Paper SAS404-2014

Examples of Logistic Modeling with the SURVEYLOGISTIC Procedure

Rob Agnelli, SAS Institute Inc.

ABSTRACT

Logistic regression is a powerful technique for predicting the outcome of a categorical response variable and is used in a wide range of disciplines. Until recently, however, this methodology was available only for data that were collected using a simple random sample. Thanks to the work of statisticians such as Binder (1983), logistic modeling has been extended to data that are collected from a complex survey design that includes strata, clusters, and weights.

Through examples, this paper provides guidance in using PROC SURVEYLOGISTIC to apply logistic regression modeling techniques to data that are collected from a complex survey design. The examples relate to calculating odds ratios for models with interactions, scoring data sets, and producing Receiver Operating Characteristic (ROC) curves. As an extension of these techniques, a final example shows how to fit a generalized estimating equations (GEE) logit model.

INTRODUCTION

Logistic regression is a powerful technique for predicting the outcome of a categorical response variable and is used in a wide range of disciplines. Until recently, however, this methodology was available only for data that were collected using a simple random sample. Thanks to the work of statisticians such as Binder (1983), logistic modeling has been extended to data that are collected from surveys that have complex designs and that include strata, clusters, and weights. The SURVEYLOGISTIC procedure enables you to apply this methodology for binary, ordinal, and nominal responses.

Through examples, this paper demonstrates the use of PROC SURVEYLOGISTIC to obtain some of the more common statistics of interest with a binary response variable. The first example shows how to get odds ratio estimates for variables that are involved in an interaction. Next, you see how to create ROC curves based on a full sample and a holdout sample. The third example details how to score a data set based on a new sample. Finally, you are shown a special application of PROC SURVEYLOGISTIC to nominal, repeated measures data.

COMPUTING ODDS RATIO ESTIMATES IN THE PRESENCE OF AN INTERACTION

Because logit models involve modeling the probability of an event, it is often of interest to make comparisons of the probabilities for two different groups. Although this comparison cannot be done directly with a logit model, it is rather easy to compare the odds ($\frac{p}{(1-p)}$) of two groups by using the odds ratios, because the ratios are directly related to the parameters of the model itself. PROC SURVEYLOGISTIC reports the odds ratios as part of its default output for any main effect in the model that is not involved in an interaction. However, the procedure does not report odds ratios when a variable is involved in an interaction. The odds ratio is not reported because, if a variable is involved in an interaction, then there is not just a single odds ratio estimate. Rather, there will be several odds ratios—one for each level of the interacting variable. In order to get each of these odds ratios, it is necessary to use the LSMEANS statement with the ODDSRATIO and DIFF options. Let us look at the following example to see how this is done.

The National Survey of Family Growth (NSFG) is conducted by the National Center for Health Statistics to obtain data from respondents age 15 and older about family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. The survey is based on a stratified, cluster design with unequal weighting. The design variables are SEST for the stratum and SECU for the cluster variable. The weight variable is WGTQ1Q16. All relevant information from the 2006-2010 survey is available on the "National Survey of Family Growth" web page (Center for Disease Control and Prevention 2013).

This example uses the data for 2010, which is contained in two separate files—one file for men and one file for women. The data can be read into a SAS® data set using the following code:

```
proc format;
value EVRMARRY
      0 = 'Never Married'
      1 = 'Ever Married';
value RELIGION
      1 = 'No Religion'
      2 = 'Catholic'
```

```
      3 = 'Protestant'
      4 = 'Other Religions' ;
value TE_2BF
      1 = 'Yes'
      Other = 'No';


run;

data FemResp;
infile 'c:\sgf\2006_2010_FemResp.dat' LRECL=6251;
input  AGE_R 13-14 EVRMARRY 24 INTACT18 85 RELIGION 6117 WGTQ1Q16  6150-6167
SECU  6222 SEST  6223-6225 ;
gender='Female';
run;

data maleresp;
infile "c:\sgf\2006_2010_male.dat" LRECL=4543;
input  AGE_R 14-15 EVRMARRY 25 INTACT18 82 RELIGION 4414 WGTQ1Q16  4446-4463
SECU  4518 SEST  4519-4521 ;
gender='Male';
run;

data combine;
set femresp maleresp;
format RELIGION RELIGION.
       EVRMARRY EVRMARRY.
       INTACT18 TE_2BF.;
run;
```

One question of particular interest to researchers using the NSFG database is whether the respondent was ever married (EVRMARRY). By fitting a logistic regression model that is adjusted for the complex survey in PROC SURVEYLOGISTIC, you can predict the probability that a person has ever been married based on their gender (GENDER), religion (RELIGION), whether their parents had remained married until the respondent was 18 (INTACT18), and the respondent's age (AGE_R). By using the STRATA, CLUSTER, and WEIGHT statements, PROC SURVEYLOGISTIC can adjust for the complex survey design.

```
proc surveylogistic data=combine;
strata sest;
cluster secu;
class gender(ref='Male') religion(ref='No Religion') intact18(ref='No')/param=glm;
model evrmarry=gender|religion intact18 age_r;
weight wgtq1q16;
lsmeans gender*religion /oddsratio cl diff;
slice gender*religion / sliceby=gender diff=control('Male' 'No Religion') oddsratio
cl;
run;
```

Of particular importance is the use of the PARAM=GLM option in the CLASS statement. The default effects coding (PARAM=EFFECT) does not allow for the use of the LSMEANS statement. You must use the GLM coding in order to produce the LS-means and odds ratio estimates.

The odds ratios are produced in Output 1 for the two main effects, INTACT18 and AGE_R.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
INTACT18 Yes versus No	1.185	1.061	1.324
AGE_R	1.221	1.209	1.232

Output 1. Odds Ratio Estimates and 95% Confidence Limits

The odds of being married are 1.19 times higher for those who grew up in intact households than for those that did not grow up in intact households. Also, for each year over the age of 18, the odds that they have been married are 1.22 times higher.

The output from the DIFFS option in the LSMEANS statement includes all possible pairwise comparisons of the interaction (on the logit scale). This usually means that you get many more point estimates than you are interested in. So it is expedient to limit the output. You can do so by using the SLICE statement with the DIFF=CONTROL and SLICEBY options. In this case, the interest is in comparing each of the religious affiliation groups against "No Religion" for both males and females. By adding GENDER to the SLICEBY= option and using the DIFF=CONTROL("Male" "No Religion") option, only the odds ratios that you are interested in will be reported.

Output 2 shows the odds ratios comparing each of the religious groups against no religion for both females and males.

Gender	RELIGION	_RELIGION	Odds Ratio	Lower Confidence Limit for Odds Ratio	Upper Confidence Limit for Odds Ratio
Female	Catholic	No Religion	1.333	1.079	1.646
Female	Other Religions	No Religion	2.279	1.722	3.017
Female	Protestant	No Religion	1.454	1.196	1.767
Male	Catholic	No Religion	1.234	0.953	1.597
Male	Other Religions	No Religion	2.316	1.608	3.334
Male	Protestant	No Religion	1.905	1.528	2.377

Output 2. Odds Ratio Estimates of Religious Affiliation for Women and Men and the 95% Confidence Limits

Of particular note is the fact that non-Christian religious groups ("Other Religions") have odds of having been married between 2.27 and 2.36 more than non-religious persons for women and men, respectively. The odds ratio comparing Catholic men with men that have no religious affiliation is 1.23. However, the 95% confidence limits (0.95, 1.59) show that these two groups are not significantly different because the confidence interval contains 1.

The Type III Analysis of Effects table from this model (see Output 3) indicates that the effect for the respondent's age (AGE_R) is highly significant (p <0.0001).

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	1	129.0197	<.0001
RELIGION	3	70.6414	<.0001
gender*RELIGION	3	8.2632	0.0409
INTACT18	1	9.0169	0.0027
AGE_R	1	1673.2984	<.0001

Output 3. Type III Analysis of Effects

The legal minimum age of consent for marriage in 48 of the 50 states is 18. (Some states have parental consent laws that allow 16-year-olds to marry.) Because the survey includes respondents that are under 18 years old, the significance of age might be inflated. By creating and assigning a format that divides the respondents into two groups based on whether they are under 18 or not, you can examine the effect that the under-18 age group might be having on the analysis. The results from PROC SURVEYFREQ (see Row Percent in Output 4) show that approximately 20% of the respondents who were never married are under the age of 18, which might have a profound effect on the results.

```
   /*This format divides the continuous variable AGE_R into two groups.*/
proc format;
value AGEFF
15-17 = 'Under 18'
18-high = '18 and Over' ;
run;

proc surveyfreq data=combine;
format age_r ageff.;
strata sest;
cluster secu;
weight wgtq1q16;
tables evrmarry*age_r/row nofreq nowt;
run;
```

The SURVEYFREQ Procedure

Data Summary

Number of Strata	56
Number of Clusters	152
Number of Observations	22682
Sum of Weights	123882324

Table of EVRMARRY by AGE_R

EVRMARRY	AGE_R	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
Never Married	Under 18	10.0445	0.3168	19.7854	0.6559
	18 and Over	40.7229	0.7958	80.2146	0.6559
	Total	50.7674	0.7733	100.000	
Ever Married	Under 18	0.0137	0.0079	0.0278	0.0160
	18 and Over	49.2189	0.7742	99.9722	0.0160
	Total	49.2326	0.7733	100.000	
Total	Under 18	10.0582	0.3174		
	18 and Over	89.9418	0.3174		
	Total	100.000			

Output 4. PROC SURVEYFREQ Results

To examine whether this finding is significantly affecting the model, you can fit a separate logistic model for those respondents who are more than 18 years of age. This is what is commonly referred to in complex survey parlance as a *subgroup* or *domain analysis*. However, the approach is different from the approach that is used when you have a simple random sample. You cannot simply delete the observations or use a WHERE statement because you need to adjust for the fact that the subgroups were not part of the sample design. (Thus, the sample sizes for each of the groups is a random variable.) This is done in PROC SURVEYLOGISTIC by using the DOMAIN statement.

```
proc surveylogistic data=combine;
domain age_r;
format age_r ageff.;
strata SEST;
cluster SECU;
class gender(ref='Male') religion(ref='No Religion') intact18(ref='No')/param=glm;
```

```
model evrmarry=gender|religion intact18 age_r;
weight WGTQ1Q16;
lsmeans gender*religion /oddsratio cl diff;
slice gender*religion / sliceby=gender diff=control('Male' 'No Religion') oddsratio
cl;
run;
```

PROC SURVEYLOGISTIC now produces output for three separate models. The first model is the same one that was fit above with both age groups in it. The second model is just for those who are under age 18, and the third model is for those age18 and older. It is this last model in which you are interested. With respect to this model, Output 5 shows that age is still highly significant (although the chi-square statistic is smaller). However, the interaction between gender and religion and the main effect for whether the family was intact at age 18 all have lower *p*-values (although it is doubtful whether these changes are significant enough to change any conclusions that might be drawn from the model).

Domain Analysis for Domain AGE_R=18 and Older

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	1	125.0100	<.0001
RELIGION	3	76.0161	<.0001
gender*RELIGION	3	8.5141	0.0365
INTACT18	1	7.5133	0.0061
AGE_R	1	1311.9877	<.0001

Output 5. Type III Analysis of Effects for Respondents Age 18 and Older

COMPUTING ROC CURVES FOR PREDICTIVE ACCURACY

As is the case with most predictive modeling procedures, being able to get a measure of the predictive accuracy of your model is a concern. In the context of a logistic regression, measuring predictive accuracy is often done via a receiver operating characteristic or ROC curve. For a model with high predictive accuracy, the ROC curve rises quickly, and, subsequently, the area under the curve will be closer to the maximum of 1. Conversely, a model that has low predictive accuracy rises slowly and has a small amount of area under the curve that is closer to the minimum of 0.5. Let us look at an example.

The Medical Expenditure Panel Survey (MEPS) is a collection of large-scale surveys that are conducted by the Agency for Healthcare Research and Quality. The MEPS collects data about how families and individuals, their medical providers, and employers across the United States use health services. One area of particular interest to researchers is the percentage of respondents who have health insurance coverage. For this example, let us consider the data for 1999. You can download the data in SAS transport format from the Medical Expenditure Panel Survey website (U.S Department of Health and Human Services 2013a) and read it into SAS using the following code:

```
proc format;
value povcat9h
      1 = 'POOR'
      2 = 'NEAR POOR'
      3 = 'LOW'
      4 = 'MIDDLE'
      5 = 'HIGH'
      ;
value inscov9f
      1,2 = 'Insured'
      3 = 'Uninsured'
      ;
run;
```

```
libname mylib '';
filename in1 'c:\SGF\H38.SSP';

proc xcopy in=in1 out=mylib import;
run;

data meps;
   set mylib.H38;
   format povcat99 povcat9h. inscov99 inscov9f.;
   keep INSCOV99 TOTEXP99 POVCAT99 VARSTR99 VARPSU99 PERWT99F;;
run;
```

The MEPS has a complex sample design that includes both stratification (VARSTR99) and clustering (VARPSU99). The sampling weights (PERWT99F) are adjusted for nonresponse and ranked with respect to population control totals from the Current Population Survey. The data set contains the variable INSCOV99, which indicates whether a person did or did not have insurance coverage in 1999. It also contains two variables that can be used to predict the probability that a person had health coverage—TOTEXP99 represents the total health care expenditure in 1999, and POVCAT99 divides family income with reference to the poverty line (based on family size and composition) into five categories.

The goal is to measure the predictive accuracy of this model through the use of an ROC curve. PROC SURVEYLOGISTIC does not produce ROC curves directly. It can calculate predicted probabilities for a model, which can then be used in the LOGISTIC procedure to produce an ROC curve.

The model is fit using the following PROC SURVEYLOGISTIC code:

```
proc surveylogistic data=meps;
stratum VARSTR99;
cluster VARPSU99;
weight PERWT99F;
class POVCAT99;
model INSCOV99 = TOTEXP99 POVCAT99/expb;
output out=pred_ds p=phat;
run;
```

The parameter estimates and odds ratio estimates are shown in Output 6.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	1.6075	0.0645	620.2929	<.0001	4.990
TOTEXP99		1	0.000236	0.000073	10.5679	0.0012	1.000
POVCAT99	HIGH	1	1.0830	0.0673	258.7814	<.0001	2.954
POVCAT99	LOW	1	-0.3970	0.0612	42.0120	<.0001	0.672
POVCAT99	MIDDLE	1	0.3336	0.0567	34.5601	<.0001	1.396
POVCAT99	NEAR POOR	1	-0.5745	0.0888	41.8848	<.0001	0.563

Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
TOTEXP99			1.000	1.000	1.000
POVCAT99 HIGH	versus	POOR	4.610	3.739	5.683
POVCAT99 LOW	versus	POOR	1.049	0.863	1.276
POVCAT99 MIDDLE	versus	POOR	2.179	1.799	2.639
POVCAT99 NEAR POOR	versus	POOR	0.879	0.693	1.114

Output 6. Parameter Estimates and Odds Ratio Estimates

Examine the last column of the parameter estimates table, Exp (Est). Notice that the Exp(Est) is the same as the odds ratio for TOTEXP99 but not for each of the poverty (POV) categories. This is often a source of confusion and stems from the fact that the default effects coding (PARAM=EFFECT) is used for variables that are listed in the CLASS statement. Rather than comparing each level to a reference, effects coding compares each nonreference level to the mean. The odds ratios, however, are invariant to your choice of parameterization and always represent a comparison of each level to the reference.

The OUTPUT statement creates a new SAS data set called WORK.PRED_DS with a variable called PHAT to represent the predicted probability. This data set is used as input to PROC LOGISTIC while the ROC statement with the PRED= option creates the ROC.

```
proc logistic data=pred_ds;
weight PERWT99F;
baseline_model:model INSCOV99= ;
roc 'Surveylogistic Model' pred=phat;
ods select ROCOverlay ROCAssociation;
run;
```

Although PROC LOGISTIC does not calculate correct standard errors for data from a complex survey design, you can use it to produce the ROC curve because the predicted probabilities that are being used as input are corrected for the design effect. However, you must specify a MODEL statement to identify the response variable. In addition, be sure to use only those observations that were used in fitting the model to create the plot and area under the curve.

You might notice the following note in the PROC SURVEYLOGISTIC output:

```
Note:   1053 observations having nonpositive frequencies or weights were excluded
        since they do not contribute to the analysis.
```

To remove these same 1053 observations from the ROC curve, you must also add the WEIGHT statement in PROC LOGISTIC.

By fitting an intercept-only model, you will get a plot of an ROC curve that rises at a 45-degree angle and has an area of 0.5. You can use this plot as a comparative tool because it represents a random allocation of responses and nonresponses (which would be analogous to flipping a coin). The ODS SELECT statement limits the output so that only the ROC curves and area under the curves are produced.

Figure 1 shows an overlay of the two curves.

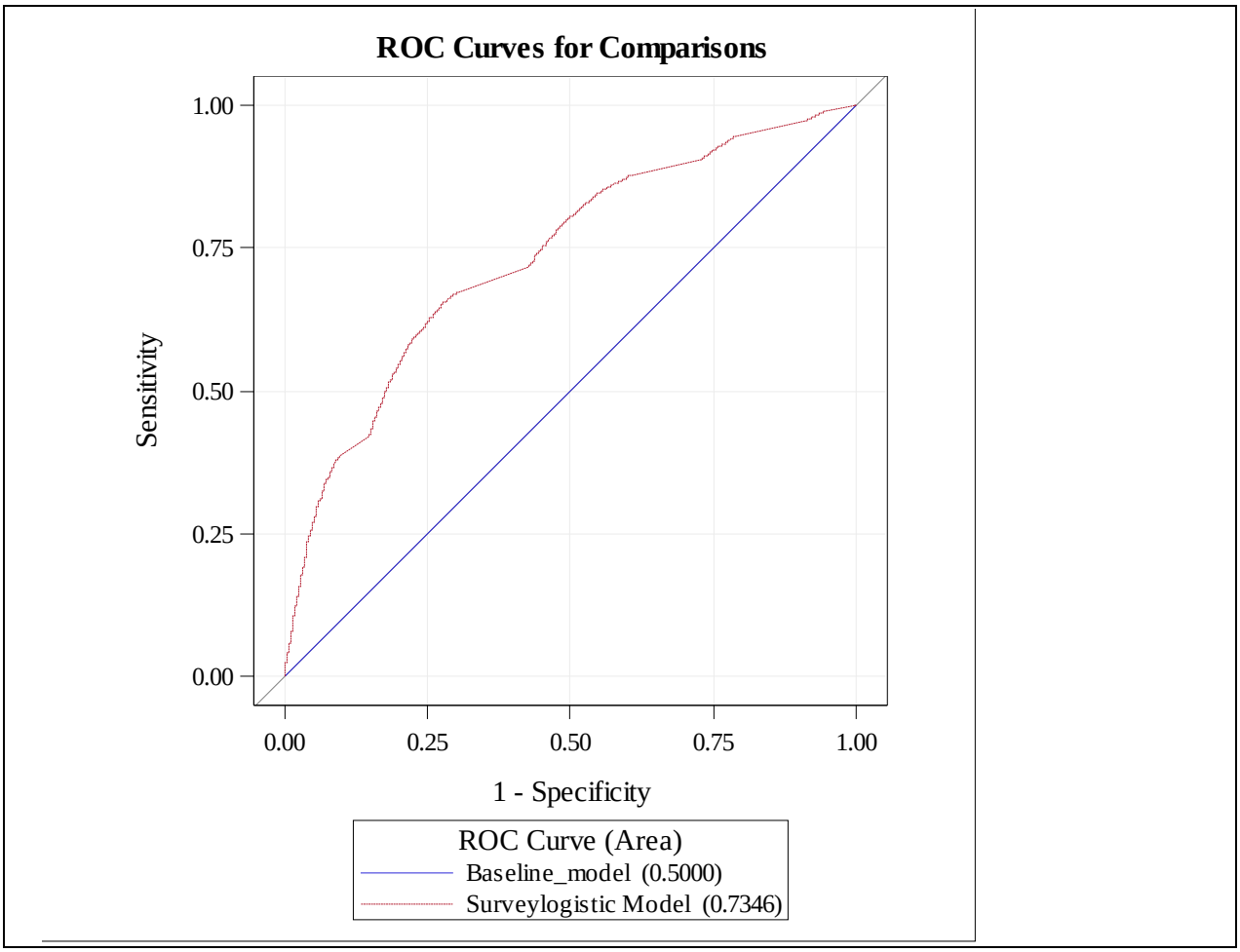

Figure 1. Overlay of ROC Curves

The curve for the fitted model rises rather quickly, which suggests that the model does a fairly good job of predicting the response. The area under the curve is 0.7400 with a 95% confidence interval of 0.7312 to 0.7488. (See Output 7.)

ROC Association Statistics

		Mann-Whitney					
		Standard	95% Wald		Somers' D		
ROC Model	Area	Error	Confidence Limits		(Gini)	Gamma	Tau-a
Baseline_mode	0.5000	0	0.5000	0.5000	0	.	0
Surveylogistic Model	0.7400	0.00449	0.7312	0.7488	0.4800	0.4865	0.1125

Output 7. Area under the ROC Curve Statistics

In order to remove any potential bias in determining the predictive accuracy of the model using the area under the ROC curve statistic, it is often suggested that you compute the area with independent data that were not used to fit the model. To that end, you can use a holdout sample to compute the area under the ROC curve using the SURVEYSELECT procedure.

The following code produces a 10% holdout sample:

```
proc surveyselect data=meps rate=.1 outall out=rand_meps seed=123;
run;
```

The OUTALL option in the PROC SURVEYSELECT statement includes all observations from the input data set in the output data set. By default, the output data set includes only those units selected for the sample. When you specify the OUTALL option, the output data set includes all observations from the input data set and also contains a variable that indicates each observation's selection status. The variable SELECTED=1 for an observation that is selected for the sample and equals 0 for an observation that is not selected. The RATE=.1 selects a 10% sample. Recalling what was said previously about the necessity of using the DOMAIN statement when you perform analyses on unplanned subgroups, to run the model without the holdout sample, you must use SELECTED as the DOMAIN variable.

```
proc surveylogistic data=rand_meps;
domain selected;
stratum VARSTR99;
cluster VARPSU99;
weight PERWT99F;
class POVCAT99;
model INSCOV99 = TOTEXP99 POVCAT99;
output out=pred_ho_ds(where=(domain='Selection Indicator=0'& selected=1)) p=phat;
run;
```

Because PROC SURVEYLOGISTIC will produce three models, it is necessary to use a WHERE statement to capture the correct predicted values. The model that you want to use to predict values consists of those observations that were not selected as part of the holdout sample (DOMAIN='Selection Indicator=0'). However, you want to keep only the predicted values for those who were part of the holdout sample (SELECTED=1).

Once again, the output data set is used as input to PROC LOGISTIC, and an ROC curve is produced.

```
proc logistic data=pred_ho_ds;
weight PERWT99F;
Baseline_model:model INSCOV99= ;
roc 'Model predictions of the Hold Out Sample' pred=phat;
ods select ROCOverlay ROCAssociation;
run;
```

Figure 2 and Output 8 show an ROC curve that is very similar to the original sample. In addition, the area under the ROC curve is slightly less, at 0.7330, and has a corresponding 95% confidence limit of 0.7046 to 0.7613.

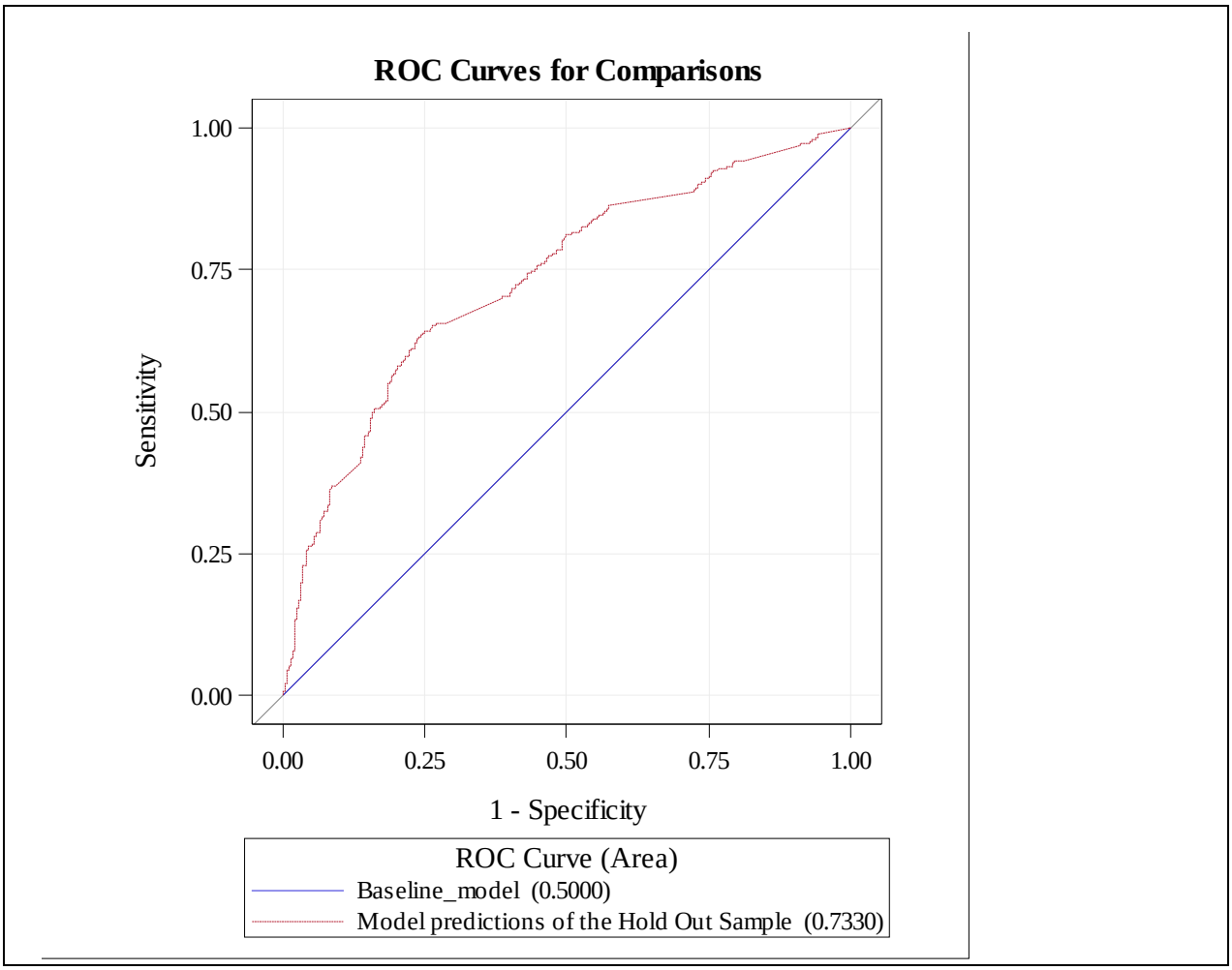

Figure 2. The ROC Curve for the Holdout Sample

ROC Association Statistics

		Mann-Whitney				
		Standard	95% Wald		Somers' D		
ROC Model	Area	Error	Confidence Limits		(Gini)	Gamma	Tau-a
Baseline_model	0.5000	0	0.5000	0.5000	0	.	0
Hold Out Sample	0.7330	0.0145	0.7046	0.7613	0.4659	0.4723	0.1052

Output 8. Area under the ROC Curve Statistics

SCORING A LOGISTIC MODEL

When you develop a logistic model, it is often of interest to apply the coefficients of your model to a future data set to get predicted probabilities. You can do this rather easily using the STORE statement in PROC SURVEYLOGISTIC with the PLM procedure. PROC PLM uses item stores that are created by some of the modeling procedures to read in the results from a specific model that was fit and stored through the STORE statement. These item stores can then be recalled for use in the future. This can be particularly useful in scoring additional data sets, especially for a procedure like PROC SURVEYLOGISTIC, which does not have any built-in functionality for scoring data.

Recall that in the example above, a model was fit using insurance coverage data from 1999. Now let us apply those coefficients to the data from the year 2000 and create a classification table to see how accurately the model predicts. You can download the data for the year 2000 from the Medical Expenditure Panel Survey website (U.S. Department of Health and Human Services 2013b).

Similar to how the DATA step was used in the previous example, you can read in the 2000 data using the following DATA step:

```
libname mylib '';
filename in1 'c:\SGF\H50.SSP';
proc xcopy in=in1 out=mylib import;
run;
data meps2000;
set mylib.H50;
   /*Rename the variables in the model so they can be used in the scoring
     equation.*/
inscov99=inscov00;
totexp99=totexp00;
povcat99=povcat00;
varstr99=varstr00;
varpsu99=varpsu00;
perwt99f=perwt00f;
format povcat99 povcat9h. inscov99 inscov9f.;
keep INSCOV99 TOTEXP99 POVCAT99 VARSTR99 VARPSU99 PERWT99F;;
run;
```

The same model above is fit to the data from 1999, but the STORE statement is added to store the parameter estimates. The STORE statement saves the context and results of the model to an item store called MEPS_MODEL, which includes, among other things, the parameter estimates.

```
proc surveylogistic data=meps;
stratum VARSTR99;
cluster VARPSU99;
weight PERWT99F;
class POVCAT99;
model INSCOV99 = TOTEXP99 POVCAT99;
store meps_model;
run;
```

PROC PLM is then invoked. The RESTORE= option reads in the item store that is to be processed, and the SCORE statement creates a SAS data set called SCORED_2000. Because the PREDICTED keyword is included, as well as the ILINK option which requests that the inverse logit link be applied to the linear predictor, the data set contains the predicted probabilities.

```
proc plm restore=meps_model;
score data=meps2000 out=scored_2000 predicted/ ilink;
run;
```

Although this data set contains predicted probabilities for each of the observations for the year 2000, it does not actually classify the observations into events or nonevents. In order to do this, you need to create a new variable that is based on a probability cutpoint. Typically, this cutpoint is based on some a priori knowledge about the proportion of the population with health insurance. Based on the data from 1999, PROC SURVEYFREQ shows that approximately 89% of the population has health insurance. This statistic can be saved to a SAS data set using an ODS OUTPUT statement and a WHERE statement that keeps only the percentage of insured respondents. The resulting data set has a single observation with only the insured and their percentage.

```
proc surveyfreq data=meps;
stratum VARSTR99;
cluster VARPSU99;
weight PERWT99F;
tables INSCOV99/nowt nofreq;
ods output oneway=summary_ds(where=(f_inscov99='Insured'));
run;
```

You can now use this percentage to classify each of the observations as Insured or Uninsured based on their predicted probabilities. Because this data set contains only one observation, you can perform a one-to-many merge with the data set that contains the predicted values. Taking advantage of the fact that variables that are read from input data sets are retained across observations, you can perform the merge by executing the SET statement to read the single observation data set on the first iteration of the DATA step only. The classification variable PRED is then created using an IF/THEN statement that compares the predicted probability with the cutpoint. Finally, the FREQ procedure computes the classification table.

```
data scored_2000_ctable;
set scored_2000;
if _n_=1 then set summary_ds;

cutpoint=percent/100;
if (predicted>=cutpoint) then pred='Insured  ';
else pred='Uninsured';

label pred='Predicted Response'
      Inscov99='Observed Response';
run;
proc freq data=scored_2000_ctable;
tables pred*inscov99;
run;
```

Output 9 shows the resulting classification table.

```
                        Table of pred by inscov99

                pred(Predicted Response)
                          inscov99(Observed Response)

                Frequency ,
                Percent   ,Insured ,Uninsured ,  Total
                          ,        ,          ,
                ffffffffffffffffffffffffffffff
                Insured   ,  11182 ,      721 ,  11903
                          ,  44.56 ,     2.87 ,  47.43
                ffffffffffffffffffffffffffffff
                Uninsured ,  10145 ,     3048 ,  13193
                          ,  40.42 ,    12.15 ,  52.57
                ffffffffffffffffffffffffffffff
                Total        21327       3769    25096
                             84.98      15.02   100.00
```

Output 9. The Classification Table for the Year 2000

Notice that the model correctly classified 56.71% (44.56%+12.15%) while having a false negative rate of 40.42% and a false positive rate of 2.87%.

FITTING A GEE-TYPE MODEL TO MULTINOMIAL DATA

Clusters of correlated observations can happen outside of the context of a complex survey design and are commonly referred to as *repeated measures designs*. In these settings, an experimental unity is randomly selected from the population and measurements are collected from these units on a number of occasions. Until recently, there was a limitation on analyzing repeated measures categorical data Thanks to the work of Liang and Zeger (1986), the generalized estimating equations (GEE) methodology was developed to handle this type of analysis. Within SAS, the GENMOD procedure can apply this methodology through the use of the REPEATED statement.

One limitation of PROC GENMOD is that it cannot model an unordered multinomial or generalized logit model, which is commonly used for nominal response variables. An alternative is to use PROC SURVEYLOGISTIC with the CLUSTER statement and the default Taylor series method for the variance. While this approach uses conventional maximum likelihood estimation of the parameter estimates (as opposed to the quasi-likelihood approach used by

PROC GENMOD's GEE algorithm), it will produce standard errors that adjust for the repeated measures. This approach should produce results that are very similar to those calculated using the GEE algorithm.

One of the assumptions of the GEE algorithm is that the number of clusters is large. Therefore, a significant bias can be introduced in situations where the number of clusters is small. To combat this bias, Morel (1989) developed a small-sample bias correction that is available with the VARADJUST=MOREL option in the MODEL statement. This is shown in the following example.

A nationwide chain of athletic clubs is trying to determine their clientele's exercise preference during peak hours in order to properly allocate the number of support staff (personal trainers and class instructors) at the different times of the day in their new locations. They sample 12 of their gyms during the morning and evening hours and capture the number of people who are using either the weight or cardio machines or who are taking a class.

The following statements create the data set WORK.EXERCISE and request the analysis. The LINK=GLOGIT option in the MODEL statement forms the generalized logits, and the CLUSTER statement identifies that the measurements taken within each of the 12 gyms are correlated. Because of the relatively few number of clusters (12), the Morel adjustment is used by invoking the VARADJUST=MOREL option in the MODEL statement.

```
data exercise;
input gym exercise$ time$;
datalines;
1 weight morning
1 weight morning
1 weight morning
1 weight morning
1 weight morning
…
2 weight morning
2 weight morning
2 weight morning
2 weight evening
2 weight evening
2 cardio evening
2 cardio evening
2 cardio evening
2 cardio evening
2 cardio evening
2 cardio evening
2 cardio evening
2 class morning
2 class morning
2 class morning
…
;

proc surveylogistic data=exercise;
cluster gym;
class time;
model exercise=time/link=glogit varadjust=morel;
run;
```

Summary information about the model, the variance estimation technique, and the response variable is shown in Output 10.

Model Information

Data Set	WORK.EXERCISE
Response Variable	exercise
Number of Response Levels	3
Cluster Variable	gym
Number of Clusters	12
Model	Generalized Logit
Optimization Technique	Newton-Raphson
Variance Adjustment	Morel
Upper Bound ADJBOUND	0.5
Lower Bound DEFFBOUND	1

Variance Estimation

Method	Taylor Series
Variance Adjustment	Morel
Upper Bound ADJBOUND	0.5
Lower Bound DEFFBOUND	1

Number of Observations Read	482
Number of Observations Used	482

Response Profile

Ordered Value	Exercise	Total Frequency
1	cardio	161
2	class	182
3	weight	139

Logits modeled use exercise='weight' as the reference category.

Output 10. Model Summary Information

The Type III Anaylsis of Effects reported in Output 11 shows that there is a significant difference across the logits between morning and evening peak times (p=0.0204). Note that there are 2 denominator degrees of freedom (DF) because there are two generalized logits being modeled: cardio versus weight and class versus weight.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Time	2	7.7858	0.0204

Output 11. Type III Analysis of Effects for Time

The odds ratio estimates are presented in Output 12.

```
                              Odds Ratio Estimates

                                          Point       95% Wald
Effect                         Exercise  Estimate   Confidence Limits

time evening versus morning    cardio     2.637   1.224       5.684
time evening versus morning     class     0.968   0.434       2.160
```

Output 12. Odds Ratio Estimates

For the patrons of the gym, the odds of using the cardio machine are 2.637 (95%CL 1.224, 5.684) times more than for those who use the weight training machines in the evening as opposed to the morning. Likewise, comparing class attendance to weight machine use, the patrons have practically equal odds of using them in the evening versus the morning (point estimate 0.968 with 95%CL 0.434, 2.160).

Although the effect was significant for TIME across the logits, it appears that the effect itself is also different between the two logits based on the estimates for the two odds ratios. This hypothesis can be tested by using a TEST statement. In order to apply the TEST statement when you have CLASS variables and a generalized logit model, you must understand the naming convention that is used by the procedure. The rules that the procedure uses are discussed in "The LOGISTIC Procedure: Input and Output Data Sets" (SAS Institue Inc. 2013). It is perhaps easiest to run the equivalent model in PROC LOGISTIC and create an OUTEST= data set. Then, running the CONTENTS procedure on that data set reveals the variable names.

```
proc logistic data=exercise outest=parms noprint;
class time;
model exercise=time/link=glogit;
run;

proc contents data=parms;
run;
```

The PROC CONTENTS output in Output 13 shows that the two effects are called timeevening_cardio and timeevening_class.

```
                            The CONTENTS Procedure

          Alphabetic List of Variables and Attributes

  # Variable       Type  Len Label

  5 Intercept_     Num   8   Intercept: exercise=cardio
    cardio
  6 Intercept_     Num   8   Intercept: exercise=class
    class
 10 _ESTTYPE_      Char  4   Estimation Type
  1 _LINK_         Char  8   Link Function
  9 _LNLIKE_       Num   8   Model Log Likelihood
  4 _NAME_         Char  8   Row Names for Parameter Estimates
                             and Covariance Matrix
  3 _STATUS_       Char 11   Convergence Status
  2 _TYPE_         Char  8   Type of Statistics
  7 timeevening_   Num   8   Time Evening: exercise=cardio
    cardio
  8 timeevening_   Num   8   Time Evening: exercise=class
    class
```

Output 13. PROC CONTENTS Output Revealing the Variable Names for the TEST Statement

By adding the TEST statement to the previous PROC SURVEYLOGISTIC statements, you see that the time effect is actually different between the two logits. Output 14 shows a *p*-value of 0.0185.

```
proc surveylogistic data=exercise;
cluster gym;
class time;
model exercise=time/link=glogit varadjust=morel;
time_effect_between_logits:test timeevening_cardio=timeevening_class;
run;
```

```
                        Linear Hypotheses Testing Results

                                Wald
Label                           Chi-Square    DF    Pr > ChiSq

time_effect_between_logits      5.5468        1     0.0185
```

Output 14. Results from the TEST Statement

CONCLUSION

Logistic regression models based on data from a complex survey sample can be readily fit using PROC SURVEYLOGISTIC. While PROC SURVETLOGISTIC does not have direct options to produce odds ratio estimates when there is an interaction, ROC curves, and score data sets, these can be implemented rather easily using some existing options and output data sets. Although PROC GENMOD does not offer a GEE-type model for nominal data, by applying the Taylor series approximation of the variance, PROC SURVEYLOGISTIC is able to fit this type of model.

REFERENCES

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

Center for Disease Control and Prevention. 2013. "National Survey of Family Growth" web page. Available at www.cdc.gov/nchs/nsfg/nsfg_2006_2010_puf.htm.

Liang, K.-Y. and Zeger, S.L. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, 73:13–22.

Morel, J.G. 1989. "Logistic Regression under Complex Survey Designs," *Survey Methodology*, 15:203–223.

SAS Institute Inc. 2013. "The LOGISTIC Procedure: Input and Output Data Sets." *SAS/STAT® 9.3 User's Guide*. Available at: support.sas.com/documentation/onlinedoc/stat/930/statug.pdf.

U.S. Department of Health and Human Services. 2013a. "Medical Expenditure Panel Survey" web page. Available at www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-038.

—. 2013b. "Medical Expenditure Panel Survey" web page. Available at meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-050.

ACKNOWLEDGMENTS

The author gratefully acknowledges Phil Gibbs and David Schlotzhauer for their invaluable assistance and feedback during the writing of this paper.

RECOMMENDED READING

SAS Institute Inc. 2013. *SAS/STAT® User's Guide*. Available at support.sas.com/documentation/onlinedoc/stat/930/statug.pdf.

CONTACT INFORMATION

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

Rob Agnelli
SAS Institute Inc
300 SAS Campus Drive
Cary, NC, 27513
Work Phone: 919-677-8008
E-mail: support@sas.com
Web: support.sas.com