic simulations, our methods had better power than simple analyses treating X as continuous or categorical. The most powerful approach expanded X using restricted cubic splines; power was similar between our method and isotonic regression (although isotonic regression had inflated type I error rate, so presumably its power would be smaller had we accounted for model-selection uncertainty). Finally, when the true relationship between X and Y conditional on Z was non-monotonic, our test statistics had poor power. This was expected, as our methods assume monotonicity. Splines and treating X as categorical had higher power under this simulation scenario. Isotonic regression also had inflated power, although given that isotonic regression assumes monotonicity and the U-shaped effect of our simulation scenario, we believe many of these are false positives.

It is worth noting that power was comparable between all of our test statistics (T_1, T_2, and T_3) regardless of how we computed the p-value (empirically or asymptotically). Figure 1 shows the distribution of asymptotic p-values under the null for T_1, T_2, and T_3; all three are highly correlated, the residual-based test-statistics (T_2

and T_3) particularly so.

We also evaluated the type I error rate of our method at smaller sample sizes ($n = 50, 100$). As expected, empirical-based p-values were more accurate than their asymptotic counterparts, although both approaches yielded type I error rates close to the nominal 5% level (Table 2).

We also performed a simulation where the model for $P(X|Z)$ was incorrectly specified. Specifically, we generated data as before ($n = 500$) except that X was generated with $\beta^X = 0, 1, 2, 3$ for $l = 1, 2, 3, 4$, respectively. We then performed the analyses assuming a proportional odds model (i.e., constant β^X) for $P(X|Z)$. The type I error rates were under control (Table 2), and the power under the alternative simulation scenarios was also similar to that reported in Table 1 (data not shown).

Finally, we also performed additional simulations with different numbers of categories for Y and X, and saw results similar to those reported here.

5. EXAMPLES

5.1. Anisometropic amblyopia

Amblyopia is a leading cause of acquired monocular visual impairment. Anisometropic amblyopia typically presents later than other types of amblyopia. Because improvement in visual acuity with amblyopia treatment depends on the age at which treatment begins, earlier detection of children with anisometropic amblyopia is desired. In a photoscreening program of preschool children, anisometropia (≥1D difference in refractive power between the eyes in any meridian) was detected on 974 preschool children (Leon et al. 2008). Anisometropia magnitude (difference in spherical equivalent <1D, 1 to <2D, 2 to <4D, and ≥4D) was measured along with age and visual acuity, which is used to define amblyopia levels (severe, moderate, mild, and no amblyopia). There is interest in testing the association between anisometropia and amblyopia while adjusting for the effect of age.

Investigators carried out ordinal logistic regression with amblyopia as outcome (Y) and anisometropia (X) and age (Z) as continuous input variables (Leon et al. 2008) and found highly significant association between amblyopia and anisometropia after adjustment for age ($p < 10^{-20}$). Applying the methods described in this paper with ordinal logistic regression as models for $P(Y|Z)$ and $P(X|Z)$ led to an even smaller p-value. However, p-value comparisons at this magnitude are essentially meaningless as they lead to the same conclusion. Such significant results are partly due to the large sample size of the study. The comparison between statistical methods would be more meaningful if the sample size had been smaller. Thus, we generated 50 data sets of 50 subjects randomly selected from the 974 children, and compared the results of our method and the method used by the investigators (Figure 2). Computing p-values based on the asymptotic distribution of the test statistic, T_1 yielded smaller p-values than the method used in the original study analysis. However, when p-values were calculated based on empirical distributions, they were similar to the p-values of the method used in the original analysis. This is consistent with our simulation results with $n = 50$ where the asymptotic p-values for T_1 had slightly inflated type I error rates. Results were similar for T_2. In contrast, p-values based on the asymptotic distribution of T_3 tended to be slightly larger than their empirical counterparts, also consistent with our simulation results for $n = 50$ (Table 2).

The residuals of our method using all subjects are plotted in Figure 3. Note that the four levels of anisometropia are still well separated in the residual plot, while the four levels of amblyopia are overlapping. This indicates that age has a relatively weaker association with anisometropia than with amblyopia.

5.2. Cervical neoplastic stage and condom use

Cervical specimens for 150 non-pregnant HIV-infected women in Lusaka, Zambia were collected (Parham et al. 2006). Based on cytological analysis, 36 specimens were categorized as normal, 35 as low-grade squamous intraepithelial lesions, 49 as high-grade squamous intraepithelial lesions, and 30 as squamous cell carcinoma. The women also reported condom use: 53 women reported never using condoms, 60 rarely, 13 almost always, and 24 always. Researchers were interested in testing for an association between cervical stage and condom use. Kendall's rank correlation tau was estimated as -0.03 (using R function `cor.test` with argument `method = "kendall"`), which was not statistically different from zero (p-value=0.64). However, CD4 T-cell count, age, education, and marital status have been linked to cervical abnormalities and may be associated with condom use, so researchers wanted to test for an association between cervical stage and reported condom use after adjusting for these variables. Our method was applied resulting in p-values ranging from 0.76 to 0.91, depending on the test statistic employed, indicating that there was insufficient evidence to conclude that cytological abnormalities were associated with condom use after adjusting for CD4 T-cell count, age, education, and marital status. For sake of comparison, treating condom use as a continuous variable (1, 2, 3, or 4) and putting it in an ordinal logistic model assuming linearity and expanding using restricted cubic splines with 3 knots yielded p-values of 0.52 and 0.48, respectively. When condom use was treated as a categorical variable, the p-value was 0.66. The residuals of our method are plotted in Figure 3.

6. DISCUSSION

We have developed a new method for testing for associations between two ordinal categorical variables while adjusting for other continuous or categorical variables. In our approach, we separately fit the two ordinal variables on the other covariates using multinomial models such as ordinal logistic regression and then built test statistics based on the predicted probability distributions for these two variables. For our three test statistics, we described approaches to calculate p-values based on either empirical or asymptotic distributions. Our methods are simple to implement and simulations showed our new tests are powerful to detect monotonic associations between two ordinal variables while appropriately adjusting for the effects of other covariates.

In the process of constructing test statistics, we defined a new concept of residual for ordinal outcome variables. This residual can be calculated for any multinomial regression model as long as the outcome variable is ordinal. In our definition of residuals, we assigned scores $+1$, -1, and 0, reflecting the direction of comparison between the observed and expected outcomes; positive (negative) residuals imply the observed level is higher (lower) than the expected. Our residuals are consistent with concordance-discordance statistics such as Kendall's tau and Goodman and Kruskal's gamma, which similarly compare the direction between observations, but make no assumption regarding the magnitude of the distance between ordered categories. Our definition results in one residual per subject, which is therefore useful for constructing test statistics. Our residuals may also be useful in other ways (e.g., diagnostics), which we are currently studying.

Other types of residuals, such as deviance and Pearson residuals, have been defined for logistic regression (Agresti 2002), and they can be extended to multinomial outcomes. However, deviance residuals ignore the order information, and Pearson residuals result in multiple values for each subject when there are more than two levels. Hence, the utility of these residuals for our purposes is not readily apparent. For proportional odds models, McCullagh (1980) described a different concept of residual. However, this residual is defined for each level of the multinomial outcome variable and is "always positive and thus does not indicate the direction of departure of the observed values from the fitted values."

Our method has some features which may be undesirable in certain scenarios.

First, if one of the two ordinal variables can be designated as the outcome variable, a traditional regression analysis with the other variable as a predictor can model interactive effects between the predictor variable and the covariates. In our method, both ordinal variables are treated equally, avoiding the need to pick one as the outcome, but we therefore assume no interaction effects exist between the predictor and other covariates. Second, our method requires explicit modeling of the relationship between Y and $\boldsymbol{Z}$ and between X and $\boldsymbol{Z}$. The consistency of our results therefore depends on correct specification of these models. It should be recognized that our method is applicable for any (even different) multinomial regression models of Y on $\boldsymbol{Z}$ and X on $\boldsymbol{Z}$. And our limited simulations suggest results are fairly robust to misspecifications of these models, although this warrants further investigation. Note that a single regression analysis of Y on X and $\boldsymbol{Z}$ does not require modeling the relationship between X and $\boldsymbol{Z}$, but requires explicit specification of the pattern of effects of X and $\boldsymbol{Z}$ on Y. Third, our approach is designed for testing the relationship between two ordinal variables while adjusting for other covariates. When the primary interest is the relationship between an ordinal outcome and a continous or categorical variable while adjusting for an ordinal covariate, our methods may not be useful.

Although our presentation focused on hypothesis testing, our test statistics are to some extent interpretable and thus may be used to measure the magnitude of association between Y and X conditional on $\boldsymbol{Z}$. The test statistic T_1 captures the discrepancy in gamma (the difference in probability of concordant and discordant random pairs) between the observed distribution for (Y, X) and the expected distribution under the null. The test statistic T_2 is the correlation coefficient between the residuals of models $Y|\boldsymbol{Z}$ and $X|\boldsymbol{Z}$.

When an ordinal variable has only two levels, it is a binary variable. For a binary outcome variable, logistic regression is often used for association testing. For a binary input variable, treating it as continuous or categorical would result in the same model. It is interesting to evaluate how our method performs when Y or X or both are binary variables. If X is binary, our simulations showed that our approach yields similar results to ordinal logistic regression with X treated as a categorical variable (data not shown). If Y is binary, our approach yields results consistent with ordinal logistic regression treating X as the outcome variable and Y as a categorical predictor. Finally, if both Y and X are binary, our approach yields results consistent with logistic regression with X as a dichotomous predictor.

Another possible direction of research is to extend the weighted average of stratum-specific association measures described in the Introduction to continuous or multivariable $\boldsymbol{Z}$. The weighted average approach currently requires grouping a continuous or multivariable $\boldsymbol{Z}$ into discrete categories, and computing concordance and discordance only for subject pairs falling in the same category. An alternative is to score concordance and discordance for all subject pairs, but to weight the scores according to the similarity in $\boldsymbol{Z}$ between subjects. We are working to evaluate the performance of this approach.

Finally, our methods suggest a potential solution to the general problem in regression when the input variable of interest is ordinal. As we stated in the Introduction, in any regression analysis, ordinal predictor variables have to be treated as continuous or categorical variables, imposing a linearity assumption in the former and ignoring order information in the latter. We have defined individual-level residuals for ordinal variables and developed a residual-based method for testing correlation between ordinal Y and X. We believe a similar approach can be developed when Y is another type of variable as long as its individual-level residuals can be calculated. We are also working to evaluate this approach for various variable types for Y.

Regression analysis with ordinal input variables has been difficult to deal with

appropriately. Our method will be useful for testing for association between two ordinal variables while adjusting for other covariates. Our simulation and data analysis code is available as supplemental materials and also at
http://biostat.mc.vanderbilt.edu/OrdinalRegression .

Supplemental Materials

Title: Code for data analysis and simulations

R-functions for data analysis: `COBOT-analysis.r` contains R-functions for analyzing data using proportional odds models for $P(Y|\boldsymbol{Z})$ and $P(X|\boldsymbol{Z})$. COBOT stands for conditional ordinal-by-ordinal test.

R-functions for simulations: `COBOT-simulation.r`.

Other simulation results: `ordinal-otherresults.pdf`

References

Agresti, A. (1977), "Considerations in measuring partial association for ordinal categorical data," *Journal of the American Statistical Association*, 72, 37–45.

Agresti, A. (2002), *Categorical Data Analysis* (2nd ed.), Hoboken, New Jersey: John Wiley.

Aitchison, J., and Silvey, S. D. (1957), "The generalization of probit analysis to the case of multiple responses," *Biometrika*, 44, 131–140.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference under Order Restrictions*, London: John Wiley.

Davis, J. A. (1967), "A partial coefficient for Goodman and Kruskal's gamma," *Journal of the American Statistical Association*, 62, 189–193.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall.

Farewell, V. T. (1982), "A note on regression analysis of ordinal data with variability of classification," *Biometrika*, 69, 533–538.

Goodman, L. A. (1959), "Partial tests for partial taus," *Biometrika*, 46, 425–432.

Goodman, L. A. (1979), "Simple models for the analysis of association in cross-classifications having ordered categories," *Journal of the American Statistical Association*, 74, 537–552.

Goodman, L. A., and Kruskal, W. H. (1954), "Measures of association for cross classifications," *Journal of the American Statistical Association*, 49, 732–764.

Hawkes, R. K. (1971), "The multivariate analysis of ordinal measures," *American Journal of Sociology*, 76, 908–926.

Kendall, M. G. (1948), *Rank Correlation Methods*, Hafner: New York.

Läärä, E., and Matthews, J. N. S. (1985), "The equivalence of two models for ordinal data," *Biometrika*, 72, 206–207.

Leon, A., Donahue, S. P., Morrison, D. G., Estes, R. L., and Li, C. (2008), "The age-dependent effect of anisometropia magnitude on anisometropic amblyopia severity," *Journal of AAPOS*, 12, 150–156.

Mantel, N. (1963), "Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure," *Journal of the American Statistical Association*, 58, 690–700.

Mantel, N. (1966), "Models for complex contigency tables and polychotomous dosage response curve," *Biometrics*, 22, 83–95.

McCullagh, P. (1980), "Regression models for ordinal data," *Journal of the Royal Statistical Society*, Ser. B, 42, 109–142.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, Mass: Addition-Wesley.

Parham, G. P., Sahasrabuddhe, V. V., Mwanahamuntu, M. H., Shepherd, B. E., Hicks, M. L., Stringer, E. M., and Vermund, S. H. (2006), "Prevalence and predictors of squamous intraepithelial lesions of the cervix in HIV-infected women in Lusaka, Zambia," *Gynecologic Oncology*, 103, 1017–1022.

Ramsay, J. O. (1988), "Monotone regression splines in action," *Statistical Science*, 3, 425-441.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley & Sons.

Schemper, M. (1991), "Non-parametric partial association revisited," *The Statistician*, 40, 73–76.

Somers, R. H. (1962), "A new asymmetric measure for ordinal variables," *American Sociological Review*, 27, 799–811.

Stefanski, L. A., and Boos, D. D. (2002), "The calculus of M-estimation," *The American Statistician*, 56, 29–38.

Torgerson, W. S. (1956), "A non-parametric test of correlation using rank orders within subgroups," *Psychometrika*, 21, 145–152.

Table 1: Type I Error Rate and Power (%)

Analysis method	Simulation scenarios			
	Null	Linear	Non-linear	Non-monotonic
T_1 Empirical	5.0	86.0	57.5	7.4
Asymptotic	4.8	85.4	56.4	7.0
T_2 Empirical	5.0	85.8	57.7	6.9
Asymptotic	4.6	85.9	57.8	7.0
T_3 Empirical	5.0	85.9	58.0	7.0
Asymptotic	4.9	85.2	57.0	6.6
X linear	4.9	87.4	52.4	5.7
X categorical	5.1	70.3	52.5	28.5
Isotonic	7.1	77.5	57.4	21.9
Splines	4.8	79.9	60.0	34.8

Note: $n = 500$

Table 2: Type I Error Rate for Small Sample Sizes and Violation of Model Assumptions (%)

Analysis method	$n = 50$	$n = 100$	Misspecified $P(X\|Z)$
T_1 Empirical	4.7	4.4	5.2
Asymptotic	6.0	4.8	4.9
T_2 Empirical	5.1	4.7	5.2
Asymptotic	7.0	5.6	5.3
T_3 Empirical	5.6	5.0	5.0
Asymptotic	4.0	4.1	5.2
X linear	5.3	4.8	5.1
X categorical	4.2	4.5	5.2
Isotonic	7.1	7.0	7.1
Splines	5.2	4.6	5.2

Figure 1: Comparison of p-values of T_1, T_2, and T_3 under the null. Top row are all p-values on log-scale. Bottom row contains p-values between 0 and 0.1. Sample size is 500.

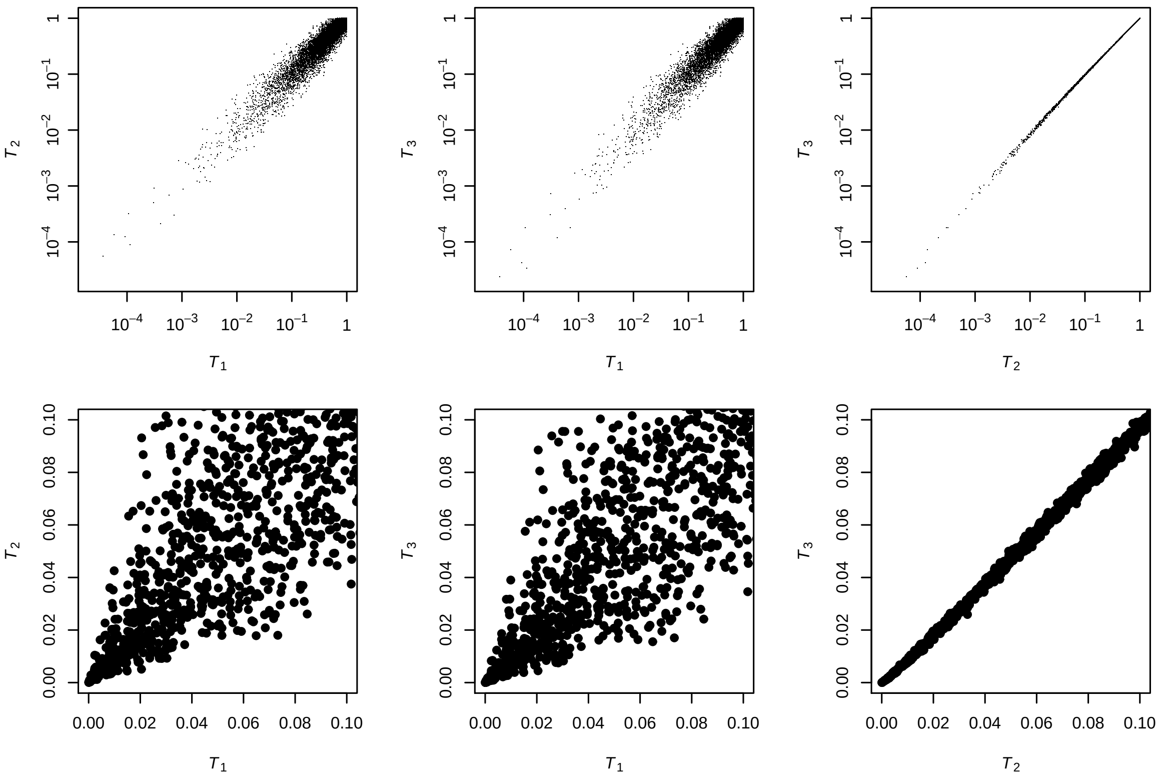


Figure 2: Results of amblyopia data analysis. The x-axis is the p-value based on ordinal logistic regression with anisometropia treated as a continuous variable. The y-axis is the p-value using T_1. The left and right plots contain p-values based on the asymptotic and empirical distributions of T_1, respectively. Each point represents results using a random sample of 50 children.

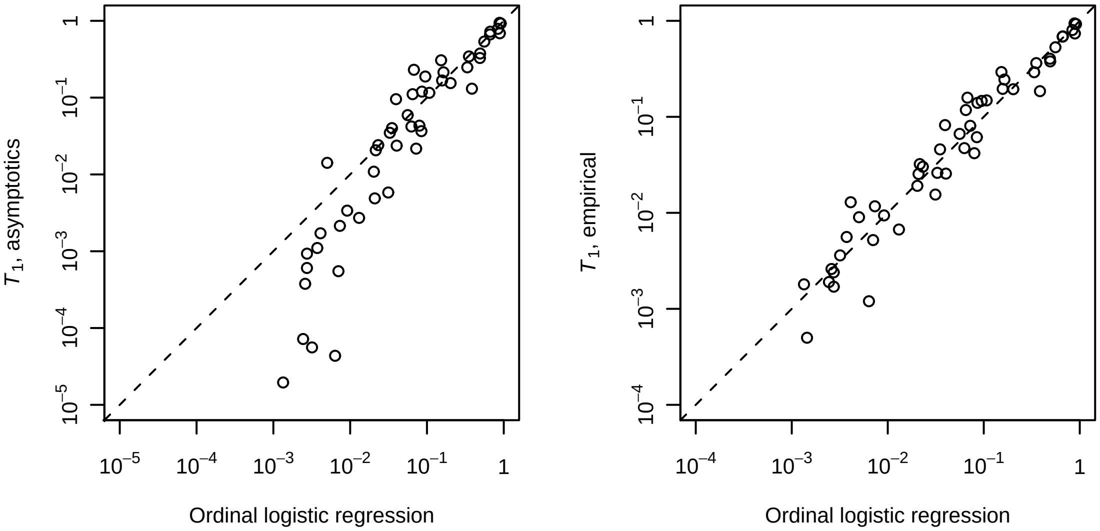
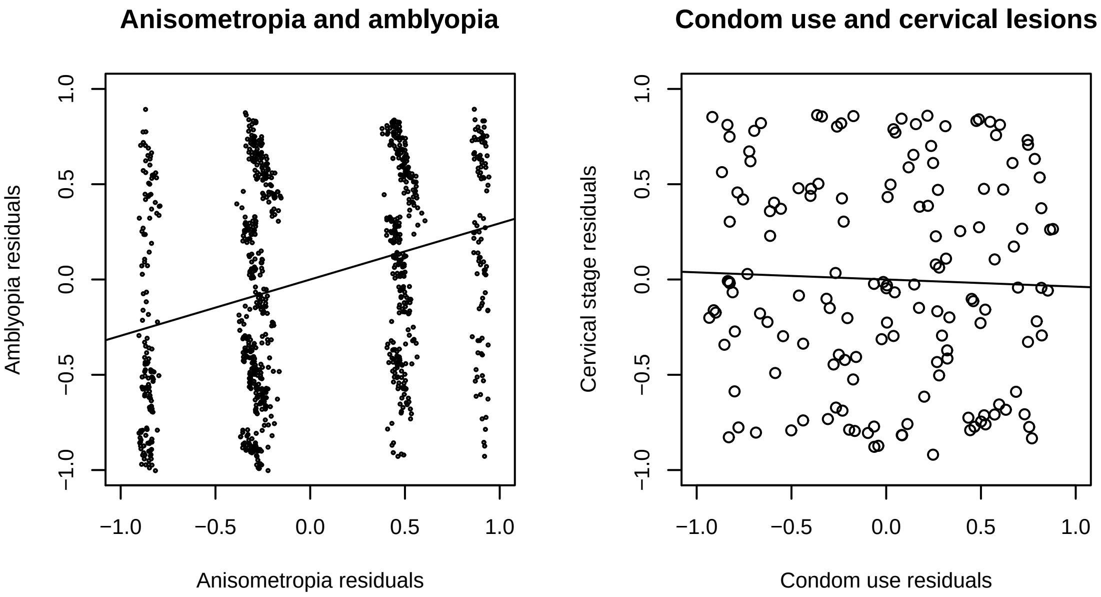


Figure 3: Residuals for the two data sets. The lines are fitted linear regression lines. Some jittering is applied so that the dots are not on top of others. Note the correlation between residuals in the amblyopia example and the lack of correlation in the cervical lesions example.


Anisometropia and amblyopia
Amblyopia residuals
Anisometropia residuals
Condom use and cervical lesions
Cervical stage residuals
Condom use residuals