Appl. Statist. (1992)
41, *No.* 2, *pp.* 287–297

Analysis of Ordinal Paired Comparison Data

By ALAN AGRESTI†

University of Florida, Gainesville, USA

[Received December 1989. Revised November 1990]

SUMMARY

Two types of model are discussed for paired comparisons of several treatments using ordinal scales such as (A ⋘ B, A ≪ B, A < B, A = B, A > B, A ≫ B, A ⋙ B), where A ⋘ B denotes strong preference for treatment B over treatment A, A ≪ B denotes moderate preference for B, A < B denotes weak preference for B, A = B denotes no preference, and so forth. For the binary scale (A < B, A > B), special cases of the models using logit transforms simplify to the Bradley–Terry model. When the same raters compare each pair of treatments, one can allow within-rater dependence by fitting the models with constrained maximum likelihood.

Keywords: Bradley–Terry model; Constrained maximum likelihood; Logit models; Log-linear models; Probit model; Repeated measures; Wilcoxon signed rank test

1. Introduction

This paper considers experiments in which responses are categorical measurements resulting from pairwise comparisons of treatments. For instance, wine-tasters might compare several brands of chardonnay wine in pairwise taste tests, in each comparison indicating which brand is preferable. Or, fashion designers might make pairwise comparisons of fabrics, indicating which is softer to the touch.

In a comparison of treatments h and i, let $Y_{hi} = 1$ represent preference for i and $Y_{hi} = 2$ represent preference for h. Bradley and Terry (1952) proposed a model having the logit representation

$$\log\{P(Y_{hi}=1)/P(Y_{hi}=2)\} = \mu_i - \mu_h$$

for which $P(Y_{hi}=1) = \pi_i/(\pi_i+\pi_h)$, where $\pi_k = \exp\mu_k$. This logit model has equivalent representations using the quasi-symmetry and quasi-independence models (Fienberg and Larntz, 1976; Imrey *et al.*, 1976). David (1988) presented a good survey of this model and others for analysing paired comparison data.

Rao and Kupper (1967) generalized the Bradley–Terry model to allow ‘no preference’ or ‘tie’ in a comparison. Glenn and David (1960) and Davidson (1970) presented related models. All three papers illustrated their models with data from an experiment described by Fleckenstein *et al.* (1958), in which 30 secretaries made pairwise comparisons of five brands of typewriter ribbon.

Some applications permit comparisons by using a more refined ordinal scale.

†*Address for correspondence*: Department of Statistics, Fourth Floor Little Hall, University of Florida, Gainesville, FL 32611, USA.

0035–9254/92/41287 $2.00

TABLE 1
Fit of models for pairwise comparisons of typewriter ribbons†

Pair	*Results for the following scale of preference for brands h and i:*						
(h, i)	*Strong for i*	*Moderate for i*	*Mild for i*	*No preference*	*Mild for h*	*Moderate for h*	*Strong for h*
1, 2	4 (2.1)	4 (5.5)	0 (2.6)	5 (4.9)	5 (3.1)	8 (7.9)	4 (3.7)
	(1.9)	(5.4)	(2.8)	(5.4)	(3.4)	(7.9)	(3.3)
	(1.9)	(6.0)	(2.8)	(5.2)	(3.2)	(7.7)	(3.3)
1, 3	5 (5.2)	12 (9.6)	4 (3.3)	6 (4.6)	0 (2.1)	2 (3.9)	1 (1.3)
	(4.9)	(9.6)	(3.4)	(4.8)	(2.2)	(3.8)	(1.2)
	(5.2)	(9.2)	(3.2)	(4.7)	(2.4)	(4.3)	(1.0)
1, 4	0 (0.7)	2 (2.4)	2 (1.5)	2 (3.9)	5 (3.3)	15 (11.2)	4 (7.0)
	(0.8)	(2.8)	(1.7)	(4.0)	(3.2)	(10.7)	(6.8)
	(0.6)	(3.1)	(1.9)	(4.2)	(3.1)	(10.0)	(7.1)
1, 5	4 (3.2)	4 (7.2)	5 (3.0)	4 (5.0)	2 (2.8)	8 (6.2)	3 (2.6)
	(2.8)	(7.1)	(3.2)	(5.5)	(3.0)	(6.1)	(2.2)
	(2.9)	(7.4)	(3.1)	(5.2)	(2.9)	(6.3)	(2.2)
2, 3	6 (6.2)	9 (10.6)	3 (3.3)	4 (4.2)	1 (1.8)	4 (2.9)	3 (0.9)
	(6.3)	(10.5)	(3.2)	(4.2)	(1.8)	(3.0)	(0.9)
	(6.4)	(9.7)	(3.2)	(4.4)	(2.1)	(3.5)	(0.8)
2, 4	2 (1.0)	4 (3.2)	1 (1.9)	6 (4.3)	4 (3.4)	8 (10.3)	5 (5.9)
	(1.1)	(3.6)	(2.1)	(4.6)	(3.4)	(9.9)	(5.3)
	(0.9)	(3.8)	(2.2)	(4.6)	(3.2)	(9.5)	(5.8)
2, 5	5 (4.1)	5 (8.4)	6 (3.2)	6 (4.9)	1 (2.5)	4 (5.0)	3 (1.9)
	(3.7)	(8.4)	(3.4)	(5.3)	(2.7)	(4.9)	(1.7)
	(3.8)	(8.2)	(3.2)	(5.1)	(2.7)	(5.4)	(1.6)
3, 4	0 (0.3)	1 (1.1)	0 (0.9)	3 (2.7)	1 (2.9)	13 (12.4)	12 (9.7)
	(0.4)	(1.4)	(0.9)	(2.5)	(2.3)	(11.0)	(11.6)
	(0.2)	(1.5)	(1.1)	(2.9)	(2.4)	(10.2)	(11.7)
3, 5	1 (1.5)	6 (4.3)	3 (2.2)	4 (4.7)	3 (3.3)	11 (9.2)	2 (4.8)
	(1.4)	(4.2)	(2.4)	(5.0)	(3.4)	(9.2)	(4.4)
	(1.3)	(4.8)	(2.5)	(4.9)	(3.2)	(8.8)	(4.6)
4, 5	6 (7.4)	12 (11.4)	5 (3.3)	3 (3.7)	1 (1.4)	3 (2.2)	0 (0.6)
	(7.4)	(11.0)	(3.1)	(3.8)	(1.6)	(2.5)	(0.7)
	(8.0)	(10.1)	(3.0)	(3.9)	(1.8)	(2.7)	(0.5)

†Source of data, Fleckenstein *et al.* (1958); the first value for each pair is the fit from the adjacent categories logit model, the second is from the cumulative logit model and the third is from the cumulative probit model.

Comparisons of treatments h and i might use a scale such as $(h \lll i, h \ll i, h < i, h = i, h > i, h \gg i, h \ggg i)$, where $h \lll i$ indicates strong preference for i over h, $h \ll i$ indicates moderate preference for i, $h < i$ indicates mild preference for i, $h = i$ indicates no preference, and so forth. In fact, this was the scale reported by Fleckenstein *et al.* (1958) for comparisons of typewriter ribbons. Table 1 exhibits their data. This paper adapts two types of model for ordinal responses (Agresti, 1990) to analyse paired comparison data such as Table 1. We also discuss fitting the models by using constrained maximum likelihood to allow within-rater dependence when the same raters compare each pair of treatments.

2. Cumulative Link Model

Let I denote the number of treatments, and let J denote the number of categories in the ordinal response scale. For a randomly selected rater, let $P(Y_{hi}=j)$ denote the probability that comparison of treatments h and i results in response j, where $j=1$ denotes the least favourable response for h and $j=J$ denotes its most favourable response. We assume that the scale is symmetric, in the sense that $Y_{hi}=j$ is equivalent to $Y_{ih} = J-j+1$, for all j.

We first specify a model by generalizing arguments given by Glenn and David (1960) and Anderson and Philips (1981). For comparison of treatments h and i, suppose that there is an underlying continuous response variable Y^*_{hi}. Let $\alpha_1 < \alpha_2 < \ldots < \alpha_{J-1}$ denote cutpoints such that $Y_{hi}=j$ when Y^*_{hi} falls between α_{j-1} and $\alpha_j, j=1, \ldots, J$, where $\alpha_0 = -\infty$ and $\alpha_J = \infty$. We assume that Y^*_{hi} can be expressed as $Y^*_{hi} = Y_h - Y_i$, where Y_t represents an underlying rating of treatment t, $t=1, \ldots, I$. Finally, we assume that there are treatment merit parameters $\{\mu_t\}$ such that $(Y_h - \mu_h, Y_i - \mu_i)$ has the same distribution for each treatment pair. Then, $Z = Y_i - \mu_i - (Y_h - \mu_h)$ has the same distribution for each pair and

$$Y_{hi} = j \text{ is equivalent to } \alpha_{j-1} - (\mu_h - \mu_i) < Z < \alpha_j - (\mu_h - \mu_i).$$

Let F denote the cumulative distribution function of Z. Since Z has the same distribution as $-Z$, F is symmetric about 0, satisfying $F^{-1}(\pi) = -F^{-1}(1-\pi)$ for all $0 \leqslant \pi \leqslant 1$. Paired comparisons satisfy the model

$$F^{-1}\{P(Y_{hi} \leqslant j)\} = \alpha_j - (\mu_h - \mu_i), \qquad j=1, \ldots, J-1. \tag{2.1}$$

Important special cases for F are the logistic and the standard normal models. For the logistic model, F^{-1} is the logit link and equation (2.1) is a cumulative logit model. For the normal model, F^{-1} is the probit link. In its general form, we refer to equation (2.1) as a *cumulative link* model. We assume an absence of an effect of the order in which two treatments are compared or listed, in that $P(Y_{hi} \leqslant j) = P(Y_{ih} \geqslant J-j+1)$. The symmetry of F implies that $\alpha_j = -\alpha_{J-j}, j=1, \ldots, J-1$, so that $\Sigma\alpha_j = 0$ and $\alpha_{J/2} = 0$ when J is even. Model identifiability requires a constraint such as $\Sigma\mu_a = 0$ or $\mu_1 = 0$.

For the logit link, a simple interpretation of model (2.1) follows from the expression

$$\log\left\{\frac{P(Y_{hi} \leqslant j)/P(Y_{hi} > j)}{P(Y_{ih} \leqslant j)/P(Y_{ih} > j)}\right\} = 2(\mu_i - \mu_h).$$

The odds that treatment i is rated better than treatment h by at least some fixed amount are $\exp\{2(\mu_i - \mu_h)\}$ times the odds that h is rated better than i by at least that amount. Equivalence of the I treatments corresponds to $\mu_1 = \mu_2 = \ldots = \mu_I$.

For the logit link, model (2.1) simplifies to the Bradley–Terry model when $J=2$ and to the Rao–Kupper model allowing ties when $J=3$. Tutz (1986) and Cox and Snell (1989) used this model. For the probit link, model (2.1) with $J=2$ is called the Thurstone–Mosteller model (Mosteller, 1951), and when $J=3$ it simplifies to a model allowing ties proposed by Glenn and David (1960). The cumulative link model (2.1) is a special case of a model described by McCullagh (1980) (see also Walker and Duncan (1967)) in which treatment effects are the same for each cutpoint j for cumulative probabilities.

3. Adjacent Categories Logit Model

The model of form (2.1) also makes sense when we apply the link with adjacent response probabilities, rather than cumulative probabilities. Conditional on response

j or $j+1$, we could assume that an underlying random variable $Z + \mu_i - \mu_h$ determines the preference, where Z has the same distribution for each pair of treatments and each pair of responses. For the logit link, the model in this case is

$$\log\{P(Y_{hi}=j)/P(Y_{hi}=j+1)\} = \alpha_j - (\mu_h - \mu_i), \qquad j=1, \ldots, J-1. \quad (3.1)$$

Assuming that there is no effect relating to the order in which raters observe treatments, this logit is the same as $\log\{P(Y_{ih}=J-j+1)/P(Y_{ih}=J-j)\}$, so that $\alpha_j = -\alpha_{J-j}, j=1, \ldots, J-1$.

Model (3.1) satisfies

$$P(Y_{hi}=j)/P(Y_{ih}=j) = \exp\{(J+1-2j)(\mu_i - \mu_h)\}. \quad (3.2)$$

When $J=7$, if $\lambda = \exp\{2(\mu_i - \mu_h)\}$ is the odds that treatment i is mildly preferred to treatment h (instead of h being mildly preferred to i), then λ^2 is the odds that i is moderately preferred (instead of h being moderately preferred), and λ^3 is the odds that i is strongly preferred. For a fixed set of categories, model (3.1) is simpler to interpret than the cumulative link model (2.1). The interpretation refers directly to an odds for a given outcome, rather than an odds ratio for two groupings of outcomes. Models (3.1) and (2.1) share the property that Y_{hi} is stochastically higher than Y_{ih} when $\mu_i < \mu_h$. Like the cumulative logit model, equation (3.1) simplifies to the Bradley–Terry model when $J=2$.

Model (3.1) treats each pair of adjacent responses identically. More generally, we could permit a positive 'distance' d_j between responses j and $j+1$, where $d_j = d_{J-j}$ for $j=1, \ldots, J-1$, such that $Z + d_j(\mu_i - \mu_h)$ has the same distribution for each pair of responses. A small distance diminishes influences of treatment effects. This leads to a generalization of model (3.2) satisfying

$$P(Y_{hi}=j)/P(Y_{ih}=j) = \exp\{(v_{J-j+1} - v_j)(\mu_i - \mu_h)\} \quad (3.3)$$

where $\{v_j\}$ are monotone scores satisfying $\{v_{j+1} - v_j = d_j\}$. Without loss of generality, we can let $\{v_j\}$ satisfy $v_j = -v_{J-j+1}$.

For a sample, let $n_{(hi)j}$ denote the number of times that response j occurs in comparing treatment h with treatment i, for each $h<i$ and $j=1, \ldots, J$. Let $\{m_{(hi)j}\}$ denote expected frequencies for the $\binom{I}{2} \times J$ contingency table in which each row displays comparisons for a particular pair of treatments. Model (3.3) is equivalent to the log-linear model

$$\log m_{(hi)j} = \mu + \lambda^X_{(hi)} + \lambda^Y_j + v_j(\mu_h - \mu_i) \quad (3.4)$$

where, for all j, $\lambda^Y_j = \lambda^Y_{J-j+1}$. This model is a special case of the *row effects* model introduced by Goodman (1979).

Let

$$n_{(hi)+} = \sum_j n_{(hi)j},$$

$$n_{(+)j} = \sum\sum_{h<i} n_{(hi)j},$$

$$n_{(k)j} = \sum_{a>k} n_{(ka)j} + \sum_{a<k} n_{(ak)J-j+1}$$

and

$$n_{(k)+} = \sum_j n_{(k)j}.$$

Then $n_{(k)j}$ is the number of times that response j is made in comparing treatment k with all other treatments. Suppose that we treat the $n_{(hi)+}$ comparisons of treatments h and i on the J-category scale as independent multinomial trials, and suppose that comparisons of different pairs of treatments are independent. Then cell counts in different rows of $\{n_{(hi)j}\}$ are independent multinomial samples, and the likelihood equations for model (3.4) are

$$\hat{m}_{(hi)+} = n_{(hi)+}, \qquad \text{for all } h < i,$$

$$\hat{m}_{(+)j} + \hat{m}_{(+)J-j+1} = n_{(+)j} + n_{(+)J-j+1}, \qquad j = 1, \ldots, J,$$

$$\sum_j v_j \hat{m}_{(k)j} = \sum_j v_j n_{(k)j}, \qquad k = i, \ldots, I.$$

With the first equation, the last set of equations implies that for the scores $\{v_j\}$ the mean response when treatment k is compared with other treatments is the same for the observed and fitted data. Thus, like model (2.1), model (3.3) gives a way of describing location shifts among treatments. In fact, for this model, the maximum likelihood (ML) estimates $\{\hat{\mu}_k\}$ have the same order as these sample means.

The last set of likelihood equations results from differentiating the log-likelihood with respect to $\{\mu_k\}$. Let $M_k = \Sigma_j v_j n_{(k)j}$, $k = 1, \ldots, I$. For the hypothesis of no treatment effects (H_0: $\mu_1 = \ldots = \mu_I$), let $(\pi_1, \ldots, \pi_J)$ with $\pi_j = \pi_{J-j+1}$ denote the response distribution for each pair of treatments. Under H_0, $E(M_k) = 0$, $\text{var}(M_k) = n_{(k)+} \Sigma v_j^2 \pi_j$ and $\text{cov}(M_h, M_k) = -n_{(hk)+} \Sigma v_j^2 \pi_j$. When there is the same number of observations for each pair of treatments, the correlation for each (M_h, M_k) pair is $-(I-1)^{-1}$, and the efficient score statistic for testing H_0 then has the simple form

$$(I-1) \Sigma M_k^2 / 2 \Sigma v_j^2 n_{(+)j}. \tag{3.5}$$

Its asymptotic null distribution is χ^2 with $I-1$ degrees of freedom.

4. Fitting the Models

For the independent multinomial sampling model, we can obtain ML fitting of the cumulative link model (2.1) by using an iterative routine described by McCullagh (1980). For logit and probit links, this can be implemented in SAS by using the procedure LOGISTIC. We can obtain ML fitting of model (3.3) with standard Newton–Raphson routines for log-linear models or the equivalent logit models. This can be implemented, for instance, in SAS by using procedure CATMOD and in GLIM.

For the cumulative link or adjacent categories logit models, we can test the goodness of fit by comparing observed counts $\{n_{(hi)j}\}$ with fitted values $\{\hat{m}_{(hi)j}\}$ for the model, by using a likelihood ratio statistic G^2. The residual degrees of freedom for the χ^2-distribution are $\binom{I}{2}(J-1) - \{(I-1) + (J-1)/2\}$ when J is odd and $\binom{I}{2}(J-1) - \{(I-1) + (J-2)/2\}$ when J is even. For either model, we can test the hypothesis of equivalent treatments by the reduction in the likelihood ratio statistic compared with the simpler model having $\mu_1 = \ldots = \mu_I$, based on $I-1$ degrees of freedom.

5. Example

Table 1 displays ML fitted values for logit model (3.1) applied to the typewriter ribbon data. The table is sparse, but the model seems to fit well, with $G^2=48.2$ based on 53 degrees of freedom. Table 2 shows the ML estimates for the model, with the constraint $\Sigma\,\hat{\mu}_k=0$. For ribbons 3 and 4, for instance, conditional on the event that one of them is mildly preferred, $\exp[2\{0.270-(-0.340)\}] = 3.4$ is the estimated odds that ribbon 3 is mildly preferred to ribbon 4. Similarly, $(3.39)^2 = 11.5$ is the estimated odds that ribbon 3 is moderately preferred to ribbon 4, and $(3.39)^3=38.9$ is the estimated odds that ribbon 3 is strongly preferred to ribbon 4. The likelihood ratio statistic for testing $\mu_1 = \mu_2 = \ldots = \mu_5$ equals 60.7, and the efficient score statistic (3.5) equals 60.4, both based on 4 degrees of freedom. Using the estimated covariance matrix of $\{\hat{\mu}_k\}$, we performed Bonferroni multiple comparisons for the 10 pairs of ribbons. For any overall confidence level between 0.60 and 0.96, ribbon 3 ranks best, ribbon 4 ranks worst, and ribbons (5, 1, 2) are all worse than ribbon 3 and better than ribbon 4 but not significantly different from each other.

We also used the more general form (3.3) of this logit model for unequal interval scores. Rather than assigning scores, we could treat the $\{v_j\}$ themselves as parameters. The related association model (3.4) is no longer log-linear but is a special case of a log-multiplicative (row–column) model introduced by Goodman (1979). When we fitted this model subject to the constraint $\{v_j = -v_{J-j+1} \text{ all } j\}$ and achieved identifiability by setting $v_7 = -v_1 = 3.0$, we obtained estimated scores $\{-3, -2.32, -1.93, 0, 1.93, 2.32, 3\}$ and treatment effects $\{0.041, -0.053, 0.238, -0.304, 0.079\}$. The estimated odds of mild, moderate and strong preference of treatment 3 to treatment 4 are then 8.1, 12.4 and 25.9. This model has residual $G^2=45.6$, based on 51 degrees of freedom, so the fit is not better than that provided by the simpler model.

TABLE 2
Parameter estimates for models fitted to the typewriter ribbon data†

Parameter	*Estimates for the following models:*		
	Adjacent categories logit	*Cumulative probit*	*Cumulative logit*
α_1	−0.85	−1.38	−2.40
α_2	0.83	−0.49	−0.83
α_3	−0.54	−0.22	−0.37
μ_1	0.042 (0.040)	0.058 (0.076)	0.117 (0.129)
μ_2	−0.050 (0.040)	−0.088 (0.076)	−0.196 (0.130)
μ_3	0.270 (0.046)	0.494 (0.079)	0.887 (0.138)
μ_4	−0.340 (0.050)	−0.607 (0.080)	−1.048 (0.141)
μ_5	0.078 (0.041)	0.143 (0.076)	0.240 (0.130)
G^2	48.2	54.8	49.8
Degrees of freedom	53	53	53
Test of homogeneity‡	60.7	77.7	82.7

†Estimated asymptotic standard errors are given in parentheses.
‡4 degrees of freedom.

Table 1 also displays ML fitted values for the cumulative logit and probit models. These models also fit quite well. Table 2 shows their ML estimates. The fits and the estimates are qualitatively similar to those for model (3.1). The probit fit gives $\hat{\mu}_3 - \hat{\mu}_4 = 1.10$. For an assumed underlying normal distribution of preference, we estimate that the difference in means between treatments 3 and 4 is about a standard deviation of the difference scale. The latent variable Z in the construction for cumulative link models in Section 2 has standard deviation 1 for the probit (inverse standard normal) link and standard deviation $\pi/3^{1/2} = 1.8$ for the logit (inverse logistic) link. Both models fit Table 1 well, so if we divide the parameter estimates for the cumulative logit model by 1.8 we obtain estimates similar to those for the cumulative probit model.

6. Alternative Estimation for Dependent Samples

The standard assumption in ML fitting of Bradley–Terry models and their extensions is that different ratings of the same pair or different pairs of treatments are statistically independent. For data such as Table 1, this may not be valid, since the same 30 secretaries rated each pair of typewriter ribbons. Different ratings by the same secretary may be statistically dependent. If the separate ratings by each secretary were available (they are not), we could investigate this dependence by fitting models in a different manner.

Let $p = \binom{I}{2}$ be the number of pairs of treatments. When all n raters evaluate each pair, the frequencies of the joint ratings can be displayed in a J^p contingency table. Models discussed in this paper apply to the one-dimensional margins of that table. Suppose that there is within-rater dependence of ratings but the one-way margins of this table are treated as independent samples. From arguments in Liang and Zeger (1986), ML estimates of model parameters are still consistent, if the model holds. However, the estimates of the standard errors are not valid.

To allow potential within-rater dependence, we fit models to the margins of the joint ratings table, treating cell counts in that table as multinomial or independent Poisson variates, i.e. we maximize the likelihood subject to the constraint that the model holds for the one-way margins. To do this, we used a Fisher scoring algorithm adapted from Aitchison and Silvey (1958) and Haber (1985) to obtain ML cell fitted values, and then used the delta method to obtain the ML model parameter estimates and their estimated covariance matrix.

Using notation from Aitchison and Silvey (1958) and Haber (1985), we let $\hat{\boldsymbol{\Theta}}$ denote the constrained ML estimate of the cell expected frequencies $\boldsymbol{\Theta}$ in the table of joint ratings. The marginal expected frequencies are $\mathbf{A}\boldsymbol{\Theta}$ for an appropriate matrix $\mathbf{A}$ of 0s and 1s. Assume independent Poisson sampling for cell counts in the joint ratings table, and suppose that the model of interest for the marginal distributions has form $\mathbf{C}\log(\mathbf{A}\boldsymbol{\Theta}) = \mathbf{X}\boldsymbol{\beta}$. The model corresponds to constraints $\mathbf{U}'\mathbf{C}\log(\mathbf{A}\boldsymbol{\Theta}) = \mathbf{0}$, where $\mathbf{U}$ has columns that span the space orthogonal to that spanned by the columns of $\mathbf{X}$. The information matrix for Poisson sampling and the matrix of derivatives of the constraint equations are

$$\mathbf{B} = \operatorname{diag}(\hat{\boldsymbol{\Theta}})^{-1},$$
$$\mathbf{H} = \mathbf{A}' \operatorname{diag}(\mathbf{A}\hat{\boldsymbol{\Theta}})^{-1}\mathbf{C}'\mathbf{U},$$

where diag( ) denotes the diagonal matrix with elements of the vector on the main

TABLE 3
Data from the soft drink testing experiment†

A–B	*A–C*	*B–C*	*A–B*	*A–C*	*B–C*	*A–B*	*A–C*	*B–C*	*A–B*	*A–C*	*B–C*	*A–B*	*A–C*	*B–C*
1	1	2	2	1	1	2	3	3	3	3	4	4	3	3
1	1	4	2	1	5	2	3	4	3	4	1	4	3	4
1	2	2	2	1	5	2	3	4	3	4	2	4	3	4
1	2	3	2	1	5	2	4	3	3	4	4	4	3	5
1	2	3	2	2	1	2	4	4	3	5	2	4	4	2
1	2	3	2	2	3	2	4	4	3	5	4	4	4	2
1	2	3	2	2	3	2	4	5	4	1	1	4	4	4
1	2	5	2	2	4	3	1	2	4	1	4	4	5	2
1	4	4	2	2	5	3	2	1	4	1	4	5	1	3
1	4	5	2	2	5	3	2	3	4	2	2	5	2	1
1	4	5	2	3	2	3	2	3	4	2	5	5	5	1
1	5	4	2	3	2	3	3	3	4	3	2	5	5	3
1	5	5												

†A ≡ Coke, B ≡ Classic Coke, C ≡ Pepsi.

diagonal. Aitchison and Silvey (1958) showed that $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is asymptotically normal, with estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ given by

$$\mathbf{P} = \mathbf{B}^{-1}\{\mathbf{I} - \mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{B}^{-1}\}.$$

Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\log(\mathbf{A}\hat{\boldsymbol{\theta}})$, the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\,\mathrm{diag}(\mathbf{A}\hat{\boldsymbol{\theta}})^{-1}\mathbf{A}\mathbf{P}\mathbf{A}'\,\mathrm{diag}(\mathbf{A}\hat{\boldsymbol{\theta}})^{-1}\mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

To illustrate this dependence analysis, we consider Table 3, which refers to a soft drink tasting experiment in which each subject made pairwise comparisons of Coke, Classic Coke and Pepsi. The 53 subjects were graduate students at the University of Florida taking courses (autumn 1989) in categorical data analysis or statistics for psychologists. The experiment used the rating scale ($h \gg i$, $h > i$, $h = i$, $h < i$, $h \ll i$). In Table 3, the response sequence (1, 2, 5), for instance, means that the subject rated Coke much better than Classic Coke, Coke better than Pepsi and Classic Coke much worse than Pepsi. The joint ratings for the 53 subjects produce counts in a contingency table having $5^3 = 125$ cells. This table is highly sparse. The one-way margins to which the model applies are less sparse, with counts ranging from 2 to 19.

Table 4 shows comparisons for ML fitting of the adjacent categories logit model (3.1) to these data. The drinks were ranked in the order (Coke, Pepsi, Classic Coke), but there is evidence of a difference only between Coke and Classic Coke. For these data, estimated asymptotic standard errors for estimated differences in parameters are similar whether we treat the three ratings by each subject as independent or dependent observations. The estimated differences are smaller for the dependence analysis, however, and that analysis shows less evidence of treatment effects. Though the two sets of estimates must be similar for large n when the model holds, they can be quite different when some one-way marginal counts are small or when the model fits poorly. Since the sample size is relatively small, it is improper to make generalizations from the results in Table 4, but they do show that it may be unwise blindly to assume independence of repeated ratings in paired comparison models.

TABLE 4
Comparisons from ML fitting of the adjacent categories logit model to Table 3†

Comparison	*Results for the following samples:* *Independent samples*	 *Dependent samples*
$\mu_1-\mu_2$	0.258 (0.095)	0.182 (0.092)
$\mu_1-\mu_3$	0.148 (0.093)	0.102 (0.086)
$\mu_2-\mu_3$	−0.110 (0.093)	−0.080 (0.088)
Test of homogeneity $\mu_1=\mu_2=\mu_3$‡	7.79	3.41

†Estimated asymptotic standard errors are given in parentheses.
‡2 degrees of freedom.

Fitting a model by constrained ML is sometimes impractical, because the joint ratings table may be huge. For instance, the table for the typewriter ribbon analysis would have $7^{10} = 2.8 \times 10^8$ cells. Since the asymptotic covariance of the sample marginal logits (or other links for the one-way margins) is determined by the $\binom{p}{2}$ two-way marginal tables rather than by the entire p-way table, it is often feasible to obtain constrained estimates by using a weighted least squares analysis for such tables. The weighted least squares analysis seems to have reasonable validity when the one-way marginal totals mostly exceed 5, as recommended for related analyses described by Koch *et al.* (1977). Such an analysis can be performed with SAS (procedure CATMOD).

Alternative ways to handle potential dependence are to obtain a robust estimate of the covariance matrix of the independence estimates that is consistent even when there is dependence, or to estimate the model parameters and to obtain their covariance estimates under some assumed correlation structure for within-rater dependence. This might be done by generalizing methods that Liang and Zeger (1986) and Lipsitz *et al.* (1990) proposed for dependent responses with logit models. A topic for future research is to compare methods of handling possible dependence.

7. Other Comments

Semenya *et al.* (1983) and Agresti *et al.* (1987) have described other ways of applying Bradley–Terry models to ordered categorical responses. Their models, which apply to standard one-way layout or regression problems rather than paired comparisons, describe the probability of concordance for pairs of settings of explanatory variables.

Scheffé (1952) proposed an analysis of variance for ordinal paired comparisons, based on assigning monotone scores $\{v_j\}$ to the response scale and using ordinary least squares. Scheffé's model deals directly with estimating mean responses for the treatments for that assigned scale. A simple version of his model implies that

$$\sum_j v_j P(Y_{hi}=j) = \mu_h - \mu_i$$

for all h and i, where $v_j = -v_{J-j+1}$ for all j. For fixed $\{v_j\}$, Scheffé's analysis gives the same ranking of the treatments as does logit model (3.3). Scheffé's analysis treats the observed response as normally distributed with constant variance, rather than categorical. Alternatively, we could recognize the categorical nature of the scale by fitting his model by using ML or weighted least squares, assuming multinomial sampling.

The analysis of treatment effects with logit model (3.3) also has connections with nonparametric methods for matched pairs. For pairwise comparisons of I treatments, Mehra (1964) proposed a generalization of the Wilcoxon signed rank statistic. The efficient score statistic (3.5) for model (3.3) with signed mid-rank scores for $\{v_j\}$ is simply Mehra's test generalized to allow for ties and falls in a class of statistics discussed by Ghosh (1973). For $I=2$, the score test of treatment effect with the Bradley–Terry model for a binary scale is the sign test. For large J, we expect the efficiency of treatment comparisons based on model (3.3) relative to the Bradley–Terry model to be approximated by the efficiency of Mehra's test relative to the test using the Bradley–Terry model, at least when the scores $\{v_j\}$ are highly correlated with signed mid-rank scores. In a balanced case, Mehra noted that this efficiency is the same as that of the Wilcoxon test relative to the sign test. For many standard underlying distributions (e.g. normal, logistic), the efficiency gain may be considerable.

The Bradley–Terry model has been generalized in various ways to account for possible departures from basic assumptions. Many modifications apply in a straightforward manner to models discussed here. For instance, to account for an effect relating to the order in which a rater observes treatments, we could follow Fienberg (1979) and Cox and Snell (1989), p. 160, and add a parameter that induces a shift in the cutpoints, so that for all $h < i$

$$F^{-1}\{P(Y_{hi} \leqslant j)\} = F^{-1}\{P(Y_{ih} \geqslant J-j+1)\} + \delta.$$

Acknowledgements

I thank Joseph B. Lang for writing a program for the constrained ML analysis reported in Section 6, Dr Malay Ghosh for the Mehra (1964) and Ghosh (1973) references, and my classes for participating in the soft drink tasting experiment. This research was partially supported by National Institutes of Health grant GM 43824.

References

Agresti, A. (1990) *Categorical Data Analysis*, sects 8.2, 9.4 and 9.5. New York: Wiley.

Agresti, A., Schollenberger, J. and Wackerly, D. (1987) Models for the probability of concordance in cross-classification tables. *Qual. Quant.*, **21**, 49–57.

Aitchison, J. and Silvey, S. D. (1958) Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.*, **29**, 813–828.

Anderson, J. A. and Philips, P. R. (1981) Regression, discrimination and measurement models for ordered categorical variables. *Appl. Statist.*, **30**, 22–31.

Bradley, R. A. and Terry, M. B. (1952) Rank analysis of incomplete block designs: I, The method of paired comparisons. *Biometrika*, **39**, 324–345.

Cox, D. R. and Snell, E. J. (1989) *Analysis of Binary Data*, 2nd edn. London: Chapman and Hall.

David, H. A. (1988) *The Method of Paired Comparisons*, 2nd edn. London: Griffin.

Davidson, R. R. (1970) On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *J. Am. Statist. Ass.*, **65**, 317–328.

Fienberg, S. E. (1979) Log linear representation for paired comparison models with ties and within-pair order effects. *Biometrics*, **35**, 479–481.

Fienberg, S. E. and Larntz, K. (1976) Log linear representation for paired and multiple comparisons models. *Biometrika*, **63**, 245–254.

Fleckenstein, J., Freund, R. A. and Jackson, J. E. (1958) A paired comparison test of typewriter carbon papers. *Tappi*, **41**, 128–130.

Ghosh, M. (1973) On a class of asymptotically optimal nonparametric tests for grouped data II. *Ann. Inst. Statist. Math.*, **25**, 109–122.

Glenn, W. A. and David, H. A. (1960) Ties in paired-comparison experiments using a modified Thurstone–Mosteller model. *Biometrics*, **16**, 86–109.

Goodman, L. A. (1979) Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Statist. Ass.*, **74**, 537–552.

Haber, M. (1985) Log-linear models for correlated marginal totals of a frequency table. *Communs Statist. Theory Meth.*, **14**, 2845-2856.

Imrey, P. B., Johnson, W. D. and Koch, G. G. (1976) An incomplete contingency table approach to paired-comparison experiments. *J. Am. Statist. Ass.*, **71**, 614–623.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehnen, R. G. (1977) A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133–158.

Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1990) Using the jackknife to estimate the variance of regression estimators from repeated measures studies. *Communs Statist. Theory Meth.*, **19**, 821–845.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

Mehra, K. L. (1964) Rank tests for paired-comparison experiments involving several treatments. *Ann. Math. Statist.*, **35**, 122–137.

Mosteller, F. (1951) Remarks on the method of paired comparisons: I, The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**, 3–9.

Rao, P. V. and Kupper, L. L. (1967) Ties in paired-comparison experiments: a generalization of the Bradley–Terry model. *J. Am. Statist. Ass.*, **62**, 194–204.

Scheffé, H. (1952) An analysis of variance for paired comparisons. *J. Am. Statist. Ass.*, **47**, 381–400.

Semenya, K., Koch, G. G., Stokes, M. E. and Forthofer, R. N. (1983) Linear model methods for some rank function analyses of ordinal categorical data. *Communs Statist. Theory Meth.*, **12**, 1277–1298.

Tutz, G. (1986) Bradley–Terry–Luce models with an ordered response. *J. Math. Psychol.*, **30**, 306–316.

Walker, S. H. and Duncan, D. B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika*, **54**, 167–179.