ALTERNATIVES TO COSTLY MAINFRAME LOGISTIC REGRESSION ADJUSTING FOR COMPLEX SURVEY DESIGN

Barbara Lepidus Carlson, Steven B. Cohen, and Alan C. Monheit, Agency for Health Care Policy and Research
Barbara Lepidus Carlson, 2101 East Jefferson Street, Suite 500, Rockville, MD 20852

Key Words: RTILOGIT, SUDAAN, SURREGR, microcomputer
The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Health Care Policy and Research is intended or should be inferred. The authors would like to acknowledge Dr. James Reschovsky for his methodological assistance.

1. Introduction

In order to correctly interpret regression analyses for data resulting from a national survey with a complex sample design, the standard errors must take into account the effect of such a design and correctly deal with the sampling weights. A complex design refers to a design which deviates from simple random sampling, often incorporating methods such as stratification, clustering, and disproportionate sampling. Most of the commonly used statistical computing packages (SAS, SPSS, BMDP) assume a simple random sample and, although they can provide weighted parameter estimates, the associated standard error estimates they produce are often too small, yielding confidence intervals that are too narrow and anticonservative hypothesis tests (i.e., rejecting the null hypothesis when in fact it should not have been).

Several software packages exist which can correctly take the design effect into consideration. These software packages now exist for mainframes, minicomputers, and microcomputers. Most have the capacity to run weighted least squares and logistic regression analysis. Unfortunately, logistic regression analysis even under simple-random-sampling assumptions can be quite expensive to run on a mainframe computer, given the iterative and computing-intensive nature of its calculations. A package which also adjusts for both sampling weights and the complex nature of the sample design requires additional computing time and cost.

Analysts working on the 1987 National Medical Expenditure Survey (which has a stratified multi-stage area probability design) address critical health care policy issues, involving economic, sociological, and behavioral analyses which are often of a complex multivariate nature. More specifically, many of these analyses focus on dependent variables that are categorical in nature with two or more classifications. The application of appropriate logistic or multinomial logistic regression procedures on mainframe computers, that adjust for survey design complexities, is often characterized by expensive computer runs with charges exceeding $1,000. Worse yet, some computer runs will run out of the allotted CPU time before completing the analysis and still run up high computing costs. As a consequence of the frequency of application of these logistic regression analyses for hypothesis testing and estimation of model parameters, and their associated expense, there is great appeal in considering cost-effective analytical alternatives.

The dilemma of choosing between an appropriate modeling strategy with significant computing costs versus a less costly yet less appropriate approach arises quite often, particularly with respect to weighted logistic regression analysis. The purpose of this paper is to explore alternatives to mainframe logistic regression analysis with an adjustment for the complex survey design, and examine their usefulness as well as their drawbacks and limitations in capacity.

Alternative 1. Running weighted least squares regression on the mainframe, with software adjusting for the complex survey design; i.e., abandoning the logistic model, while still running a mainframe package which correctly addresses the complex nature of the survey.

Alternative 2. Running weighted logistic regression on the microcomputer, with software correctly addressing the complex nature of the design; i.e., abandoning the mainframe environment, while still running a logistic model and adjusting for the complex design.

Alternative 3. Running weighted logistic regression on the microcomputer using a software package that does not initially adjust for the complex design, and then adjusting the standard errors by the square root of the design effect[1].

Several regression models will be evaluated using the original methodology and the three alternatives in an attempt to test the limits of these methods and the conditions under which they are most and least useful. The effects of the number of observations, the number of independent variables in the model, and the nature of the dichotomous dependent variable on the outcomes will be assessed. Parameter estimates and their standard errors will be examined and compared among the four methods. In addition, computing costs (for mainframe runs) and computing times will be compared. Programming statements and sample output are available from the authors.

2. Background

The software packages being used in this analysis are: SAS, RTILOGIT, SURREGR, and SUDAAN, the last three of which are products of the Research Triangle Institute (RTI). SAS release 5.18 on the mainframe and SAS release 6.04 on the microcomputer were used (SAS Institute, 1985, SAS Institute, 1990). Only the PC version of SUDAAN (version 6.10) was used (Research Triangle Institute, 1991), and will be referred to as "PC SUDAAN" in this paper to avoid confusion with its mainframe counterpart. A decision was made not to use the mainframe version of SUDAAN (which incorporates RTILOGIT (Shah et al., 1984) and SURREGR (Holt, 1977), as well as other earlier RTI packages), since past experience indicates that the current version is slower and more expensive to run than its predecessors (Carlson and Cohen, 1991).

In order to run RTILOGIT, the data must first be run through the SAS version 5 supplementary logistic regression procedure, PROC LOGIST (Harrell, 1986), whose estimates are then read into the RTI procedure along with subsampling and weight variables. The mainframe SAS portion of an RTILOGIT run was clearly the more computing intensive of the two parts, and therefore will not be used alone as an alternative to the RTILOGIT. It was decided not to use other widely-used packages, such as those produced by Iowa State University (PC CARP) and Westat, Inc. (WESLOG), since these packages would have required extra effort and/or cost on the part of the analysts; i.e., they either did not accept SAS data or required replicate weights. Others had no PC counterparts.

The three RTI packages use a first-order Taylor Series expansion to approximate the variance. As stated previously, SAS operates under the assumption of a simple random sample, and therefore uses standard variance calculations.

2.1 A Look at the Proposed Alternatives

Alternative 1. Weighted least squares

While a least squares regression model applied to a dichotomous dependent variable approximately predicts its proportion, p, and a logistic regression model predicts ln(p/(1-p)), if the parameters resulting from a logistic model are transformed to predict the probability of an outcome of interest, the two models are comparable. While having a dichotomous dependent variable violates assumptions for least squares regression, it is generally accepted (Greene, 1990, Neter et al., 1983) that a least squares model approximates the results of a logistic model if the dependent variable proportion is not close to zero or one.

Alternative 2. Microcomputers

The use of microcomputers for logistic regression analysis of data resulting from surveys with complex sample designs is becoming increasingly feasible. Even with large datasets, the microcomputer packages designed for complex survey data analysis have been previously found to be useful, as long as the regression models are not too large (in terms of the number of independent variables) (Carlson and Cohen, 1991). Once the PC packages are equipped to make use of expanded, extended, or virtual memory, this may in fact be the best alternative to the expensive mainframe runs. When the packages do successfully run, the processing time tends not to be unduly lengthy.

Alternative 3. Post-analysis design effect adjustment

When using data resulting from a survey with a complex sample design to estimate population parameters, it is not advisable to analyze the data without the sampling weights, even when the sampling strata variables are included in the model (Skinner et al., 1989, DuMouchel and Duncan, 1983). There would still be complexities of the design unaccounted for in that approach. Use of the sampling weight will yield approximately unbiased parameter estimates of those found in the population, and correcting for the different stages of sampling will make the estimated standard errors of the parameter estimates more accurate. However, a less expensive alternative than using the specialized software, which adjusts for the complex design, is to run weighted logistic regression analyses under simple random sampling assumptions (in SAS, for example), and then to correct the variances by an average design effect for a set of related statistics (Cox and Cohen, 1985).

3. Methods

3.1 Procedures

After the models of interest were decided upon, two analysis files were created from several source SAS data files, keeping only the variables of interest, one for each of the two subsamples to be used in this evaluation. The data were sorted by stratum and primary sampling unit (PSU), the first two subsampling levels, as required by the RTI software. It is also required that there be at least two PSUs per stratum. Since this was already the case for our two subsamples, no collapsing of strata was necessary.

The dependent variables were coded as 0,1 variables. Any nominal independent variables with more than two categories were turned into dummy variables, omitting a reference category from the model. The SAS data files and SAS programs were downloaded from the mainframe to the microcomputer using Procomm Plus terminal emulation software with the Kermit protocol and then read into the PC version of SAS. Prior to downloading, the SAS version 5 data files were converted to SAS transport files and then downloaded via binary protocol. For the two variables in the models with missing values, any observations with a missing value were imputed with the modal value of that variable. Only one data file needed to be downloaded from the mainframe to the PC, since the smaller of the two analysis files is a subset of the larger one. Downloading the data file took 5.25 hours and cost about thirteen dollars during discount hours. However, both time and cost are highly variable between different hardware and software configurations.

The original method plus the three alternative methods will be used on each of eight different models. Two subpopulations of different sizes, persons less than age 65 (n=28,726) and persons between ages 45 and 64 (n=5,958), will be used to evaluate the effect of sample size on the packages' efficiency and capacity. Two different sets of independent variables were chosen such that one model has a relatively large number of variables (17) and one model has a comparatively small number (10). Two different dependent variables ("did person have any hospitalizations in 1987" and "did person have any dental visits in 1987") were chosen such that the proportions of the population with the 1-value (indicating utilization) are approximately .1 and .4. This yields the eight different combinations. All regression analyses were run with an intercept in the model.

The standard method to which the three alternatives will be compared is the use of RTILOGIT software. This is a mainframe computing package that runs weighted logistic regression analysis and correctly adjusts for the complex nature of the sample design. A two-step process is necessary in order to obtain the results from RTILOGIT. The first step is to use the SAS supplementary logistic regression procedure, PROC LOGIST, with the option to normalize the weights (i.e., adjust them so that they sum to the unweighted number of observations) and to save the results in an output file. The next step is to run RTILOGIT using this output file, specifying the same model along with the weight and subsampling variables.

The first alternative method is to run weighted least squares regression with these dichtomous dependent variables, using a package which adjusts for the complex survey design: SURREGR. The specification of the SURREGR procedure is straightforward, primarily a model statement followed by the specification of the weight and the first two subsampling levels: stratum and primary sampling unit. While using this method with a dichotomous dependent variable clearly violates assumptions necessary for least squares regression, this method does have potential utility for exploratory model-fitting and determining correlates for imputation classification variables. Such imputation techniques involve regression analyses to determine variables which correlate with the variable of interest as well as those related to nonresponse for that variable, then using significant factors common to both models to form classifications within which donor records can be used to impute values to recipients with missing values.

In order to compare these parameter estimates to those resulting from logistic models, a transformation was made to the logistic parameter estimates from RTILOGIT, PC SAS, and PC SUDAAN. The logistic parameter estimates were transformed to marginal probabilities (derivatives of the probabilities), or change effects, as follows. The transformations were done for each of the eight models separately and were done differently for continuous, dichotomous, and dummy variables. The method for continuous variables involved computing the instantaneous rate of change, and was based on the formula for the

proportion $p = 1/(1+e^{-(\beta 0+\beta 1x1+\ldots+\beta kxk)})$. To transform βi, each observation was run through the model obtained from the weighted logistic regression analysis, yielding a value p_i. The weighted mean value of p_i across all observations, $\bar{p}_i$, was obtained for each model. The adjustment of the logistic parameter estimates βi to marginal probabilities was accomplished by multiplying them by $\bar{p}_i(1-\bar{p}_i)$ (Maddala, 1985). For the two models on hospitalizations for the middle-aged population, this factor was .088; for the non-elderly population, .080. For the two models on dental visits for the middle-aged population, this factor was .247; for the non-elderly population, .245.

For discrete values, a rate of change (which we will take the liberty of calling a "marginal probability") was computed as follows. For each dichotomous independent variable, two new variables were created, which resulted from running each observation through the obtained logistic regression model, but setting one variable at a time equal to zero for everyone, then setting the same variable equal to one. A weighted mean difference of the predicted values was obtained, which became the marginal probability. A similar method was used for dummy variables, but setting values for each set of related dummy variables simultaneously, and subtracting by the mean predicted value for the omitted reference category.

The second alternative is to run SUDAAN's PROC LOGISTIC on the microcomputer. Unlike RTILOGIT on the mainframe, PC SUDAAN can run weighted logistic regression analysis in only one step. Once again, the specification is straightforward, with a model statement followed by the weight and nest (subsampling levels) statements. It should be noted that this new version of PC SUDAAN does make use of extended memory, unlike versions prior to version 6.

The third alternative is to run SAS weighted logistic regression on the microcomputer, PROC LOGISTIC, and then adjust the standard errors (i.e., multiply them by the square root of the design effect). When running the weighted logistic regressions, the weights must first be normalized via PROC and DATA steps, since the weight-normalizing option available in version 5 SAS is not available in version 6 (the only version available for the PC environment). One must first find the unweighted (n) and weighted (N) number of observations included in the model (i.e., with no missing values for any of the regression variables). A new normalized weight is calculated as the original weight times n/N. In addition, the signs of the parameter estimates for SAS have to be reversed, since version 6 of SAS presents relationships with respect to the lowest value of the dependent variable, which is the 0 value here, not the 1 value.

Several strategies for design effect adjustments were explored in order to find the optimal approximations of the true design effects obtained from the RTILOGIT output. It was decided that the design effect be obtained by running weighted least squares regression models with and without adjusting for the complex design, which are less expensive and computing-intensive than the logistic regression runs. This was done here using the same dependent variables: any hospitalizations and any dental visits. Each parameter estimate will have a variance resulting from each of the two methods. A design effect for each variable from each model is computed by creating a ratio with the numerator being the variance resulting from the PROC SURREGR or (PC) SUDAAN's PROC REGRESS and the denominator being the squared standard error resulting from a weighted PROC REG in SAS. For this paper, the regression analyses needed for design effect calculations were done on the mainframe computer, costing between $1.16 and $7.68 for each model (during discounted hours); however, these regression runs could have been done on the PC instead in an estimated two to fourteen minutes per model.

The median design effect across all models was 1.253. The lowest value was 0.599, indicating a smaller variance when the complexities were accounted for, and the highest value was 4.306, indicating that ignoring the design effect would have clearly yielded anticonservative results.

Once computed for each independent variable for each corresponding model, the square root of the design effect was multiplied by the standard error resulting from the SAS weighted logistic models to yield the standard errors found on Table 4. On Table 5, the χ^2 statistics were likewise adjusted by dividing the Wald χ^2 statistic by the design effect.

The four different methods will be evaluated with respect to effiency and accuracy. Efficiency is measured in terms of computing time and computing cost, reflected in Tables 1 and 2. The three alternatives will be compared to the standard RTILOGIT run. Computing time and cost for the mainframe automatically appear on the printout. These are measured in terms of CPU seconds and total dollars. The mainframe runs were carried out during discounted hours (evenings, weekends), which yields a sixty-percent discount over regular hours. These discounted dollars are presented here. The computing costs at this facility are a function of CPU time, I/O count, number of tape drives used, and region used, and are presented here to give a sense of the magnitude of the cost differences as well as rough dollar estimates for other similar IBM mainframe systems. The microcomputer executions have no costs other than initial software and hardware purchasing costs, although one could factor in the costs of downloading the data (thirteen dollars in this instance, but highly variable between methods and systems).

Microcomputer execution time for PC SUDAAN is measured in terms of the (rounded) number of minutes from the submission of the job until completion. SAS on the personal computer automatically records CPU time. Although PC execution time is measured in terms of minutes and mainframe execution time in terms of seconds, a direct comparison between the two modes is inappropriate. One must keep in mind that elapsed time may in fact be faster from the submission of the job to the receipt of the printout on the PC, unless one has a mainframe computer with print facilities on site. CPU time on both mainframes and PCs is system-specific, and can be altered on a given PC by changing hard disk and caching specifications. As with computing costs, the times are presented to give a sense of the magnitude of the differences between the packages.

The primary outcome of interest, however, is accuracy. For each of the eight models and each of the four methods, the estimated regression parameters and their standard errors will be compared, in addition to their significance in the model. Significance of each parameter estimate in the model is determined by an F statistic for RTILOGIT, PC SUDAAN, and SURREGR, and by a Wald χ^2 statistic (subsequently adjusted by the design effect) for the PC SAS. The F statistic for PC SUDAAN is Satterthwaite-adjusted.

3.2 Computing Environment

The mainframe computer used is an IBM 3090 Model 300J located at the National Institutes of Health in Bethesda, Maryland. It runs under the OS/MVS/ESA operating system.

The microcomputer used is an AST brand IBM-compatible personal computer (Bravo 486/25), with an 80486 microprocessor and 4 mb RAM. It has a 200 mb hard drive (configured as one drive under the MS-DOS version 5.0 operating system) running at 25 mHz and an Intel 80387-compatible math co-processor built into the microprocessor.

3.3 The Survey

The Household component of the 1987 National Medical Expenditure Survey[2] is a national probability sample of the civilian, noninstitutionalized U.S. population. The household survey component was designed to provide statistically unbiased national estimates of health care utilization, expenditures, and access to care, and health insurance coverage for calendar year 1987. To provide focused estimates of subpopulations of particular policy concern, the Household Survey oversampled the elderly, those with difficulties in performing activities of daily living, poor and low-income families, and the black and Hispanic minorities (Edwards and Berlin, 1989).

The Household Survey (HHS) sample design can be characterized as a stratified multi-stage area probability design with three stages of sample selection: (1) selection of PSUs (counties or groups of contiguous counties) (2) selection of area segments within PSUs; (3) selection and screening of dwelling units within segments (Cohen et al. 1991). The total round one HHS sample ultimately comprised 36,400 individuals in roughly 15,000 households.

3.4 The Data

The independent variables chosen for the models were primarily sociodemographic in nature. For the smaller model, age, sex, race, census region, marital status, and self-perceived health status were used to predict any hospitalizations or any dental visits. The larger model contained these variables as well as those pertaining to SMSA status (urbanicity), insurance coverage, the presence of functional limitations, family income, and education.

As stated previously, the two analysis files contained 5,958 and 28,726 observations. On the PC, that translated into 1.5 mb and 7.2 mb, respectively, in SAS format. While the models contained either 10 or 17 independent variables, the files themselves contained almost 50 variables, including the sampling weight variable, and variables describing the sampling strata and primary sampling units, as well as the variables used in creating those used in the final models.

4. Results

4.1. Efficiency

4.1.1. Time

As can be seen in Table 1, using the specialized software for weighted least squares regression (SURREGR) took much less CPU time than using the specialized software for weighted logistic regression (RTILOGIT). The RTILOGIT runs took eleven to sixteen times longer to run than the SURREGR. In addition, the RTILOGIT runs for the larger models with the larger number of observations could not complete computations in the allotted (default) CPU time of 100 seconds.

When using the PC, the time measured is in terms of minutes, as opposed to the mainframe's measurement in terms of seconds. Here we can see that the actual execution time of the regression models was twice as fast for the PC SAS compared to PC SUDAAN. The earlier version of PC SUDAAN (version 5.53) could not make use of extended memory, and could not analyze the largest four of the eight models due to software memory constraints.

One must keep in mind, however, the extra execution time needed to run the two extra regression analyses per model necessary for computing the design effects used in adjusting the standard errors from the PC SAS logistic regressions. Other than actual execution time, there is also extra effort involved in using the PC SAS alternative: one must calculate a normalized weight and adjust the original weight accordingly, reverse the signs of the parameter estimates, calculate the design effect ratios for each variable, and adjust the standard errors and/or test statistics by the design effect.

For all of the packages, the CPU time increased with sample size and increased with model size. The larger model (17 variables) with the smaller number of observations (5,958) usually ran in less than half the time of the smaller model (10 variables) with the larger number of observations (28,726). There was no discernible and consistent difference in execution time between the two different dependent variables.

Only two of the three logistic regression packages being evaluated show the number of iterations needed for the logistic model to converge. PC SAS does not have this as part of the output or log. For the six models that ran in both RTILOGIT and PC SUDAAN, the models converged in one or two less iterations in PC SUDAAN, using default specifications. The hospitalization models converged in five iterations in PC SUDAAN, while the dental visit models converged in three or four iterations.

4.1.2. Cost

Table 2 is presented to give a sense of the magnitude of the costs of estimating these regression models on the two mainframe packages. These are roughly proportional to CPU time. One could estimate that the cost for the two RTILOGIT runs which ran out of allotted CPU time would have been about $115 at the discount rate ($288 at the normal rate). Again, the costs were between ten and sixteen times higher for the RTILOGIT compared with the SURREGR. The PC packages have no costs per execution, but may require some one-time initial costs of software and hardware purchases.

When one adjusts for the size of the model and the number of observations, it is apparent that the incremental cost per observation decreases in SURREGR as the number of observations increase, while the incremental cost in RTILOGIT remains fairly constant. With respect to the number of variables in the model, for both mainframe packages, the incremental cost per variable increases with every variable added to the model, particularly for RTILOGIT.

4.2. Accuracy

4.2.1 Parameter Estimates

When comparing the marginal probabilities resulting from the weighted logistic packages to the regression coefficients from the weighted least squares package (Table 3), one might have the initial impression that the two strategies yield quite similar point estimates of the regression coefficients. Except for values hovering around zero, the coefficients have the same directionality. Note that the three weighted logistic packages, RTILOGIT, PC SUDAAN, and PC SAS, yielded exactly the same parameter estimates as each other, as expected.

For the most part, there does not appear to be a consistent pattern with respect to the size of the effects for the logistic models compared to the least squares model, when one examines the number of observations or number of variables in the model. If one examines the smaller model for dental visits with the middle-aged (smaller) population, the relative absolute differences in the effects are, on

average, less than two percent when comparing the SURREGR model to the three logistic models. The smaller model for hospitalizations, with the middle-aged population, had discrepancies more on the order of seventeen percent. There does appear to be a closer correspondence between the two methods for the dependent variable with the proportion around .4 versus the variable with the proportion around .1, as theory would suggest. For models comparable with respect to file size and model size, the dental models had on average anywhere from one-tenth to two-thirds the magnitude of the discrepancy found in the hospitalization models.

4.2.2 Standard Errors

In Table 4 we compare the standard errors among the logistic packages. The SURREGR standard errors are not examined here since they are not comparable to those obtained in the logistic models. Here we look at how well the design effect adjustment to the PC SAS standard errors provided estimates similar to those obtained from the RTILOGIT and PC SUDAAN software. The PC SAS standard errors were multiplied by the square root of the design effect on a variable-by-variable, model-by-model basis. Note that RTILOGIT and PC SUDAAN yielded the exact same estimates of the standard errors, as well as test statistics, for the parameter estimates, since they are derived by the same formulae.

In examining model size and file size, there appears to be no pattern whereby one of the methods yields larger standard errors than the other. The average relative absolute differences ranged from 1.3%, for the dental visit model with the smaller number of variables and larger number of observations, to almost ten percent, for the hospitalization model with the smaller number of variables and smaller number of observations. However, as was found with the parameter estimates, the dental visit models yielded smaller disprepancies (in absolute value) between the adjusted PC SAS and the RTI packages than did the hospitalization models. The average relative absolute difference for dental models ranged between eighteen and thirty percent the size of the comparable hospitalization models.

4.2.3 Significance

The significance of each of the parameter estimates in each model was evaluated (Table 5) at $\alpha = .01$ (**) and $\alpha = .05$ (*). It should be reiterated that the standards to which the other packages and strategies should be compared are RTILOGIT and PC SUDAAN, which yield the optimal parameter estimates and associated estimated standard errors. There were only three disparities found. For the larger dental visit model with the larger population, having any ADL (functional) limitations was not significant at the $\alpha = .05$ level, but was significant at this level using the adjusted SAS method (not shown).

In the larger dental visit model for the smaller population, having public insurance was significant at the $\alpha = .05$, but not $\alpha = .01$, level; however, the two alternate methods were significant at the $\alpha = .01$ level. In the larger hospitalization model with the smaller population, have public insurance was significant at the $\alpha = .01$ level; however, the two alternate methods were more conservative, being significant only at the $\alpha = .05$ level.

5. Discussion

Using a weighted least squares regression package which correctly adjusts for design complexities yielded similar effects to those found in the marginal probabilities from the weighted logistic models. Differences were found when the dependent variable had a proportion close to .5 versus close to .1. The differences between the two methods, although not glaring, were not negligible in some cases. While thirty-five of the 108 parameter estimates had no difference out to three decimal places, four parameter estimates were off by as much as 50 to 60 percent and four were off by between 33 and 42 percents.

When moving to the PC environment, there are three possible disadvantages. First, one must download the data. Depending on hardware and software facilities for data transfers, downloading may not be a trivial issue. Second, the analysis runs can take a long time when done on a PC. The authors found, however, that the runs for these models were not excessively long. Third, one must be sure to have a well-suited hardware environment, including a large hard disk, sufficient RAM, and a mathematical co-processor.

Using the SAS with design effect adjustment approach is clearly an option to consider, but required extra computing resources as well as additional effort on the part of the researcher. As stated previously, one must run two extra regression models per model in order to obtain the variances which go into the design effect ratio. In SAS version 6 on the PC or mainframe, one must manually determine the factor by which the weight is normalized and then adjust the weight accordingly (although this is not necessary in SAS version 5 on the mainframe, which has a NORMWT option). In version 6, with a 0,1 dependent variable, one must remember to reverse the signs of the parameter estimates, or recode the 0's to 2's, in order to interpret the relationships in the expected manner. The calculation of the design effects and the adjustment of the standard errors and test statistics (and corresponding p-values) resulting from the SAS weighted logistic regression is straightforward, but can be time-consuming and tedious.

Even so, the results will only be an approximation to the results obtained from the appropriate software. The computed design effects, based on similar but unidentical models, ranged from 0.599 to 4.306, but were generally around the median value of 1.253. In many cases, while the unadjusted standard errors would have been much smaller than when correctly computed, the adjusted standard errors did not always come out optimally, sometimes too small, sometimes too large.

In actual practice, if one is using a least-squares approach with a dichotomous dependent variable, one might want to consider two more adjustments. One would be an adjustment for heteroscedasticity, which could be done by dividing the sampling weight by $\sqrt{(\hat{p}(1-\hat{p}))}$ and using this as the new weight. In addition, values predicted by the least squares regression equation may fall out of the 0,1 range. In those cases, it might make sense to force values less than 0 to be 0, and likewise force values greater than 1 to have the value 1.

6. Summary

In the absence of a sufficiently powerful microcomputer, with increased memory, adequate disk storage space, and a mathematical co-processor, it seems reasonable to use weighted least squares regression, while accounting for the complex design, on either the PC or mainframe for preliminary analysis and model exploration. Using a design effect adjustment with a simple-random-sample based logistic regression analysis is also an option, but requires quite a bit of extra effort; however, in the absence of the specialized survey data analysis software, this is worth pursuing. For final parameter estimates and associated

standard errors when the dependent variable is dichotomous, it is advisable that either the extra expense in running weighted logistic regression while correctly accounting for the complex sample design be incurred, or that appropriate microcomputing software be purchased so that the process can be completed on the PC.

7. References

Carlson BL and Cohen SB (1991). Evaluation of the Efficiency of Using Personal Computers for Regression Analysis on Complex Survey Data. American Statistical Association, 1991 Proceedings of the Section on Survey Research Methods.

Cohen S, DiGaetano R, and Waksberg J (1991). Sample Design of the 1987 Household Survey (AHCPR Publication No. 91-0037). National Medical Expenditure Survey Methods 3, Agency for Health Care Policy and Research. Rockville, MD: Public Health Service.

Cox BG and Cohen SB (1985). Methodological Issues for Health Care Surveys, New York: Marcel Dekker, Inc. (p. 358-9)

DuMouchel WH and Duncan GJ (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. Journal of the American Statistical Association, 78:383.

Edwards W and Berlin M (1989). Questionnaires and Data Collection Methods for the Household Survey and Survey of American Indians and Alaska Natives (DHHS Publication No. (PHS) 89-3450). National Medical Expenditure Survey Methods 2, National Center for Health Services Research and Health Care Technology Assessment. Rockville, MD: Public Health Service.

Greene WH (1990). Econometric Analysis. New York: Macmillan (p. 666)

Harrell FE (1986). The LOGIST Procedure. In SUGI Supplemental Library User's Guide, Version 5 Edition. Cary, NC: SAS Institute, Inc.

Holt MM (1977). SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Research Triangle Park, NC: Research Triangle Institute (revised April 1982 by B.V. Shah).

Maddala GS (1985). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press. (p. 23)

Neter J, Wasserman W, and Kutner MH (1983). Applied Linear Regression Models. Homewood, IL: Richard D. Irwin, Inc. (pp. 356-67)

Research Triangle Institute (1991). Software for Survey Data Analysis (SUDAAN) Version 5.30. Research Triangle Park, NC: Research Triangle Institute.

SAS Institute, Inc. (1985). SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: SAS Institute, Inc. (Release 5.18).

SAS Institute, Inc. (1990). SAS Technical Report P-200, SAS/STAT Software: CALIS and LOGISTIC Procedures, Release 6.04, Cary, NC: SAS Institute, Inc.

Shah BV, Folson RE, Harrell FE, and Dillard CN (1984). Survey Data Analysis Software for Logistic Regression. Research Triangle Park, NC: Research Triangle Institute.

Skinner CJ, Holt D, and Smith TMF (1989). Analysis of Complex Surveys. Chichester: John Wiley and Sons. (pp. 76-9)

7. Tables (Complete tables are available from the authors.)

Tables 1 and 2. Execution Time and Mainframe Computing Costs

			Execution Time				Mainframe Computing Costs		
			RTILOGIT (CPU secs)	SURREGR (CPU secs)	PC SUDAAN (minutes)	PC SAS (minutes)	RTILOGIT (dollars*)	SURREGR (dollars*)	RTILOGIT/ SURREGR
Ages 45-64 (n=5,958)	Smaller model (10 indep. vars.)	Any hospitalizatns. (p=.1)	19.78	1.62	4	1.60	11.05	0.91	12.14
		Any dental visits (p=.4)	17.01	1.61	3	1.40	9.50	0.91	10.44
	Larger model (17 indep. vars.)	Any hospitalizatns.	45.79	2.83	5	2.22	25.55	1.59	16.07
		Any dental visits	46.26	2.82	4	2.23	25.81	1.58	16.34
Ages 0-64 (n=28,726)	Smaller model	Any hospitalizatns.	94.73	6.91	18	8.20	52.86	3.76	13.66
		Any dental visits	80.40	6.89	14	7.17	44.87	3.86	11.62
	Larger model	Any hospitalizatns.	**	11.89	25	11.20	**	6.65	ø
		Any dental visits	**	11.89	22	11.42	**	6.65	ø

*Discounted 60% (evening, weekend rates) **Abended at 100.71 CPU seconds, $56.18

Tables 3, 4, and 5 (partial: Ages 45-64, Larger model). Parameter Estimates, Standard Errors, and Significance

	Parameter Estimates and Significance						Standard Errors			
	Any Hospitalizations			Any Dental Visits			Any Hospitalizations		Any Dental Visits	
	RTILOGIT PC SUDAAN	PC SAS	SURREGR	RTILOGIT PC SUDAAN	PC SAS	SURREGR	RTILOGIT PC SUDAAN	PC SAS (adj.)	RTILOGIT PC SUDAAN	PC SAS (adj.)
Age	.001	.001	.001	.002	.002	.002	.009	.009	.005	.005
Sex	.000	.000	-.002	.070**	.070**	.067**	.102	.100	.061	.061
Hispanic	-.015	-.015	-.015	.029	.029	.034	.174	.172	.156	.150
Black, non-Hisp.	.019	.019	.020	-.110**	-.110**	-.104**	.128	.141	.090	.088
Midwest	.009	.009	.009	-.013	-.013	-.018	.148	.151	.086	.087
South	.011	.011	.013	-.040*	-.040*	-.039*	.142	.140	.089	.090
West	-.005	-.005	-.002	-.027	-.027	-.029	.180	.169	.092	.093
Health Status	.064**	.064**	.067**	-.056**	-.056**	-.049**	.065	.075	.045	.044
Prev. Married	.005	.005	.006	-.008	-.008	-.009	.130	.134	.083	.083
Never Married	.025	.025	.025	.088**	.088**	.089**	.251	.280	.146	.144
Moderate MSA	-.005	-.005	-.005	.015	.015	.016	.127	.124	.076	.077
Non-MSA	.011	.011	.012	-.012	-.012	-.013	.134	.136	.102	.103
Priv. Insurance	.053**	.053**	.060**	.169**	.169**	.150**	.171	.156	.122	.110
Public Insurance	.031**	.031*	.044*	.085*	.085**	.076**	.200	.233	.161	.142
Any ADL Limit'ns	.134**	.134**	.206**	.086*	.086*	.084*	.215	.327	.176	.168
Family Income	-.000	-.000	-.000	.000**	.000**	.000**	.000	.000	.000	.000
Education	-.001	-.001	-.001	.037**	.037**	.038**	.018	.018	.013	.013
Mean relative absolute difference	.141			.084			.079		.024	

*p<.05, **p<.01 Significance based on F statistics, except for PC SAS which was based on χ^2 statistics.

Note: Logistic coefficients were transformed to marginal probabilities in order to compare with least-squares coefficients. Values shown as .000 are not exactly equal to zero.

1. A design effect is the ratio of the variance that one obtains when correctly adjusting for the complex design over the variance that one obtains under simple random sampling assumptions. It is a measure of the effect of the complex design on the variance of a survey estimate.

2.Sponsored by the Agency for Health Care Policy and Research, formerly the National Center for Health Services Research and Health Care Technology Assessment.