Statistical Methods in Medical Research 1996; **5**: 311–329

Innovative strategies using SUDAAN for analysis of health surveys with complex samples

Lisa M LaVange Quintiles, Inc., Research Triangle Park, North Carolina, USA, **Sally C Stearns** Department of Health Policy and Administration, University of North Carolina, Chapel Hill, North Carolina, USA, **Jennifer E Lafata** Center for Clinical Effectiveness, Henry Ford Health System, Detroit, Michigan, USA, **Gary G Koch** Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA and **Babubhai V Shah** Research Triangle Institute, Research Triangle Park, North Carolina, USA

Large-scale health surveys provide a wealth of information for addressing problems in health sciences research. Designed for multiple purposes, these surveys frequently have large sample sizes and extensive measurements of demographic and socioeconomic characteristics, risk factors, disease outcomes and health care service use and costs. Complex features of the sampling design typically employed to select the survey sample, coupled with the vast amount of information available from the survey database, underlie issues that must be addressed during data processing and analysis. Numerous articles in the literature have focused on the debate of whether or not, and how, to control for features of the sample design during data analysis. Traditional statistical methods for simple random samples and the software that accompanies them have historically not had the capacity to account for the survey design. Recent advancements in statistical methodology for survey data analysis have greatly expanded the analytical tools available to the survey analyst. Commercial software packages that incorporate these methods offer the analyst convenient ways for applying such tools to large survey databases in an easy and efficient manner. We present an overview of analysis strategies for survey data and illustrate their application via the SUDAAN software system. Examples for analyses are provided through data from two large US health surveys, the National Health Interview Survey and the Longitudinal Study of Aging. Questions of both a cross-sectional and longitudinal nature are addressed. The examples involve logistic regression, time-to-event analysis, and repeated measures analysis.

1 Introduction

Large-scale health surveys constitute a potentially valuable resource for the field of health sciences research. Data from health surveys can be used to investigate relationships between risk factors and disease, track health care costs and utilization among subgroups of the population, shed light on community intervention studies and address numerous other research questions of interest. Because of the costs associated with identifying, locating and collecting data from a representative sample of subjects, surveys are frequently designed to accomplish multiple purposes simultaneously, and are consequently characterized by large sample sizes with extensive measurements taken on each subject. Data items may include demographic and socioeconomic

Address for correspondence: Lisa M LaVange, Quintiles, Inc., PO Box 13979, Research Triangle Park, NC 27709-3979, USA.

0962-2802(96)116RA

characteristics, risk factors, disease outcomes, behavioural outcomes, knowledge and attitudes about health risks, and health care events and costs.

As government resources face increasingly competing demands, emphasis is placed on maximizing the use of existing survey databases for research questions in lieu of launching new data collection efforts. Consequently, efficient and easily implemented methods for complex survey analysis are in demand. As a result of advances both in statistical methodology and software development for complex survey analysis, the choices for analysis strategies available to the survey researcher have been greatly expanded in the past few years. We review some of these developments and demonstrate their application with a commercially available and widely used software system written explicitly for survey data analysis, SUDAAN.[1] Illustrative analyses for data from two US health surveys are presented for the application of logistic regression, time-to-event analysis, and repeated measures analysis of survey data.

Health surveys conducted in the USA and used as the source for varied research endeavours include the second National Health and Nutrition Examination Survey (NHANES II),[2] the National Health Interview Survey (NHIS),[3] and the Longitudinal Study of Aging (LSOA).[4] The NHANES II is the second in a series of surveys conducted by the National Center for Health Statistics (NCHS) that provides a vast array of data on nutritional intake and biomedical assessments. Data from the NHANES I and II have been the basis of research dealing with important public health concerns such as the relationship between blood lead and blood pressure in adults,[5] the relationship between dietary fat and breast cancer,[6] and the relationship between body iron stores and cancer.[7,8] Korn and Graubard[9] provide a comprehensive review of over 50 epidemiological studies based on the three NHANES and the Hispanic Health and Nutrition Examination Survey (HHANES).[10]

The NHIS is another NCHS-sponsored survey conducted yearly for the general US population; it provides data on personal, sociodemographic, and health characteristics and health services utilization. Data from the 1989 NHIS form the basis for our first illustrative analysis with the scenario being activity limitations in the 65 years and over population. The LSOA, also sponsored by NCHS, is a follow-up survey of the population aged 70 years and over, as represented in the 1984 NHIS. Data from the LSOA conducted in 1986, 1988 and 1990 form the basis for two illustrative analyses presented here with the scenario being predictors of hospitalization and death among the elderly.

2 Background and motivation

Large-scale surveys are characterized by the use of complex probability sampling for the selection of subjects that are representative of the target population. The sample design may employ stratification, multiple stages of selection, clustering, and/or unequal probabilities of selection.[11,12] Each of these features requires consideration during data analysis when the population from which the subjects are sampled is the primary target for inference. Traditional methods of statistical analysis may not apply in that the assumption of independent and identically distributed sampling units

drawn from an infinite population is implicit in their use. These methods form the basis of most widely used statistical software packages. When applied to survey data, misleading results can occur.

Aspects of the sample design also impact on the data structures associated with survey databases. Data items identifying the sample design, such as stratification factors and sampling weights, provide information necessary for the accurate computation of point estimates and variances and should therefore be included on the database for each sample unit. If multiple stages of sample selection are employed, sample unit identification and substratification variables corresponding to each stage of selection may also be required. Descriptive and inferential variables may be defined at different levels of sample selection as well. For example, a common area household survey design involves selection of counties or groups of contiguous counties as the primary sampling units, blocks of households as the second stage units, and households or persons as the third stage units. For such a design, variables measured at the county, neighbourhood, household and/or person level may all be part of the database. For panel or repeat surveys, such as the LSOA, multiple assessments for each subject may add another level of complexity to the analysis file structure. It is easy to see how survey databases quickly develop complex structure. Analysis strategies should anticipate a large number of variables as well as potentially large sample sizes. Software designed to provide friendly implementation with large databases is especially useful in this setting.

Early work in survey data analysis focused on estimating linear and ratio statistics for subdomains of the target population. Examples include means, proportions, and survival rates. The use of Horvitz–Thompson estimators[13] has historically been advocated as a way of eliminating the bias due to unequal probabilities of selection when estimating linear and ratio statistics. These estimators incorporate the sampling weights, typically defined as the inverse of the probabilities of selection with possible adjustment for nonresponse or poststratification, and are unbiased for the corresponding population parameters in large samples. When estimating variances and covariances of linear and ratio statistics, not only weighting, but clustering as well must be taken into account. Because intra-cluster variation usually differs from inter-cluster variation, managing them as similar could lead to understatement or overstatement of the extent of the sampling variability present in the estimates. A direct approach based on Taylor series approximations[14,15] and simulation methods such as balanced repeated replication (BRR)[15–18] have been widely used for variance estimation.

Hypothesis testing in the context of linear models fit to survey data was the focus of important work in the 1970s. Weighted least squares (WLS) methods were developed for estimating parameters of a linear model fit to a set of linear or ratio statistics.[19] Wald statistics were proposed for assessing the goodness of fit of the models. Application of WLS in this setting called for consistent estimates of the variances and covariances of the set of statistics. Taylor series, BRR, or another suitable approach could be employed for this purpose.

Methods for fitting linear regression models directly to observation level data were also developed.[20] Approximately unbiased estimates of the variance covariance matrix of the estimated regression coefficients were developed via application of Taylor series

methods. Hypothesis testing in the context of survey linear regression was the subject of an important paper by Shah *et al.*[21] The authors proposed a transformed *F*-statistic based on the Wald chi-squared test for this purpose and presented simulation results for its performance relative to the National Assessment of Educational Progress (NAEP) sampling design. The 'variance' or 'denominator' degrees of freedom for the *F*-statistic was defined as the number of primary sampling units (PSUs) minus the number of first stage strata.

Software for estimating linear and ratio statistics, linear regression parameters, and their respective variances was developed in conjunction with this early work. The SESUDAAN and SURREGR programs developed at the Research Triangle Institute and the SUPERCARP program developed at Iowa State University were commonly used for this purpose. The SAS CATMOD procedure provided WLS estimation capabilities for survey data in the early 1980s. Numerous applications of the methods and software subsequently appeared in the medical literature.[5,22,23]

A landmark paper in 1983 laid the framework for more complex analyses of survey data.[24] Using estimating equation methodology, Binder developed pseudo-likelihood estimators of nonlinear model parameters and design consistent estimators of the variance covariance matrix for the parameter estimates. These methods are applicable to a variety of analysis strategies including logistic regression. Software for survey logistic regression was soon to follow.[25–27]

Proportional hazards regression for survey data analysis was introduced for discrete time to event models in the 1980s.[28] Binder later extended the estimating equation methodology to allow for continuous time-to-event data.[29] Software for both models was implemented in the SUDAAN system in 1992.

Related issues for survey data analysis are discussed in Skinner *et al.*[30] and Rao and Bellhouse.[31] These texts provide comprehensive reviews of milestones achieved in the field, excellent overviews of the various methods available, and rationales for selecting a particular analysis strategy. In addition, Skinner *et al.*[30] offer an in-depth discussion of design effects and how they may be used to assess the impact of the sample design on analysis.

Parallel to the methodological development in the area has been the publication of numerous articles discussing the advantages and disadvantages of taking the complex sampling design into account during data analysis.[32–37] The focus of these papers is the determination of the appropriate population of interest, be it a superpopulation underlying the finite sampling framework or the survey population. Korn and Graubard[9,38] provide excellent discussions of this issue and offer practical guidelines to the analyst for dealing with situations in which accounting for the sample design may adversely affect the performance of the estimators for a question with scope beyond the survey population.

Concurrent with the application of estimating equation methodology in survey data analysis has been the widespread use of generalized estimating equations (GEE) in the biomedical sciences.[39–41] Although developed under a different premise, i.e. quasi-likelihood theory applied to longitudinal settings, similarities exist between the two approaches. For example, survey logistic regression applied under the assumption of with replacement sampling of PSUs yields similar results to GEE logistic regression of repeated binary measurements when the subject is treated as the PSU. LaVange *et al.*[42]

and Graubard and Korn[43] provide discussion of the similarities between the two methodologies, and Koch and Paquette[44] describe the applicability of both to studies in dental research. The importance of this link lies in the fact that advances in GEE methods for biomedical data may be applicable to survey data with slight modifications. Examples include methods for incorporating correlational models in survey logistic regression and methods for fitting proportional odds regression to survey data.

3 Methods and software

Accounting for features of the complex sampling design during data analysis requires special attention and usually calls for software designed specifically for that purpose. Stratification, when applied appropriately, can reduce the variances of the parameter estimates, while unequal weighting and clustering typically have the opposite effect, i.e. increased variances. Ignoring the complex sampling design through application of standard methods and software will usually result in both variances and probability levels of hypothesis tests that are too small; significance may be cited erroneously.

Many statistical software packages (e.g. SAS, SPSS, BMDP) allow for a weight variable to be specified in some procedures. The resultant weighted estimates are unbiased (or asymptotically unbiased) for the finite population parameters, such as means, proportions, ratios, and linear and logistic regression parameters. However, the associated variances and *P*-values for hypothesis tests about the parameters are incorrect for design effects different from one. Software packages providing the ability to correctly compute variances and hypothesis tests under complex sampling fall into one of three categories according to the methods employed: Taylor series, BRR, and jackknife methods. SUDAAN[1] and PC Carp[45] are the most widely used packages based on Taylor series methods. WesVarPC[46] is a comprehensive package employing BRR, and CPLX[47,48] is available for jackknife variance estimation.

The advantage of software based on Taylor series methods is that minimal design information is required for execution. The disadvantage is that analytical derivations are required for each estimation procedure offered, as the linearization differs from statistic to statistic. In contrast, programs based on BRR and jackknife methods require additional input at the time of execution, namely the definitions of the replicate samples or half-samples, but do not require analytical derivations for each different estimator. All three methods yield variance estimators that are consistent for the population variances in large samples. It is generally thought that BRR and jackknife variance estimates have less bias in small samples than those computed via Taylor series approximations.[15]

WesVarPC is a software package providing replicate variance estimates for complex surveys and is available from Westat, located in Washington, DC, USA. Information about WesVarPC is available via the World Wide Web. Procedures are available for estimating variances of ratios, proportions, odds ratios and other statistics in multiway contingency tables. Linear and logistic regression modelling capabilities are also provided. Adjustments to the sampling weights for nonresponse and poststratification are properly accounted for in the variance approximations. WesVarPC is a flexible Windows-based software system representing enhancements of the earlier user-written

SAS procedures WESVAR, WESREG and WESLOG, developed for mainframe computers in the 1980s.

SUDAAN is a comprehensive software package for survey data analysis available commercially from the Research Triangle Institute in Research Triangle Park, NC, USA. Early versions of the SUDAAN system, first available in the 1970s, consisted of user-written SAS procedures for mainframe computers. In the mid-1980s, the software was completely redesigned and offered as a stand-alone system on a variety of computing platforms.[49] Current versions of the software (Version 6.4 and higher) include SAS callable procedures as well as stand-alone procedures that read SAS or ASCII input files.[50]

Variance estimation in SUDAAN is based on Taylor series linearization methods. Different approaches are required for explicitly and implicitly defined estimators. For estimators that can be explicitly defined as (possibly nonlinear) functions of linear sample statistics (e.g. ratios, totals, means, proportions and linear regression parameters), variances are computed according to a method described by Woodruff.[14] Briefly, a linearized variable is defined for a particular statistic by forming the first-order Taylor series approximation of the deviation of the estimator from its expected value. The variance of the estimator is then computed by substituting the linearized values into the variance formula appropriate for the sample design specified.

For estimators defined as implicit functions of sample statistics, such as those for logistic and proportional hazards regression parameters, a different approach is required. The parameter estimates in this setting are computed as solutions to weighted analogues of likelihood equations and are referred to as pseudo-likelihood estimators. Binder[24,29] proposed design consistent estimates of the variances of the statistics using Taylor series expansions of the estimating equations. This approach is referred to as estimating equation methodology.

Variance formulas are currently available in SUDAAN for the following design options:

- with replacement sampling at the first stage
- without replacement sampling with equal probabilities at the first stage
- without replacement sampling with unequal probabilities at the first stage.

For any of these options, subsequent stages of sampling may be specified, with or without replacement, with equal probabilities of selection. In addition, stratification may be included at any stage of selection. Other options exist for handling of missing sample units and for computation of post-stratified estimates and their variances.

In terms of software design, the linearization, which is dependent on the statistic, is separated from the variance computation, which is dependent on the sample design but applies to any linearization. As new procedures are added, linearizations are developed and programmed, but the variance formula modules remain intact, thus making for an easily expandable software system.

For tests of hypotheses concerning the population parameters of interest, Wald chi-squared test statistics are constructed in SUDAAN based on the design consistent estimators of variance. Alternative test statistics based on Satterthwaite adjustments to the simple random sampling variance–covariance matrix as well as several F approximations are also offered. These alternative tests have been shown to perform better than the Wald chi-squared test in situations where the number of first stage

sampling units is small relative to the degrees of freedom for the hypothesized contrast.[51–53]

Three descriptive procedures are available in SUDAAN: CROSSTAB for cross tabulations of survey variables and tests of association; RATIO for general ratio estimation, including poststratified and standardized estimates; and DESCRIPT for mean, total and quantile estimation. Four modelling procedures are available: REGRESS for linear regression; LOGISTIC for ordinary logistic regression; MULTILOG for multinomial logistic regression and proportional odds regression (Version 7 and higher); and SURVIVAL for continuous and discrete time proportional hazards modelling.

In addition to being widely used for survey data, SUDAAN has seen new interest as a package offering GEE methodology for analysis of repeated measures data structures from experimental studies, particularly clinical trials.[54,44] As a result, options in Version 7 include the capability for specifying equal correlation models in logistic regression and proportional odds regression.

4 Example 1: National Health Interview Survey

Data from the 1989 NHIS form the basis for our first example. The goal of this analysis was to use existing databases to estimate the prevalence of activity limitation among the noninstitutionalized elderly for US states and counties. With the use of county identifiers provided by NCHS for this research, county level variables from the 1990 Area Resource File (ARF)[55] were linked to the NHIS database for use in the modelling.

4.1 Sample design

As mentioned previously, the NHIS is a nation-wide survey sponsored annually by NCHS. It is designed to measure personal, sociodemographic, and health characteristics of the US population. The NHIS employs a stratified, multistage probability sample of civilian noninstitutionalized households.[3] At the first stage of sample selection, 198 PSUs were selected from a list of 1900 counties, small groups of contiguous counties and metropolitan statistical areas (MSAs). The largest of these PSUs were sampled with certainty ($N = 52$), while the remainder were grouped into 73 strata. Within each first stage stratum, two PSUs were selected with probability proportional to population size, yielding a sample of 198 first stage units.

Second stage sampling units corresponded to area segments containing approximately four to eight households each. The 1980 census data were used to define the area segments, with allowances made for new construction. The third stage of sampling consisted of selecting households within the area segments. For those households selected, interviews were conducted with a designated household respondent. The sample size for the subpopulation aged 65 years and over of the 1989 NHIS was approximately 22,000, with 89% responding.

4.2 Application

The example analysis presented here is part of a larger research effort reported elsewhere[56] that undertook small area estimation of the prevalence of activity

limitation among the noninstitutionalized elderly. Although the NHIS sample is designed to produce unbiased estimates at the national- or regional-level, it does not, in general, support the estimation of subregional characteristics. As a solution, synthetic estimation methods were applied to NHIS data linked with ARF data to produce state- and county-level estimates of the proportion of elderly with activity limitation.

4.3 Methods

The first step for applying synthetic estimation methods consisted of fitting survey logistic regression models to activity limitation indicators on the NHIS supplemented with county-level variables from the ARF. The model based predicted probabilities were then extrapolated to calculate estimates of activity limitation for the small areas of interest. Because of the extrapolation process, all explanatory variables included in the logistic regression model fit at the national-level had to be available for the small area population as well. Predictors corresponding to survey respondent characteristics were consequently limited to age, sex, and race. Because of the linkage to the ARF, a number of socioeconomic characteristics of the population residing in the county, as well as measures of the local health care supply, were also available for inclusion in the models.

Survey logistic regression models were fit to individual-level data from the NHIS supplemented with county level variables from the ARF. The dependent variable for the model presented here was a binary variable indicating individuals who were limited in their major activity versus individuals who were not limited in any way. Variable selection procedures were applied to reduce the initial set of predictors. Those remaining in the final model were age (as midpoints of intervals for 65–69, 70–74, 75–79, 80–84, and 85 years and older), age squared, race (Black, Hispanic or Other race versus White), sex, and two county-level covariates: the percentage of the population unemployed and per capita income. Although the latter are available as continuous variables on the ARF, they were collapsed into three categories reflecting low, moderate and high levels to simplify the calculation of the synthetic estimates.

Features of the NHIS sample design included stratification, multiple stages of selection, unequal probabilities of selection, and nonresponse adjustments to the weights. Standard logistic procedures that allow for the specification of weights, such as SAS PROC LOGISTIC, will produce accurate estimates of the logistic regression coefficients and odds ratios. However, to have variance estimation and significance levels accurately account for clustering in sample selection, survey regression methods were required. The SUDAAN LOGISTIC procedure was employed to fit the desired model. The sample design was approximated by a with replacement sampling scheme for nonself-representing PSUs. For self-representing PSUs (i.e. those sampled with certainty), the second stage unit was treated as the PSU and with replacement sampling applied at that stage, treating all area segments as existing in a single stratum. Although SUDAAN allows for full specification of multistage sample designs, surveys such as the NHIS rarely include all of the information required to calculate variances accordingly due to confidentiality, thereby making it necessary to approximate the design in an appropriate fashion. Appendix A contains program statements illustrating this specification of the NHIS design in SUDAAN.

Table 1 Survey logistic regression results for example 1: NHIS

Variable	Beta	Standard error	*P*-value	Design effect*
Age category	0.69	0.09	<0.001	1.13
Age squared	−0.02	0.12	0.271	1.13
Sex	0.25	0.05	<0.001	1.04
Race	−0.62	0.08	<0.001	1.25
Unemployment rate	0.11	0.04	0.039	1.61
Per capita income	−0.21	0.04	<0.001	1.69
Intercept	−2.17	0.25		1.32

*Design effects were computed as the ratio of the standard error computed in SUDAAN LOGISTIC to the standard error computed in SAS LOGISTIC with weights normalized to the sample size.

To assess the effects of the complex sampling scheme on the analysis results, design effects were computed as the ratio of the survey logistic variances to variances obtained via weighted logistic regression in SAS. For the SAS analysis, the weights were scaled to sum to the sample size rather than the population size. The amount that the design effect ratios exceed one is a measure of the inefficiency resulting from the sample design (primarily due to clustering). Note that the design effects presented here differ from those computed in SUDAAN in that the denominator of the latter assumes simple random sampling of individuals from a finite population and is based on Taylor series approximations while the former assumes a binomial model and is based on maximum likelihood methods.

4.4 Results

Table 1 presents results from the survey logistic analysis using SUDAAN. All of the hypothesized predictors except for age squared were significant at the 0.05 level. As the table illustrates, the design effects for all of the estimated parameters, including the intercept, ranged from 1.04 to 1.69, with an average value of 1.31. This indicates that test statistics reported from the simple logistic regression procedure are inflated by approximately 30%. Yet, given the high level of statistical significance under simple random sampling ($P < 0.007$ for all effects using the SAS LOGISTIC procedure), even when this inflation factor is considered, the main effects in the model remain statistically significant. Note that this is true even if one considers the community variables, whose test results are inflated by over 60%.

5 Example 2: Longitudinal Study of Aging

The data for this example come from the Longitudinal Study of Aging (LSOA), which was designed to measure changes in functional status and living arrangement in a cohort of older Americans.[4] Records of persons who provided the necessary identifying information were linked with both the National Death Index (NDI)[57] and Medicare claims files from 1984 through 1991.

5.1 Sample design

The LSOA was based on the Supplement on Aging to the 1984 NHIS, so it is characterized by essentially the same complex survey design features as the NHIS.

The LSOA is comprised of a nationally representative sample of 7527 individuals who were living in the community and were 70 years or older in 1984. The baseline interview was conducted in the person's home; reinterviews in 1986, 1988 and 1990 were conducted by telephone and mail. Family members were interviewed if the sample person was unable to participate in the survey. Due to budgetary restrictions, a representative subsample of only 5151 persons was reinterviewed in 1986. That subsample of 5151 persons is used for the application described here.

5.2 Application

The longitudinal nature of the LSOA enables research questions to be addressed that are not possible with cross-sectional surveys such as the NHANES or NHIS. In particular, analyses of time-to-event data or repeated measures data are possible. For the first application, we conducted a survival analysis of the time until death as a function of factors measured during the 1984 baseline survey of the LSOA. For the second application, we used multiple observations per person (i.e. the reinterviews as well as the 1984 interview) to analyse factors associated with the likelihood of having a hospitalization in the calendar quarter immediately following an interview. The conceptual motivation for both analyses was to assess the importance of having one or more hospitalizations during the year prior to the interview as a predictor of subsequent death or further hospital use.

5.3 Methods

Three different types of variables from the LSOA were used in each model: demographic characteristics, health status measures, and geographic indicators. For both analyses we also used information on provider supply measures (hospital beds, medical doctors, and nursing home (NH) beds) from the ARF;[55] these measures were merged on either by MSA or by aggregated geographic region. Unlike the previous example, county level identifiers were not available for this analysis; hence county level predictors were not used. All explanatory (covariate) variables are dichotomous indicator variables except for the Activity of Daily Living (ADL) score, the Instrumental Activity of Daily Living (IADL) score, and provider supply measures. The ADL and IADL scores were included in both linear and quadratic form to allow for nonmonotonic relationships between the scores and the dependent variables. The first two columns of Table 2 provide a list of the variables and the unweighted mean value of the variables (or proportions for indicator variables) for the 5024 persons used in the analysis. (Data for 127 individuals were not used in the analysis due to missing covariate information.)

As with the previous example, features of the LSOA design include stratification, multiple stages of selection, and unequal weighting. Consequently, survey analysis procedures in SUDAAN were used to fit proportional hazards regression models and repeated logistic regression models to the data. The sample design was approximated by a with replacement design for purposes of variance estimation in SUDAAN. As a result, only the stratum and PSU identifiers and analysis weights were required for the design specification. Although the LSOA sample derives from the 1984 NHIS sample, a slightly different designation for strata and PSUs was used for this example than for example 1. Recall in the earlier example that second stage units within self-

Table 2 Survival analysis results† for example 2: LSOA

Explanatory variable	Proportion or mean[2]	Hazard ratio	Beta	Standard error	Design effect
Demographics					
Age 75–79 (age 70–74 omitted)	0.255	1.33*	0.286	0.0689	1.202
Age 80–84	0.246	1.79*	0.58	0.0651	0.873
Age 85 and up	0.149	2.26*	0.815	0.0752	0.897
Male	0.360	1.82*	0.599	0.0624	1.387
White	0.880	1.21*	0.187	0.0918	0.999
Income $20,000 and up	0.222	0.93	−0.069	0.0647	1.066
Private health insurance	0.690	0.89	−0.117	0.0661	1.373
Medicaid eligible	0.060	0.89	−0.118	0.1179	1.203
Living alone	0.373	0.98	−0.019	0.0569	1.017
Health status measures					
In poor health (self-assessment)	0.124	1.40*	0.336	0.0769	1.148
Hospitalized in prior year	0.218	1.42*	0.349	0.0685	1.485
ADL score (range 0–21)	1.412	1.03	0.031	0.0223	0.935
ADL squared	13.962	1.00	−0.001	0.0013	1.010
IADL score (range 0–18)	2.347	1.16*	0.144	0.0227	1.018
IADL squared	24.397	1.00*	−0.005	0.0014	1.099
Geographic location					
Metropolitan area	0.623	1.24	0.217	0.1133	1.248
Urban area	0.182	1.27*	0.238	0.0966	1.260
North central region	0.257	0.97	−0.035	0.0888	0.932
South region	0.343	0.84*	−0.178	0.0761	0.877
West region	0.177	1.08	0.078	0.0859	1.240
Provider supply measures					
1989 hospital beds/1000 Medicare enrollees	4.668	1.06	0.054	0.0294	1.304
1984 medical doctors/1000 population	1.744	0.91	−0.095	0.0726	1.241
1986 NH beds/1000 Medicare enrollees	56.754	1.00	0.000	0.0030	1.127

*$P < 0.05$.
†Dependent variable: months alive following 1984 baseline interview (mean 64.4 months, with 63.4% of observations censored in 1991; $n = 5024$).
‡All explanatory variables are dichotomous variables except for ADL score, IADL score, and provider supply measures.

representing PSUs were treated as first stage units for purposes of variance calculations. For this example, the self-representing PSUs were paired together to form pseudo-strata, and a with replacement variance formula was applied. Both of these options are valid alternatives in dealing with first stage units sampled with certainty.

5.4 Survival analysis results

The dependent variable for the survival model was the number of months following the baseline interview until death or censoring (due to the end of the observation period). The mean number of months was 64.4, and 63.4% of the sample was still alive at the end of 1991 (the point through which the National Death Index records were available). All explanatory variables were measured in 1984, as SUDAAN Version 6.4 does not allow for time-varying covariates in proportional hazards regression. We fit a Cox proportional hazards continuous time model using the SURVIVAL procedure in SUDAAN. The results from the survival model estimation are in the last four columns in Table 2. The hazard ratio (third column) indicates the hazard of dying during an interval (month) given that the individual was alive at the beginning of the month.

The results show that the hazard of dying increased significantly (at $P < 0.05$) with age and with IADL difficulty. The hazard of dying was also significantly greater for males, whites, persons in poor self-assessed health, persons hospitalized at least once during the year prior to the interview, and persons living in urban areas (relative to rural areas). The hazard of dying was significantly lower for persons living in the south (relative to persons living in the northeast).

The fourth and fifth columns of Table 2 provide the beta-coefficient estimates (used to get the hazard ratios) and the design-based standard errors. The last column provides the design effect, computed here as the ratio of the design consistent variance from SUDAAN to the variance computed from weighted SRS methods. Most of the design effects exceed one (as expected), and some of the design effects are as high as 1.4 or 1.5. For example, having private health insurance in 1984 was not significantly related to the hazard of dying ($P = 0.078$) once the adjustment for the complex survey design was made, but the relationship would have been statistically significant at $P < 0.05$ if the analysis had been done using the survey weights but without the adjustment for cluster sampling in the complex survey design.

5.5 Repeated measures analysis and results

The dependent variable in the repeated measures model equalled one if the person was hospitalized in the quarter immediately following the interview and equal to zero otherwise. The dependent variable was determined using the Medicare claims files, so the 19% of the sample that could not be matched correctly to the claims files (due to missing or invalid identification numbers) had to be excluded from the analysis. A total of 4060 persons completed the 1984 interview and had valid Medicare data. Some of these individuals did not complete the subsequent interviews (either due to death or loss to follow-up). The final analysis file consisted of 13,248 interview observations, with approximately 32% of the sample members having one interview, 28% of the sample members having two interviews, 24% of the sample members having three interviews and 16% of the sample members having four interviews. Slightly over 7% of the interviews were followed by a hospitalization for the respondent during the next quarter. In the repeated measures analysis, the weight for each interview of a sample member was their original weight at baseline.

As shown in the first column in Table 3, the explanatory variables included both fixed (e.g. sex) and time-varying (e.g. hospitalization in the previous year) predictors. Indicator variables for the year of interview were included in addition to the set of predictors considered for the survival analysis. We fit a repeated logistic regression model via SUDAAN LOGISTIC. Note that under the with replacement design specification, repeated measurements are handled as if they corresponded to another stage of clustering in the sample design. The analysis dataset contains multiple observations per respondent for each respondent in the PSU. The correlations present among these repeated measurements per respondent are incorporated into the variance estimation in the same manner as correlations among multiple respondents in the same PSU. An aspect of the model fitting that requires caution is the inherent assumption that the specified model is similarly applicable to the observed data and the missing data for unavailable interviews, i.e. the extent to which the data are missing completely at random is a potential issue.

Table 3 Repeated measures logistic regression results† for example 2: LSOA

Explanatory variable	Odds ratio	Beta	Standard error	Design effect
Intercept	0.03*	−3.462	0.2940	0.845
Time-invariant demographics				
Male	1.27*	0.241	0.0768	1.085
White	1.29	0.252	0.1182	1.424
Demographics or health status measured in 1984 only				
Income $20,000 and up	1.00	0.002	0.0918	1.122
Private Health Insurance	0.98	−0.023	0.0768	0.937
In poor health (self-assessment)	1.24	0.218	0.1125	1.186
Time-varying demographics				
Age 75–79 (age 70–74 omitted)	0.93	−0.076	0.1257	1.356
Age 80–84	1.09	0.083	0.1202	1.224
Age 85 and up	1.14	0.132	0.1237	1.154
Medicaid eligible	1.10	0.093	0.1149	1.221
Living alone	1.04	0.043	0.0890	1.282
Living in a nursing home	1.30	0.265	0.1901	1.241
Time-varying health status measures				
Hospitalized in prior year	1.83*	0.602	0.0730	0.988
ADL score (range 0–21)	1.05	0.050	0.0298	1.313
ADL squared	1.00	−0.003	0.0016	1.351
IADL score (range 0–18)	1.16*	0.152	0.0304	1.412
IADL squared	0.99*	−0.006	0.0016	1.260
Geographic location (time-varying if person moved)				
Metropolitan area	1.02	0.015	0.1139	0.682
Urban area	1.17	0.160	0.1261	1.144
North central region	1.05	0.050	0.1294	1.007
outh region	0.97	−0.031	0.0990	0.768
West region	1.02	0.020	0.1096	0.721
Provider supply measures‡				
1989 hospital beds/1000 Medicare enrollees	1.03	0.026	0.0256	0.514
Medical doctors/1000 population	1.00	−0.001	0.0629	0.529
1986 NH beds/1000 Medicare enrollees	1.00	−0.004	0.0041	0.974
Time indicators (relative to 4060 interviews in 1984)				
1986 (3714 interviews)	0.98	−0.024	0.1003	1.186
1988 (3149 interviews)	1.14	0.132	0.1205	1.526
1990 (2325 interviews)	1.07	0.071	0.1138	1.106

*$P < 0.05$.

†Dependent variable: hospitalization in the quarter following interview (n = 13,248). Percent of persons hospitalized in quarter following interview: 7.31%

‡The only provider supply measure with time-variation was medical doctors per 1000 population.

Appendix B contains SUDAAN program statements illustrating the design specification for the LSOA analysis. Note that the 'subpopulation' statement was used to include only those sample members with valid Medicare data. Subsetting the analysis file prior to running SUDAAN can result in erroneous results if entire PSUs are missing. In order to correctly compute variances, the program requires that complete information on the design be available on the analysis file.[58] The use of the subpopulation statement guarantees this to be the case.

The second column in Table 3 gives the estimated odds ratios. Persons who were males, hospitalized at least once in the prior year, or had generally greater IADL impairment were significantly more likely ($P < 0.05$) to have a hospitalization in the

quarter following the interview than were persons without these characteristics. The odds of having a hospital stay in the next quarter decreased slightly at very high levels of IADL impairment, as indicated by the negative beta-coefficient on the squared term. Controlling for all the other covariates, having a hospital stay in the year prior was associated with a greater increase in the odds of having a hospital stay in the next quarter than any other variable. The last column in Table 3 gives the design effect, which exceeds one for all variables and is as high as 1.5. Four additional variables (being White, the self-assessment of poor health, and the two ADL terms) would have been statistically significant at $P < 0.05$ if the standard errors had not been adjusted for the effect of the survey design.

In summary, both of the examples using the LSOA show that adjusting for the effects of the complex sample design on the standard errors leads to more conservative assessments of the statistical significance of most parameters. In particular, variables with a large design effect might not satisfy a particular criterion of statistical significance (e.g. $P < 0.05$) once the clustering in sample selection is accounted for (although the estimated relationships may still be of interest from a policy perspective).

6 Discussion

We have provided illustration of only a few of the analytical tools now available for complex survey data analysis, with particular emphasis on those applicable to research questions of interest for health surveys. Although we focused much of the discussion on methods available in the SUDAAN software system and presented results only from that system, other packages are referenced that are widely available, offering user-friendly analytical procedures and alternative variance estimation methodologies.

When the target population of interest is fairly close to the survey population, these methods are most useful. In such cases, the variance estimation procedure employed should reflect the sample design as accurately as possible. Due to confidentiality considerations, information required to compute variances according to a specific design may not be available and approximations must be used. This was the case for the example analyses presented here. In order to reflect without replacement sampling, population counts of sampling units within each stratum or substratum are required. Unequal probabilities of selection require knowledge of the joint probabilities of selection among the sample units in addition to the analysis weights. For large national surveys with some first stage units selected with certainty, full specification of such sampling information could result in loss of confidentiality. Consequently, US surveys are frequently released with only limited design information and analysis weights accompanying the data.

With replacement variance approximations require only first stage strata and PSU identifiers in addition to the analysis weights; consequently, these approximations are frequently employed (see examples 1 and 2). Such approximations are generally thought to be conservative in that the estimated variances will be slightly larger than those computed under the actual sample design.[21]

Analyses focusing on survey populations often involve fairly straightforward questions due to the fact that inference is being made about the finite population

parameters and not about some other hypothetical population. Controlling for confounding, for example, is not typically necessary in examining relationships among finite population parameters. The design-based estimates of such relationships are consistent for those occurring in the population sampled, making the issue of confounding moot.

On the other hand, when inference is to a broader target population than that being surveyed, accounting for the full survey design may not be necessary to address the research questions of interest (see, for example, Koch and Beck[34]). In fact, situations exist where design based estimation results in a significant loss of efficiency, and it is recommended that the design be ignored.[38] A cautious perspective is necessary in these situations because hypothetical assumptions serve as the principal basis for how well the sample represents the target population rather than the survey design.

The decision concerning design versus model based analysis strategies may best be made by considering the impact of the design characteristics in the context of the specific research question being addressed. For example, the use of a finite population correction factor[11] in computing variances for without replacement sampling will usually result in a reduction of variance. However, if inference is to be made to any extension of the survey population, such as the population that would be surveyed in another year, then taking advantage of that variance reduction might not be appropriate. Similarly, stratified sampling often results in lower variances. However, if inference is to a broader population than that targeted, it may be more appropriate to include stratification variables as covariates in the analysis.

For data analyses involving modelling, the design features requiring the most consideration are weighting and clustering. Including sampling weights in the appropriate estimating equations will eliminate bias in the estimation process. Accounting for clustering, at least at the PSU level, will result in more accurate variance approximations. Any bias resulting from assuming with replacement sampling will be in the conservative direction. To avoid being overly conservative, it may be possible to apply innovative modelling strategies to account for the sampling design.[59–61]

Acknowledgements

For the LSOA analysis, we would like to thank Mary Grace Kovar for conceptual contributions, Thomas Walke for programming assistance, and the Milbank Memorial Fund for financial support.

References

1 Shah BV, Barnwell BG, Hunt PN, LaVange LM. *SUDAAN user's manual (release 5.50).* Research Triangle Park, NC: Research Triangle Institute, 1991.

2 McDowell A, Engel A, Massey JT, Maurer K. Plan and operation of the Second Health and Nutrition Examination Survey 1976–80. *Vital and Health Statistics* 1981; **1**(15).

3 National Center for Health Statistics. Design and estimation for the National Health Interview Survey, 1985–94. *Vital and Health Statistics* 1989; **2**(110).

4 Kovar MG, Chyba M, Fitti JE. The Longitudinal Study of Aging: 1984–1990. *Vital and Health Statistics* 1992; **1**(28).

5 Harlan WR, Landis JR, Schmouder RL, Goldstein NH, Harlan LC. Blood lead and blood pressure. *Journal of the American Medical Association* 1985; **253**: 530–34.

6 Jones DY, Schatzkin A, Green SB *et al.* Dietary

fat and breast cancer in NHANES I Epidemiologic Follow-up Study. *Journal of the National Cancer Institute* 1987; **79**: 465–71.
7 Stevens RG, Jones DY, Micozzi MS, Taylor PR. Body iron stores and the risk of cancer. *New England Journal of Medicine* 1988; **319**: 1047–62.
8 Yip R, Williamson DF. Body iron stores and risk of cancer [To the Editor]. *New England Journal of Medicine* 1989; **320**: 1012.
9 Korn EL and Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. *American Journal of Public Health* 1991; **81**: 1166–73.
10 Plan and operation of the Hispanic Health and Nutrition Examination Survey 1982–84. *Vital and Health Statistics* 1985; **1**(19).
11 Kish L. *Survey Sampling.* New York: John Wiley, 1965.
12 Cochran WG. *Sampling Techniques*, 3rd edn. New York: John Wiley, 1977.
13 Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**: 663–85.
14 Woodruff RS. A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* 1971; **66**: 411–14.
15 Kish L, Frankel MR. Inference from complex samples (with discussion). *Journal of the Royal Statistical Society Series B* 1974; **36**: 1–37.
16 Kish L, Frankel MR. Balanced repeated replications for standard errors. *Journal of the American Statistical Association* 1970; **65**: 1071–94.
17 Bean J, Schnack GA. An application of balanced repeated replication to the estimation of variance components. *Proceedings of the Section on Social Statistics* Part II. Washington, DC: American Statistical Association, 1977; 938–42.
18 Wolter KM. *Introduction to Variance Estimation.* New York: Springer, 1985.
19 Koch GG, Freeman DH, Freeman JL. Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* 1975; **43**: 59–78.
20 Fuller WA. Regression analysis for sample surveys. *Sankhya Series C* 1975; **37**: 117–32.
21 Shah BV, Holt MM, Folsom RE. Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute* 1977; **47**: 43–57.
22 Freeman DH, Freeman JL, Brock DB, Koch GG. Strategies in the multivariate analysis of data from complex surveys. II. An application to the United States Health Interview Survey. *International Statistical Review* 1976; **44**(3): 317–30.
23 Landis JR, Lepkowski JM, Eklund SA, Stehouwer SA. A statistical methodology for analyzing data from a complex survey: the first National Health and Nutrition Examination Survey. *Vital and Health Statistics* 1982; **2**(92).
24 Binder DA. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 1983; **51**: 279–92.
25 Shah BV, Folsom RE, Harrell FE, Dillard CN. *Survey data analysis software for logistic regression.* Research Triangle Park, NC: Research Triangle Institute, 1984.
26 LaVange LM, Iannachione VG, Garfinkel SG. An application of logistic regression methods to survey data: predicting high cost users of medical care. *Proceedings of the Section on Survey Research Methods.* Washington, DC: American Statistical Association, 1986.
27 Morel JG. Logistic regression under complex survey designs. *Survey Methodology* 1989; **15**: 203–23.
28 Chambless LE, Boyle KE. Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics, Theory and Methods* 1985; **14**: 1377–92.
29 Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika* 1992; **79**: 139–47.
30 Skinner CJ, Holt D, Smith TMF eds. *Analysis of complex surveys.* New York: John Wiley, 1989.
31 Rao JNK, Bellhouse DR. The history and development of the theoretical foundations of survey based estimation and statistical analysis. In: *Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions*, Gail MH, Johnson NL eds. Alexandria, VA: American Statistical Association, 1989.
32 Koch GG, Gillings DB, Stokes ME. Biostatistical implications of design, sampling and measurement to the analysis of health science data. *Annual Review of Public Health* 1980; **1**: 163–225.
33 Koch GG, Gillings DB. Inference, design based vs model based. In: *Encyclopedia of statistical sciences*, Johnson NL, Kotz S eds. New York: John Wiley, 1983; **4**: 84–8.
34 Koch GG, Beck JD. Statistical methodologies useful for the analysis of data from risk assessment studies. *Journal of Public Health Dentistry* 1992; **52**(3): 146–67.
35 Pfeffermann D, Holmes DJ. Robustness considerations in the choice of a method of

inference for regression analysis of survey data. *Journal of the Royal Statistical Society Series A* 1985; **148**: 268–78.
36 Pfeffermann D, Smith TMF. Regression models for grouped populations in cross-section surveys. *International Statistical Review* 1985; **53**: 37–59.
37 Sarndall CE. Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics* 1978; **5**: 27–52.
38 Korn EL, Graubard BI. Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society Series A* 1995; **158**(2): 263–95.
39 Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
40 Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–30.
41 Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**: 1049–60.
42 LaVange LM, Keyes LL, Koch GG, Margolis PA. Application of sample survey methods for modeling ratios to incidence densities. *Statistics in Medicine* 1994; **13**: 343–55.
43 Graubard, BI, Korn EL. Regression analysis with clustered data. *Statistics in Medicine* 1994; **13**: 509–22.
44 Koch GG, Paquette DW. Design and statistical considerations in periodontal clinical trials. *Journal of Periodontal Research* 1996; **1**(1).
45 Fuller WA, Kennedy W, Schnell D, Sullivan G, Park HJ. *PC-CARP*. Ames, IA: Statistical Laboratory, Iowa State University, 1986.
46 Rust K. Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics* 1985; **1**: 381–97.
47 Fay RE. A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association* 1985; **80**: 148–57.
48 Fay RE. CPLX: Contingency table analysis for complex designs, program documentation. US Bureau of the Census, 1989.
49 LaVange LM, Shah BV, Barnwell BG, Killinger JF. SUDAAN: a comprehensive package for survey data analysis. In: Liepins GL, Uppuluri VRR eds. *Data quality control*. New York: Marcel Dekker, 1991.
50 Shah BV, Barnwell BG, Bieler GS. *SUDAAN user's manual (release 7.0)*. Research Triangle Park, NC: Research Triangle Institute, 1996.
51 Shah BV, Folsom RE, LaVange LM, Boyle KE, Wheeless SC, Williams RL. *Statistical methods and mathematical algorithms used in SUDAAN*. Research Triangle Park, NC: Research Triangle Institute, 1993.
52 Thomas DA, Rao JNK. A Monte-Carlo study of exact levels of goodness-of-fit statistics under cluster sampling. *Proceedings of the Section on Survey Research Methods*. Washington, DC: American Statistical Association, 1984.
53 Korn EL, Graubard BI. Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni *t* statistics. *American Statistician* 1990; **44**: 270–76.
54 LaVange LM, Koch GG. Analysis of repeated measures studies with multiple regression methods for sample surveys. Manuscript presented at the Drug Information Association annual meeting, Washington, DC, June, 1994.
55 US Department of Health and Human Services, Bureau of Health Professions, Office of Data Analysis and Management. *Area resource file (ARF) system: information for health resources planning and research*. Washington, DC: US Government Printing Office, ODAM4-87, 1989.
56 Lafata JE, Koch GG, Weissert WG. Estimating activity limitation in the noninstitutionalized population: a method for small areas. *American Journal of Public Health* 1994; **84**(11): 1813–17
57 Bilgrad R. *National death index user's manual*. Hyattsville, MD: Public Health Service, 1990.
58 Graubard BI, Korn EL. Survey inference for subpopulations. *American Journal of Epidemiology* 1996; **144**: 102–6.
59 Pfeffermann D, LaVange LM. Regression models for stratified multi-stage cluster samples. In: Skinner CJ, Holt D, Smith TMF eds. *Analysis of complex surveys*. New York: John Wiley, 1989: 237–60.
60 Pfeffermann D. The role of sampling weights when modeling survey data. *International Statistical Review* 1993; **61**(2): 317–37.
61 Graubard BI, Korn EL. Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research* 1996; **5**: 263–81.

Appendix A: SUDAAN program statements for example 1: NHIS logistic regression

```
PROC LOGISTIC DATA=HIS FILETYPE=SAS DESIGN=WOR;
NEST STRATUM PSU SUB SEGMENT/ MISSUNIT;
TOTCNT POPPSU _ZERO_ _MINUS1 _ _ZERO_;
WEIGHT WTF;
SUBGROUP SEX2 RACE1 POV1;
LEVELS 2 2 3;
MODEL BED_IND=AGE1 SEX1 RACE1 POV1 HLTH1;
TEST SATADJF WALDF;
SETENV LABWIDTH = 16 COLWIDTH=8 MAXIND=4;
TITLE "EX 1: MULTIPLE LOGISTIC REGRESSION";
OUTPUT /FILENAME=LOGIS FILETYPE=ASCII VARIANCE=ALL
 COVFILE=LGST3COV REPLACE CHECKFMT=F2.0;
```

Note: In the above design specification, the variable POPPSU is set to '0' for self-representing PSUs to indicate stratification at the first stage, and set to '−1' for nonself-representing PSUs to indicate with replacement sampling at the first stage. The keyword _MINUS1_ on the TOTCNT statement indicates with replacement sampling at the second stage for self-representing PSUs.

Appendix B: SUDAAN program statements for example 2: LSOA survival analysis

```
PROC SURVIVAL DATA=SURV FILETYPE=SAS DESIGN=WR EST_NO=5200;
NEST STRATUM PSEUDPSU;
WEIGHT WGT86;
MODEL MONTHS=
 AGE7579 AGE8084 AGE85UP
 MALE WHITE PHOSP PRIVHI84 MCAID PHEALTH
 ADL ADL2 IADL IADL2
 ALONE INC20KUP METRO URBAN NCENT SOUTH WEST
 HBPOP84 MDPOP NHBPE86;
EVENT TOTDIE;
PRINT /BETAFMT=F8.3 SEBETAFMT=F8.5 HRFMT=F8.3;
PRINT DEFT/; PRINT HR/;
```

SUDAAN program statements for example 2: LSOA repeated measures logistic regression

```
PROC LOGISTIC DATA=REP FILETYPE=SAS DESIGN=WR EST_NO=16000;
NEST STRATUM PSEUDPSU;
WEIGHT WGT86;
```

```
SUBPOPN MCCOV2 = 1 /NAME='PERSONS WITH MEDICARE CLAIMS
ONLY';
MODEL PROBHOS3=
 AGE7579 AGE8084 AGE85UP
 MALE WHITE PHOSP PRIVHI84 MCAID PHEALTH
 ADL ADL2 IADL IADL2
 ALONE NH INC20KUP METRO URBAN NCENT SOUTH WEST
 HBPOP84 MDPOP NHBPE86 YEAR86 YEAR88 YEAR90;
PRINT /BETAFMT=F8.5 SEBETAFMT=F8.5;
PRINT DEFT/;
```