Annu. Rev. Public Health. 1997. 18:51–82

CATEGORICAL DATA ANALYSIS IN PUBLIC HEALTH

John S. Preisser
Section on Biostatistics, Department of Public Health Sciences, Bowman Gray School of Medicine, Winston-Salem, North Carolina 27157

Gary G. Koch
Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27599-7400

KEY WORDS: exact inference, conditional logistic regression, proportional odds, generalized estimating equations, weighted least squares

ABSTRACT

A greater variety of categorical data methods are used today than 15 years ago. This article surveys categorical data methods widely applied in public health research. Whereas large sample chi-square methods, logistic regression analysis, and weighted least squares modeling of repeated measures once comprised the primary analytic tools for categorical data problems, today's methodology is comprised of a much broader range of tools made available by increasing computational efficiency. These include computational algorithms for exact inference of small samples and sparsely distributed data, conditional logistic regression for modeling highly stratified data, and generalized estimating equations for cluster samples. The latter, in particular, has found wide use in modeling the marginal probabilities of correlated counted, binary, and multinomial outcomes. The various methods are illustrated with examples including a study of the prevalence of cerebral palsy in very low birthweight infants and a study of cancer screening in primary care settings.

INTRODUCTION

Public health research is frequently concerned with the relationship of categorical response variables with one or more explanatory variables. A categorical response variable may have two or more possibly ordered categories. For example, in a study of the factors related to cerebral palsy in premature neonates,

0163-7525/97/0510-0051$08.00

the outcome of interest is whether an infant has developed cerebral palsy by one year of age. Categorical data methods are used to describe trends in the rate of cerebral palsy in infants born at a particular North Carolina hospital. In a statewide study of cancer screening in primary care practices in Colorado, interest is in the physician and patient factors which predict the extent of breast cancer screening that women receive. The outcome for each woman takes a value of 0, 1, or 2 for the number of types of screening procedures (mammogram and clinical breast exam) received in the past year. The statistical analysis should consider that the level of one woman's cancer screening may be correlated to that of another patient in the same medical practice. An array of categorical data methods are available for analyzing such data. Only 15 years ago, much of it was either nonexistent, not widely known, or not easily implemented because computational advances lagged behind methodological ones.

Historically, emphasis has been placed on large sample chi-square methods for the analysis of contingency tables (43). In more recent years, with the advent of increasingly powerful computing, there has been an emergence of interest in exact inference (3). Exact methods have enabled estimation and hypothesis testing for tables with small samples and tables of moderate size with highly skewed marginal distributions for which large sample approximations are not justified. For tables in which sample size requirements are met, Mantel-Haenszel methods (47) may be employed. Less widely known are variations of these randomization methods that adjust for continuous covariates in the spirit of nonparametric analysis of covariance (40).

Unconditional or ordinary logistic regression, introduced in its contemporary form by Cox (18), remains one of the more commonly employed statistical methods and provides a means of simultaneously assessing the relationships of covariates to a dichotomous response. For responses with more than two categories, extensions of unconditional logistic regression (42) include the proportional odds model (61) for an ordered response. In the cerebral palsy example, a three-level outcome is death before one year, cerebral palsy, or healthy status at one year of age. The proportional odds model may be applied to the cancer screening study, but the estimation procedure should take into account the correlation of patient responses within the same practice. If the response is nominal, polychotomous logistic regression may apply. This paper emphasizes the odds ratio as a measure of association; however, other measures may be of interest such as the relative risk sometimes modeled in epidemiology (73).

In public health, data sometimes result from highly stratified random samples where there are many strata but only a few observations per stratum (65). Conditional logistic regression, traditionally used in case-control studies, applies to such data when interest is in a factor that varies within stratum, such as treatment in multi-center clinical trials with only a few patients per center. It also

applies to cluster samples of small size such as multi-period crossover trials by assuming independence conditional upon subject specific parameters. In some logistic or conditional logistic regression problems, the sample size may be too small, or the data too sparse to justify large sample methods. In these situations, exact logistic regression provides estimates, tests, and confidence intervals for parameters that may be nonestimable otherwise.

This paper does not cover loglinear models (8, 38, 39) explicitly, but we note their historically important role in categorical data analysis, especially in the field of sociology (15). These methods are used to model counts and to describe the associations among a collection of categorical variables without necessarily defining a particular subset as the response variables. Their model parameters have interpretations conditional on the values of the other variables and they do not extend easily to cluster sample situations (53).

In this review we emphasize marginal models in contrast to transitional and random effects models whose parameters have different interpretations (see concluding remarks). Marginal models are concerned with the distribution of the response averaged over the possible values of the other variables, and their regression coefficients have the same interpretation as coefficients from a cross-sectional analysis. They receive wide use in public health research where interpretations are often made on the effect of certain exposures, conditions, interventions, or risk factors on a response that may be a specific disease or health behavior. For independent data, marginal modeling has been unified for different types of responses into the theory of generalized linear models (22, 62, 68). These models, which include linear regression, logistic regression and poisson regression, share similar fitting algorithms and large sample theory.

In more recent years, nonlikelihood-based large sample methods have been developed to analyze data from studies that involve cluster samples. The generalized estimating equations (GEE) procedure (51, 52, 92) can handle large numbers of clusters of varying size and is used to fit a broad range of models including loglinear and logistic models for correlated poisson, binary, and multinomial outcomes. For example, GEE can be used to fit a proportional odds model to the cancer screening data while accounting for the intracorrelation of patients within a medical practice. Prior to GEE, correlated categorical data were frequently analyzed by weighted least squares (WLS) methods (41, 44). WLS, however, is limited to situations where all predictors are categorical with sufficient sample sizes per cross-classification for the asymptotic properties to hold. Nevertheless, WLS is still useful for modeling such response functions as rank measures of association and incidence densities.

In choosing an appropriate statistical method for a categorical data situation, one should consider the measurement scale of the response variable as well as the form in which the data are best characterized. Categorical data methods

apply to situations when the response variable is nominal or ordinal. A nominal scale only indicates whether a subject belongs to a particular category. Ordinal scales, on the other hand, assign values to the categories ranking them from poorest to best. Sometimes, however, aggregate measures, such as counts, incidence densities, or grouped survival times, are of interest. We distinguish between data represented in contingency table form to those in case record form to more readily identify appropriate statistical methods. Note, however, that the distinction between these two categorical data structures is tenuous, at best, because case record data can be conceptualized as a contingency table, albeit with sparse or zero cells, and vice versa, a table can be expressed as case records. Table 1 presents some useful statistical methods for categorical data.

This paper reviews and illustrates categorical data methods commonly used today, with particular emphasis on developments of the past 15 years. The sections are organized according to different categorical data situations that typically arise in public health research. A more comprehensive review, from a perspective of categorical data modeling in the past fifty years, is provided by Imrey et al (37). Due to the enormity of our subject, we do not attempt an exhaustive review of the literature, but often refer to other reviews or books that document the many contributions to the field. The next section reviews methods for contingency tables with special emphasis on exact inference. Unconditional, ordered, conditional, and exact logistic regression are reviewed in the third section. GEE and WLS are considered in the following two sections, respectively. The final section mentions other contributions and comments on future directions of categorical data analysis in public health.

CONTINGENCY TABLE METHODS

The statistical analyses of complex data sets often involve a combination of nonparametric and model-based testing and estimation procedures. This section discusses inference for contingency tables, especially for small samples. Later sections focus on model-based inference for case record data. Statistical

Table 1 Statistical methods for categorical data analysis in public health

Data structure	Small samples	Large samples
Contingency tables	Fisher's exact test and extensions	Chi square and Mantel-Haenszel methods
Case record and simple random samples	Exact logistic regression	Logistic regression, proportional odds models and extensions of logistic regression
Cluster samples	Exact logistic regression for stratified models	Generalized estimating equations and weighted least squares

analyses of contingency tables involve the analysis of two-way tables for the assessment of significant association between two variables, and the analysis of sets of two-way tables for testing conditional independence of two variables given additional variables or for testing homogeneity of odds ratios across strata formed by the additional variables. Statistics exist for situations in which both the row (exposure or factor) and column (response) variable are nominal, when only the column variable is ordinal, or when both row and column variable are ordinal. These methods involve statistics that assign scores, often integers or ranks, to the categories of ordinal variables.

The role of Mantel-Haenszel methods in public health research is well documented (11, 41, 42, 47). Historically, several related procedures were proposed that combined information relating a primary factor and a response variable across a set of 2×2 contingency tables, most notably the statistic by Mantel & Haenszel (60). This statistic and its extensions continue to receive wide use, partly because they require few assumptions. The only requirement is randomization of subjects to levels of the factor either explicitly as in randomized controlled clinical trials or implicitly by hypotheses or from conditional distribution arguments for observational data. They require only that the overall sample size be reasonably large for asymptotic results to hold. Thus they are applicable to many situations, including matched case-control studies in which tables represent matched sets of only a few subjects. A potential limitation is that generalizations of results to a target population require nonstatistical arguments of the representativeness of the study subjects to a cross-section of individuals in a target population because, in a strict statistical sense, the conclusions apply only to the study sample. Such representativeness is also an inherent and unverifiable assumption of model-based methods. An overview of Mantel-Haenszel methods is given by Kuritz et al (47).

When large sample methods cannot be justified, owing either to small samples or highly skewed observed table margins, exact methods are employed. Conditional exact methods, which we call exact methods for convenience, are based on the enumeration of a reference set of tables with the margins fixed to the totals observed in the data. P-values are determined by summing probabilities associated with tables from the reference set identified as more extreme than the table observed. These are tables with probabilities not exceeding the one observed or having some meaningful test statistic identifying a deviation from the null hypothesis greater than the deviation represented by the observed table. Monte Carlo methods are used when the sample size is too small for asymptotic inference, but too large for calculation of all probabilities in the complete permutation distribution required for exact inference. The Monte Carlo method consists of repeated sampling from the reference set and computing an unbiased point estimate and confidence interval for the exact p-value based on the sample. All of the analyses presented in this section were performed with the statistical

package StatXact (83), the most comprehensive software available for performing exact significance tests and confidence interval estimation. Agresti (3) gives a comprehensive review of exact inference.

To facilitate discussion of these methods some notation is established. Let $h = 1, 2, \ldots, q$ index the strata and let n_{hij} denote the number of subjects in the sample who are jointly classified as belonging to the i-th factor level (row), $i = 1, \ldots, s$, the j-th level of the response variable (column), $j = 1, \ldots, r$, and the hth stratum. The resulting $s \times r$ contingency table for the hth stratum is displayed in Table 2. Nonparametric inference for a set of q tables is based on conditioning on appropriate marginal totals. Fixing row margins $(N_{h1\cdot}, \ldots, N_{hs\cdot})$ and column margins $(N_{h\cdot 1}, \ldots, N_{h\cdot r})$ in a two-way table from a simple random sample induces the noncentral multivariate hypergeometric distribution, which is the foundation for both large sample asymptotic-based Mantel-Haenszel methods and exact inference for small samples based upon permutation distributions.

The Analysis of 2 × 2 Tables

Exact inference for the 2×2 table was proposed by Fisher (3). Letting θ denote the odds ratio, and H_0 and H_1 the null and research hypothesis, respectively, the one-sided p-value for the Fisher's exact test for H_0: $\theta = 1$ vs H_1: $\theta > 1$ is obtained by summing the probabilities corresponding to tables in which the sample odds ratio is at least as large as observed, or equivalently those tables whose cell count for the first row and first column is at least as large as n_{11}. Fisher showed that conditioning on the row and column margins from the observed table with cell counts $(n_{11}, n_{12}, n_{21}, n_{22})$ gives the probability of observing $n_{11} = t$ as

$$P(n_{11} = t | N, N_{1\cdot}, N_{\cdot 1}, \theta) = \frac{\binom{N_{1\cdot}}{t}\binom{N - N_{1\cdot}}{N_{\cdot 1} - t_{11}}\theta^t}{\sum_u \binom{N_{1\cdot}}{u}\binom{N - N_{1\cdot}}{N_{\cdot 1} - u}\theta^u}, \qquad 1.$$

Table 2 Observed contingency table for stratum h

	Response levels				
Factor levels	1	2		r	Total
1	n_{h11}	n_{h12}	...	n_{h1r}	$N_{h1\cdot}$
2	n_{h21}	n_{h22}	...	n_{h2r}	$N_{h2\cdot}$
.	...	...	...	...	...
s	n_{hs1}	n_{hs2}	...	n_{hsr}	$N_{hs\cdot}$
Total	$N_{h\cdot 1}$	$N_{h\cdot 2}$		$N_{h\cdot r}$	N_h

where the index of summation, u, ranges from the maximum of 0 and $N_{1\cdot} + N_{\cdot 1} - N$ to the minimum of $N_{1\cdot}$ and $N_{\cdot 1}$, the possible values for n_{11} for the given marginal totals. Under H_0, Expression 1 with $\theta = 1$ is the hypergeometric distribution. A two-sided p-value is calculated by summing the probabilities of tables from the reference set whose probabilities are no larger than the probability of the observed table.

Exact inference can also provide point estimates and confidence intervals for θ. The conditional maximum likelihood (CML) estimator is the value of θ that maximizes Expression 1 given the observed table. For tables with a zero cell the unconditional maximum likelihood (ML) estimator of θ given by $(n_{11}n_{22})/(n_{12}n_{21})$, and the CML estimator do not exist, although one can obtain a median unbiased estimate (34). In such situations, exact confidence intervals can be obtained while the asymptotic ones do not exist. Because exact methods tend to be conservative owing to the discrete distribution of the p-value, exact tests and confidence intervals are often determined excluding half the probability of the observed table, the so-called mid-p adjustment (3).

Example: osteogenic sarcoma In a study of nonmetastatic sarcoma, investigators were interested in determining the predictors of disease within three years (DIS) (29). The data for 46 patients appear in Table 3. Associations are examined for 2×2 tables formed by the response DIS and each predictor, any osteoblastic pathology (AOP), low level of lymphocytic infiltration (LI), and male gender (MALE), respectively, where, in the table, Y = yes and N = no. Because some cells have zero or small counts the Fisher's exact test is applied. Note that for the test of association of LI with DIS, the 2×2 table formed by collapsing over AOP and MALE has cell counts $n_{11} = 10, n_{12} = 0, n_{21} = 19$, and $n_{22} = 17$. Setting $\theta = 1$ in Expression 1, the probability under H_0 of observing this table is .0049, which is the one-sided p-value for testing H_0:

Table 3 Osteogenic sarcoma data*

			DIS	
AOP	LI	MALE	N	Y
N	N	N	3	0
N	N	Y	4	0
N	Y	N	5	0
N	Y	Y	5	4
Y	N	N	2	0
Y	N	Y	1	0
Y	Y	N	3	2
Y	Y	Y	6	11

*Reproduced from Mehta & Patel (65).

$\theta = 1$ vs H_1: $\theta > 1$ since n_{11} observed was the largest possible. The complete set of possible values of n_{11} for the given marginal totals range from 0 to 10 giving a reference set of 11 tables. To calculate the two-sided p-value, the tables whose probabilities are less than .0049 are identified. These are tables $n_{11} = 0$, with P = .00005, $n_{11} = 1$ with P = .0002, and $n_{11} = 2$ with P = .0024. Summing these probabilities with the one observed gives a two-sided Fisher's exact test p-value of .0075. Because $n_{12} = 0$, the asymptotic-based confidence interval for θ involving LI does not exist. An exact mid-p adjusted interval, however, is given in Table 4 with a general summary of results. Note that the presence of each covariate is positively associated with disease.

Tests of Marginal Homogeneity in 2 × 2 Tables

In the previous section, marginal totals in the 2×2 table were conditioned upon to obtain inference on the odds ratio, θ. In some situations, however, it is the parameters of the marginal distributions of each variable that are of interest, and not θ. Studies of matched pairs or repeated measures are concerned with tests of marginal homogeneity, H_0: $\pi_{1\cdot} = \pi_{\cdot 1}$ vs H_1: $\pi_{1\cdot} \neq \pi_{\cdot 1}$, where $\pi_{1\cdot}$ is the probability of observing the first level of the row variable, and $\pi_{\cdot 1}$ is the probability of observing the first level of the column variable. These hypotheses address questions concerning the equality of matched cases and controls with respect to an exposure variable, or equality of an outcome at two different times. The statistic, $(n_{12} - n_{21})/(n_{12} + n_{21})$, is based on the discordant pairs, or those that do not agree. The large sample version of the test is due to McNemar (64) and is described in Fleiss (27). Agresti (3) describes an exact approach for small samples.

The Analysis of s × r Tables

For $s \times r$ tables with sufficiently large samples, Mantel-Haenszel statistics having asymptotic chi-square distributions under the null hypothesis are often employed. For small samples, conditioning on the margins yields the noncentral multivariate hypergeometric distribution, an extension of Expression 1, from which probabilities of tables are derived. When rows and columns are nominal, the Freeman-Halton test provides a generalization of the Fisher's exact test. Its

Table 4 P-values and 95% confidence intervals for osteogenic sarcoma data

	p-Value		95% CI for odds ratio	
Predictor	Exact	Asymptotic	Exact	Asymptotic
AOP	.032	.022	(1.19, 19.33)	(1.20, 17.63)
LI	.0075	.0042	(2.99, ∞)	Undefined
MALE	.026	.021	(1.24, 44.02)	(1.17, 31.64)

p-value is defined as the null probability of the set of tables having probability no greater than that of the observed table. Alternatively, the p-value can be based upon a meaningful statistic such as Pearson's chi-squared statistic, $X^2 = \sum\sum[n_{ij} - \hat{m}_{ij}]^2/\hat{m}_{ij}$, where $\hat{m}_{ij} = N_{i.}N_{.j}/N$. The exact p-value is determined by summing the probabilities of those tables in the reference set with X^2 at least as great as the one observed. This differs from the large sample Pearson chi-squared test which compares X^2 to its asymptotic distribution. It can also give very different results than the Freeman-Halton test, because tables that are more contradictory to H_0 according to X^2 need not be less likely (3). For this reason, exact p-values based on a test statistic such as X^2 are sometimes preferred. This is the main principle in exact methods, that exact p-values are derived with respect to a reference set of test statistics that otherwise might be compared to chi-square values if the asymptotic assumptions were justified.

Exact p-values are obtained in similar ways for tables in which the column or both the row and the column variable are ordinal. When only the column variable is ordinal, hypothesis tests can be carried out by general linear rank statistics. For tables with two rows, these take the form $T = \sum_{j=1}^{r} w_j n_{1j}$, where the $\{w_j\}$ are column scores. This statistic compares two multinomial populations with ordered responses, or c populations, each generating a binary response for which the same binomial distribution is the null hypothesis versus a trend alternative for the response rates. The latter finds application, for example, in dose-response studies and usually has integer scores, $w_j = j - 1$, assigned. Arbitrary scores yield permutation tests, although a common choice are ranks (or midranks for ties) giving the Wilcoxon Rank Sum Test. The Kruskal-Wallis statistic generalizes the Wilcoxon statistic for comparing ordered response distributions in more than two multinomials. When both variables are ordinal, arbitrary scores assigned to rows and columns give general linear by linear association statistics for which exact or asymptotic p-values are available. These statistics include Spearman and Pearson–type correlation statistics, and Jonckheere-Terpstra statistics (83), the latter reducing to the Wilcoxon Rank Sum Test for $2 \times J$ tables. In a table with integer scores for both the row and column variable, an exact estimate is available for the common local odds ratio corresponding to any two adjacent rows and any two adjacent columns (2).

Example: outcomes in newborns: infant mortality and cerebral palsy. Table 5 presents data on 216 very low birthweight (VLBW) infants who had birthweight between 500 and 1350 grams, and were born in the same obstetric referral center between 7/1/88 and 6/30/92 to a mother residing in Forsyth County, North Carolina. The goal is to determine whether trends in infant mortality and cerebral palsy at one year of age (adjusted for due date) are statistically significant or likely due to chance. Some studies have reported an increase

Table 5 Outcomes in very low birthweight infants

	Males			Females		
Year	H	C	D	H	C	D
1988–89	17	5	12	9	2	6
1989–90	15	4	10	20	1	8
1990–91	16	1	15	19	0	7
1991–92	14	2	5	27	0	1

in the rate of CP in the 1980s for surviving infants explained by an increasing survival of VLBW infants (6). The outcome variable is the status of the infant at the one-year follow-up visit: H = healthy status or possible cerebral palsy, C = definite diagnosis of cerebral palsy, and D = died before the one-year follow-up visit. A definite diagnosis of CP occurred when both an attending physician and physical therapist agreed on the diagnosis. There were only 4 cases of possible CP, so these were combined with healthy infants or those without a definite diagnosis of CP. To apply the trend test to mortality, columns H and C are combined, forming the response survived versus died. The set of 4 binomial distributions for each gender are tested for equality versus a trend alternative in response rates. Sample sizes are sufficiently large to justify asymptotic based p-values for the linear rank statistic with integer scores, known as the Cochran-Armitage test (83). Two-sided p-values for males and females and overall (ignoring gender) are .79, .011, and .038, respectively. Thus, there is a statistically significant decreasing trend for females but not for males, perhaps because of a high rate in 1990–1991. The second question is whether the rate of cerebral palsy decreased for those infants who survived to follow-up. A trend test is applied to the data in Table 6, but the p-values are determined by exact methods because of zero and low counts of cerebral palsy. The exact p-values for males, females, and overall are .27, .017, and .010, respectively. CML estimates (and exact 95% confidence intervals) of the equal local odds ratios are 1.4 (0.78, 2.8), 6.4 (1.09, 276), and 2.0 (1.15, 3.77), respectively, favoring a higher CP rate in earlier years. In sum, there is a statistically significant decreasing trend for females in the rate of CP, but not for males.

The Analysis of Sets of 2 × 2 Tables

TESTS OF CONDITIONAL INDEPENDENCE AND ESTIMATION OF THE COMMON ODDS RATIO Agresti (3) describes Birch's test, an exact version of the Mantel-Haenszel test for q 2×2 tables. The p-value for H_0: $\theta = 1$ vs H_1: $\theta > 1$ is the null probability that $T = \sum_h n_{h11}$ is at least as large as observed, given the observed marginal totals $\{N_{h1.}, N_{h2.}, N_{h.1}, N_{h.2}, h = 1, \ldots, q\}$. Whereas the Mantel-Haenszel test is directed at the alternative hypothesis that a weighted

average of the stratum-specific odds ratios differs from 1, Birch's test addresses the extent to which the strata have a homogeneous odds ratio that is greater than one. For one stratum, Birch's test reduces to Fisher's exact test.

If H_0 is rejected, the average factor-response association may be estimated by the Mantel-Haenszel estimator if the sample size is sufficiently large (47). Alternatively, conditioning on the strata totals, and assuming a common odds ratio, θ, the CML estimator of θ, and exact confidence intervals can be obtained. The estimate of θ is obtained by maximizing the conditional likelihood of the joint distribution of $\{n_{111}, \ldots, n_{q11}\}$ which is a product of q terms of the type given in Expression 1. If T attains its minimum or maximum possible value, asymptotic confidence intervals for the Mantel-Haenszel estimator do not exist, but one-sided exact intervals can be obtained.

TESTS OF HOMOGENEITY OF ODDS RATIO The exact test of conditional independence involves the odds ratio being constant across strata. Zelen (93) gives an exact test for homogeneity of odds ratios. For large samples, the likelihood ratio test (1) provides a test of the stratum by factor interaction in a model with the main effects. Tarone (85) has provided a correction to the asymptotically not strictly valid "Breslow-Day" test (11).

Example: outcomes in newborns: comparing rates of cerebral palsy A possible explanation for the decline in CP is the introduction of surfactant replacement therapy in 1990 to prevent or ameliorate lung disease in some newborns (10). This suggests examining the odds ratio for CP and pre-1990 vs post-1990 (formed by combining column 1 with 2 and column 3 with 4 in Table 6) to determine if the rate of CP is different before versus after the introduction of surfactant. Since Table 6 suggests that the prevalence of CP is greater in males than females, gender is considered as a stratification variable. Zelen's exact test for a constant odds ratio gives p-value $= .49$, suggesting that the odds ratios are homogeneous. The exact p-value for the test of conditional independence is .026, so the rate of CP before 1990 is significantly greater than the rate after 1990. Asymptotic and exact results are given in Table 7.

Table 6 Number of infants (out of those who survived) with cerebral palsy

	Year			
	1988–89	1989–90	1990–91	1991–92
Males	5/22	4/19	1/17	2/16
Females	2/11	1/21	0/19	0/27

Table 7 Results for relationship of pre/post 1990 with cerebral palsy

	Gender adjusted		Unadjusted	
	Exact	Asymptotic	Exact	Asymptotic
Test of independence	.026	.022	.013	.009
Odds ratio estimate	4.26	4.18	4.94	4.98
95% Confidence Interval	(1.20, 19.83)	(1.12, 15.61)	(1.26, 28.46)	(1.41, 22.7)

The Analysis of Sets of $2 \times r$ Tables

Sets of $2 \times r$ tables present a special case for exact inference, since the analysis of sets of $s \times r$ tables for $s > 2$ is generally not computationally feasible at the present state of computing (3). A flexible class of statistics is given by the stratified linear rank statistic $T = \sum_{h=1}^{q} \sum_{j=1}^{r} w_{hj} n_{h1j}$, which is a generalization of the linear rank statistic summed over strata. For large samples they correspond to Mantel-Haenszel methods, but exact inference is also possible. For an ordinal response with integer scores, a CML estimate is available in StatXact (83) for the equal odds ratio assumed constant within tables and across strata. Hilton & Mehta (35) provide exact sample size calculations for ordinal data.

Example: outcomes in newborns: joint consideration of CP and death The variable, outcome at follow-up visit (OFUV), has three categories displayed in Table 5. Joint consideration of CP and death is of interest because the CP rate is affected by the mortality rate. The stratified linear rank statistic is used to test the association of pre/post 1990 with OFUV, adjusting for gender. Using Wilcoxon (rank) scores, the exact two-sided p-value is .085, compared to the asymptotic p-value of .079. So there is suggestive evidence that the outcomes of VLBW infants improved after 1990.

Nonparametric Analysis of Covariance for Sets of $s \times r$ Tables

As stated in the previous section, exact methods are generally not available beyond the $q \times 2 \times r$ case. For large samples, the Mantel-Haenszel statistic extends to $q \times s \times r$ tables, where q is the number of tables equal to the number of levels resulting from the cross-classification of stratifying variables. Since the landmark paper in 1959, these extended Mantel-Haenszel methods have been developed to test hypotheses involving average effects of the primary factor on the distribution of the response variable, adjusted for the potential stratum effects. Such statistics exist when both factor and response are nominal, only the response is ordinal, or both are ordinal. Kuritz et al (47) discuss these methods including their application to repeated measures situations.

Mantel-Haenszel methods may be adapted to test the independence of two categorical variables given additional covariates that may be categorical or

continuous. The technique combines the rank analysis of covariance method of Quade (77) with the randomization model framework of extended Mantel-Haenszel statistics to carry out nonparametric comparisons between groups, after adjusting for the effects of the covariates. The methodology described in Koch et al (40, 41) and illustrated in Stokes et al (84) is as follows: First, the column variable (response) and the continuous covariate are ranked; then the ranked response is regressed, via a simple linear regression model, on the ranked covariate to produce residuals. The Mantel-Haenszel test defines the groups as row categories and the residuals as the column (outcome) variable, using the calculated values of the residuals for column category scores.

Example: outcomes in newborns: comparing rates of cerebral palsy Since birthweight (BW) is an important determinant of CP (6), rank analysis of covariance is used to test the association for OFUV and pre/post 1990 adjusting for BW. First both OFUV and BW are ranked. Then residuals are obtained from the linear regression of the ranks of OFUV on the ranks of BW. A linear rank statistic is determined with pre/post 1990 defining rows and the residuals as scores for columns. The p-value .066 is suggestive for differences in the rates of outcomes for infants, adjusting for birthweight. To adjust for gender and BW, the ranking and regression steps are carried out for each gender. Then a stratified linear rank statistic is applied with gender as the stratification variable, and with pre/post 1990 as the row variable and residuals as the column variable. The nonsignificant p-value of .14 may indicate that the differences in outcomes for infants before versus after 1990 are partly explained by differences in the distributions of gender and BW. In the next section, logistic regression is applied to try to further explain the differences in CP and mortality rates.

DEVELOPMENTS IN LOGISTIC REGRESSION

Logistic regression is used to describe the relationship between a categorical response variable and a set of explanatory variables that may include both categorical and continuous variables. We review here unconditional logistic regression for dichotomous responses. Extensions of logistic regression to multi-category responses are discussed below in particular, the proportional odds model for ordinal responses. Applications of conditional logistic regression are presented including case-control studies as well as lesser known applications for highly stratified data with small sample sizes per stratum. We also discuss exact logistic regression for small sample sizes or sparsely distributed data. In the remainder of this article, models are presented for categorical data, with notation introduced as needed for case record data.

Unconditional Logistic Regression

Unconditional logistic regression remains a widely used and effective tool for assessing the magnitude of the effect of explanatory variables on response probabilities. Let the response variable $y_i = 1$ with probability π_i for the ith individual if the outcome of interest is observed, and 0 otherwise. Also suppose that all individuals' responses are statistically independent and that π_i may depend on a set of p explanatory variables, $x_{i1}, \ldots x_{ip}$, through an unknown regression parameter vector, $\beta = (\beta_1, \ldots, \beta_p)$, and γ, an intercept parameter for a reference population. To show this dependence, write $\pi_i(\beta_A)$, where $\beta_A = (\gamma, \beta)$ is the augmented parameter vector that includes the intercept. The joint distribution of the y_i, $i = 1, \ldots, N$ is

$$\prod_{i=1}^{N} \pi_i(\beta_A)^{yi}[1 - \pi_i(\beta_A)]^{1-yi}. \qquad 2.$$

Viewed as a function of the unknown β_A with the data fixed, Expression 2 is the likelihood function. The logistic model for π_i is

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \eta_i = \gamma + x_i'\beta, \qquad 3.$$

where η_i is called the linear predictor and $x_i'\beta$ is notation for $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. The odds ratio of a positive response for a unit increase in x_k is $\exp(\beta_k)$. The ML estimate $\hat{\beta}_A = (\hat{\gamma}, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is obtained by inserting Expression 3 into Expression 2 and differentiating with respect to β_A to obtain the likelihood score equations

$$\sum_{i=1}^{N} x_{Ai}[(y_i - \pi_i(\beta_A)] = 0, \qquad 4.$$

where $x_{Ai}' = (1, x_{i1}, \ldots, x_{ip})$, the augmented covariate vector for the ith individual. The solution to Equation 4 requires iteration since it is nonlinear in β_A. For sufficiently large samples $\hat{\beta}_A$ has an approximate multivariate normal distribution. Inverting the Fisher information matrix yields the covariance matrix for $\hat{\beta}_A$ from which standard errors are extracted to construct confidence intervals for the true parameters β_k, or tests of significance based on Wald tests. For the public health data analyst who is mainly interested in applying logistic regression, Hosmer & Lemeshow (36) provide a very readable text.

Logistic Regression for Ordinal Responses

The proportional odds model (61) specifies a cumulative logit link to relate p covariates to an ordinal response that takes the values, $j = 1, \ldots, r$ with corresponding multinomial probabilities for the ith individual, $\pi_{i1}, \pi_{i2}, \ldots \pi_{ir}$,

and $\sum_{j=1}^{r} \pi_{ij} = 1$. Let θ_{ig} denote the odds of observing the gth category or above for the ith individual, that is

$$\theta_{ig} = \frac{P(Y_i \geq g)}{P(Y_i < g)} = \frac{\pi_{ig} + \pi_{i(g+1)} + \cdots + \pi_{ir}}{\pi_{i1} + \pi_{i2} + \cdots \pi_{i(g-1)}}.$$

The proportional odds model is

$$\text{logit}(P(Y_i \geq g)) = \log(\theta_{ig}) = \gamma_g + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \qquad 5.$$

for $g = 2, \ldots, r$, where the intercept terms account for different logodds and probabilities that the response is at least as great as g in the reference population. Parameter interpretations are such that no matter how the cut off dichotomizes the ordinal response, the odds ratio does not change. Interpretations proceed in global terms for the odds of a higher response compared to a lower. This assumption can be checked with the score test for the proportional odds assumption (74), which is part of the output of SAS PROC LOGISTIC. If the assumption is contradicted, generalized logits may be modeled. These describe the odds of making each response relative to a reference response, giving $r - 1$ distinct sets of parameter estimates. Also known as polychotomous logistic regression, this approach has been applied to many areas, including public health (23, 48). Sample size calculations for Expression 5 are given by Whitehead (89).

Example: expanded data set for outcomes in newborns O'Shea et al (70) investigated the rate of CP over time according to the availability of surfactant replacement therapy. We consider the CP rate from 7/1/86 to 6/30/94, dividing time into two four-year intervals. The first interval, 7/1/86 to 6/30/90, corresponds to the pre-surfactant years, and 7/1/90 to 6/30/94 corresponds to the post-surfactant years. The proportional odds model is applied to OFUV using data on 394 very low birthweight infants, which include the 216 babies in Table 5. The model assumes that the effect of any covariate measured by an odds ratio, θ_{i3}, comparing the outcome of death (D) to alive at follow-up (H or C) will be equal to the odds ratio, θ_{i2}, comparing an unfavorable outcome (D or C) to H. Of interest is whether the difference in OFUV between the two periods can be explained by gender and BW (per 100 grams). All predictors were significant at the .0001 level. Table 8 shows that after adjusting for gender and BW, there remains a significant improvement in the outcomes of infants over time consistent with the nonsignificant trend from the nonparametric analysis of covariance, based on the smaller data set. This analysis shows that the occurrence of an unfavorable outcome compared to a more favorable one had 2.57 times higher odds before 1990 than after 1990 in the population of VLBW babies studied. The improvement seen for the larger data set can perhaps be explained by surfactant replacement therapy, but this information was not available for individual babies in this analysis. The results should be scrutinized

Table 8 Results of proportional odds model for outcome in infants

	Parameter estimate	Standard error	Odds ratio
Intercept 1	2.09	0.49	—
Intercept 2	2.56	0.50	—
Pre90	0.95	0.24	2.57
Male	0.95	0.24	2.57
BW (100 g)	−0.46	0.055	0.63

more closely, however, because a score test p-value of 0.012 rejects the proportional odds assumption. As a way to evaluate the general findings, separate ordinary logistic regressions were fit to the two odds of interest, θ_{i3}, and θ_{i2}. For both models, all explanatory variables were significant with p-values less than .01, and respective parameter estimates of PRE90 were 0.78 and 1.04. In other words, the occurence of death before the one-year follow-up visit had 2.18 times higher odds before 1990 as after, and the occurence of death or CP versus healthy outcome had 2.83 times higher odds before 1990 as after 1990 after adjusting for potential differences in the distributions of gender and birthweight. An alternative approach to analysis uses polychotomous logistic regression to model the odds of D versus H, and C versus H. This approach would be especially useful if the goal were to estimate the response probabilities of OFUV for different subgroups of newborns. Stokes et al (84) describe the use of SAS PROC CATMOD for polychotomous logistic regression.

Conditional Logistic Regression

Historically, conditional logistic regression has been applied to matched case-control studies in which a subject known to have the event of interest (case) is paired, or matched with a person who does not have the event (control) (11). Data are collected, retrospectively, to determine whether the case and control were exposed to certain risk factors. For example, Mack et al (57) studied women in a retirement community to determine if there was an association between the use of estrogen and the incidence of endometrial cancer. Cases were matched to controls with the same age and marital status, and who were living in the same community at the time of the diagnosis of the case. Information was collected on obesity, hypertension, gallbladder disease history, and nonestrogen drug use. The stratified logistic model for $\pi_{ij} = P(y_{ij} = 1 \mid x_{ij})$, where $y_{ij} = 1$ if the jth subject in the ith matched pair has the event and x_{ij} is the vector of risk factors, is

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_i + x'_{ij}\beta, \qquad 6.$$

where γ_i is a stratum-specific parameter for the ith matched set, and β is a p dimensional vector of parameters common across all strata. In a matched case-control study the stratum refers to the matched set. Unconditional ML estimation discussed above would give incorrect results because there are too many parameters to estimate in relation to the sample size. A solution can be obtained for matched pairs by conditioning on the set of $\{y_{i1} + y_{i2}\}$ to eliminate the γ_i. The resulting conditional likelihood is maximized to provide CML estimates of factors that differ within stratum. To implement the model for matched pairs, each is transformed to a single observation with explanatory variables having values equal to the differences between the corresponding values for the case and the control. The outcome variable has the value 1 for all paired observations. An unconditional logistic regression model without intercept is fitted to the newly created covariates and artificial response. Conditional logistic regression reduces to McNemar's test when a single dichotomous explanatory variable is used. A related but more general method is needed for studies with n cases and m controls. It involves a likelihood like that for proportional hazards models for time-to-event data. Stokes et al (84) illustrate the use of SAS PROC PHREG for n cases and m controls, and SAS PROC LOGISTIC for the 1:1 endometrial cancer study.

Conditional logistic regression applies to other highly stratified data where primary interest is in within-stratum factors. Multicenter clinical trials with many centers and small numbers of patients per center can have centers as the strata. Center effects are conditioned out and inference is made on within-center treatment comparisons. Conditional logistic regression also applies to some correlated data situations. In multiperiod cross-over trials with 2 or more treatments, conditional maximum likelihood removes the subject-specific effect to make inference on the within-subject treatment effects. Stokes et al (84) provide such an illustration. Alternatively, generalized estimating equations (90) (described below) provide more comprehensive modeling, which also accounts for explanatory variables at the subject level.

Example: schizophrenia and birth complications In a case-control study looking at the role of birth complications in patients with schizophrenia, several siblings from each of 7 families were studied. Each individual was classified as either normal (NM) or with schizophrenia (SH). The goal was to determine if there was a positive relationship between schizophrenia and a birth complications (BC) index that took a value somewhere between 0 (uncomplicated birth) to 15 (severely complicated birth). The data appear in Table 9. In Expression 6, γ_i denotes the ith family effect, and β represents the effect of birth complications on schizophrenia. The conditional maximum likelihood estimate for β is 0.325, with asymptotic p-value (Wald-type) $= .053$. The odds

Table 9 Schizophrenia and birth complications data*

Family ID	BC index	Number of siblings SH	NM	Family ID	BC index	Number of siblings SH	NM
1	15	1	0	3	1	0	1
1	7	0	1	4	2	1	0
1	6	0	1	4	0	0	4
1	5	0	1	5	6	0	1
1	3	0	2	5	3	1	0
1	2	0	3	5	0	1	0
1	0	0	1	6	3	0	1
2	2	1	0	6	0	1	2
2	0	0	1	7	6	1	0
3	9	1	0	7	2	0	1
3	2	0	1				

*Data reproduced from Mehta & Patel (65).

ratio estimate is 1.384 with a 95% asymptotic-based confidence interval of (0.996, 1.924).

Exact Inference for Logistic Regression

When sample sizes are too small or data too sparsely distributed, exact methods of estimation and inference can be applied for models in the previous three sections. The principle of conditioning on observed margins of tables to obtain exact tests for contingency tables can be extended to logistic regression. Inference is based on exact permutational distributions of the sufficient statistics that correspond to the regression parameter of interest, conditioning on the sufficient statistics of the remaining parameters fixed at their observed values. Exact logistic regression simplifies to exact tests for contingency tables because observed row and column margins correspond to certain sufficient statistics. The conditioning principle applies to both the unstratified model (Expression 3) and the stratified model (Expression 6) so that the same algorithms are applied to both in the computations. For unstratified binary data, the $p \times 1$ vector of sufficient statistics for β, with elements $t_1, \ldots, t_p$, is $t = \sum_{i=1}^{n} y_i x_i$. The conditional distribution of t is obtained by conditioning on $m = \sum_{i=1}^{n} y_i$, the sufficient statistic for γ. Exact inference for β requires fixing some of the sufficient statistics at their observed values and varying those of interest over their permissible ranges. If interest is in a single parameter, say β_1, then inference is based on the distribution of $P(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p)$, which may be represented by $f(t_1|\beta_1)$. An exact p-value for H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$ is given by the conditional probabilities test as described by Mehta & Patel (65). A confidence interval can be obtained by inverting the test. Maximizing

$f(t_1|\beta_1)$ yields a conditional maximum likelihood estimate for β_1, if it exists; otherwise an alternative estimate is the median unbiased estimate that is always defined.

In the stratified binary model, exact inference is carried out in an analogous fashion as discussed by Mehta & Patel (65). The approach extends the conditional likelihood method discussed above in that instead of conditioning only on the sufficient statistic for the stratum-specific parameters, the sufficient statistics for the other parameters are conditioned on as well, to obtain the conditional distribution of the sufficient statistic of the parameter of interest. Exact inference for the polychotomous logistic regression is considered by Hirji (33).

Example: osteogenic sarcoma The data in Table 3 are reconsidered to assess the simultaneous relationship of the predictors with disease through the unstratified logistic regression model with covariates AOP, LI, and MALE. Because all of the diseased subjects had LI present, LI is a perfect predictor, and the usual ML logistic regression analysis cannot be used. However, it is possible to estimate β_{LI} and provide a confidence interval and p-value using exact inference. Table 10 gives two-sided p-values, model-adjusted estimates, and confidence intervals for odds ratios for each predictor obtained using LogXact (56). Unlike the unadjusted results p-values in Table 4, the simultaneous consideration of the three predictors reveals that LI is nearly significant, but AOP and MALE are not.

Example: schizophrenia and birth complications In the previous discussion of this example, the total number of cases with schizophrenia in each family were conditioned upon to obtain CML estimates. The p-value and confidence interval presented were based upon large sample properties of the conditional likelihood. However, exact inference may be more appropriate because there were only 29 family members. The exact p-value is 0.017 and the 95% exact confidence interval for the odds ratio is (1.02, 2.10). The results change from not significant at the 0.05 level to significant when one switches from asymptotic to exact inference.

Table 10 Exact logistic regression results for osteogenic sarcoma data

Parameter	p-value	Odds ratio	95% CI
AOP	.15	3.18	(0.60, 20)
LI	.061	6.59	(0.85, ∞)
MALE	.12	4.70	(0.70, 56)

GENERALIZED ESTIMATING EQUATIONS

The previous section described logistic regression analysis in which observations were assumed independent and parameter estimation was achieved by maximum likelihood or exact methods. In many categorical data situations, however, independence assumptions may not hold, and ignoring the intracluster correlation would lead to incorrect variance estimates for the regression parameters (52). This section discusses generalized estimating equations (51) that extend unconditional logistic regression and other generalized linear models (68) to the analysis of cluster sample data. We provide a brief summary of the theory of generalized linear models (GLMs) as background for a general discussion of GEE. The final sections discuss GEE methods for correlated binary and multinomial responses, respectively.

Background

GLMs are a general class of models for independent observations. A GLM has three components: (*a*) a random component that defines the form of the distribution for the *i*th observation belonging to the exponential family of distributions; this family includes, among others, the binomial, poisson, and normal distributions, corresponding to logistic, loglinear, and linear regression, respectively; (*b*) the systematic component given by the linear predictor, $\eta_i = x'_{Ai}\beta_A$, for the *i*th observation; and (*c*) the link function, g(·), which links the random and systematic components by the equation, $\eta_i = g(\mu_i)$. The marginal distribution of y_i has mean and variance

$$\mathrm{E}(y_i) = \mu_i, \quad \mathrm{g}(\mu_i) = \eta_i = \gamma + x'_i\beta, \quad \mathrm{var}(y_i) = \mathrm{v}(\mu_i)\phi, \qquad 7.$$

where v(·) is the variance function, γ is the intercept, β is a $p \times 1$ vector of regression coefficients, and ϕ is the scale parameter, either known or to be estimated. The regression coefficients, β_A, are estimated by solving the maximum likelihood score equations for β_A:

$$\sum_{i=1}^{N} D'_i \,\mathrm{var}(y_i)^{-1}(y_i - \mu_i(\beta_A)) = 0, \qquad 8.$$

where $D_i = \partial\mu_i/\partial\beta_A$. For the logistic regression model in Expression 3, $\mu_i = \pi_i$, $\mathrm{v}(\mu_i) = \mu_i(1-\mu_i)$, the variance of a bernoulli random variable, $\phi = 1$, and $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$. Alternative generalized linear models for a binary response result from using other link functions such as the probit and complementary log-log link. Like the logit link, these have the attractive quality that all possible values of the linear predictor in $(-\infty, +\infty)$ correspond to values of π_i in (0,1). For the logit link, Equation 8 reduces to Equation 4 because $\partial\mu_i/\partial\eta_i = \mu_i(1-\mu_i)$, and applying the chain rule from calculus gives

$D_i = \mu_i(1 - \mu_i)x'_{Ai}$. A book by McCullagh & Nelder (62) is considered the primary reference. Dobson (22) has a shorter introductory text. Healy (32) provides an introduction to GLIM (71), a software package for fitting GLMs.

Note that Expression 8, which in general terms is called an estimating equation because its solution gives an estimate of β_A, is completely specified by the mean and variance of y_i given in Expression 7. Without further distributional assumptions it would be called a quasi-score and its integral a quasi-likelihood (62, 75, 88). For binomial responses or counts, a quasi-likelihood analysis may be carried out in which ϕ is estimated. Such an analysis allows for overdispersion ($\phi > 1$), which occurs when variation in the responses exceeds that expected under binomial or poisson sampling (50).

GEE: Examples and Theory

Like GLMs, generalized estimating equations apply to continuous responses as well as discrete responses, including counts, binomial, or multinomial responses. GEEs may be viewed as an extension of GLMs from random sampling to cluster sampling as they account for the dependence of observations within clusters. In other words, GEE applies to regression models with multivariate responses, or a cluster of responses whose elements correspond to correlated (usually positively) observations. Three general regression formulations are commonly used in public health.

1. The response vector corresponds to repeated measures taken on subjects, and parameter estimates are obtained for the marginal means in a single regression model. In a prospective study of coronary heart disease risk factors (28), young adults were followed annually for seven years to describe and compare blood pressure patterns among educational, ethnic, and gender groups (55a). An individual defines a cluster, and his or her blood pressure at a given time is an observation.

2. In a cross-sectional study, the response vector may correspond to a cluster of distinct subjects, such as patients in a medical practice. Then, a single regression model relates covariates to a patient's response. In a state-wide study of cancer screening in Colorado, patient records of 1798 women over 50 years of age were sampled from 132 physicians in as many primary care practices to determine the level of preventive services being provided (16, 58). Practices define clusters and patient charts are observations. The data are discussed below.

3. When the response vector consists of different outcome variables, separate regression models may be fit to each, yet taking the correlation among responses into account through a multivariate procedure such as GEE may

provide more efficient estimation of regression parameters. In the management of dizziness in primary care settings, the response vector consists of three clinical decisions: whether or not to require medication, order laboratory tests, and whether to refer for further testing (82). A cluster is comprised of a patient's 3 binary outcomes.

Some notation and general theory for GEE is presented before special consideration is given to binary and ordinal responses. Let n_i equal the number of observations in the ith cluster, and let K be the total number of clusters of possibly varying size. Clusters are indexed by i and observations are indexed by t. Let $y_i = (y_{i1}, \ldots, y_{in_i})'$ be a vector of n_i outcome values, and $X_i = (x_{i1}, \ldots, x_{in_i})'$ a $n_i \times p$ matrix of cluster-and/or observation-level covariate values for the ith cluster. Estimates of β_A are obtained by solving the generalized estimating equations

$$\sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \mu_i(\beta_A)) = 0 \qquad 9.$$

where D_i is now an $n_i \times (p+1)$ matrix, $V_i = A_i R_i(\alpha) A_i$ is the working variance matrix of y_i, and $A_i = \text{diag}\{v^{1/2}(\mu_{it})\}$ is a $n_i \times n_i$ diagonal matrix. Furthermore, $R_i(\alpha)$ is a $n_i \times n_i$ working correlation matrix that depends on an unknown parameter vector α. GEE estimation involves alternating between iteratively reweighted least squares estimation for β_A and method of moments estimation for ϕ and α. For a large number of clusters, the GEE estimate of β_A has an approximate multivariate normal distribution (51). In other words, valid inference requires a sufficiently large sample size in terms of the number of clusters. This is in contrast to conditional logistic regression applied to cluster samples (with clusters conditioned as strata) where the sample size requirements are in terms of the number of observations, albeit at the price of the stronger assumption of conditional independence given cluster-specific parameters. The "robust" or "sandwich" variance estimate of $\hat{\beta}_A$ is given by

$$\left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1} \left\{ \sum_{i=1}^{K} D_i' V_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' V_i^{-1} D_i \right\} \times \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1}, \qquad 10.$$

where β_A, ϕ, and α are replaced by their estimates. Diagonal estimates of Expression 10 are the variances of individual parameter estimates, and off-diagonal elements are their estimated covariances. Unlike model-based counterparts,

they are robust in the sense that they are valid even if $R_i(\alpha)$ is mis-specified, assuming correct specification of the model for the mean (51). Like quasi-score equations, Expression 9 only depends on means and second moment parameters (variances and covariances) and does not require the specification of the response vector's full multivariate distribution. They extend quasi-likelihood by introducing additional parameters, α, to describe the nature of the intra-cluster correlation. However, the power of statistical tests for β_A can be increased by carefully modeling $\mathrm{R}_i(\alpha)$: (*a*) in repeated measures such as the blood pressure study, a first-order autoregressive correlation gives a larger correlation to observations that are closer in time as opposed to those that are temporally farther apart; (*b*) in the cancer screening study, an exchangeable correlation is used in which any two women in the same practice are assumed to have equal correlation; and (*c*) in the dizziness study there are 3 distinct correlations to estimate, one for each outcome pair. Alternatively, the odds ratio may be used as a measure of association (54).

The Analysis of Correlated Binary Data

The GEE procedure can be used to fit logistic models to binary responses from cluster samples. The mean and variance of the tth binary response in the ith cluster is given by

$$\mathrm{E}(y_{it}) = Pr(y_{it}) = \pi_{it}, \quad \mathrm{logit}(\pi_{it}) = \eta_{it} = \gamma + x'_{it}\beta,$$
$$\mathrm{v}(\pi_{it}) = \pi_{it}(1 - \pi_{it}), \qquad 11.$$

where $\mathrm{v}(\pi_{it})$ is the variance of y_{it}. The regression parameter vector, β, may address variation across both clusters and observations. Expression 11 assumes that β is the same for each outcome, although for the dizziness example we would write β_t instead and obtain estimates of the coefficients for each outcome.

Example: cancer screening in primary care practices As part of the Partners Project (16, 58), patient and physician characteristics were examined for their associations with mammography screening. Let $y_{it} = 1$ if the tth patient in the ith practice received a mammogram in the past year, and 0 otherwise. A logistic regression model is fit with the covariates patient age (three categories: 50–59, 60–69, 70–74) and smoking status, physician gender, whether the physician practices family medicine or internal medicine, and whether the patient made a health maintenance visit (HMV) in the previous year. An exchangeable correlation structure is assumed where the pairwise correlation is the same for any two women in the same practice. Twenty-two women were sampled per practice, but cluster sizes vary slightly due to occasional missing data. Parameter estimates and their standard errors are given in Table 11 (*a*); the estimate of the exchangeable correlation was .037. Patients who made a HMV

Table 11 GEE results for cancer screening data

	(*a*) Mammography			(*b*) Ordinal outcome		
Parameter	Estimate	Standard error	Odds ratio	Estimate	Standard error	Odds ratio
Intercept 1	−1.40	0.22	—	−2.57	0.21	—
Intercept 2	—	—	—	−1.06	0.20	—
Hmv	2.28	0.14	9.80	3.17	0.15	23.8
Internal med.	0.15	0.17	1.16	0.23	0.16	1.26
Female physician	0.36	0.18	1.43	0.52	0.16	1.68
Patient nonsmoker	0.45	0.15	1.56	0.32	0.12	1.38
Patient (60, 69) yrs	−0.04	0.12	0.97	0.03	0.12	1.03
Patient 70 + yrs	−0.19	0.17	0.82	−0.12	0.15	0.89

had 9.8 times higher odds to get a mammogram than those who did not receive a HMV. Female physicians had 1.43 times higher odds as male physicians to order mammograms, and nonsmoking patients had 1.56 times higher odds to get a mammogram than smokers. Physician specialty and patient age were not significantly associated with the level of mammography screening after adjusting for the other predictors.

The Analysis of Correlated Ordinal Data

The GEE approach can be applied to correlated ordinal data by extending the proportional odds model in Expression 5. Modifying Expression 5, the model becomes

$$\log(\theta_{itg}) = \gamma_{0g} + \beta_1 x_{it1} + \cdots + \beta_p x_{itp} \qquad 12.$$

where θ_{itg} is the odds of observing a response in the gth category or above for the tth member of the ith cluster. The estimating equations are generalizations of Expression 9. As in GEE for binary outcomes, analogous working correlation structures are possible, but R_i is more complicated because there is dependency between outcome levels of the ordinal response (within observations), as well as across observations. In general, additional computational burden results with increasing cluster sizes because R_i is commonly inverted at each iteration. An independence working correlation may be used to alleviate this concern, although some efficiency may be lost.

For correlated nominal outcomes, GEE may be applied to generalized logit models. In work sampling studies, for example, repeated observations are made on a subject who is categorized into performing one of a finite number of tasks. These tasks often form nonordered categories. Miller et al (67) apply GEE methods to a medical setting in which at any given time during a

doctor/patient session the doctor might be observed with the patient, with the patient's chart, or engaged in another task. Other developments include GEE influence diagnostics (76) and second-order generalized estimating equations that model marginal means and odds ratios, simultaneously (31, 53).

Example: cancer screening in primary care practices A second analysis concerns the level of adherence to recommendations of the National Cancer Institute for annual mammogram and clinical breast exam. Each woman is assigned a score from 0 to 2 indicating the number of exams received, with a 2 indicating that both mammogram and clinical breast exam were received in the past year. A working exchangeable correlation structure is used (55). Parameter estimates and their standard errors are given in Table 11(*b*). Patients who made a HMV had 23.8 times higher odds to have a higher level of cancer screening than women who did not make a HMV. The other predictors had effects similar in magnitude as when only mammogram was considered.

WEIGHTED LEAST SQUARES

Historically, marginal modeling of discrete repeated measures was investigated by Koch & Reinfurt (45) and Koch et al (44), who extended the weighted last squares (WLS) methodology of Grizzle et al (30). The WLS method fits models to a large class of response functions, including marginal proportions, marginal logits, cumulative logits, and mean scores. Although GEE methods apply to many analyses for observational units, WLS is particularly useful for fitting models to aggregate quantities such as rank measures of correlation and incidence densities, as discussed below. The procedure applies noniterative generalized least squares to the response functions of interest, using their observed covariance matrix as weights. For sufficiently large samples, the response functions have an approximate multivariate normal distribution. Tests of hypotheses can be carried out on pertinent linear combinations of the response functions. The WLS method requires that subjects be grouped into moderately large subpopulations defined by identical covariate values. Both GEE and WLS are nonlikelihood or semiparametric methods that use weighted least squares algorithms to fit models for marginal response functions. GEE accounts for the correlation structure to gain efficiency, whereas WLS may be viewed as a special case of GEE using an empirically estimated covariance matrix (66).

Rank Measures of Correlation for Ordinal Responses

In some clinical trials, patients receive a randomly allocated dosage level of a drug and their outcome status is determined at two or more visits. In studies where both the dose and the outcome (none, some, marked) are ordinal, rank measures of association may be modeled using WLS procedures to determine

the extent of treatment effects and treatment by time interactions (13, 42). This approach is especially useful if there is uncertainty about the goodness of fit of the specific structure imposed by models for ordinal data such as a proportional odds assumption. Carr et al (13) describe WLS applied to Goodman-Kruskal rank correlation coefficients, sometimes called gamma coefficients, which make minimal assumptions about the true nature of the association except for ordinality of the treatment and response. Analyses are based on hypothesis tests applied to linear functions of the correlations with specifications for treatment effects and interactions of treatment and center. Other background variables, in addition to center, such as baseline measurement are easily incorporated in the analysis.

Carr et al (13) consider a clinical trial of 111 patients who had respiratory illness. In each of two centers patients were randomly assigned one of two treatments (placebo = 0, active = 1). The response variables were ordinal classifications (0 = terrible, 1 = poor, 2 = fair, 3 = good, 4 = excellent) of the health status of each patient at each of four successive visits after the initiation of treatment. The baseline status of patients was also evaluated according to the same ordinal classification. Inference is based on the gamma coefficients at each visit within each center, and their estimated covariance matrix. For clinic 1, the gammas based on 56 patients were .148, .467, .292, and .242 for visits 1, 2, 3, and 4, respectively. Similarly, the gammas for clinic 2 based on 55 patients were .421, .647, .559, and .433. The results indicate that the treatment $\times$ visit interaction is significant ($p = .044$) for the combined centers. The interaction is due to the stronger extent of better response for active treatment at visits 2 and 3 than visits 1 and 4. They found no treatment by center interaction.

The Modeling of Incidence Densities

Weighted least squares methods may also be used to model correlated incidence densities. Lavange et al (49) consider repeated measures that describe the effect of passive exposure to tobacco smoke on lower respiratory illness (LRI) in children. They present an analysis of 158 children exposed to passive tobacco smoke compared to 136 unexposed. Interest was in comparing the incidence rates for different groups of children. A complication in the data is that the at-risk periods for each child vary randomly. They use WLS to model incidence densities for which the covariance structure is determined by sample survey methods.

In the first stage of the WLS analysis, subgroup-specific incidence densities and their covariance matrix are estimated, the latter by a first-order Taylor series approximation. The incidence density (ID) for a particular subgroup is defined as the ratio of the number of illness episodes to the total person-time at risk. These estimates are used to compare marginal rates of LRI between

the exposed and unexposed groups. The SUDAAN software package (81) was used to estimate incidence densities and test contrasts among them. In the second stage of analysis, WLS is applied using SAS PROC CATMOD to fit models to these densities in order to assess the main effects and interactions among the various risk factors. For the exposed and unexposed groups, the IDs were 1.24 and 0.65, respectively, for exposed and unexposed groups. A test of no differences between LRI for the exposed and unexposed groups had a Wald p-value of .001, indicating that the exposed group had a significantly higher rate of illness.

In another approach, they fit a poisson model using GEE to counts representing the number of episodes of illness per child at each period. This approach estimates average rates by using an offset term in the model to account for time at risk in each period. The GEE procedure has the advantage of applying to more comprehensive models than WLS, including those with continuous covariates. Alternatively, the WLS approach is attractive because direct estimates of adjusted incidence densities for risk factors of interest can be provided. Moreover, no distributional assumptions are required, only random sampling of primary sampling units (persons).

CONCLUDING REMARKS

Categorical data analysis encompasses a much broader array of techniques today than only 15 years ago due to many methodological and computational advances. While chi-square and unconditional logistic regression methods still comprise the staple diet of public health researchers, methods for ordinal response data are receiving wider use. More recently, fast algorithms for exact inference such as provided by StatXact (83) have encouraged the inferential analysis of small data sets. Analyses of complex data sets characterized by cluster samples, continuous and categorical covariates, and cluster- and observation-level covariates, are no longer restricted to naive analyses based upon aggregating information at the cluster level. While the generalized estimating equations approach has replaced weighted least squares in many analyses of correlated discrete data, the latter is still useful for a wide variety of response functions.

Another approach to modeling correlated categorical data comes from a sample survey perspective (7, 39, 49, 81). Used widely in the analysis of very large sample survey data sets, but applicable in epidemiology and other disciplines, survey regression weights observations according to the sampling design, and provides estimators like GEE. For simple random samples, the results correspond to those obtained with a working independence assumption in GEE.

Some authors propose maximum likelihood methods for correlated discrete data (5, 25, 94), which has certain advantages and disadvantages compared to

GEE. In particular, maximum likelihood applies to a wider range of missing data assumptions (4, 26), although the GEE procedure has been modified to increase their scope (78). Finally, GEE is better suited for analyzing large numbers of clusters of varying size and makes fewer distributional assumptions.

In addition to marginal models, random effects and transitional models address cluster samples of discrete data in public health research (21, 52, 69, 72, 90). These three classes of models differ with respect to parameter interpretations. In random effects models, regression parameters have cluster-specific interpretations, and the surplus of parameters are reduced by assuming underlying probability distributions, usually normal distributions, specified up to a few unknown parameters (12, 17, 63, 86). In transitional models, the estimated regression parameters have interpretations that are conditional on the values of other outcomes, such as measurements at previous times, or the responses of other family members (9, 79, 91).

Bayesian statistical methods are also used in public health (24). In particular, Empirical Bayes methods have been employed in estimating disease incidence and mortality rates (14). These procedures are useful when diseases are rare, and the poisson variation is overdispersed. It allows non-zero estimation of very low rates corresponding to zero observed numbers of deaths. The Empirical Bayes estimate of the rate is a weighted average of a local rate and the global rate of all the regions. Others have described binary response models in epidemiology when an exposure variable is measured or classified with error (46, 80, 87). Categorical methods also apply to grouped survival data (20, 42), including extensions of Mantel-Haenszel methods (19, 59). Stokes et al (84) illustrate these methods as well as poisson regression for categorized time-to-event data using the SAS system.

While many of the methods discussed in this paper have only appeared in the past 15 years, the coming years will witness their wider dissemination among public health researchers. Computational advances will continue expanding the scope of exact methods, and complex data sets will more commonly be analyzed with statistical methods that incorporate various patterns of dependency in the data.

ACKNOWLEDGMENTS

The work of the first author was partially supported by a grant from the Agency for Health Care Policy and Research (HS 06992). We would like to acknowledge the support of Stuart Cohen, Maureen McClatchey, and Brent Shelton in regard to the cancer screening data, and Michael O'Shea who provided the cerebral palsy data. We especially thank Bahjat Qaqish for many helpful discussions.

Literature Cited

1. Agresti A. 1990. *Categorical Data Analysis*. New York: Wiley & Sons
2. Agresti A, Mehta CR, Patel NR. 1990. Exact inference for contingency tables with ordered categories. *J. Am. Stat. Assoc.* 85:453-458
3. Agresti A. 1992. A survey of exact inference for contingency tables. *Statist. Sci.* 7:131–77
4. Baker SG, Laird NM. 1988. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Am. Stat. Assoc.* 83:62–69
5. Balagtas CC, Becker MP, Lang JB. 1993. Marginal modeling of categorical data from crossover experiments. *Appl. Statist.* 44:63–77
6. Bhushan V, Paneth N, Kiely JL. 1993. Impact of improved survival on very low birth weight infants on recent secular trends in the prevalence of cerebral palsy. *J. Pediatrics* 91:1094–100
7. Binder DA. 1983. On the variance of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* 51:279–92
8. Bishop YMM, Fienberg SE, Holland PW. 1975. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press
9. Bonney GE. 1986. Regressive logistic models for familial disease and other binary traits. *Biometrics* 42:611–25
10. Bose C, Corbet A, Bose G, Garcia-Prats J, Lombardy L, et al. 1990. Improved outcome at 28 days of age for very low birthweight infants treated with a single dose of a synthetic surfactant. *J. Pediatrics* 117:947–53
11. Breslow NE. 1996 Statistics in epidemiology: the case-control study. *J. Am. Stat. Assoc.* 91:14–28
12. Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear models. *J. Am. Stat. Assoc.* 88:9–25
13. Carr GJ, Hafner KB, Koch GG. 1989. Analysis of rank measures of association for ordinal data from longitudinal studies. *J. Am. Stat. Assoc.* 84:797–804
14. Clayton DG, Kaldor J. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43:671–81
15. Clogg CC. 1981. New developments in latent structure analysis. In *Factor Analysis and Measurement in Sociological Research*, ed. DM Jackson, EF Borgatta, pp. 215–46. Beverly Hills, CA: Sage
16. Cohen SJ, McClatchey MW, Wolfe P, Shelton BJ, Archer PG. 1996. Health maintenance visit as the determinant of cancer screening by primary care physicians. *J. Gen. Int. Med.* 11(Suppl. 1):121 (Abstr.)
17. Conaway M. 1990. A random effects model for binary data. *Biometrics* 46:317–28
18. Cox DR. 1958. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B* 20:215–42
19. Cox DR. 1972. Regression models and life tables. *J. R. Stat. Soc. Ser. B* 34:187–220
20. Deddens JA, Koch GG. 1988. Survival analysis: grouped data. In *Encyclopedia of Statistical Sciences*, ed. A Kotz, NL Johnson, 9:129–34. New York: Wiley & Sons
21. Diggle PJ, Liang KY, Zeger SL. 1994. *Analysis of Longitudinal Data*. Oxford: Oxford Univ. Press
22. Dobson AJ. 1990. *An Introduction to Generalized Linear Models*. London: Chapman & Hall
23. Elston JM, Koch GG, Weissert WG. 1990. Regression-adjusted small area estimates of functional dependency in the noninstitutionalized American population age 65 and over. *Am. J. Public Health* 81:335–43
24. Etzioni RD, Kadane JB. 1995. Bayesian statistical methods in public health and medicine. *Annu. Rev. Public Health* 16:23–41
25. Fitzmaurice GM, Laird NM. 1993. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 80:141–51
26. Fitzmaurice GM, Molenberghs G, Lipsitz SR. 1995. Regression models for longitudinal binary responses with informative drop-outs. *J. R. Stat. Soc. Ser. B* 57:691–704
27. Fleiss JL. 1981. *Statistical Methods for Rates and Proportions*. New York: Wiley & Sons
28. Friedman GD, Cutter GR, Donahue RP,

Hughes GH, Hulley SB, et al. 1988. Cardia: study design, recruitment, and some characteristics of the examined subjects. *J. Clin. Epidemiol.* 41:1105–16
29. Goorin AM, Perez-Atayde A, Gebhardt M, Andersen JW, Wilkinson RH, et al. 1987. Weekly high-dose methotrexate and doxorubicin for osteosarcoma. *J. Clin. Oncol.* 5:1178–84
30. Grizzle JE, Starmer CF, Koch GG. 1969. Analysis of categorical data by linear models. *Biometrics* 25:489–504
31. Heagerty SL, Zeger SL. 1996. Marginal regression models for clustered ordinal measurements, *J. Am. Stat. Assoc.* 91:1024–36
32. Healy MJR. 1990. *Glim: An Introduction*. Oxford: Oxford Univ. Press
33. Hirji KF. 1992. Exact distributions for polytomous data. *J. Am. Stat. Assoc.* 87:487–92
34. Hirji KF, Tsiatis AA, Mehta CR. 1989. Median unbiased estimation for binary data. *Am. Stat.* 43:7–11
35. Hilton JF, Mehta CR. 1993. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* 49:609–16
36. Hosmer DW, Lemeshow S. 1989. *Applied Logistic Regression*. New York: Wiley & Sons
37. Imrey PB, Koch GG, Preisser JS. 1996. The evolution of categorical data modeling: a biometric perspective. In *Advances in Biometry: Fifty Years of the Biometric Society*, ed. P Armitage, HA David, pp. 89–113. New York: Wiley & Sons
38. Imrey PB, Koch GG, Stokes ME 1981. Categorical data analysis: some reflections on the log linear model and logistic regression, part I. *Int. Stat. Rev.* 49:265–83
39. Imrey PB, Koch GG, Stokes ME 1982. Categorical data analysis: some reflections on the log linear model and logistic regression, part II. *Int. Stat. Rev.* 50:35–64
40. Koch GG, Amara IA, Davis GW, Gillings DB. 1982. A review of some statistical methods for covariance analysis of categorical data. *Biometrics* 38:563–95
41. Koch GG, Carr GJ, Amara IA, Stokes ME, Uryniak TJ. 1990. Categorical data analysis. In *Statistical Methodology in the Pharmaceutical Sciences*, ed. DA Berry, pp. 391–475. New York: Marcel Dekker
42. Koch GG, Edwards S. 1988. Clinical efficacy trials with categorical data. In *Biopharmaceutical Statistics for Drug Development*, ed. KE Peace, pp. 403–51. New York: Marcel Dekker
43. Koch GG, Gillings DB, Stokes ME. 1980. Biostatistical implications of design, sampling and measurement to the analysis of health science data. *Annu. Rev. Public Health* 1:163–225
44. Koch GG, Landis JR, Freeman JL, Freeman DH, Lehnen RG. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 33:133–58
45. Koch GG, Reinfurt DW. 1971. The analysis of categorical data from mixed models. *Biometrics* 27:157–73
46. Kohlmeier L, Bellach B. 1995. Exposure assessment error and its handling in nutritional epidemiology. *Annu. Rev. Public Health* 16:43–59
47. Kuritz SJ, Landis JR, Koch GG. 1988. A general overview of Mantel-Haenszel methods: applications and recent developments. *Annu. Rev. Public Health* 9:123–60
48. Lafata JE, Koch GG, Weissert WG. 1994. Estimating activity limitation in the non institutional population: a method for small areas. *Am. J. Public Health* 84:1813–17
49. Lavange LM, Keyes LL, Koch GG, Margolis PA. 1994. Application of sample survey methods for modelling ratios to incidence densitites. *Stat. Med.* 13:343–55
50. Liang KY, McCullagh P. 1993 Case studies in binary dispersion. *Biometrics* 49:623–30
51. Liang KY, Zeger SL. 1986 Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22
52. Liang KY, Zeger SL. 1993 Regression analysis for correlated data. *Annu. Rev. Public Health* 14:43–68
53. Liang KY, Zeger SL, Qaqish BF. 1992. Multivariate regression models for categorical data. *J. R. Statist. Soc. Ser. B* 54:3–40
54. Lipsitz SR, Laird NM, Harrington DP. 1991. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* 78:153–60
55. Lipsitz SR, Kim K, Zhao L. 1994. Analysis of repeated categorical data using generalized estimating equations. *Stat. Med.* 13:1149–63
55a. Liu KL, Ruth KJ, Flack JM, Jones-Webb R, Burke G, et al. 1996. Blood pressure in young blacks and whites: relevance of obesity and lifestyle factors in determining differences. *Circulation* 93:60–66
56. LogXact. 1992. *Software for Exact Logistic Regression*. Cambridge, MA: Cytel Software

57. Mack TM, Pike MC, Henderson BE, Pfeffer RI, Gerkins VR, et al. 1976. Estrogens and endometrial cancer in a retirement community. *N. Engl. J. Med.* 294:1262–67
58. Main DS, Cohen SJ, DiClemente CC. 1995. Measuring physician readiness to change in cancer screening: preliminary results. *Am. J. Prev. Med.* 11:54–58
59. Mantel N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo. Rep.* 50:163–70
60. Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22:719–48
61. McCullagh P. 1980. Regression models for ordinal data (with discussion). *J. R. Stat. Soc. Ser. B* 42:109–42
62. McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. London: Chapman & Hall. 2nd ed.
63. McCulloch CE. 1994. Maximum likelihood variance components estimation for binary data. *J. Am. Stat. Assoc.* 89:330–35
64. McNemar Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12:153–57
65. Mehta CR, Patel NR. 1995. Exact logistic regression: theory and examples. *Stat. Med.* 14:2143–60
66. Miller ME, Davis CS, Landis JR. 1993. The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics* 49:1033–44
67. Miller ME, James MK, Langefeld CD, Espeland MA, Freedman JA. 1996. Some techniques for the analysis of work sampling data. *Stat. Med.* 15:607–18
68. Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *J. R. Stat. Soc. Ser. A* 135:370–84
69. Neuhaus JA, Kalbfleisch JD, Hauck WW. 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* 59:25–35
70. O'Shea TM, Klinepeter KL, de Regnier RA, Dillard RG, Goldstein D, et al. 1993. Decreased risk of cerebral palsy (CP) with increasing use of surfactant therapy. *Pediatric Res.* (Part 2) 269A (Abstr.)
71. Payne CD. 1986. *The GLIM Manual*. Release 3.77. Oxford:NAG
72. Pendergast JF, Gange SJ, Newton MA, Lindstrom MJ, Palta M, et al. 1996. A survey of methods for analyzing clustered binary response data. *Int. Stat. Rev.* 64:89–118
73. Prentice RL, Farewell VT. 1986. Relative risk and odds ratio regression. *Annu. Rev. Public Health* 7:35–58
74. Peterson B, Harrell FE. 1990. Partial proportional odds models for ordinal response variables. *Appl. Stat.* 39:205–17
75. Preisser JS, Koch GG, Shockley WW. 1993. Statistical considerations for cross-sectional data relating to tracheal reconstruction over time. *J. Biopharm. Stat.* 3:167–83
76. Preisser JS, Qaqish BF. 1996. Deletion diagnostics for generalised estimating equations. *Biometrika* 83:551–62
77. Quade D. 1967. Rank analysis of covariance. *J. Am. Stat. Assoc.* 62:1187–200
78. Robins JM, Rotnitzky A, Zhao LP. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* 90:106–21
79. Rosner B. 1984. Multivariate methods in ophthalmology with applications to other paried-data situations. *Biometrics* 40:1025–35
80. Satten G, Kupper L. 1993. Inferences about exposure-disease associations using probability-of-exposure information. *J. Am. Stat. Assoc.* 88:200–8
81. Shah BV, Barnwell BG, Hunt PN, Lavange LM. 1992. *SUDAAN User's Manual Release 6.0.* Research Triangle Park, NC
82. Sloane PD, Dallara J, Roach C, Bailey KE, Mitchell M, et al. 1994. Management of dizziness in primary care. *J. Am. Board Fam. Pract.* 7:1–8
83. StatXact. 1996. *StatXact: Statistical Software for Exact Nonparametric Inference*, Version 3. Cambridge, MA: Cytel Software
84. Stokes ME, Davis CS, Koch GG. 1995. *Categorical Data Analysis Using the SAS System.* Cary, NC: SAS Institute
85. Tarone RE. 1985. On heterogeneity tests based on efficient scores. *Biometrika* 72:91–95
86. Ten Have TR. 1996. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* 52:473–91
87. Thomas D, Stram D, Dwyer J. 1993. Exposure measurement error influence on exposure-disease relationships and methods for correction. *Annu. Rev. Public Health* 14:69–93

88. Wedderburn RWM. 1974. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 61:439–47
89. Whitehead J. 1993. Sample size calculations for ordered categorical data. *Stat. Med.* 12:2257–71
90. Zeger SL, Liang KY. 1992. An overview of methods for the analysis of longitudinal data. *Stat. Med.* 11:1825–39
91. Zeger SL, Qaqish BF. 1988. Markov regression models for time-series: a quasi-likelihood approach. *Biometrics* 44:1019–31
92. Zeger SL, Liang KY. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121–30
93. Zelen M. 1971. The analysis of several 2 2 contingency tables. *Biometrika* 58:129–37
94. Zhao LP, Prentice RL. 1990. Correlated binary regression using a quadratic exponential model. *Biometrika* 77:642–48