Paper 1938-2018

Frequencies, Unequal Variance Weights, and Sampling Weights: Similarities and Differences in SAS®

Robert M. Lucas, Robert M. Lucas Consulting, Fort Collins, CO, USA

ABSTRACT

There is confusion among many SAS users on the similarity and differences in how SAS uses frequencies, sampling weights, and unequal variance weights in estimating parameters and their variances. This paper will describe the calculation details for each and compare the results using several SAS procedures. The author will also give advice on which is appropriate for different situations.

INTRODUCTION

Using frequencies or weights in SAS procedures may affect point estimates, their standard errors, and sample sizes. Consequently, the p-values for statistical tests of model parameters may change. Improper use of frequencies or weights may lead to inaccurate tests which can result in incorrect inferences or inferior models for prediction.

A challenge to understanding the impact of frequencies or weights is that their impact varies depending upon the model type and the estimators used in specific procedures. For example, the effect of weights on standard error estimates differ between linear model procedures and logistic regression procedures. The survey data analysis procedures use standard error estimators that are designed to accommodate sample designs more complex than simple random samples so they can yield different estimates of standard errors than the non-survey procedures, even for simple random samples.

To compare the effect of weights and frequencies, we used three sampling schemes: a simple random sample, a stratified sample with proportional allocation, and a stratified sample with equal allocation.

The definitions of a frequency, unequal variance weight, and sampling weight are given below:

- A **frequency** is the number of observations in a sample where all of the variables have identical values. Use the FREQ statement to name the variable that contains frequencies (SAS/STAT® 2017, p. 3738).
- An **unequal variance weight** is a weight that is proportional to the inverse of the variance of the error term in a linear model (SAS/STAT® 2017, p. 3772).
- A **sampling weight** is the inverse of the probability that the observations was selected into the sample.

The STATS option on the SURVEYSELECT procedure PROC statement includes the probability of selection and the sampling weight in the output data set (SAS/STAT® 2017, p. 9727).

UNEQUAL VARIANCE WEIGHTS

Weighted Least Squares

The equation for the general linear model is

$$Y = X\beta + \epsilon$$

where Y is the response vector, X the design matrix, β the fixed effect parameters, and ε the random error term. The assumptions are that the errors are independent, E[ε]=0 and Var[ε]=σ^2I (Graybill 1976 p. 176)

Weighted least squares (WLS) finds the values of the model parameters that minimize the objective function

$$\sum_{i=1}^{n} w_i (\hat{y}_i - yi)^2 \quad .$$

The solution is given by

$\hat{\beta}$=(X'WX)$^{-1}$ X'WY

where W is a diagonal matrix with the w_is in the WLS objective function on the diagonal. The solution is invariant to the scale of the weights. Ordinary least squares (OLS) is a special case when all weights equal one.

The variance of $\hat{\beta}$ is

$Var(\hat{\beta}) = \frac{\sigma^2}{(n-p)} (X'WX)^{-1}$.

Weighted least squares estimates with weights proportional to the inverse of the error variance are Best Linear Unbiased Estimates (BLUE) (Graybill 1976, p. 218, SAS/STAT® 2017, p. 3772). Hence the term, unequal variance weights.

IMPLEMENTING WEIGHTED LEAST SQUARES.

You can use code like what is shown below to evaluate the constant variance assumption:

```
title  "PROC GLM SRS";
proc glm data=work.&sampdata;
title2 "Output Residuals";
  class c1;
  model IntResp=c1 x1 x2 x3/solution;
  output out=glmres r=residuals;
run;
quit;
title2;
title;

title "Boxplot of Residuals";
proc boxplot data=work.glmres;
  plot residuals*c1;
run;

title "Mean, variance and Standard Deviation of Residuals";
proc means data=work.glmres mean var std;
  class c1;
run;
title;
```

The distribution of the residuals is shown below:

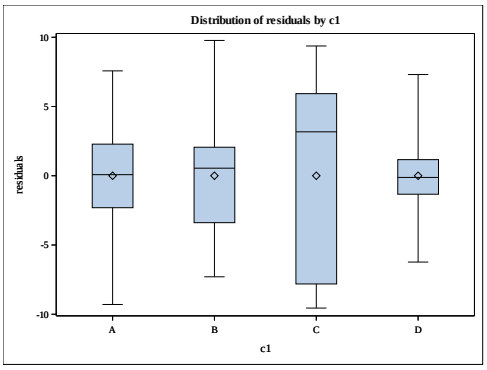

Figure 1. Boxplot of Residuals by C1

Certainly, the constant variance assumption is questionable. The estimates of the residual variance by C1 are given below:

Levels of C1	Number of Observations	Variance of Residuals
A	124	11.44
B	21	23.29
C	10	56.62
D	45	6.56

Table 1. Variance of Residuals by C1

Using the inverse of the variances above for each observation, you can fit the WLS estimates with the code below:

```
title "GLM SRS Wt=1/Var";
proc glm data=work.&sampdata;
  class c1;
  model IntResp= c1 x1 x2 x3/solution;
  weight VarInv;
  output out=VarUnEquWtRes  r=residuals;
run;
```

A more modern approach to accounting for unequal variances is to use the MIXED procedure with a REPEATED statement. PROC MIXED uses generalized least squares (GLS) to estimate the regression parameters (SAS/STAT® 2017, p.6339). WLS is a special case of GLS. For example:

```
title "PROC MIXED SRS, with Repeated Statement";
title2 "Group by C1";
proc mixed data=work.&sampdata;
  class c1;
  model IntResp = c1 x1 x2 x3/solution ;
  repeated c1/ group=c1 type=vc r;
run;
title2;
title;
```

The TYPE=VC specifies a variance component covariance matrix structure. The GROUP=C1 specifies how to group the observations to calculate the variance components.

Weighted Least Squares Results

The parameter estimates and their standard errors are tabled below:

Parameter		PROC GLM OLS Estimates	PROC GLM WLS Estimates	PROC MIXED Estimates
Intercept		100.156	100.441	100.450
c1	A	0.475	0.509	0.509
c1	B	-2.739	-2.740	-2.742
c1	C	-4.249	-3.881	-3.869
c1	D	0.000	0.000	0.000
x1		-0.407	-0.794	-0.808
x2		0.900	0.917	0.918
x3		-0.022	-0.034	-0.035

Table 2. Comparison of OLS, WLS, and GLS Parameter Estimates

Parameter		PROC GLM OLS Std. Err.	PROC GLM WLS Std. Err.	PROC MIXED Std. Err.
Intercept		0.7422	0.6028	0.5994
c1	A	0.6757	0.5224	0.5187
c1	B	1.1567	1.2671	1.2680
c1	C	1.3518	2.4431	2.5826
c1	D	.	.	.
x1		0.2903	0.2806	0.2804
x2		0.0449	0.0415	0.0414
x3		0.0311	0.0323	0.0323

Table 3. Comparison of OLS, WLS, and GLS Standard Error Estimates

You can see that the WLS and PROC MIXED estimates agree more with each other than with the OLS estimates. The WLS and PROC MIXED estimates are more accurate because they both account for the fact that the variance of the error term is not constant.

The table below compares the variance of the OLS and WLS residuals by the levels of C1 with the variance component estimates from PROC MIXED:

Levels of C1	PROC GLM Residual Variance	PROC GLM WLS Residual Variance	PROC MIXED Variance Component Estimates
A	11.445	11.293	11.480
B	23.286	23.106	23.538
C	56.620	63.223	63.922
D	6.564	6.355	6.433

Table 4. Comparison of OLS, WLS, and PROC MIXED Residual Variances

For C1=C, the WLS and PROC MIXED estimates agree but the OLS estimate is much different.

FREQUENCIES AND SAMPLING WEIGHTS IN LINEAR MODELS

SAMPLING SCENARIOS

The SURVEYSELECT procedure is used to select all samples. Comparisons are made for three sampling scenarios:

- **Scenario 1,** simple random sample (SRS) of 200 from a population of 1,000,000.
- **Scenario 2**, stratified random sample (STRS) of 200 proportionally allocated to the four levels of categorical variable C1.
- **Scenario 3**, stratified random sample of 200 with an equal allocation of 50 observation to the levels of C1.

For Scenarios 2 and 3, within each stratum, the cases were selected with a SRS. Consequently, the sampling weights are equal within each stratum but differ among the strata. The sampling weights and counts are shown below:

Levels Of C1	Scenario 1: SRS		Scenario 2: STRS, Proportional Allocation		Scenario 3: STRS, Equal Allocation	
	Weights	Count	Weights	Count	Weights	Count
A	5,000.00	124	4,973.74	126	12,533.82	50
B	5,000.00	21	4,824.68	19	1,833.38	50
C	5,000.00	10	4,918.88	17	1,672.42	50
D	5,000.00	45	4,950.48	40	3,960.38	50
Sum	1,000,000.00	200	1,000,000.00	202	1,000,000.00	200

Table 5. Sample Counts and Weights by the Levels of C1

The Scenario 1 weights are integers so comparing results based on frequencies or weights is straight-forward. The Scenario 2 sample size is greater than 200 because the SAMPRATE= option is used. PROC SURVEYSELECT rounds up the non-integer allocation based on the sampling rate to calculate the sample size for each stratum. The sampling weights are all slightly less than 5000 and are noninteger values. Because the population counts vary substantially by the levels of C1, the equal allocation weights are much different because the population distribution for the levels of C1 are not uniform.

To compare the effects of weights and frequencies on linear models, we fit models using four procedures, SURVEYREG, GLM, HPREG, and DMREG for each of the three sampling scenarios (The DMREG procedure is used in the SAS® Enterprise Miner™ Regression Node to fit linear and logistic regression models). Models are fit ignoring the weights, using a WEIGHT statement, and using a FREQ statement. PROC SURVEYREG does not have a FREQ statement, while PROC DMREG does not have a WEIGHT statement.

SAMPLE SCENARIO 1 COMPARISONS

The parameter estimates and their standard errors are summarized in the table below:

Parameter	Parameter Estimates	Std. Err. PROC SURVEYREG	Std. Err. Ignoring Weights or with Weights	Std. Err. Using Weights as Frequencies
Intercept	99.366	0.6365	0.7422	0.0103
C1=A	1.773	0.5200	0.6757	0.0094
C1=B	-4.262	1.3929	1.1567	0.0161
C1=C	-1.892	2.3586	1.3518	0.0188
C1=D	0.000	.	.	.
X1	-0.949	0.3462	0.2903	0.0040
X2	0.891	0.0443	0.0449	0.0006
X3	0.036	0.0281	0.0311	0.0004

Table 6. Parameter Estimates and Their Standard Errors for Scenario 1 Sample

Since the sampling weights are all equal and integers, the point estimates for all four procedures are equal regardless of the analysis scenario.

The PROC SURVEYREG standard error estimates are different from the other procedures because PROC SURVEYREG uses a different estimator. The SAS survey procedures use design-based estimators for finite populations based on a Taylor series linearization by default (SAS/STAT® 2017, p. 9666). Other variance estimators are available in the survey procedures. The other procedures use the model-based estimator for weighted least squares,

$$Var(\hat{\beta}) = \frac{\sigma^2}{(n-p)} (X'WX)^{-1}$$

where σ^2 is replaced by its estimate. The scale of the weights cancel out in the variance calculation thus they do not affect the estimated standard errors. The WEIGHT statement does not affect the sample size, n. The sample size, n, is the sum of the frequencies when a FREQ statement is used, so the divisor in the equation above is 1,000,000-p, not 200-p. The much greater divisor substantially decreases the standard error estimate.

Lower standard errors increase the magnitude of the test statistics, thus decreasing the p-values. The larger sample size increases the degrees of freedom for test, also decreasing p-values.

SAMPLING SCENARIO 2 COMPARISONS

The parameter estimates are compared in the table below:

Parameter	Unweighted	Weighted, PROCs SURVEYREG, GLM, and HPREG	Using Weights as Frequencies, PROCS GLM and HPREG	Using Weights as Frequencies, PROC DMREG
Intercept	99.64869	99.64974	99.64974	99.64974
C1=A	1.14732	1.14776	1.14776	1.14776
C1=B	-3.50870	-3.50737	-3.50738	-3.50737
C1=C	-4.69037	-4.68986	-4.68986	-4.68986
C1=D	0.00000	0.00000	0.00000	0.00000
X1	-0.67583	-0.67360	-0.67361	-0.67360
X2	0.91566	0.91536	0.91536	0.91536
X3	-0.00122	-0.00118	-0.00118	-0.00118

Table 7. Parameter Estimates for Scenario 2 Sample

Since the sampling weights are nearly equal, the unweighted and weighted parameter estimates differ only slightly. The GLM and HPREG procedures truncate frequencies to integers causing small differences between the weighted and frequency-based parameter estimates. The weights are large, about 5,000, so the truncation only has a minor impact. PROC DMREG does not truncate frequency values so its parameter estimates agree with the weighted estimates.

The table below compares the standard error estimates:

Parameter	"Best" Estimate, SURVEYREG with Weights and Strata	Weighted, PROCs GLM and HPREG	Using Weights as Frequencies, PROCS GLM, HPREG, and DMREG
Intercept	0.7349	0.9029	0.0126
C1=A	0.6556	0.8496	0.0119
C1=B	1.7187	1.3343	0.0186
C1=C	2.2635	1.3341	0.0186
C1=D	.	.	.
X1	0.2742	0.3581	0.0050
X2	0.0372	0.0442	0.0006
X3	0.0047	0.0191	0.0003

Table 8. Comparison of Standard Error Estimates for the Scenario 2 Sample

The PROC SURVEYREG procedure result's are labeled "Best" because they account for all the features of the sample design, unequal weights, and stratification. PROC SURVEYREG uses a design-based estimator that accounts for the stratified sample when the STRATA statement is included. The survey procedures estimate the overall variance by the weighted sum of the within-strata variances, thus excluding the among strata variation (Cochran 1963, p. 90). The other procedures cannot account for the stratification.

The standard errors using the weights as frequencies are much smaller because the sum of the frequencies is used as the sample size. The PROC DMREG estimates are slightly different than the PROC GLM or PROC HPREG estimates, but not to the number of decimal places displayed. This is because PROC DMREG does not truncate frequency values.

SAMPLING SCENARIO 3 COMPARISONS

In this scenario, the distribution of the sample by the levels of C1 is much different than the population, therefore the unweighted and weight parameters estimates will differ more than seen in Scenario 2.

Parameter	Unweighted	Weighted using PROCS SURVEYREG, GLM, or HPREG	Using Weights as Frequencies, PROC DMREG	Using Normalized Weights or Frequencies
Intercept	98.9344	99.3659	99.3659	99.3659
C1=A	1.5965	1.7733	1.7733	1.7733
C1=B	-4.5895	-4.2620	-4.2620	-4.2620
C1=C	-2.0602	-1.8920	-1.8920	-1.8920
C1=D	0.0000	0.0000	0.0000	0.0000
X1	-0.6878	-0.9488	-0.9488	-0.9488
X2	0.8985	0.8907	0.8907	0.8907
X3	0.0469	0.0357	0.0357	0.0357

Table 9. Comparison of Parameter Estimates for the Scenario 3 Sample

All the weighted estimates, as well as using weights as frequencies, agree to the thousandth decimal place. Conversely, the unweighted and weighted estimates do not agree. Some differences are relatively large; for example, the weighted parameter estimate for X1 is almost 40 percent larger in magnitude than the unweighted estimate

The normalized weights or frequencies estimates agree with the non-normalized estimates because the parameter estimates are invariant to the scale of the weights (You should not normalize frequencies except when using the DMREG procedure or other SAS® Enterprise Miner™ procedures that do not truncate frequency values to integers).

Example code for normalizing weights is shown below:

```
**  NormWt.sas;
**  Create a macro variable &sampsize that is number of rows in the table;
proc sql;
  select count(*) into: sampsize
  from work.&sampdata;
quit;
**  Create macro variable &sumwts that is the sum of the sampling weights
    or frequencies in the table;
proc sql;
  select
  sum(samplingweight) into: sumwts
  from work.&sampdata;
quit;
**  Calculate the Normalized sampling weight by scaling the sampling
    weights so that their sum equals the number of rows in the table;
data work.&sampdata;
  set work.&sampdata;
  NormWt=&sampsize*samplingweight/&sumwts;
run;
```

The table below shows the benefits of normalizing the frequencies for the DMREG procedure:

Parameter	"Best" Estimate, SURVEYREG with Weights and Strata	Weighted, PROCs GLM and HPREG	Using Weights as Frequencies, PROC DMREG	Using Normalized Weights or Frequencies. PROC DMREG
Intercept	0.852722	0.994414	0.013815	0.994414
C1=A	0.732171	0.892862	0.012404	0.892862
C1=B	1.166428	1.436102	0.019951	1.436102
C1=C	1.458550	1.409362	0.019580	1.409362
C1=D	.	.	.	.
X1	0.393228	0.296307	0.004116	0.296307
X2	0.044397	0.053197	0.000739	0.053197
X3	0.025453	0.044420	0.000617	0.044420

Table 10. Comparison of Standard Error Estimates for Scenario 3 Sample

The PROC SURVEYREG estimates are labeled “Best” because they incorporate both the unequal weights and the stratification. Because the normalized frequencies now sum to the sample size of 200, the standard error estimates from PROC DMREG now match the weighted estimates produced by PROC GLM and PROC HPREG, and more closely agree with the design-based estimates produced by PROC SURVEYREG.

FREQUENCIES AND SAMPLING WEIGHTS IN LOGISTIC REGRESSION MODELS

SAMPLING SCENARIOS

Comparisons are made for three different sampling scenarios:

- **Scenario 4**, a simple random sample (SRS) of 2000 from a population of 1,000,000,
- **Scenario 5**, a stratified random sample (STRS) of 2000 proportionally allocated to the four levels of categorical variable C1,
- **Scenario 6**, a stratified random sample of 2000, with an equal allocation of 500 observation to the levels of C1.

For Scenarios 4 and 5, within each stratum, the cases are selected with a SRS. Consequently, the sampling weights are equal within each stratum but differ among the strata.

The sample counts and weights are summarized below:

.Levels Of C1	Scenario 4 SRS		Scenario 5 STRS, Proportional Allocation		Scenario 6 STRS, Equal Allocation	
	Weights	**Count**	**Weights**	**Count**	**Weights**	**Count**
A	500.00	1270	499.75	1254	1,253.38	500
B	500.00	191	498.20	184	183.34	500
C	500.00	140	497.74	168	167.24	500
D	500.00	399	498.79	397	396.04	500
Sum	1,000,000.00	2000	1000000.00	2003	1,000,000.00	2000

Table 11. Sample Counts and Weights by the Levels of C1

The Scenario 4 weights are integers, therefore comparing results based on frequencies or weights is straight-forward. The Scenario 5 sample size is greater than 2000 because the SAMPRATE= was used. PROC SURVEYSELECT rounds up the non-integer allocation based on the sampling rate to calculate the sample size for each stratum. The sampling weights are all slightly less than 500 and noninteger values.

Because the population counts vary substantially by the levels of C1, the equal allocation weights are much different because the population distribution of the levels of C1 are not uniform.

To compare the effects of weights and frequencies on logistic regression, we fit models using four procedures, SURVEYLOGISTIC, LOGISTIC, HPLOGISTIC, and DMREG. Models are fit ignoring the weights, using a WEIGHT statement, and using a FREQ statement. PROC SURVEYLOGISTIC does have a FREQ statement, but it is not considered. PROC DMREG does not have a WEIGHT statement.

SAMPLING SCENARIO 4 COMPARISONS

Because PROC HPLOGISTIC does not have an option to use effects coding for categorical variables, less than full-rank reference coding is used for all the procedures to facilitate comparisons. Coding does not affect predicted values or odds ratios, but it does affect the parameter estimates for the intercept and levels of categorical variables and their interpretation.

The parameter estimates are summarized in the table below:

Parameter	PROC SURVEY-LOGISTIC	PROC LOGISTIC	PROC HPLOGISTIC	PROC DMREG
Intercept	-2.1922	-2.1922	-2.1921	-2.1920
C1=A	-0.3286	-0.3286	-0.3285	-0.3286
C1=B	-0.3830	-0.3830	-0.3830	-0.3830
C1=C	-0.5646	-0.5646	-0.5646	-0.5649
C1=D	0.0000	0.0000	0.0000	0.0000
X1	-0.1996	-0.1996	-0.1997	-0.1995
X2	-0.0308	-0.0308	-0.0308	-0.0308
X3	0.0046	0.0046	0.0046	0.0046

Table 12. Parameter Estimates for the Scenario 4 Sample

Estimating the parameters in a logistic regression model requires an iterative numerical algorithm. Minor differences in parameter estimates may arise among the different procedure because of differences in the details of the numerical algorithms. The small differences seen above are due to the numerical details, and not because of ignoring the weights, using the weights, or using the weights as frequencies.

The standard error estimates are tabled below:

Parameter	PROC SURVEYLOGISTIC Ignoring Weights	PROC SURVEYLOGISTIC Using Weights	PROCS LOGISTIC, HPLOGISTIC, and DMREG Ignoring Weights	PROCS LOGISTIC, HPLOGISTIC, and DMREG Using Weights or Frequencies
Intercept	0.2920	0.2920	0.2885	0.0129
C1=A	0.2545	0.2545	0.2582	0.0116
C1=B	0.4827	0.4827	0.4735	0.0212
C1=C	0.5087	0.5087	0.5037	0.0225
C1=D	.	.	.	.
X1	0.1409	0.1409	0.1398	0.0063
X2	0.0168	0.0168	0.0178	0.0008
X3	0.0092	0.0092	0.0118	0.0005

Table 13. Standard Error Estimates for Sample Scenario 4

The SURVEYLOGISTIC procedure's standard error estimates are not affected by the scale of the weights. However, this is not true for the other procedures.

PROC SURVEYLOGISTIC uses a design-based estimator, with Taylor Linearization as the default. (SAS/STAT® 2017, p. 9387). Other variance estimators are available in the survey procedures, and none are affected by the scale of the weights.

The other procedures use the inverse of the Fisher Information Matrix (SAS/STAT® 2017, p. 5564 and McCullagh and Nelder, 1983. p. 470.), a model-based estimator.

The way that weights or frequencies are incorporated into the likelihood function increase the magnitude of the Fisher Information Matrix, thus the inverse is smaller in magnitude, resulting in smaller estimates of the standard errors. Consequently, the standard error estimates are not invariant to the scale of the weights, unlike the linear modeling procedures.

The Fisher Information Matrix is complicated for logistic regression in general. However, to obtain some insight, the matrix reduces to a scaler for an intercept only model. For a simple random sample from a Bernoulli distribution of size n with parameter p, the inverse of the Fisher Information equals p(1-p)/n.

In the example, when weights are ignore the variance is p(1-p)/2000. When a WEIGHT or FREQUENCY statement is used, the sum of the weights or the frequencies is used for the sample size so the variance of the weighted estimate is p(1-p)/1,000,000.

SAMPLING SCENARIO 5 COMPARISONS

Details of the comparisons are not presented. As seen in Scenario 2, because the weights are nearly equal, parameter estimates show only minor differences among unweighted, weighted, or using the weights as frequencies. As seen in Scenario 4, using weights or frequencies yields substantially smaller estimates of the standard errors.

SAMPLING SCENARIO 6 COMPARISONS

In this scenario, the distribution of the sample by the levels of C1 is much different than the population. Consequently, the weight parameters estimates may differ greatly from the unweighted estimates as seen in the table below:

Parameter	Unweighted	Using Weights, or Frequencies
Intercept	-2.283	-2.126
C1=A	0.039	0.076
C1=B	-0.349	-0.238
C1=C	-0.518	-0.458
C1=D	0.000	0.000
X1	-0.081	-0.208
X2	-0.030	-0.026
X3	0.003	-0.007

Table 14. Comparison of Parameters Estimates for the Scenario 6 Sample

Because the sampling distribution is substantially different than the population distribution, the weighted and unweighted estimates have relatively large differences in some of the parameter estimates. For example, the weighted parameter estimate for X1 is more than twice as large in magnitude as its unweighted estimate. Therefore ignoring the sampling weights will result in substantially biased estimates

The standard errors are tabled below:

Parameter	"Best" Estimates, PROC SURVEYLOGISTIC Using Weights and Strata.	PROCS LOGISTIC, HPLOGISTIC, and DMREG Using Weights or Weights as Frequencies	PROCS LOGISTIC, HPLOGISTIC, and DMREG Using Normalized Weights
Intercept	0.3079	0.0123	0.2743
C1=A	0.2775	0.0110	0.2460
C1=B	0.3545	0.0213	0.4754
C1=C	0.3204	0.0212	0.4742
C1=D	.	.	.
X1	0.1531	0.0060	0.1330
X2	0.0272	0.0009	0.0210
X3	0.0262	0.0010	0.0232

Table 15. Comparison of Standard Errors for Scenario 6 Sample

The PROC SURVEYLOGISTIC standard error estimates are labeled “Best” because the procedure uses a designed-based estimator that accounts for the unequal weights and stratification. The standard errors using weights or frequencies produced by the other procedures are much smaller. Normalizing the weights or frequencies give standard error estimates much closer to the “Best”

PROC LOGISTIC has a NORMALIZE option on the WEIGHT statement. The other procedures do not have a NORMALIZE option so the weights or frequencies must be normalized like shown in the example code in Scenario 3 above.

PROC LOGISTIC has a STRATA statement but it is designed for matched set analysis (SAS/STAT® 2017, p. 5555.), not to account for stratified sampling.

STRATIFYING BY A BINARY RESPONSE

SAMPLING SCENARIO

In the population, the distribution of the BinResp variable is 5.05% events and 94.95 % nonevents

It is common practice in predictive modeling to over-represent the rare events in the development sample. **Scenario 7** is a stratified random sample of 2000, with an equal allocation of 1000 observation to each level of BinResp. The sample counts and sampling weights are tabled below:

BinResp	Count	Sampling Weight
0	1000	949.459
1	1000	50.541
Sum	2000	1,000,000

Table 16. Sample Counts and Weights by the Levels of BinResp

The sum of the sampling weights is over all of the observations in the sample and estimates the size of the population.

Four procedure results are compared: SURVEYLOGISTIC, LOGISTIC, HPLOGISTIC, and DMREG.

SAMPLING SCENARIO 7 COMPARISONS

For this scenario, the parameter estimates and their standard errors are compared for the analytical approaches, ignoring the sampling weights and using the sampling weights. The weights are used as frequencies in PROC DMREG.

Parameter	Unweighted	Weighted, PROCS SURVEYLOGISTIC, LOGISTIC, and HPLOGISTIC	Using Weights as Frequencies, PROC DMREG	Using Normalized Weights or Frequencies
Intercept	0.466	-2.481	-2.481	-2.481
C1=A	-0.263	-0.269	-0.269	-0.269
C1=B	-0.220	-0.215	-0.215	-0.215
C1=C	0.166	0.121	0.121	0.121
C1=D	0.000	0.000	0.000	0.000
X1	0.128	0.133	0.133	0.133
X2	-0.025	-0.025	-0.025	-0.025
X3	-0.007	-0.005	-0.005	-0.005

Table 17. Comparison of Parameter Estimates for Scenario 7 Sample

The weighted parameter or frequency estimates agree to the thousandth decimal place. The unweighted estimates are similar to the weighed estimates except for the intercept term. The intercept is different because the unweighted percent of event cases is 50, but the weighted percent is 5.05.

The range of the predicted probabilities is dramatically affected by the magnitude of the intercept, as seen in the figures below:

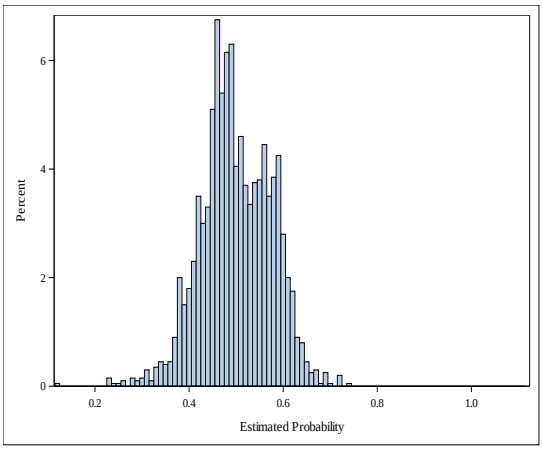

Figure 2. Distribution of Unweighted Predicted Probabilities

The unweighted predicted probabilities are distributed around the unweighted mean of 0.50.

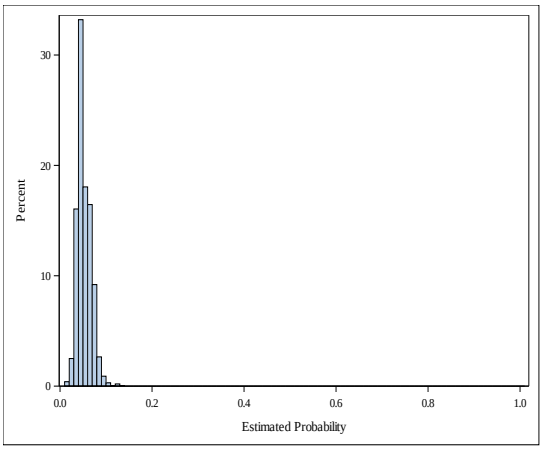

Figure 3. Distribution of the Weighted Predicted Probabilities

The weighted predicted probabilities are distributed around the weighted mean of 0.0505.

By adjusting the unweighted intercept appropriately, you can get predicted probabilities comparable to the weighted predictions. You can make the adjustment in SAS® Enterprise Miner™ with the Prior Probabilities Tab in the Target Profiler. (SAS® Enterprise Miner™2017, p. 212).

In the LOGISTIC and HPLOGISTIC procedures you can use the OFFSET= option on the MODEL statement. By choosing the offset value equal to $log(((1-\rho)\pi)/((1-\pi)\rho))$ where ρ equals the proportion of event cases in the sample and π equals the proportion of event cases in the population, the predicted probabilities are adjusted to the scale of the population mean.

The standard errors are compared in the table below:

Parameter	"Best" Estimates, PROC SURVEYLOGISTIC Using Weights and Strata.	Unweighted, PROCs LOGISTIC HPLOGISTIC, and DMREG	PROCs LOGISTIC, HPLOGISTIC, and DMREG Using Weights or Frequencies	Using Normalized Weights or Frequencies
Intercept	0.114	0.123	0.012	0.266
C1=A	0.116	0.114	0.011	0.253
C1=B	0.195	0.191	0.020	0.445
C1=C	0.198	0.198	0.019	0.420
C1=D	.	.	.	.
X1	0.046	0.044	0.004	0.095
X2	0.008	0.007	0.001	0.018
X3	0.005	0.006	0.001	0.013

Table 18. Comparison of Standard Errors for Scenario 7 Sample

The PROC SURVEYLOGISTIC estimates are labeled “Best” because they incorporate the unequal weights and stratification. The unweighted estimates are very similar to the “Best” estimates. As seen before, the un-normalized weights or frequencies are much smaller. The estimates based on the normalized weights or frequencies tend to be about twice as large as the “Best” estimates.

CONCLUSIONS

The impact of weights or frequencies varies by type of estimate, types of procedures, and class of models. Important differences and recommendations are given below.

UNEQUAL VARIANCE WEIGHTS

When the random error variances are not equal, you can use weighted least squares to obtain Best Linear Unbiased Estimates (BLUE) for linear regression parameters. You should use PROC MIXED with a REPEATED statement instead because PROC MIXED simultaneously estimates the regression parameters and variance components.

FREQUENCIES AND SAMPLING WEIGHTS

Regression Coefficients

Both linear and logistic regression coefficients are invariant to the scale of the weights. If the sampling weights are substantially different among the sample observations, you should do a weighted analysis to get more accurate estimates of population parameters. Because PROC DMREG does not have a weight statement, you should normalize the weights so they sum up to the sample size and use the normalized weights as frequencies.

Standard Errors of Regression Coefficients

Statistical Inference Applications

The survey procedures use design-based estimators that account for complex features of a sample design including unequal weights and stratification. For designs more complex than a simple random sample, you should use a survey procedure to obtain the "best" estimates of the standard errors of regression coefficients and their test for significance.

Predictive Modeling Applications

The survey procedures do not have automated methods for model selection because they are design for finite population inferences. Automated model selection is necessary in many data mining/predictive modeling applications. Below we give recommendations for doing automated model selection with weights.

Linear Regression Models

Standard error estimates are invariant to the scale of the weights. Frequencies may substantially reduce the estimates because the sum or the frequencies is used as the sample size for estimating the standard errors and calculating degrees of freedom for test of significance. Never use weights as frequencies except in SAS Enterprise Miner™ nodes or procedures that do not have a WEIGHT statement. You should use the normalized weights as frequencies.

Logistic Regression Models.

Standard error estimates are **not** invariant to the scale of the weights.

When you stratify by an independent variable and do not do proportional allocation, you should normalize the weights so that they sum up to the sample size. In PROC DMREG, you should use the normalized weights as frequencies.

When you stratify by a binary response variable and do not do proportional allocation, you have two choices. The first is use normalized weights as described above. The second is to adjust the predicted probabilities in Enterprise Miner™ with its adjustment for prior probabilities feature or use the OFFSET= option in PROC LOGISTIC or PROC HPLOGISTIC.

REFERENCES

Cochran, W.G. 1963. *Sampling Techniques, 2nd Edition*, New York: John Wiley & Sons, Inc.
Graybill, Franklin A. 1976. *Theory and Applications of the Linear Model*. North Scituate, Massachusetts: Duxbury Press
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models, Second Edition*. New York: Chapman & Hall
SAS Institute Inc. 2017. SAS® Enterprise Miner™ 14.3 Reference Help. Cary, NC: SAS Institute Inc.
SAS Institute Inc. 2017. SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

I thank my son, Henry Trey Lucas, for his technical review and helpful comments.

CONTACT INFORMATION

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

Robert M. Lucas, PhD
Robert M Lucas Consulting
robertlucas1972@gmail.com