Paper SA-07-2012

Modeling Complex Survey Data

Taylor Lewis, University of Maryland, College Park, MD

ABSTRACT

Using PROC REG or PROC LOGISTIC to model data emanating from a complex sample survey may lead one to erroneous inferences. Instead, the survey's design features should be incorporated by employing one of the PROCs prefixed by SURVEY, such as PROC SURVEYREG or PROC SURVEYLOGISTIC. After a brief discussion of the features constituting a "complex" survey, the syntax and output of these procedures is demonstrated. Many of the theoretical and conceptual subtleties to modeling data from a finite population as opposed to the more familiar simple random sample setting are noted. Examples are drawn from a real, publicly-available survey data set.

BACKGROUND ON COMPLEX SURVEY DESIGN FEATURES

Before launching into the details and examples of modeling complex survey data, this paper begins with a discussion of what constitutes "complex" survey data. Specifically, four distinct features are discussed: finite population corrections, clustering, stratification, and unequal weights. There is no loss of generality placing this material at the front of the paper, because the syntax to account for one or more of these features is identical across all SURVEY PROCs (e.g., PROC SURVEYREG or PROC SURVEYLOGISTIC).

In most introductory statistics courses covering topics such as analysis of variance (ANOVA), regression, or general linear models, the implied data generating mechanism is simple random sampling with replacement, possibly from an infinite or hypothetical population. Under that paradigm, we are allowed to assume that data (especially the outcome variable) are independently and identically distributed, or i.i.d for short. In contrast, survey researchers tend to select samples without replacement from finite, or enumerable, populations, and simple random sampling is the exception rather than the rule. Alternative sample designs can yield efficiencies in certain circumstances, but they are most commonly adopted out of necessity or to save on data collection costs.

For sake of an example, assume a sample of n = 1,000 high schoolers is taken at random from some school's population of N = 5,000 high schoolers, and the key variable of interest y is some measure of mathematical aptitude, perhaps from a standardized test. We know the sample mean $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is an unbiased estimate of $\bar{Y}$, the true population mean, or the average test score across all high schoolers. If the sample is selected with replacement, meaning each of the 5,000 students in the population can be selected multiple times, the sample variance of this sample mean is $\mathrm{var}(\bar{y}) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)}$. If the sample is selected without replacement, however, the variance formula is modified to $\mathrm{var}(\bar{y}) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)}\left(1-\frac{n}{N}\right)$. In other words, by sampling without replacement the variance is reduced by a factor of the sampling rate—in this case, 20%.

The term (1 – n/N) is called the *finite population correction*, or FPC, and it appears in nearly all estimator variance formulas, not just that of the sample mean. Notice that as the sampling fraction approaches 1, the variance tends to 0, which makes sense. Another way of interpreting this is that, as the proportion of the population in the sample grows, uncertainty in a sample-based estimate decreases. In the most extreme case of a census ($n = N$), the FPC is 0 and there is no variance. The sample-based estimate defaults to the given population quantity.

To incorporate the FPC, one must use the TOTAL= or RATE= options in the PROC statement. SAS determines the sample size, n, from the SURVEY PROC's input data set, but relies on the user to specify the population size, N. Alternatively, users can specify the sampling rate, n/N, directly via the RATE= option. If neither of these options appears, the SURVEY PROC simply ignores the FPC.

The second feature is *clustering*, which occurs when the unit sampled is actually a cluster of population units. For instance, let us assume the desired mathematics aptitude score can only be obtained from an in-person examination which must be conducted on school grounds. Instead of directly sampling students from a roster and contacting each separately, it may be more efficient to utilize the students' homeroom, one of a comprehensive set of classroom locations in which all students begin the first hour of their school day.

Suppose the population of N = 5,000 high schoolers can be partitioned into C = 200 homerooms (i.e., clusters) of 25 students each. An alternative method of sampling n = 1,000 students is to sample c = 40 homerooms and administer the aptitude test to all 25 students therein. The clustering should be accounted for, however, by specifying this homeroom identifier variable in the CLUSTER statement of the respective SURVEY PROC.

There is no mandate to sample all units within a cluster. For instance, we could have achieved the n = 1,000 sample size by selecting c = 100 clusters (homerooms) in the first-stage, but selecting only 10 students at random from each homeroom in the second stage. This is an example of a multi-stage sampling design in which the primary sampling units (PSUs) are homerooms and the secondary sampling units (SSUs) are students. It is worth emphasizing, however, that only the PSU identifier should be specified in the CLUSTER statement. When SAS sees two variables in the CLUSTER statement, it assumes the combination of the two defines a PSU, which can result in an unduly low variance estimate. Specifying only the PSU corresponds to the ultimate cluster assumption (p. 67 of Heeringa et al., 2010) that is frequently used to simplify variance calculations.

The third distinctive feature of complex survey data is *stratification*, which arises when PSUs are allocated into one of a mutually exclusive and exhaustive set of groups, or *strata* (singular: stratum) and an independent sample is selected within each. Whereas clustering typically decreases precision, in all but a few rare circumstances, stratification increases precision. The reason is that the overall variance consists of stratum-specific variance estimates summed over all strata. When strata are constructed homogeneously with respect to the principle outcome variable(s), there can be considerable precision gains relative to simple random sampling.

Returning to our hypothetical mathematics aptitude example, a sensible stratification variable might be grade level. If we grouped all homerooms into one of the H = 4 grade levels—ninth through twelfth—prior to independently selecting a (possibly variable) number of clusters within each, we should inform the SURVEY PROC of this grade level identifier via the STRATA statement.

Parenthetically alluded to above was how sampling rates of clusters may vary across strata. In general, when sampling rates vary across ultimately sampled units, analysts should account for this by assigning a unit-level weight equaling the inverse of that unit's selection probability. Weights are the fourth distinct feature of complex survey data. A weight can be interpreted as the number of population units a sample unit represents. For instance, if a sample unit's selection probability was one-fourth, that unit would be assigned a weight of four. That unit's survey responses represent itself and three other comparable units in the population. The numeric weight variable should be specified in the WEIGHT statement of the SURVEY PROC. If no WEIGHT statement appears, weights are implicitly assigned as 1.

A REAL-WORLD COMPLEX SURVEY: THE NATIONAL AMBULATORY MEDICAL CARE SURVEY (NAMCS)

The fictitious mathematics aptitude survey was introduced to facilitate exposition of the particular complex survey features data analysts may encounter. We now shift attention to a real-world complex survey, the National Ambulatory Medical Care Survey (NAMCS). Sponsored by the National Center for Health Statistics (NCHS), a Federal Statistical Agency within the Centers for Disease Control and Prevention (CDC), NAMCS collects data on outpatient visits to non-emergency physician's offices. That is, the ultimate sample unit is a physician visit. Examples of variables measured by the survey include diagnoses made, chronic illnesses of the patient, time spent with the physician, and medications prescribed or renewed. NCHS administers NAMCS on a yearly basis, and in addition to various tabulations and publications, NCHS releases data via a public-use microdata file. Instructions on how to download the data can be found on their website: http://www.cdc.gov/nchs/ahcd.htm. In this paper, we will demonstrate analyses using the 2009 NAMCS public-use data set.

A wise first step to understand the design elements and complex features of a survey is to consult any user documentation materials available. According to the documentation (available on-line at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc09.pdf), NAMCS has three of the features discussed in the previous section:

1. Stratification –strata are identified by distinct codes of the CSTRATM variable.
2. Clustering – each PSU is identified (within a stratum) by distinct codes of CPSUM.
3. Unequal Weighting – weights can be specified by the PATWT variable.

The NAMCS public-use file contains hundreds of variables. Aside from the key design variables mentioned above, Table 1 summarizes the manageable subset of these variables that will be used in this paper.

Table 1. Summary of NAMCS 2009 Public-Use Data Set Variables Used in this Paper.

Variable Name	Description	Coding Structure
SEX	Patient Gender	1 = Female 2 = Male
AGE	Patient Age in Years	Continuous
RACER	Patient Race	1 = White 2 = Black 3 = Other
SOLO	Indicator of Solo Physician Practice	1 = Yes 2 = No
MED	Indicator of Medication Prescribed	0 = No 1 = Yes
TIMEMD	Time Spent with Physician in Minutes	Continuous, ranging from 0 to 240
TOTCHRON	Patient's Total Number of Chronic Conditions	Count ranging from 0 to 10
MAJOR	Primary Reason for Visit	1 = New problem (<3 mos. onset) 2 = Chronic problem, routine 3 = Chronic problem, flare up 4 = Pre-/Post-surgery 5 = Preventive care (e.g. routine prenatal, well-baby, screening, insurance, general exams)

LINEAR REGRESSION

MODELING DATA COLLECTED FROM SIMPLE RANDOM SAMPLES

The traditional simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ states the expected value of continuous outcome variable *y* is a linear function of some predictor variable *x*. The betas are referred to as the model's *parameters* or *coefficients*. β_0 represents the intercept and β_1 the slope or the expected change in *y* given a one-unit increase in *x*. The term ε_i represents a deviation or *residual* of the i^{th} unit's outcome, the distance between the observed y_i and its expected value (what lies on the regression line). It is assumed the ε_i 's are distributed i.i.d. with mean 0 and some constant variance σ^2.

The terms and assumptions just described apply to the (infinite) population. Given a simple random sample of size *n*, we can calculate unbiased estimates of these parameters. We typically use "hat" notation to distinguish them from their population counterparts. For instance, the familiar least squares estimates can be calculated:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad , \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

For multiple linear regression models, those with two or more predictor variables, matrix notation is handy for simplifying the algebra required to estimate parameters. Of course, matrices can also be used for the simple linear regression case as well. Let *p* denote the number of model parameters. In the simple linear regression case, *p* = 2. The first step is to construct an (*n* x *p*) design matrix **X** consisting of a column of 1s for the intercept and a separate column for each distinct predictor variable in the model. These can be either continuous or a sequence of (*k* -1) 0/1 indicator or "dummy" variables for a categorical variable with *k* levels.

The second step is to construct an (n x 1) column vector **Y** containing the n values of the outcome variable y. Below is a visualization of what **X** and **Y** would look like for the simple linear regression case:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}$$

The same estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained by calculating $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, e.g., via PROC IML, where the superscript T signifies the matrix transpose and -1 the inverse of the given matrix. In general, the matrix $\hat{\boldsymbol{\beta}}$ will be a (p x 1) vector of "beta hats." The mean squared error term of the model $\hat{\sigma}^2$ can be obtained by squaring and summing the n entries of the (n x 1) residual matrix $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \hat{\mathbf{Y}}$ and dividing the result by $n - p$. The sum of the squared entries of **e** is called the *residual sum of squares*.

Another useful matrix is the (p x p) model parameter covariance matrix, defined as $\text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}$. For the simple linear regression case, this matrix would look like

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) \end{bmatrix}$$

Entries along the diagonal are the estimated sample variances for the model parameters while off-diagonal entries are covariances between any two model parameters. The $p(p - 1)/2$ distinct covariances are symmetric about the diagonal. As will be shown later, this matrix can be used to test hypotheses about individual parameters or evaluate whether the model can be simplified by simultaneously eliminating more than one parameter.

MODELING DATA COLLECTED FROM A COMPLEX SURVEY

Shah, Holt, and Folsom (1977) discuss how complex survey data can violate many of the underlying assumptions in standard least squares regression. In contrast to simple random sampling from an infinite population, survey data are typically collected to make generalizations about finite populations. As such, model parameters have a subtly different interpretation. Instead of a Greek beta (β), a finite population regression parameter is typically denoted with a Roman letter, e.g., B_0 or B_1. Using sample weights, the finite population parameters can be estimated by

$$\hat{B}_{1w} = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2} , \qquad \hat{B}_{0w} = \bar{y}_w - \hat{B}_{1w}\bar{x}_w$$

The two frameworks for conceptualizing model parameters can be unified through the idea of a *superpopulation* (Heeringa et al., 2010, p. 184), which stipulates that, although the N units in the survey population represent a finite set, the given regression model corresponds to some overarching superpopulation model. Put another way, one might view the finite population as itself a sample from an infinite population, acknowledging there may be variability in a particular estimate $\hat{B}$ between one realized finite population and another. Heeringa et al. (2010) point out, however, that any such variability is negligible for large populations (i.e., large N), so $\hat{B}$ effectively approximates the superpopulation parameter.

The estimates $\hat{B}_{0w}$ and $\hat{B}_{1w}$ are *weighted least squares estimates*, subscripted with a w to distinguish them from their unweighted versions. (Note, also, the distinction between $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$). These weighted estimates approximate the corresponding finite population quantities. For example, $\hat{B}_{1w}$ estimates B_1, the slope coefficient that would be computed had we conducted a full census of the population. In a similar vein, recall the minimization function in ordinary least squares regression is the residual sum of squares $RSS = \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$. In weighted least squares estimation, the function being minimized is $RSS_w = \sum_{i=1}^{n} w_i\left(y_i - (\hat{B}_0 + \hat{B}_1 x_i)\right)^2$, which can be thought of as an estimate of the residual sum of squares for the model fit using the entire finite population, or $RSS_{pop} = \sum_{i=1}^{N}\left(y_i - (B_0 + B_1 x_i)\right)^2$.

Weights are incorporated into the matrix notation by introducing an (n x n) diagonal matrix **W** in which the i^{th} unit's weight appears along the diagonal while all off-diagonal entries are 0. Specifically, the weighted least squares estimates can be calculated by $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$. This is the same form of a weighted least squares estimator sometimes used to account for violations in the homogeneity of residuals assumption in simple random sampling—see Weisberg (2005, pp. 96-97). In that context, however, weights are assigned to be inversely proportional to the estimated residual variance.

To obtain weighted least squares parameter estimates in SAS, use the WEIGHT statement within PROC SURVEYREG. These estimates will match what is output from PROC REG with a WEIGHT statement, except if there is stratification and/or clustering involved, standard errors will still be incorrect—and likely biased downward, as will be shown shortly. The standards errors from PROC SURVEYREG properly account for the complex survey features by (the default) Taylor series linearization (Fuller, 1975). It is beyond the scope of this paper to delve too deeply into the theory behind Taylor series linearization to estimate the matrix $\mathrm{cov}(\hat{\boldsymbol{\beta}})$, but suffice it to say the process is more involved than simply inserting the **W** matrix into the simple linear regression version of $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}$.

Lohr (1999) offers a concise summarization of the on-going debate amongst survey researchers as to whether weights are truly needed when modeling survey data. The *model-based* perspective, an alternative to the *design-based* perspective adopted in this paper, argues that if the model is correctly specified, the weighted and unweighted regressions both approximate the true population parameters. Thus, if one is confident the given model correctly describes the true state of affairs in the population, there is no need to use the weights, since arbitrarily variable weights can inflate variances (Kish, 1992). Pfeffermann and Holmes (1985) note, however, that using the weights provides robustness against a misspecified model, such as one missing certain influential predictor variables, and Skinner et al. (1989) assert weights can protect against biases introduced by non-ignorable sample designs, when the sample inclusion indicator is strongly related to the outcome variable.

FITTING A LINEAR REGRESSION MODEL USING DATA FROM NAMCS

Suppose one wanted to model time spent with the physician (TIMEMD) as a function of whether any medication was prescribed or renewed (MED), patient gender (SEX), age (AGE), and race (RACER), the total number of chronic conditions afflicting the patient (TOTCHRON), primary reason for the visit (MAJOR), and whether the doctor is the lone physician in the practice (SOLO). The following syntax fits this model naïvely ignoring the complex survey design features:

```
proc surveyreg data=NAMCS_2009;
  class MED SEX RACER SOLO MAJOR;
  /* use CLASS statement to treat numeric variables as categorical */
  model TIMEMD = MED SEX RACER SOLO MAJOR
                 AGE TOTCHRON
                 / solution; /* SOLUTION option needed to output parameter
estimates when CLASS statement is used */
run;
```

```
                     1) Ignoring the Complex Survey Features

                        Estimated Regression Coefficients

                                         Standard
           Parameter      Estimate          Error    t Value    Pr > |t|

           Intercept    18.8493335     0.36534171      51.59      <.0001
```

```
                    MED 0          0.2983170    0.18240145        1.64       0.1020
                    MED 1          0.0000000    0.00000000         .          .
                    SEX 1          0.1469686    0.15252028        0.96       0.3353
                    SEX 2          0.0000000    0.00000000         .          .
                    RACER 1       -0.5620540    0.31655393       -1.78       0.0758
                    RACER 2       -0.7721509    0.36956357       -2.09       0.0367
                    RACER 3        0.0000000    0.00000000         .          .
                    SOLO 1         2.8481878    0.17606163       16.18       <.0001
                    SOLO 2         0.0000000    0.00000000         .          .
                    MAJOR 1       -0.4957786    0.20935883       -2.37       0.0179
                    MAJOR 2        0.1034439    0.22806649        0.45       0.6501
                    MAJOR 3        2.3397329    0.33209039        7.05       <.0001
                    MAJOR 4       -1.8119289    0.32919691       -5.50       <.0001
                    MAJOR 5        0.0000000    0.00000000         .          .
                    AGE            0.0056041    0.00362426        1.55       0.1221
                    TOTCHRON       0.4224505    0.06633694        6.37       <.0001

NOTE: The denominator degrees of freedom for the t tests is 28275. Matrix X'X is
singular and a generalized inverse was used to solve the normal equations.
```

PROC SURVEYREG without the STRATA, CLUSTER, or WEIGHT statement is akin to fitting the model using PROC REG. One benefit of PROC SURVEYREG, however, is that a CLASS statement is allowed. With PROC REG, the onus of creating dummy variables to represent categorical predictors lies with the analyst. The CLASS statement performs this task on the fly for all variables listed. The note regarding singularity of the $\mathbf{X}^T\mathbf{X}$ matrix is intended to inform the user that one of the dummy variables for each CLASS variable was dropped. By default, the PROC drops the last category in the sort-ordered list. This explains why, for example, the parameter estimate and standard error for MAJOR category 5 are both zero. As commented in the syntax, the SOLUTION option should be specified when a CLASS statement appears; otherwise, the table of estimated regression coefficients will not be printed to the output window.

Next, the same model is fit, only now accounting for NAMCS' three complex survey features.

```
  proc surveyreg data=NAMCS_2009;
    strata CSTRATM;
    cluster CPSUM;
    class MED SEX RACER SOLO MAJOR;
    model TIMEMD = MED SEX RACER SOLO MAJOR
                   AGE TOTCHRON / solution;
  weight PATWT;
  run;
```

```
                     2) Accounting for the Complex Survey Features

                                Estimated Regression Coefficients

                                            Standard
                   Parameter      Estimate         Error    t Value    Pr > |t|

                   Intercept    18.7181878    0.79267944      23.61      <.0001
                   MED 0        -0.2098052    0.35788673      -0.59      0.5580
                   MED 1         0.0000000    0.00000000        .         .
                   SEX 1        -0.0832461    0.23038481      -0.36      0.7180
                   SEX 2         0.0000000    0.00000000        .         .
                   RACER 1      -0.0793324    0.59702281      -0.13      0.8943
                   RACER 2       0.1570886    0.73119135       0.21      0.8300
                   RACER 3       0.0000000    0.00000000        .         .
                   SOLO 1        1.4381012    0.64863855       2.22      0.0270
                   SOLO 2        0.0000000    0.00000000        .         .
                   MAJOR 1      -1.4257590    0.50914036      -2.80      0.0053
                   MAJOR 2      -1.0238478    0.56655914      -1.81      0.0713
                   MAJOR 3       1.3824545    0.77712045       1.78      0.0758
                   MAJOR 4      -2.0803614    0.59004176      -3.53      0.0005
                   MAJOR 5       0.0000000    0.00000000        .         .
                   AGE           0.0188608    0.00917546       2.06      0.0403
```

```
                  TOTCHRON      0.4216447    0.13840225       3.05      0.0024

NOTE: The denominator degrees of freedom for the t tests is 520. Matrix X'WX is
singular and a generalized inverse was used to solve the normal equations.
```

The two tables of regression parameters are quite different. For one, the estimates have changed after incorporating the weights (PATWT). Interestingly, AGE only becomes significant at the α = .05 level *after* accounting for the complex design. The standard errors are also generally higher, likely attributable to both weighting and clustering. Users can examine this variance inflation by specifying DEFF after the slash in the MODEL statement. DEFF is short for the *design effect* (Kish, 1965), defined as the ratio of the estimated variance under the complex design to the comparable variance under simple random sampling—equivalently, the ratio of standard errors squared. A design effect of 1.12, for example, implies the complex survey design's variance is 12% higher than what would have been achieved under simple random sampling.

Also, notice how the note below the parameter estimates table states the degrees of freedom is now 520 as opposed to 28275 in the PROC SURVEYREG run ignoring the complex features. This is because the degrees of freedom under a complex survey designs is actually the number of PSUs minus the number of strata, a count that is often drastically less than the number of observations minus 1 as in simple random sampling.

The column labeled "t Value" provides the parameter estimate divided by its standard error, or $\frac{\hat{B}}{\sqrt{\mathrm{var}(\hat{B})}}$. This is being referenced against a random t variable with 520 degrees of freedom. A small associated *p*-value suggests the parameter may not differ significantly from zero.

Another useful section of output is the Tests of Model Effects, shown below for both models

```
                        1) Ignoring the Complex Survey Features

                                Tests of Model Effects

                       Effect       Num DF    F Value    Pr > F

                       Model            11      41.90    <.0001
                       Intercept         1    9688.90    <.0001
                       MED               1       2.67    0.1020
                       SEX               1       0.93    0.3353
                       RACER             2       2.20    0.1109
                       SOLO              1     261.70    <.0001
                       MAJOR             4      30.59    <.0001
                       AGE               1       2.39    0.1221
                       TOTCHRON          1      40.55    <.0001

                     2) Accounting for the Complex Survey Features

                                Tests of Model Effects

                       Effect       Num DF    F Value    Pr > F

                       Model            11       5.85    <.0001
                       Intercept         1    1273.14    <.0001
                       MED               1       0.34    0.5580
                       SEX               1       0.13    0.7180
                       RACER             2       0.12    0.8894
                       SOLO              1       4.92    0.0270
                       MAJOR             4       7.67    <.0001
                       AGE               1       4.23    0.0403
                      TOTCHRON          1       9.28    0.0024

NOTE: The denominator degrees of freedom for the t tests is 520. Matrix X'WX is
singular and a generalized inverse was used to solve the normal equations.
```

The effect line labeled “Model” is the global F test for whether all parameters aside from the intercept are significantly different from zero, the same F test that would appear in “Model” row of an ANOVA table. Based on the current model’s F test, there is little support of this null hypothesis. For continuous or dichotomous predictors—those with 1 numerator degree of freedom—the “F Value” column is just the squared “t Value” columns appearing in the parameter estimates table. For polytomous variables appearing in the CLASS statement, however, the F Value column tests whether all applicable dummy variables parameters are significantly different from zero. Overall, it appears race (RACER) has little explanatory power in the model of time spent with the physician, at least in the presence of the other predictor variables.

SIMPLIFYING THE MODEL

Recall that in simple random sampling one can construct the following F test to determine whether a set of parameters in a multiple linear regression model are not significantly different from zero, suggesting those terms can be eliminated from the model:

$$F_{SRS} = \frac{\dfrac{RSS_{red} - RSS_{full}}{\text{Number of Parameters Dropped}}}{\hat{\sigma}^2_{full}}$$

where RSS_{red} is the residual sum of squares for the reduced model, the model without the candidates terms for elimination, and RSS_{full} is that for the full model. The term $\hat{\sigma}^2_{full}$ denotes the mean squared error for the full model, or RSS_{full} divided by the degrees of freedom for error. Under the null hypothesis and normally distributed errors, this statistic has an F distribution with numerator degrees of freedom equaling the number of parameters dropped and denominator degrees of freedom equaling the degrees of freedom for error of the full model.

For sake of an illustration, assume the full model is $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ and we are considering reducing it to $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$. That is, we want to test H_0: $\beta_2 = 0$ versus H_A: $\beta_2 \neq 0$. An alternative form of the F test above is to fit the full model but calculate a *contrast* by forming **C** = $\begin{bmatrix}0\\0\\1\end{bmatrix}$ and finding $F_{SRS} = (\mathbf{C}^T \hat{\boldsymbol{\beta}})^T(\mathbf{C}^T \mathrm{cov}(\hat{\boldsymbol{\beta}})\,\mathbf{C})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}})/\mathrm{rank}(\mathbf{C})$, where rank(**C**) denotes the number of linearly independent columns in the matrix **C**. The matrix algebra in this univariate case—rank(**C**) = 1—reduces to the parameter estimate squared divided by its variance, which is also just the squared univariate t Value appearing in the Estimated Regression Coefficients table. To assess significance, the F statistic should be referenced against a random F variable with numerator degrees of freedom rank(**C**) and denominator degrees of freedom that associated with the mean squared error of the model.

Though perhaps uninteresting in the univariate scenario, the **C** matrix can be expanded with additional columns to test multiple parameter hypotheses simultaneously. The F statistic has the same form, only the numerator degrees of freedom increases to reflect the number of linearly independent columns in **C**. For instance, to test H_0: $\beta_1 = 0, \beta_2 = 0$ versus H_A: $\beta_1 \neq 0, \beta_2 \neq 0$, one would assign **C** = $\begin{bmatrix}0 & 0\\1 & 0\\0 & 1\end{bmatrix}$ and compare to a random F variable with the same denominator degrees of freedom, only now with numerator degrees of freedom equal to 2. Customized contrasts such as these can be calculated within PROC SURVEYREG by using the CONTRAST statement.

Further scrutiny of the Tests for Model Effects table in the output accounting for the complex design reveals three of the predictor variables—MED, SEX, and RACER—all appear insignificant in the current model. A logical follow-up question might be whether all three variables could be dropped. The syntax below demonstrates how one can utilize the CONTRAST statement to test this hypothesis. Commas separate each column of the **C** matrix. There is no need to specify contrast coefficients for all 12 rows (intercept + 11 distinct parameters), since any variable name which does not appear has an implied contrast coefficient of 0(s). The last coefficient for each CLASS variable corresponds to the dropped dummy variable, so it may seem unnecessary to provide a value. For some reason, however, the CONTRAST statement in PROC SURVEYREG requires the excluded dummy variable (or dummy variable whose parameter estimate was set to 0) be given a coefficient of -1 to work properly.

```
proc surveyreg data=NAMCS_2009;
  strata CSTRATM;
  cluster CPSUM;
  class MED SEX RACER SOLO MAJOR;
  model TIMEMD = MED SEX RACER SOLO MAJOR
                 AGE TOTCHRON / solution;
contrast 'Test of Reduced Model' MED 1 -1, SEX 1 -1, RACER 1 0 -1, RACER 0 1 -1 ;
weight PATWT;
run;
```

```
                      Analysis of Contrasts

       Contrast                 Num DF    F Value    Pr > F

       Test of Reduced Model         4       0.15    0.9614

 NOTE: The denominator degrees of freedom for the F tests is 520.
```

The CONTRAST statement generates a new output element entitled “Analysis of Contrasts.” As with the univariate tests, a low *p*-value suggests the terms are influential in explaining TIMEMD. In this case (*p* = 0.9614), it appears little explanatory power would be lost by dropping all terms related to MED, SEX, and RACER.

Technically speaking, when the model is fit using complex survey data, the F tests reported in the Analysis of Contrast and Tests of Model Effects tables may not strictly adhere to the same F distribution as in simple random sampling. Korn and Graubard (1990) suggest a modified test statistic $F_{ADJ} = F_{SRS} * \frac{(DF - q + 1)}{DF}$, where *DF* is the degrees of freedom calculated under the complex design (# PSUs – # strata). F_{ADJ} is then referenced against a random F variable with *q* and (*DF* – *q* + 1) degrees of freedom, where *q* is the number of parameters dropped. This alternative is not yet available in PROC SURVEYREG, but it is not difficult to compute by hand. Using figures from Analysis of Contrasts portion of the output, the adjustment factor would be $\frac{(520-4+1)}{520} = \frac{517}{520} \approx 1$, so we can quickly gather $F_{ADJ} \approx F_{SRS}$. Still, the adjustment may be more sizeable when fewer degrees of freedom are available.

LOGISTIC REGRESSION

MODELING DATA COLLECTED FROM SIMPLE RANDOM SAMPLES

Fitting a logistic regression model is typically the preferred approach when the outcome variable of interest is binomial, such as a yes/no question or an indicator of the presence/absence of some characteristic. We can describe this outcome in general terminology as either an event or non-event. It may seem reasonable to feed a 0/1 variable as the outcome to a PROC that fits a linear regression model based on one or more predictor variables, such that the expected outcome is the expected probability of whatever event was coded 1. Unfortunately, there are a few problems with this approach. For one, the expected outcome (probability) may fall outside the bounds of [0, 1], which would be a nuisance for interpretation. Also problematic is that the model violates the assumption of normally distributed residuals. Regardless of the predicted outcome, the actual outcome can be only one of two values (unless the residual happens to be zero), which limits the number of distinct residuals to two as well.

The logistic regression model involves a transformation aimed at alleviating these problems. Instead of modeling the predicted probability of some event based on linear combination of predictor variable(s), parameters are linear in terms of the *logit*, or log-odds of the i^{th} unit experiencing the event. Recall the odds are defined as the ratio of two probabilities, the probability of experiencing the event and the probability of not experiencing the event.

For the single-variable setting, the logit function is defined as $\text{logit}(event \mid x_i) = \log\left(\frac{\Pr(event \mid x_i)}{1-\Pr(event \mid x_i)}\right) = \beta_0 + \beta_1 x_i$.

Notice how there is no residual term ε_i. Consequently, there are no assumptions regarding their distribution as there are in linear regression. It should be emphasized, however, that there still is the assumption that observations sampled are i.i.d, therefore naturally reflecting the population structure.

Although we are not modeling them directly, the i^{th} unit’s predicted probability given its value of *x* can be extracted by solving the logit equation for $\Pr(event \mid x_i)$, which is just $\frac{\exp(\text{logit}(event \mid x_i))}{1+\exp(\text{logit}(event \mid x_i))} = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$, where exp() denotes exponentiation of whatever falls within the parentheses. A nice feature of the logit transformation is that the predicted probability always falls within (0,1), even while model parameters can range from (-∞,∞).

The logit transformation carries with it a different interpretation of model parameters. For instance, in the single-variable model above, β_1 no longer represents the expected change in the outcome variable itself, but rather the expected change in the logit or log-odds given a one-unit change in *x*. Exponentiating this parameter returns an odds ratio, which is somewhat easier to handle.

To illustrate, suppose *x* is a 0/1 variable where 1 denotes males and 0 females and the outcome variable is whether the individual has had a heart attack. If $\beta_1 = 0.50$, the odds ratio of a heart attack for men versus women is exp(0.50) = 1.65. The colloquial interpretation is that men are 65% more likely to have experienced a heart attack. In truth, such an interpretation is more apt for the *relative risk* statistic, the ratio of the two predicted probabilities (see Section 2.3.4 of Agresti, 1996). The tendency to treat the odds ratio as if it were the relative risk may be an artifact of historical logistic regression analysis in the realm of rare epidemiological events, where much of the literature on the technique developed. In the case of a small probability of some event, the two statistics approximate each other. Note that if $\beta_1 = 0$, exp(0) = 1, which is to say odds ratio is 1 and, thus, the two groups' predicted probabilities of experiencing the event are the same.

From a sample of size *n*, the true *β*'s can be estimated and, as before, are topped with hats to signify they are sample-based estimates. As opposed to minimizing the residual sum of squares, however, the logistic regression parameters are estimated via an iterative procedure based on the method of maximum likelihood. For each of the *p* distinct columns of the design matrix, the associated maximum likelihood parameter estimates are the solution to $\sum_{i=1}^{n} x_{pi} y_i = \sum_{i=1}^{n} x_{pi} \hat{\pi}_i$, where $\hat{\pi}_i == \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_{(p-1)} x_{(p-1)i}\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_{(p-1)} x_{(p-1)i}\right)}$ is the (estimated) predicted probability of the event occurring for unit *i*. While these are typically termed *score* equations in logistic regression, the comparable equations in least squares regression, $\sum_{i=1}^{n} x_{pi} y_i = \sum_{i=1}^{n} x_{pi} \hat{y}_i$, are more often termed *normal* equations. A sensible byproduct of these criteria is that, for categorical variables, the sum of predicted probabilities matches the count of each level's successes.

As will be demonstrated shortly, the (*p* x *p*) matrix $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ is still central to testing hypotheses of individual parameters or assessing whether multiple parameters are not significantly different from zero; however, this matrix is not estimated the same way as with linear regression. We still require the (*n* x *n*) design matrix **X**, but after conducting the iterative process to find the maximum likelihood parameter estimates, one must construct an (*n* x *n*) matrix **V** with diagonal entries equal to $\hat{\pi}_i (1-\hat{\pi}_i)$ and 0s on the off-diagonals. Then, $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ = $(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}$.

MODELING DATA COLLECTED FROM A COMPLEX SURVEY

Similar arguments and conceptual subtleties apply to logistic regression parameters estimated from a model using complex survey data drawn from a finite population—for an accessible discussion, see Section 6.4 of Hosmer and Lemeshow (2000). When weights are involved, pseudo-maximum likelihood estimates are obtained by iteratively adjusting the estimated parameters until each of the *p* modified score equations $\left(\sum_{i=1}^{n} w_i x_{pi} y_i = \sum_{i=1}^{n} w_i x_{pi} \hat{\pi}_i\right)$ holds.

As with weighted least squares estimation in the finite population context, the resulting parameters can be thought of as the model parameters that would be estimated had we conducted a full census of the population. Running PROC LOGISTIC with a WEIGHT statement will give the exact same pseud-maximum likelihood estimates as PROC SURVEYLOGISTIC with a WEIGHT statement. As warned earlier, however, if stratification and/or clustering is present in the complex survey data set, standard errors will be only be correct in the SURVEYLOGISTIC output.

Binder (1983) developed the theory of Taylor series linearization to estimate $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ accounting for the complex design. Further discussion of this was deemed a bit too theoretical for the current paper, but interested readers seeking more detail can refer to that article.

FITTING A LOGISTIC REGRESSION MODEL USING DATA FROM NAMCS

Suppose the event we are interested in predicting is whether a medication was prescribed or renewed (MED). We might anticipate this outcome being a function of time spent with the physician (TIMEMD), patient gender (SEX), age (AGE), and race (RACER), the total number of chronic conditions afflicting the patient (TOTCHRON), and primary reason for the visit (MAJOR). The syntax below fits the naïve model, ignoring all complex survey features.

```
proc surveylogistic data=NAMCS_2009;
  class SEX RACER MAJOR / param=ref;
  model MED(event='1') = SEX RACER MAJOR /* categorical predictors, must appear in
CLASS statement */
                         AGE TOTCHRON TIMEMD /* continuous predictors */ ;
run;
```

By default, SAS treats the first value in the sorted list of outcome variable values as the event. This is inappropriate in the present case since MED is coded 0 when a prescription is *not* issued or renewed. The EVENT=' ' option can be specified in parentheses after the outcome variable in the MODEL statement to explicitly declare the event of interest.

All three categorical variables appear in the CLASS statement. The option PARAM=REF overrides the default effect parameterization with the reference group parameterization, which treats categorical variables much the same as is done automatically when the CLASS statement is used in PROC SURVEYREG. The indicator variable corresponding to the last level in the sorted list of distinct levels is dropped. To assign a specific reference category, specify REF='' in parentheses after the variable is listed in the CLASS statement. (The PARAM=REF option after the slash is still required.)

It is wise practice to purposefully assign the reference category since all other parameters created from the given class variable are interpreted as the change in the log-odds ratio between that category and the reference category (the omitted group). Earlier, it was mentioned that exponentiating the parameter of a 0/1 indicator variable returns the odds ratio between the group coded 1 and the group coded 0. In this case, it represents the odds ratio between the given category and the reference group. The option EXPB can be placed after the slash in the MODEL statement to append these values to the table of parameter estimates. However, for all variables not involved in an interaction, the Odds Ratios portion of the output (not shown) essentially displays this information.

```
                  1) Ignoring the Complex Survey Features

                 Analysis of Maximum Likelihood Estimates

                                        Standard          Wald
   Parameter      DF     Estimate          Error    Chi-Square    Pr > ChiSq

   Intercept       1       0.4688         0.0741       39.9988        <.0001
   SEX       1     1       0.0764         0.0286        7.1504        0.0075
   RACER     1     1       0.0123         0.0638        0.0373        0.8468
   RACER     2     1     0.000131         0.0746        0.0000        0.9986
   MAJOR     1     1       0.3395         0.0389       76.3087        <.0001
   MAJOR     2     1       0.5754         0.0422      185.8831        <.0001
   MAJOR     3     1       0.4815         0.0595       65.5123        <.0001
   MAJOR     4     1      -0.4343         0.0551       62.1484        <.0001
   AGE             1     -0.00067       0.000672        0.9823        0.3216
   TOTCHRON        1       0.3524         0.0153      530.6116        <.0001
   TIMEMD          1     -0.00256        0.00113        5.1710        0.0230
```

One difference between the Analysis of Maximum Likelihood Estimates in the PROC SURVEYLOGISTIC output and the comparable summarization given in PROC SURVEYREG is that Wald chi-square statistics are given to evaluate whether the parameter is significantly different from 0. These are calculated by squaring the estimate and dividing by the corresponding diagonal entry in the $\text{cov}(\hat{\boldsymbol{\beta}})$ matrix—equivalently, squaring the quotient of the estimate and its standard error. Since these are essentially single-parameter contrasts, it should come as no surprise the *p*-values are based on a reference chi-square distribution with 1 degree of freedom.

To fit the same model properly accounting for the stratification, clustering, and unequal weights, the PROC SURVEYLOGISTIC code above can be fleshed out with the following STRATA, CLUSTER, and WEIGHT statements

```
proc surveylogistic data=NAMCS_2009;
  strata CSTRATM;
  cluster CPSUM;
  class SEX RACER MAJOR / param=ref;
  model MED(event='1') = SEX RACER MAJOR
                         AGE TOTCHRON TIMEMD ;
weight PATWT;
run;
```

```
                 2) Accounting for the Complex Survey Features

                  Analysis of Maximum Likelihood Estimates

                                      Standard          Wald
    Parameter       DF    Estimate       Error    Chi-Square    Pr > ChiSq

    Intercept        1      0.2409      0.1821        1.7498        0.1859
    SEX       1      1      0.0786      0.0439        3.2032        0.0735
    RACER     1      1      0.2027      0.1683        1.4503        0.2285
    RACER     2      1      0.1201      0.1901        0.3993        0.5275
    MAJOR     1      1      0.3508      0.0839       17.4813        <.0001
    MAJOR     2      1      0.5717      0.0893       41.0013        <.0001
    MAJOR     3      1      0.4157      0.1273       10.6680        0.0011
    MAJOR     4      1     -0.6651      0.1665       15.9558        <.0001
    AGE              1    0.000351     0.00157        0.0500        0.8231
    TOTCHRON         1      0.3546      0.0467       57.6121        <.0001
    TIMEMD           1    0.000935     0.00262        0.1274        0.7211
```

One can observe how incorporating the weights substantively alters many of the estimates. For instance, the TIMEMD parameter changed sign and is no longer significant. Standard errors are also notably higher, as would be expected.

SIMPLIFYING THE MODEL

In the simple random sampling setting, one can construct a likelihood ratio test (p. 37 of Hosmer and Lemeshow, 2000) to determine whether a logistic regression model can be reduced by eliminating one of more terms. Specifically, -2(L_{red} – L_{full}) can be compared to a reference chi-square distribution with q degrees of freedom, where L_{red} is the log-likelihood of the reduced model, L_{full} is the log-likelihood of the full model, and q is the number of parameters up for elimination. As Hosmer and Lemeshow (2000, p. 217) later note, however, this test no longer applies when the model is fit using complex survey data, because the likelihoods for the full and reduced models are actually pseudo-likelihoods. Instead, a general Wald chi-square contrast can be formulated.

Suppose we wanted to test whether race, sex, age, and time spent with the physician could simultaneously be dropped from the model. The syntax below performs this test, drawing upon the same concepts of using the CONTRAST statement that were demonstrated in testing reduced linear regression models.

```
proc surveylogistic data=NAMCS_2009;
  strata CSTRATM;
  cluster CPSUM;
  class SEX RACER MAJOR / param=ref;
  model MED(event='1') = SEX RACER MAJOR
                         AGE TOTCHRON TIMEMD ;
weight PATWT;
contrast 'Test of Reduced Model' SEX 1, RACER 1 0, RACER 0 1, AGE 1, TIMEMD 1;
run;
```

```
                      Contrast Test Results

                                              Wald
     Contrast                  DF    Chi-Square    Pr > ChiSq

     Test of Reduced Model      5        6.1011        0.2965
```

One difference between this CONTRAST statement and that in PROC SURVEYREG is the user is not required to specify the -1 contrast coefficient for the reference category. If one does, PROC SURVEYLOGISTIC sends a note to the log stating that the coefficient was ignored. The other distinction is that the contrast result takes the form of a Wald chi-square test statistic. There are 5 degrees of freedom associated with the contrast since the matrix **C** has 5 linearly independent columns. Though it does not print by default, SAS will output this matrix if the option E is specified after the slash in the CONTRAST statement.

The Korn and Graubard (1990) recommended correction applies to the Wald chi-square contrast test statistic just as it did for the F test illustrated in PROC SURVEYREG. The key difference between the two, however, is that the F statistic is a chi-square statistic divided through by its degrees of freedom q (= rank(**C**)). Thus, to apply the correction proposed by Korn and Graubard (1990), one would compute $F_{ADJ} = W_{SRS} * \frac{(DF - q + 1)}{DF * q}$. To assess significance, this result is still compared to a reference F distribution with q numerator degrees of freedom and denominator degrees of freedom *DF*. As with PROC SURVEYREG, this modified version is not yet available in PROC SURVEYLOGISTIC. With q = 5 and *DF* = 520, the adjusted Wald test statistic would be $W_{ADJ} = 6.1011 * \frac{(520 - 5 + 1)}{520 * 5} = 1.22$ (p = .30). Both the uncorrected and correct test statistics suggest the terms associated with sex, race, age, and time spent with the physician can be dropped from the model.

MODELING DOMAINS (OR SUBSETS) OF A COMPLEX SURVEY DATASET

The models illustrated thus far have been fit using the entire survey data set. There are surely occasions where the analyst wants to restrict the model to a *domain*, or subset, of the data. For example, one may wish to restrict the model of TIMEMD to a particular characteristic(s) of the patient (e.g., patients diagnosed with osteoporosis), the visit (e.g., whether any imaging tests were performed), or even the doctor's practice (e.g., whether solo or group practice). There are scores of indicator variables in the public-use file which can be used to narrow down the analysis.

Simply filtering the original data set for the domain of interest, either in a separate DATA step or on the fly using a WHERE statement, may result in erroneous inferences. The proper technique is to specify the domain indicator variable(s) in a DOMAIN statement. There are two reasons for this. For one, supplying the full data set informs SAS of the complete survey design. It is possible certain strata and/or PSUs will be filtered out because no cases therein meet the domain criteria. This would impact the degrees of freedom. Second, subsetting the data is only appropriate when the sample size is fixed from one hypothetical sample to another. For instance, if the sample were stratified by a particular variable, such that a fixed number of sample units was selected within each, it may be plausible to subset on a particular stratum or groups of strata. This will likely not be the case, in general, so it is recommended to utilize the DOMAIN statement to ensure proper inferences.

EXAMPLE OF DOMAIN ANALYSIS IN PROC SURVEYREG

Further inspection of the NAMCS 2009 data set reveals that the 0/1variable PHYS is an indicator of whether a physician was actually seen during the visit. For visits where PHYS=0, TIMEMD=0. A physician was seen (PHYS=1) in over 97% of the visits in the data set, so focusing only on these cases likely produces a model strikingly similar to the model of time spent with the physician fit previously. Nonetheless, for purposes of demonstration, the syntax below re-fits the original model placing the variable PHYS in the DOMAIN statement.

```
proc surveyreg data=NAMCS_2009;
  strata CSTRATM;
  cluster CPSUM;
  class MED SEX RACER SOLO MAJOR;
  model TIMEMD = MED SEX RACER SOLO MAJOR
                 AGE TOTCHRON / solution;
weight PATWT;
domain PHYS;
run;
```

```
                       The SURVEYREG Procedure

                                PHYS=1

             Domain Regression Analysis for Variable TIMEMD


                        Tests of Model Effects

               Effect       Num DF    F Value    Pr > F

               Model            11       6.11    <.0001
               Intercept         1    1324.10    <.0001
               MED               1       0.01    0.9108
               SEX               1       0.01    0.9155
               RACER             2       0.09    0.9127
               SOLO              1       6.31    0.0123
```

```
                  MAJOR              4       8.34    <.0001
                  AGE                1       2.66    0.1034
                  TOTCHRON           1       8.78    0.0032

            NOTE: The denominator degrees of freedom for the F tests is 520.


                     Estimated Regression Coefficients

                                          Standard
              Parameter      Estimate          Error    t Value    Pr > |t|

              Intercept    19.0869070    0.81209027      23.50      <.0001
              MED 0        -0.0415137    0.37051957      -0.11      0.9108
              MED 1         0.0000000    0.00000000        .          .
              SEX 1        -0.0252196    0.23762720      -0.11      0.9155
              SEX 2         0.0000000    0.00000000        .          .
              RACER 1       0.2284892    0.62592120       0.37      0.7152
              RACER 2       0.3224702    0.76768898       0.42      0.6746
              RACER 3       0.0000000    0.00000000        .          .
              SOLO 1        1.7042633    0.67863716       2.51      0.0123
              SOLO 2        0.0000000    0.00000000        .          .
              MAJOR 1      -1.6787516    0.51404332      -3.27      0.0012
              MAJOR 2      -0.9468646    0.53707368      -1.76      0.0785
              MAJOR 3       1.0984575    0.75090877       1.46      0.1441
              MAJOR 4      -2.3032250    0.59317997      -3.88      0.0001
              MAJOR 5       0.0000000    0.00000000        .          .
              AGE           0.0152782    0.00936571       1.63      0.1034
              TOTCHRON      0.4110722    0.13874679       2.96      0.0032
```

When the DOMAIN statement is used, a model is fit for the entire data set as well as each distinct level of the DOMAIN statement variable. In our case, the only relevant output is that pertaining to the domain where PHYS=1, so the output presented herein is abridged accordingly. Multiple variables can be specified in the DOMAIN statement, but they will be treated separately, in turn. If the domain of interest is identified by the combination of two or more variables, separate them by an asterisk (or create a new dichotomous domain indicator in a separate DATA step to limit the extraneous output).

It is a worthwhile aside to consider what occurs behind the scenes during domain analysis, because there may be situations where the DOMAIN statement is unavailable, such as was the case for certain SURVEY PROCs in earlier versions of SAS. In essence, SAS creates a domain-specific weight equaling either (1) the original weight for observations in the domain, or (2) 0 for observations outside the domain. It may seem feasible to create this weight by hand and re-run PROC SURVEYREG with this variable in the WEIGHT statement. This will not work, however, because the SURVEY PROC immediately excludes any observations where the variable identified in the WEIGHT statement is 0—exactly what we are trying to avoid. The go-around is to assign a miniscule weight to all non-domain cases that is strictly greater than zero, such as .000000000001.

The syntax below demonstrates this technique. Note how the PROC SURVEYREG output matches what was generated for that domain (PHYS=1) when the DOMAIN statement was used.

```
data NAMCS_2009;
  set NAMCS_2009;
PATWT2=PATWT*(PHYS=1)+.000000000001;
run;
proc surveyreg data=NAMCS_2009;
  strata CSTRATM;
  cluster CPSUM;
  class MED SEX RACER SOLO MAJOR;
  model TIMEMD = MED SEX RACER SOLO MAJOR
                AGE TOTCHRON / solution;
weight PATWT2;
run;
```

```
                         Tests of Model Effects
```

```
Effect       Num DF    F Value    Pr > F

Model            11       6.11    <.0001
Intercept         1    1324.10    <.0001
MED               1       0.01    0.9108
SEX               1       0.01    0.9155
RACER             2       0.09    0.9127
SOLO              1       6.31    0.0123
MAJOR             4       8.34    <.0001
AGE               1       2.66    0.1034
TOTCHRON          1       8.78    0.0032
```

```
             Estimated Regression Coefficients

                             Standard
Parameter      Estimate         Error    t Value    Pr > |t|

Intercept    19.0869070    0.81209027      23.50      <.0001
MED 0        -0.0415137    0.37051957      -0.11      0.9108
MED 1         0.0000000    0.00000000        .         .
SEX 1        -0.0252196    0.23762720      -0.11      0.9155
SEX 2         0.0000000    0.00000000        .         .
RACER 1       0.2284892    0.62592120       0.37      0.7152
RACER 2       0.3224702    0.76768898       0.42      0.6746
RACER 3       0.0000000    0.00000000        .         .
SOLO 1        1.7042633    0.67863716       2.51      0.0123
SOLO 2        0.0000000    0.00000000        .         .
MAJOR 1      -1.6787516    0.51404332      -3.27      0.0012
MAJOR 2      -0.9468646    0.53707368      -1.76      0.0785
MAJOR 3       1.0984575    0.75090877       1.46      0.1441
MAJOR 4      -2.3032250    0.59317997      -3.88      0.0001
MAJOR 5       0.0000000    0.00000000        .         .
AGE           0.0152782    0.00936571       1.63      0.1034
TOTCHRON      0.4110722    0.13874679       2.96      0.0032
```

These concepts translate to logistic regression models, as the DOMAIN statement is also available in PROC SURVEYLOGISTIC. For brevity, no additional example of domain analysis will be demonstrated, but the syntax is identical to that in PROC SURVEYREG.

CONCLUSION

This paper introduced the four idiosyncratic features of complex surveys and how they interact with the traditional treatment and assumptions of linear and logistic regression models. The core principle to be taken away is that, when modeling data emanating from a complex survey design, one of the SURVEY PROCs should be used. Although PROC SURVEYREG and PROC SURVEYLOGISTIC were the two demonstrated in this paper, another available procedure is PROC SURVEYPHREG, which has the capability of fitting Cox proportional hazards regression models. These models are one of a class of models applicable to survival analysis, when the outcome of interest is the elapsed time before a certain event occurs. For further discussion, see Mukhopadhyay (2010) or Berglund (2011).

Currently, there are no model selection algorithms built into PROC SURVEYREG or PROC SURVEYLOGISTIC, such as forward, backward, or stepwise selection. This paper demonstrated, however, that the model parameter covariance matrix can be utilized to conduct customized hypothesis tests regarding one or more parameters. It was briefly noted the default procedure in SAS to estimate these matrices is Taylor series linearization (Fuller, 1975; Binder, 1983). Though they were not discussed in this paper, replication-based methods offer an alternative avenue for estimation. For an overview, see Rust (1985). For examples of the SAS syntax necessary to implement some of these techniques, which became available in version 9.2, see Mukhopadhyay et al. (2008).

Extensions of the logistic regression model are available in PROC SURVEYLOGISTIC just as they are in PROC LOGISTIC—namely, the cumulative logit and generalized logit models. When the outcome variable consists of more than two categories, SAS defaults to fitting a cumulative logit model, whereby there are multiple intercepts but a common slope assumed between any two categories of the outcome. In general, this is only applicable to an ordinal variable. When the outcome is nominal, the LINK=GLOGIT option after the slash in the MODEL statement should be specified to fit a generalized logit model, under which separate slope terms appear. This model consists of 2 x *p* terms, each representing the change in the logit between given outcome and some reference outcome.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York, NY: Wiley.

Berglund, P. (2011). “An Overview of Survival Analysis using Complex Sample Data,” *Proceedings from the SAS Global Forum 2011 Conference*. Available on-line at: http://support.sas.com/resources/papers/proceedings11/338-2011.pdf.

Binder, D. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, **51**, pp. 279–292.

Fuller, W. A. (1975). "Regression Analysis for Sample Survey," *Sankhyā*, **37**, Series C, Pt. 3, pp. 117–132.

Hosmer, D., and Lemeshow, S. (2000). *Applied Logistic Regression*. Second Edition. New York, NY: Wiley.

Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.

Kish (1992). “Weighting for Unequal P_i,” *Journal of Official Statistics*, **8**, pp. 183 – 200.

Korn, E., and Graubard, B. (1990). “Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni T Statistics.” *The American Statistician*, **44**, pp. 270 – 276.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Mukhopadhyay, P., An, A., Tobias, R., and Watts, D. (2008). “Try, Try Again: Replication-Based Variance Estimation Methods for Survey Data Analysis in SAS 9.2,” *Proceedings from the SAS Global Forum 2008 Conference*. Available on-line at: http://www2.sas.com/proceedings/forum2008/367-2008.pdf.

Mukhopadhyay, P. (2010). “Not Hazardous to Your Health: Proportional Hazards Modeling for Survey Data with the SURVEYPHREG Procedure,” *Proceedings from the SAS Global Forum 2010 Conference*. Available on-line at: http://support.sas.com/resources/papers/proceedings10/254-2010.pdf.

Pfeffermann, D., and Holmes, D. (1985). “Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data,” *Journal of the Royal Statistical Society*, Series A, **148**, pp. 268 – 278.

Rust, K. (1985). "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, **1**, pp. 381–397.

Shah, B., Holt, M., and Folsom, F. (1977). “Inference About Regression Models from Sample Survey Data,” *Bulletin of the International Statistical Institute*, **41**, pp. 43 – 57.

Skinner, C., Holt, D., and Smith, T. (1989). *Analysis of Complex Surveys*, New York, NY: Wiley.

Weisberg, S. (2005). *Applied Linear Regression*. Third Edition. Hoboken, NJ: Wiley.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name:	Taylor Lewis
Enterprise:	Joint Program in Survey Methodology (JPSM)
Address:	1218 LeFrak Hall
City, State ZIP:	University of Maryland, College Park, MD 20742
E-mail:	tlewis@survey.umd.edu