VARIANCES FOR POLYCHOTOMOUS LOGISTIC REGRESSION USING COMPLEX SURVEY DATA

Shelley B.Bull, University of Toronto and Linda L.Pederson,University of Western Ont.
Shelley Bull, Mt.Sinai Hosp. Research Inst. 600 University Av. Toronto ON M5G 1X5

Use of polychotomous logistic regression in the analysis of data from complex surveys requires that the parameter and variance estimates take into account the design of the survey. This paper applies results on the variances of asymptotically normal estimators from complex surveys to this model, and describes how design-based variances and hypothesis tests can be calculated. An example is given using data from a population survey of attitudes toward legislative measures to restrict smoking. Design effects determined from regressions on a single factor are compared to those determined from regressions on multiple factors.
KEY WORDS: Multinomial logistic; Design-based; Taylor linearization; Wald tests; Design effects; Categorical response.

1. INTRODUCTION

Polychotomous logistic regression is frequently the method of choice when the outcome is categorical (2 or more mutually exclusive, unordered response categories) and interest is in the relationship between the outcome and covariates. The covariates may be binary, categorical, ordinal, or continuous. The logistic regression procedure is based on the likelihood $L = \pi f(y_k|\mathbf{x}_k)$, where, given the covariate vectors $\mathbf{x}_1,\ldots,\mathbf{x}_N$, the responses y_k are conditionally independent multinomial random variables with parameters $\mu_1,\ldots,\mu_J$. It is assumed that $\log(\mu_j/[1-\Sigma\mu_j])=\mathbf{x}^T\beta_j$ where the β_j are unknown parameter vectors. There are thus J regression equations, each comparing the probability of response in category j (j = 1,2,...,J) to the probability of response in a reference category. Maximum likelihood estimates of the logistic parameters can be shown to be asymptotically normally distributed. Although less well-known than dichotomous (binary) logistic regression, considerable material on the more general model is available. (See the review by Albert and Lesaffre 1986).

Several authors have addressed the application of binary logistic regression when the assumptions of simple random sampling and/or independence of observations are untenable (Koch et al 1975, Hidiroglou and Paton 1987, Roberts et al 1987). When data have been collected in a complex survey, with stratification, clustering, or unequal selection probabilities for example, logistic regression coefficients and estimated variances that ignore these features may be misleading. Binder (1983) gave design-based methods for asymptotically normal estimates based on Taylor linearization methods for variance estimation. These methods are applicable to a class of finite population parameters that include those of the generalized linear models and the polychotomous logistic regression model.

2. MOTIVATION

A complex sample survey of attitudes toward restrictive measures against smoking was conducted in 1984 in the Province of Ontario (Pederson et al 1986a). The design of the survey was stratified multistage sampling with a total of 100 first stage clusters selected with replacement from three strata. Population, stratum, and subpopulation distributions were calculated using SUPERCARP (Hidiroglou et al 1980) to take account of the design (Pederson et al 1986b, 1987); individual sampling weights, inversely proportional to the selection probability, were applied. Secondary analysis of the data focussed on the relationship of a common set of sociodemographic characteristics to several measures of attitude. The objective was to identify subgroups of smokers and non-smokers with negative and positive attitudes. These attitude measures were categorical with 2, 3 or 4 response categories. The set of socio-demographic characteristics included binary variables (eg. sex), continuous variables (eg. age), categorical factors represented by indicator variables (eg. marital status), and ordinal variables (eg. level of education).

Although design-based methods for bivariate associations with both continuous and categorical responses are available in SUPERCARP, neither this program nor any of the major statistical packages include design-based multivariate procedures for categorical response variables. Binder's methods were therefore extended to the polychotomous model and a SAS program written to implement the procedure.

Two questions related to the effect of the design on the estimated variances of the parameter estimates were of interest:

(1) Is it necessary to adjust variances in this survey? In other words, will adjustment change our conclusions about the important associations with attitudes?

(2) Are design effects for single factor regressions (e.g. age alone) similar to those for multiple factor regressions (e.g. age controlling for other characteristics)? Since design

effects are generally larger for misspecified models, one might expect the latter to have smaller design effects than the former. To address these questions, several attitude measures were selected and comparisons were made between analyses with and without adjustment for the design. Individual sampling weights were included in both analyses.

3. DESIGN-BASED ESTIMATION AND INFERENCE

3.1 Application of Binder's theory

In this paper the approach to parameter and variance estimation given by Binder (1983) is applied to the polychotomous logistic regression model. Following his notation, the finite population parameter vec $[(\mathbf{b}_1,\ldots\mathbf{b}_J)^T]$ is estimated with B, the solution of

$$W(B) = \sum_{k=1}^{n} w_k \mathbf{x}_k \otimes [\mathbf{y}_k - \boldsymbol{\mu}(\theta_k)] = \mathbf{0}$$

where the sum is over the n individuals in the sample and w_k is the sampling weight for individual k. For a given vector of covariates $\mathbf{x}^T = (x_1,\ldots,x_p)$, (with $x_1 = 1$), it is assumed that $\theta_j = \mathbf{x}^T\mathbf{b}_j$, and that $\mu_j(\boldsymbol{\theta}) = \exp(\theta_j)/[1 + \Sigma \exp(\theta_j)]$, $j = 1,2,\ldots,J$. The response $\mathbf{y}_k$ is a vector of J indicators for the multinomial outcome. The Kronecker product is denoted by $\otimes$. The covariance matrix of B, adjusted for the design, is estimated by

$$V(B) = [J^{-1}(B)]\ \Sigma(B)\ [J^{-1}(B)].$$

J(B) is a consistent estimator of the matrix of second derivatives of the log likelihood, given

$$\text{by} \quad \sum_{k=1}^{n} [w_k \mathbf{x}_k \mathbf{x}^T_k \otimes M_k]$$

where $M_k = \text{diag}\,[\mu_1(\theta_k),\ldots,\mu_J(\theta_k)] - \boldsymbol{\mu}(\theta_k)\boldsymbol{\mu}(\theta_k)^T$. $J^{-1}(B)$ is thus the usual covariance matrix estimate obtained, for example, in the final step of Newton-Raphson iteration.

$\Sigma(B)$ is a consistent estimate of the variance of a total based on the residuals $\mathbf{r}_k = \{w_k \mathbf{x}_k \otimes [\mathbf{y}_k - \boldsymbol{\mu}(\theta_k)]\}$, $(k = 1,2,\ldots,n)$. For stratified multistage sampling with the first stage clusters selected with replacement, the formula for the variance of the estimated total for a vector of variables r is a function of the cluster totals.

3.2 Design effects and Wald tests

The design effect for a single parameter estimate b_{ij} is usually defined by the ratio of the adjusted to the unadjusted variance estimate. A large sample test of the null hypothesis $H_o: b_{ij} = 0$ is made by comparing the test statistic $\{b_{ij}/se(b_{ij})\}$ to a standard normal deviate, or equivalently by comparing the square of the test statistic to a X_1^2 deviate. The design effect will therefore also be given by the square of the ratio of the test statistic based on the unadjusted variance to the test statistic based on the adjusted variance.

In general a Wald test for linear hypotheses can be formulated as $H_o: CB = \mathbf{0}$, where C is a known d by Jp coefficient matrix, which is tested using the statistic

$$Q = B^TC^T[C\ S\ C^T]^{-1}CB.$$

S denotes a consistent estimate for the covariance matrix for B and Q is distributed as approximately X^2 with d degrees of freedom, (Koch et al 1975, Roberts et al 1987). In subsequent sections we use Q_u to denote the test statistic calculated using $S = J^{-1}$, the unadjusted covariance, and use Q_a to denote the statistic based on S = V, the adjusted covariance. A multivariate design effect is then defined as Q_a/Q_u, (Rao & Scott 1981).

In analyses using the polychotomous logistic model, joint tests of parameters from different regressions are of interest, as well as tests of parameters from a single regression. An example of the former, a test for association between the variable x_2 and response in any of the outcome categories would be formulated as $H_o: b_{21} = b_{22} = \ldots = b_{2J} = 0$ and Q would be approximately X^2 with J degrees of freedom. For the latter, a test for association between x_2 or x_3 and response in outcome category j would be formulated as $H_o: b_{2j} = b_{3j} = 0$ and Q would be X^2 with 2 degrees of freedom. Joint tests of parameters from different regressions can also be formulated with several covariates; such tests are useful in determining, for example, whether a categorical factor expressed as 2 or more indicator variables is associated with a polychotomous response.

3.3 Description of SAS program

A SAS program was developed to implement the theory given in sections 3.1 and 3.2; it is similar in structure to SAS programs described by Hidiroglou and Paton (1987) for binary logistic regression. The following summarizes the steps in the program:

Step 1: Calculation of parameter estimates and unadjusted covariance matrix J^{-1} using PROC CATMOD (SAS User's Guide: Statistics version 5) with maximum likelihood estimation and application of sampling weights.

Step 2: Calculation of the residuals $\mathbf{r}_k$ from B and $\mathbf{x}_k$ using PROC MATRIX (SAS STATS 1982, to be replaced by PROC IML in version 6), including application of the sampling weights.

Step 3: Aggregation of the residuals over clusters within strata using PROC MEANS.

Step 4: Calculation of the estimate Σ as a sum of stratum specific covariance matrices using PROC CORR and PROC MATRIX.

Step 5: Calculation of the adjusted covariance matrix V using PROC

MATRIX for matrix multiplication.
Step 6: Calculation of Wald tests of hypotheses using PROC MATRIX for matrix inversion and multiplication.

Step 1 is repeated for various models in a main program and steps 2 through 6 are included in a subroutine which is called repeatedly from the main program. With the pending release of PROC CATMOD and PROC IML in SAS/PC, this program could be adapted to run on a microcomputer. Table 1 provides a summary of CPU requirements for the SAS procedures. With multiple covariates, PROC CATMOD requires the major part of the CPU time. These requirements depend on the number of response categories, the number of variables in the model, and on the number of populations. The latter is the number of unique combinations of covariate values. When there are several continuous covariates in the model, the number of populations will usually be equal to the number of observations. The number of iterations to convergence of the maximum likelihood estimates is given in the fourth column of the table.

4. WALD TEST DESIGN EFFECTS

4.1 Description of the data

Four attitude measures in each of two subpopulations (smokers and non-smokers) were selected, including responses with 2,3, and 4 categories. Two were selected from a series of questions about restrictions on smoking in specific locations such as work-places, restaurants, and airplanes. Also selected were measures of attitude toward bans on advertising and sales of tobacco in drug stores. The distributions are reported in table 2a. Table 2b defines the sociodemographic characteristics of interest in identifying subgroups with negative and positive attitudes to smoking restrictions.

4.2 Results

Table 3 reports the regressions of attitude to smoking in workplaces in a group of 362 smokers. All nine sociodemographic characteristics are examined together. The first three columns refer to the regression comparing the 'no restrictions' response category to the 'restricted' response category, while the second three columns refer to the regression comparing the 'not permitted' response category to the 'restricted' category. The Q statistics are for the hypothesis tests H_o: $b_{p1} = 0$ and H_o: $b_{p2} = 0$. The last two columns report the Q statistics for the joint hypothesis test H_o: $b_{p1} = b_{p2} = 0$. Also in table 3 are joint hypothesis test statistics for the contribution of the categorical factors: marital status, employment status, and strata; and for the contribution of polynomials in age, education, and socioeconomic status.

The unadjusted joint hypothesis tests indicate that age and education are significantly related to attitude to restrictions in workplaces. Tests of individual parameters suggest that: older smokers are more likely to prefer prohibition than younger; part-time workers are less likely to prefer no restrictions and more likely to prefer prohibition than are those working fulltime. Those more educated tend to be more moderate in their attitude: less likely to prefer prohibition and less likely to prefer no restrictions. When the hypothesis tests are adjusted for the design (Q_a in table 3), the results for age and education are essentially unchanged, but the association with employment status is weakened.

Tables 4a and 4b summarize the design effects of joint hypothesis tests of each of the sociodemographic characteristics for the four attitude-subpopulation combinations. Underscoring of entries in these tables indicates that both unadjusted and adjusted hypothesis tests are significant at the 5% level. A down arrow to the right of the entry indicates that the unadjusted test is significant at the 5% level but the adjusted test is not. An up arrow indicates the converse. The last row in the tables includes the median values of the design effects.

In the single factor analyses there are 10 instances in which adjustment of the hypothesis tests leads to 5% level significance being lost. There is one instance in which significance is gained with adjustment. Design effects in the single factor analyses range between .74 and 3.16 with only 2 exceeding 2.00. In the multiple factor analyses, however, there are only 3 instances in which changes occurred: 2 in which 5% level significance is lost with adjustment and one in which it is gained. Design effects range between .59 and 2.65 with only 2 exceeding 1.75. The design effects appear to be larger and more variable for the single factor regressions and for the smaller group of smokers.

4.3 Summary

Generalizations from comparisons based on a single, rather moderately sized survey sample must be limited, especially since associations with the outcomes tended to be weak and power in the group of smokers may be low. In the analyses reported here, adjusting the joint hypothesis tests for the design made some difference in the single factor associations detected as significant, but little difference in the multiple factor regressions. Design effects rarely exceeded 2 and were frequently less than unity for both single and multiple factor analyses. This may be due to the use of the cross-classes of smokers and non-smokers, since cross class design effects have been found to be smaller than total

sample design effects (Verma et al 1980). In an exploratory study such as this, in which type I errors are likely to be inflated because of the large number of associations being examined, adjusting the variances for the design did not appear to be worthwhile for the multiple factor analyses. This would not necessarily be the case in a confirmatory study.

5. SUGGESTIONS FOR FURTHER RESEARCH

The methods described in this paper are large sample methods. It would be valuable to establish how large a sample is needed for the methods, particularly the hypothesis tests, to be valid. Unlike likelihood ratio statistics in which the nested model must be refit, Wald statistics have the advantage that tests of nested hypotheses can be made using the estimated covariance matrix from the more complicated model. In large samples the two tests will be equivalent. However, in small samples there may be differences in behaviour. There is a need for comparison of the properties of design based tests (for example, in a Monte Carlo simulation) for moderate sized survey samples.

Also of interest would be an examination of the effect of grouping continuous covariate values to reduce the number of populations, an approach used by Roberts et al (1987). This would be computationally more efficient but might have poorer estimation efficiency, especially in smaller samples.

ACKNOWLEDGEMENTS

This project was funded by Ontario Ministry of Health grant 01710. Lori Corrin provided programming assistance.

REFERENCES

Albert, A. and E.Lesaffre (1986). Multiple group logistic discrimination, Comp. & Maths. with Appls., 12A,209-24.

Binder, DA (1983). On the variances of asymptotically normal estimators from complex surveys, International Statistical Review, 51, 279-92.

Hidiroglou, MA., Fuller, WA., Hickman, RD. (1980). Supercarp, 6th ed. Ames, Iowa: Statistical Laboratory, Iowa State University.

Hidiroglou, MA., Paton, DG (1987). Some experiences in computing estimates and their variances using data from complex survey designs, in Applied Probability, Stochastic Processes, and Sampling Theory, eds IB MacNeill and GJ Umphrey, Boston: D. Reidel, 285-308.

Koch, GG., Freeman, DH., Freeman, JL. (1975). Strategies in the multivariate analysis of data from complex surveys, International Statistical Review, 43, 59-78.

Pederson, LL., Bull, SB., Ashley, MJ and Lefcoe, NM (1986a&b). A population survey on legislative measures to restrict smoking in Ontario: 1. Design methodology and sample representativeness, and 2. Knowledge, attitudes, and predicted behaviour, American J of Preventive Medicine, 2,307-315,316-23.

Pederson, LL., Bull, SB., Ashley, MJ. and Lefcoe, NM (1987). A population survey in Ontario regarding restrictive measures in smoking: Relationship of smoking status to knowledge, attitudes and predicted behaviour, International Journal of Epidemiology, 16, 383-391.

Rao, JNK., Scott, AJ. (1981). The analysis of categorical data from complex sample surveys: Chi-square tests for goodness of fit and independence in two-way tables, Journal of American Statistical Association, 76, 221-30.

Roberts, G., Rao, JNK., and Kumar, S. (1987). Logistic regression analysis of sample survey data, Biometrika, 74, 1-12.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the world fertility survey, JRSS-A, 143, 431-473.

Table 1 Computational Summary : CPU time in sec on IBM 4341 under CMS SAS 5.16

Number of				STEP 1	STEP 2	STEPS 3 - 6	
Resp	Var	Pop	Iter	CATMOD	MATRIX	MEANS & CORR	MATRIX
2	1	2	3	.67	.87	.77	1.22
	2	57	4	1.58	.88	.83	1.27
	2	140	4	3.52	.88	.94	1.29
	16	368	5	31.63	1.63	1.82	3.84
	16	720	3	42.70	2.70	2.24	3.92
3	1	2	5	.71	1.22	.88	1.30
	2	57	5	2.20	1.25	1.01	1.44
	2	140	5	5.03	1.24	1.00	1.39
	16	362	6	74.65	2.20	3.37	7.07
	16	718	6	150.43	3.61	4.16	7.08
4	1	2	4	.75	1.20	1.04	1.47
	2	58	4	2.39	1.24	1.20	1.67
	2	141	4	5.22	1.22	1.24	1.60
	16	364	5	114.00	2.47	5.42	18.63
	16	725	7	307.02	4.16	6.68	18.59

TABLE 2a Percentage distributions of attitudes to restrictions

Attitude:	Response categories	Sub-population distributions smokers	non-smokers
Smoking in the workplace:	not permitted	9.9	20.5
	restricted area *	65.6	70.5
	no restrictions	24.4	8.9
Smoking in stores:	not permitted	64.0	68.9
	restricted area *	23.0	21.0
	no restrictions	13.0	10.1
Tobacco adds should be forbidden:	strongly agree	12.4	23.0
	agree *	24.4	30.8
	disagree	47.1	42.1
	strongly disagree	16.1	4.0
Tobacco sales in drug stores:	should be sold *	68.7	47.8
	should not be sold	31.3	52.2

* reference category in polychotomous logistic regression

TABLE 2b Definition and coding of the covariates

Characteristic	Variable Name(s)	Coding
Respondent's sex	RSEX	1 - male 0 - female
Respondent's age	CTAGE	(age - 40)
	CTAGE2	quadratic term in CTAGE
Marital status	MS1	1 - never married 0 - other (married)
	MS2	1 - separated/divorced 0 - other (married)
	MS3	1 - widowed 0 - other (married)
Attendance at church or synagogue	CHUR	1 - more than once a month 0 - less
Level of education	CTEDUC	-2 - elementary school -1 - some high school 0 - high or trade school 1 - college or some university 2 - university degree
Years lived in the area	MOBILITY	1 - 1 year or less 2 - 2 to 3 years 3 - 4 to 9 years 4 - 10 years or more
Employment status	EMPNOT	1 - not working 0 - other (full time & self-employed)
	EMPPAR	1 - working part-time 0 - other (full time & self-employed)
Socioeconomic status	CTSES	(Blishen code - 5000)/1000
	CTSES2	quadratic term in CTSES
Strata indicators	RURAL	1 - rural 0 - other (urban without bylaws)
	URBW	1 - urban with bylaws 0 - other (urban without)

TABLE 3 Multivariate results of logistic regression of attitude to smoking in workplaces on sociodemographic characteristics (362 smokers)

Variable name	no restrictions vs. restricted			not permitted vs. restricted			Joint Wald tests	
	b_1	Q_u	Q_a	b_2	Q_u	Q_a	Q_u	Q_a
RSEX	0.611	4.52*	3.53	0.596	1.51	1.32	5.29	4.81
CTAGE	-0.023	3.57	2.31	0.024	1.76	3.18	6.42*	6.12*
CTAGE2	2.3E-4	0.09	0.12	14.7E-4	2.16	2.41	2.16	2.41
							12.89*	13.41**
MS1	-0.426	1.33	1.12	-0.467	0.45	0.47	1.57	1.30
MS2	0.037	0.01	0.01	0.135	0.04	0.04	0.04	0.04
MS3	0.778	1.17	1.21	-1.698	1.46	3.16	3.05	6.24*
							4.98	8.39
CHUR	-0.530	2.69	1.75	-0.520	0.96	1.21	3.29	3.16
CTEDUC	-0.316	5.51*	3.64	-0.136	0.20	0.25	5.54	3.74
CTEDUC2	0.019	0.04	0.03	-0.466	4.03*	6.69**	4.21	7.34*
							10.16*	12.78*
MOBILITY	0.048	0.13	0.11	0.515	2.89	1.09	2.92	1.10
EMPNOT	-0.383	1.39	1.72	0.494	0.79	0.86	2.58	2.62
EMPPAR	-0.974	3.89*	3.07	0.794	1.57	1.69	6.31*	5.62
							7.14	6.53
CTSES	-0.084	1.93	1.06	0.019	0.03	0.02	2.08	1.18
CTSES2	4.9E-4	0.00	0.00	-0.055	1.99	1.56	2.05	1.57
							4.24	3.33
RURAL	0.621	3.53	1.68	0.369	0.47	0.62	3.66	1.88
URBW	0.282	0.90	0.96	0.353	0.44	0.45	1.17	1.20
							4.05	2.67

* $p<.05$ ** $p<.01$

Table 4 Summary of joint test design effects for bivariate and multivariate logistic regressions on attitudes

Part a	Workplaces				Stores			
	Smokers		non-smokers		Smokers		non-smokers	
Factor	biv	multiv	biv	multiv	biv	multiv	biv	multiv
SEX	1.65	1.10	1.14	1.20	1.66	1.06	1.02	0.92
AGE	.89	.96	.96	.94	.90	1.03	1.14	.94
MARITAL	.91	.59	1.22	1.02	1.12	1.06	1.07	.81↑
CHURCH	1.36	1.04	1.20	1.12	1.34	1.68	1.22	1.12
EDUCATION	.81↑	.79	.90	.99	1.06	1.02	.81	.84
MOBILITY	2.95	2.65	1.06↓	1.09	.80	.63	1.12	1.41
EMPLOYMENT	1.20↓	1.09	.74	.90	3.16	1.62	1.09	1.41
SES	1.37	1.27	1.37	1.14	1.51↓	1.26	1.17	.94
STRATA	1.50	1.52	.84	.89	1.28↓	1.20↓	1.37	1.38
MEDIAN	1.36	1.10	1.06	1.02	1.28	1.06	1.12	.94

Part b	Advertising				Sales			
	Smokers		non-smokers		Smokers		non-smokers	
Factor	biv	multiv	biv	multiv	biv	multiv	biv	multiv
SEX	1.70	1.36	1.27↓	1.39	1.55	1.42	.87	1.08
AGE	.95	.93	1.06↓	.73	1.05	.91	1.02	1.09
MARITAL	1.06	.79	.88	.75	1.26↓	.97	.98	.86
CHURCH	1.59	1.46	.88	.91	1.50	1.38	1.37	1.35
EDUCATION	.92	1.02	1.08	1.08	1.27↓	.98	1.19	1.32
MOBILITY	1.32	1.00	1.15	1.16	1.12	1.09	1.25	1.18
EMPLOYMENT	1.64	1.29	1.00	1.37	.78	1.45	1.91↓	1.35
SES	1.43	1.04	1.47	1.20	1.31	1.01	1.33	1.15
STRATA	1.60	1.49	1.31↓	1.29	1.45	1.52	1.78	1.96
MEDIAN	1.43	1.04	1.08	1.16	1.27	1.09	1.25	1.18