AN APPLICATION OF LOGISTIC REGRESSION METHODS TO SURVEY DATA: PREDICTING HIGH COST USERS OF MEDICAL CARE

Lisa M. LaVange, Vincent G. Iannacchione, and Steven A. Garfinkel, Research Triangle Institute

1. Introduction

Logistic regression is a widely used tool for the statistical analysis of observed proportions or rates. Methods for fitting logistic multiple regression models have been available through standard statistical packages for the past several years. Until recently, however, this capability has not been available to the survey data analyst due to the fact that standard logistic regression methods are inappropriate for the analysis of data arising from a complex sample design. RTI has developed and is currently marketing a software package consisting of probability sampling based procedures for fitting logistic models to survey data. Estimates of the model parameters and their corresponding variance covariance matrix are produced with this software that accurately reflect the sample design. A brief review of the underlying theory for the approach taken in developing the RTI survey logistic regression software is presented in Section 2 of this paper. An application of these methods to predict high cost users of medical care based on data from the National Medical Care Utilization and Expenditure Survey is described in Section 3.

2. Background and Methods

Logistic regression analysis consists of fitting a linear logistic model to an observed proportion or rate in order to measure the relationship between the outcome variable and one or more explanatory variables. The linear logistic model (Koch and Edwards, 1985) is given by

$$\pi(\underset{\sim}{x}) = \{1 + \exp(-\alpha - \underset{\sim}{x}'\underset{\sim}{\beta})\}^{-1}, \qquad (1)$$

where $\pi(\underset{\sim}{x})$ is the expected value of the observed proportion $p(\underset{\sim}{x})$ for the subpopulation defined by by the vector of explanatory variables $\underset{\sim}{x}$; and α and $\underset{\sim}{\beta}$ are the unknown constant term and vector of regression coefficients to be estimated. The coefficients given by $\underset{\sim}{\beta}$ can also be interpreted as liner regression coefficients with respect to the log odds ratio or logit transformation in that β_j measures the extent to which an increase in x_j results in an increase in the logit, given by

$$\Psi(\underset{\sim}{x}) = \log_e\{\pi(\underset{\sim}{x}) / [1 - \pi(\underset{\sim}{x})]\} \qquad (2)$$

$$= \alpha + \underset{\sim}{x}'\underset{\sim}{\beta} \ .$$

Logistic regression is typically applied to binary outcome variables for which the product binomial probability distribution can be assumed. Examples of applications include quantal bioassay experiments in which the model specifies a dose-response relationship and epidemiologic studies concerned with certain health measures such as the proportion of people seeking medical care. The model parameters α and $\underset{\sim}{\beta}$ are usually estimated by maximum likelihood methods assuming the product binomial distribution for the binary responses. Logistic models are a special case of the more general log linear model and can thus be used for the analysis of binary responses in the framework of multidimensional contingency tables. For a complete review of methods of analysis appropriate for log linear models in general and logistic models in particular, see Imrey et al. (1981, 1982).

The distributional assumptions required for standard methods of analysis of log linear models are violated when applying those methods to complex survey data involving stratification and clustering. Several articles appearing recently in the literature have proposed design effect adjustments that could be incorporated into standard categorical data analysis programs. For log linear model analyses of categorical data arising from a complex survey design, Rao and Scott (1984) suggested an adjustment to the chi-squared test statistics based on the design effects of the multinomial responses. Scott and Wild (1985) compared weighted and unweighted estimators of logistic model parameters for stratified case control studies. They found the unweighted estimators to be more efficient when the model was true but not as robust under alternatives to the model.

Procedures have been developed at RTI for the specific problem of fitting logistic regression models to survey data such that the model parameter estimates and their variance covariance matrix accurately take the survey design into account. Solutions to the weighted likelihood equations are produced resulting in design consistent estimators of the finite population regression coefficients. An application of the Taylor series method is employed to produce an explicit form for the variance covariance matrix of the regression coefficients, as suggested by Binder (1981). These methods are outlined in the following paragraphs and are due to Folsom (Shah, Folsom, et al., 1984).

Let

$y_i \equiv$ 1 if sample member i has the attribute of interest; 0 otherwise.

$\underset{\sim}{x}_i \equiv (1, x_{1i}, \ldots, x_{qi})$; a vector of predictors or regression variables

$w_i \equiv$ sampling weight for sample member i.

For the ith sample member the following logistic model is assumed:

$$\Pr[y_i = 1 | \underset{\sim}{x}_i, \underset{\sim}{\beta}] = [1 + \exp(-\underset{\sim}{x}_i\underset{\sim}{\beta})]^{-1} = p_i(\underset{\sim}{\beta}) \qquad (3)$$

For notational convenience the intercept term is included in the vector of model parameters $\underset{\sim}{\beta}$. The weighted likelihood equations are then given by

$$\sum_{i \in s} w_i \underset{\sim}{x}_i^T p_i(\hat{\underset{\sim}{\beta}}) = \sum_{i \in s} w_i \underset{\sim}{x}_i^T y_i \qquad (4)$$

Paralleling standard unweighted logistic model analysis, the solution to (4) is obtained iteratively using the Newton Raphson method. The default starting value vector $\hat{\beta}_o$ for the first first iteration is the (q+1) element null vector.

If the convergent solution is denoted by $\hat{\beta}_*$ and $\hat{p}_{i*} = [1 + \exp(-x_i \hat{\beta}_*)]^{-1}$ then the associated information matrix is given by

$$INF_* \equiv \sum_{i \in s} x_i^T x_i w_i \hat{d}_{i*} \qquad (5)$$

with

$$\hat{d}_{i*} \equiv \hat{p}_{i*} \left(1 - \hat{p}_{i*}\right) .$$

The delta method covariance matrix estimator for $\hat{\beta}_*$ is derived from the following linearized variate vectors:

$$z_{hci} \equiv x_{hci}^T \left(y_{hci} - \hat{p}_{hci*}\right)$$

$$\equiv x_{hci}^T \hat{r}_{hci*}$$

where $\hat{r}_{hci*}$ is the residual for the i-th sample member associated with the c-th primary sampling unit (PSU) or primary cluster from primary stratum-h. Weighted accumulations z_{hc} of these linearized vectors are first formed at the PSU level. The associated between-PSU within stratum mean square matrix is then formed as follows:

$$S_z \equiv \sum_{h=1}^{L} n_h S_{hz} \qquad (6)$$

where n_h denotes the number of PSUs in stratum-h and

$$S_{hz} = \sum_{c=1}^{n_h} \left(z_{hc} - z_{h\bullet}\right) \left(z_{hc} - z_{h\bullet}\right)^T / \left(n_h - 1\right)$$

depicts the (q+1) by (q+1) matrix of sample mean squares and cross products from stratum h with

$$z_{h\bullet} = \sum_{c=1}^{n_h} z_{hc}/n_h .$$

Referring to the sample weighted information matrix from equation (3) as INF*, the delta covariance matrix for $\hat{\beta}_*$ is

$$cov_\Delta\left(\beta_*\right) = \left(INF_*\right)^{-1} S_z \left(INF_*\right)^{-1} . \qquad (7)$$

The development of a logistic regression software package for survey data proceeded in two stages. The existing SAS procedure PROC LOGIST, developed by Harrell (1984) for ordinary logistic regression analysis, was first modified to produce solutions to the weighted likelihood equations given in (4) above. The user need only specify the NORMWT option and provide a variable corresponding to the sampling weight w_i for each sample member in the call to PROC LOGIST. The procedure produces output identical to that produced for ordinary logistic regression with the exception that the regression coefficient estimates are the weighted counterparts of the ordinary estimates. It should be noted that the standard errors and chi-squared test statistics also printed do not accurately correspond to the weighted estimators since the usual distributional assumptions upon which they are based are not made in the finite population setting. If the analyst is proceeding under superpopulation assumptions, the standard application of PROC LOGIST, ignoring the sampling weight, is more appropriate.

A new SAS procedure RTILOGIT was developed to process the output of the modified PROC LOGIST for variance computations. This procedure calculates the estimated variance covariance matrix of the weighted estimators according to the Taylor series method described in (5) - (7) above. Tests of the hypotheses that the logistic regression coefficients are equal to zero are produced based on a Hotelling's T^2 type statistic that is assumed to have a transformed F distribution in repeated samples (Shah, Holt, and Folsom, 1977). The new procedure RTILOGIT makes efficient use of many of the subroutines originally developed for RTI's linear regression software package for survey data SURREGR.

3. Predicting High Cost Users of Medical Care

Under a recent contract with the Health Care Finance Administration (HCFA), RTI has been involved in extensive analyses of data from the National Medical Care Utilization and Expenditure Survey (NMCUES). One research topic explored under this contract was the identification of important predictors of the high cost use of medical care. It is estimated that the 10 percent of the U.S. population that incurred the highest medical care charges in 1980 was responsible for 75 percent of all incurred charges. Knowledge about persons who incur high medical expenses is valuable for establishing policy to control costs. With this goal, explanatory models were fit to the probability of incurring high costs that incorporated demographic, socioeconomic, and health resource supply variables.

The data used in this study were collected for the civilian, noninstitutional population of the U.S. in 1980 by the NMCUES. Comprehensive data were collected at the household and person level including charges incurred for health care, payments made by all sources, health status, income and insurance coverage. Data on community characteristics in 1980 were obtained from the Area Resource File and included median gross rent, number of hospital beds, and physician supply. The price of medical care was represented by the 75th percentile of prevailing charges to Medicare for a routine follow-up visit by a general practitioner and were obtained from the Medicare Directory of Prevailing Charges, 1980. Community characteristics and prevailing charges were merged from their respective sources to the NMCUES database by county of residence.

Separate models were fit to persons aged 17 to 64 years and persons aged 65 and older. These

two groups are generally thought to differ substantially with respect to both insurance coverage and utilization. Among the 17-64 year age group, persons in the top 10 percent of the distribution of total health care costs were classified as high cost. This 10 percent accounted for an estimated 73 percent of all charges incurred. Among the 65 and older group, persons in the top 15 percent of the cost distribution were classified as high cost. This group accounted for 75 percent of all charges.

A dichotomous variable indicating membership in the high cost group was defined for each respondent. It was decided to fit logistic regression models to these outcomes instead of linear models for two reasons. First the logistic model is less restrictive in that the assumption of a linear relationship between the outcome variable and the regressor variables is not required. Secondly, the predicted probabilities of being in the high cost group produced from the final model would necessarily lie in the valid range (0,1). Since an important aim of the modelling was to predict the probability of high cost for various subgroups of the population, this second reason was fairly important.

The set of independent variables that were considered for inclusion in the models were suggested by the literature as well as earlier descriptive analyses. The modelling proceeded in three stages. First, the ordinary stepwise logistic regression procedure PROC LOGIST was used for variable selection. Variables entered the model in order of importance as measured by the likelihood R statistic (Harrell, 1975). Only those significant at the 0.15 level remained in the model at each step. Because the variances of the model parameters are usually underestimated by ignoring the effects of the sample design, the use of the ordinary test statistics at this stage was conservative. The model resulting from the stepwise procedure was then refit using the sample weight in PROC LOGIST. The design consistent estimate of the variance covariance matrix and accompanying test statistics produced by RTILOGIT were used to further reduce the model to variables significant at the 0.05 level. A last execution of LOGIST and RTILOGIT resulted in final model parameter estimates and their standard errors for use in predicting probabilities of high cost conditional on $\underset{\sim}{x}$.

Tables 1 and 2 give the final model results respectively for each age group. In addition to the estimates and their standard errors, the level of significance for the test of the hypothesis that the coefficient is equal to zero is provided. Design consistent partial R statistics are also given that can be interpreted as measures of the contribution of each variable to the model (Folsom, 1985).

The two most significant determinants of high cost use in both age groups were measures of health status: restricted activity days and chronic conditions. Their positive coefficients indicate an increase in the odds of being in the high cost group as health status declines. Employment status was an important predictor among the 17-64 year age group, with part-time employment and unemployment being positively related to high cost. Insurance coverage was important in both age groups. As would be expected, persons with some form of coverage had higher odds of being in the high cost group than persons not covered, however the positive coefficient associated with Medicaid coverage may be less a predictor than a result of the "spend down effect," i.e. that high cost use precipitates Medicaid coverage rather than the reverse relationship.

Age, sex, and marital status were important demographic correlates among persons 17-64 years of age, while sex was the only demographic variable included in the final model for those over 65 years of age. Males were at lower odds of being high cost users than females in the younger group and at higher odds in the older group. Income entered both models either through poverty level or annual family income and was positively associated with high cost use. None of the community characteristics appeared to be important predictors of high cost among persons 65 and older. Median gross rent and physician supply were both significant in the 17-64 years of age model. The former indicates that a significant portion of high cost use can be explained by the cost of living. The significance of the physician supply variable is hard to interpret since it does not separate specialists from generalists.

4. Discussion

The availability of logistic regression software packages for survey data has increased the flexibility with which data arising from a complex sample design can be analyzed. Prior to such availability, analysts were limited to design consistent linear models or to logistic models with superpopulation assumptions. The advantages of logistic over linear models include the less restrictive assumption of a possibly non-linear relationship between the response and the regressors and the fact that predicted values will lie in the valid range for probability outcomes. In some applications, such as the fitting of dose-response curves, there are historical precedents for assuming the logistic model. One disadvantage is that the logistic model algorithm is iterative and therefore more costly than the linear model solutions, particularly when the number of observations or the number of independent variables is large. Also, the interpretation of the logistic model parameters is not as straightforward as for linear model parameters. Transforming to the log odds ratio for the outcome variable allows for a more direct interpretation of the β's.

The advantage of producing design consistent estimators and test statistics over their model based counterparts is that less restrictive assumptions are required, resulting in more robust inference. However, the design consistent estimators are not as efficient if the model truly holds.

5. REFERENCES

Binder, D. A. (1981), "On the Variances of Asymptotically Normal Estimators for Complex Surveys," Survey Methodology, Vol. 7, No. 2, 157-170.

Folsom, R. E. (1985), "Sampling Variance Estimator for the Logistic Regression Multiple R^2," Personal Communication.

Harrell Jr., F. E. (1975), "The LOGIST Procedure," SAS Supplemental Library User's Guide, 181-202.

Imrey, P. B., Koch, G. G., Stokes, M. E., and collaborators (1981, 1982), "Categorical Data Analysis: Some Reflections on the Log-Linear Model and Logistic Regression, Parts I and II," International Statistical Review, 49, 263-283, and 50, 35-63.

Koch, G. G. and Edwards, S. (1985), "Logistic Regression," Encyclopedia of Statistical Sciences, Vol. V, John Wiley and Sons, New York.

Rao, J. N. K. and Scott, A. J. (1984), "On Chi-Squred Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data," Annals of Statistics, 12, 46-60.

Scott, A. J. and Wild, C. J. (1985), "Fitting Logistic Models in Case-Control Studies," Presented at the 1985 I.S.I. meetings.

Shah, B. V., Folsom, R. E., Harrell, F. E. and Dillard, C. N. (1984), "Survey Data Analysis Software for Logistic Regression," RTI Project Report.

Shah, B. V., Holt, M. M. and Folsom, R. E. (1977), "Inference about Regression Models from Sample Survey Data," Bull. Inst. Stat. Inst., XLVII (3) 43-57.

ACKNOWLEDGEMENT

The authors wish to thank Gerald Riley of the Health Care Finance Administration for his contributions to the analysis of high cost users presented in this report. The authors also wish to thank Herbert Silverman of the Health Care Finance Administration and Larry Corder of RTI for their helpful comments while these analyses were being conducted.

Table 1. Results of the Reduced-Model (0.05) Logistic Regression for Persons between 17 and 64 Years of Age (Dependent Variable: High Cost Indicator)

Variable	Logistic Regression Coefficient	Standard Error	Level of Significance	Partial Correlation Coefficient (R)
Intercept	-2.6405	0.2412	-	-
No. of Restricted Activity Days	0.0253	0.0011	<0.0001	0.248
No. of Chronic Conditions	0.2175	0.0182	<0.0001	0.135
Male*	-0.3354	0.0762	0.0002	-0.047
Not Insured*	-1.2226	0.1921	<0.0001	-0.071
Employment:**				
Employed Part Time	0.6441	0.0830	<0.0001	0.087
Unemployed, In Labor Force	0.8068	0.1887	0.0015	0.046
Unemployed, Retired	0.4549	0.1806	0.0241	0.024
Unemployed, Other	0.6284	0.1111	<0.0001	0.062
Died During Survey*	1.8679	0.5283	0.0001	0.037
Age	-0.0213	0.0032	<0.0001	-0.075
Never Married*	-0.3999	0.0972	0.0001	-0.044
Fair Perceived Health Status**	0.3321	0.1034	0.0037	0.033
Poverty Level	0.0005	0.0002	0.0034	0.033
Number of MDs & DOs Per 100,000 Population	-0.0029	0.0008	0.0089	-0.036
SMSA Central City**	0.2657	0.0793	0.0132	0.035
1980 Median Gross Rent	0.0026	0.0001	0.0280	0.022

Proportion of Log-Likelihood Explained by Model (R^2): 0.220
Overall Model Level of Significance: <0.0001
*Dichotomous variables for which the reference level is the opposite of the specified level.
**Omitted levels for multi-level categorical variables are: excellent and good perceived health status; SMSA remainder and Non-SMSA urban.
Source: National Medical Care Utilization and Expenditure Survey, 1980

Table 2. Results of the Reduced-Model (0.05) Logistic Regression for Persons 65 Years of Age and Older (Dependent Variable: High Cost Indicator)

Variable	Logistic Regression Coefficient	Standard Error	Level of Significance	Partial Correlation Coefficient (R)
Intercept	-4.9567	0.2924	-	-
No. of Restricted Activity Days	0.0096	0.0011	<0.0001	0.206
No. of Chronic Conditions	0.2723	0.0285	<0.0001	0.234
Died During Survey*	2.3840	0.4146	<0.0001	0.138
Male*	0.5473	0.1542	0.0007	0.080
Functional Limitations	0.1222	0.0369	0.0045	0.074
Private Health Insurance*	0.8245	0.1984	0.0002	0.097
Medicaid Coverage*	0.6932	0.2010	0.0011	0.078
Annualized Family Income - 1980	0.00001	0.000005	0.0123	0.043

Proportion of Log-Likelihood Explained by Model (R^2): 0.250

Overall Model Level of Significance: <0.0001

*Dichotomous variables for which the reference level is the opposite of the specified level.