Paper 264-28

Estimating Standard Errors for CLASS Variables in Generalized Linear Models Using PROC IML

David J. Pasta, DMA Corporation, Palo Alto, CA
Miriam G. Cisternas, MGC Data Services, Carlsbad, CA

ABSTRACT

Analysts who work frequently with linear models using the SAS System® often use the LSMEANS statement in PROC GLM or PROC MIXED to obtain "adjusted means." The absence of a clear analogue to these least-squares means for nonlinear or generalized linear models can be an obstacle for both the analyst and the eventual audience of the analysis. This paper shows how to use PROC IML to estimate standard errors of CLASS variable effects from a logit model and from a linear model on log-transformed data. The macro used is designed for logit or log-transformed linear models used in the analysis of health care claims data but it is easily adapted to other contexts. The log-transformation example involves a linear model on the logarithm of a continuous variable (health care costs) for which the effect of interest was the difference in cost, not the difference in the logarithm of cost. The retransformation of log(cost) back to the original units incorporates Duan's (JASA 1983) "smearing estimate" to compensate for the bias that usually arises in the retransformation. A partial check on the correctness of the PROC IML code can be obtained by comparing results from PROC IML with results from PROC GLM or PROC MIXED in the linear case. Once the principles are understood and the basic PROC IML manipulations are coded, it is easy to adapt the method to other studies and to other models. An earlier version of this paper was presented at the 6th Annual Western Users of SAS Software conference (Pasta, Cisternas, and Williamson, 1998).

INTRODUCTION

Most data analysts and most audiences of statistical analyses are comfortable with linear models to some degree. In its simplest form, the linear model for comparing two groups is obtained by calculating the mean value for each group and subtracting one mean from the other. In the context of a randomized experiment with one "untreated" (or placebo) group and one "treated" group, this mean difference is the unadjusted estimate of the treatment effect.

More complicated linear models designed to assess treatment effects include linear regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA). For these models, it has become common to compute least-squares means, also called adjusted means, for the untreated and treated groups. Those adjusted means and their difference, which is the adjusted estimate of the treatment effect, are often reported along with their standard errors or associated confidence intervals. These values can be interpreted as the estimated mean values after adjusting for the effect of the other variables included in the linear model.

As sophisticated statistical methodology becomes more and more widespread, in part because of the proliferation of the associated statistical software, various extensions to the general linear model have become increasingly commonplace. One extension to the general linear model is to perform a nonlinear transformation on the response variable before estimating an ordinary linear model. For example, in studies that analyze health care costs, it has become common to employ logarithmic transformations of the cost data.

Another form of extension to the general linear model is the generalized linear model as popularized by Nelder and Wedderburn (1972). In the generalized linear model, a nonlinear function of the response variable is assumed to be linearly related to a set of predictor variables. The choice of the so-called link function that specifies the relationship between the response variable and the predictors determines the name given to the analysis. When the logit function is used, defined as logit(p)=log(p/(1-p)), the analysis is called logit regression or more commonly logistic regression. When the link function is the inverse of the cumulative normal distribution, the analysis is called a normit regression model or a probit regression model.

A practical difficulty of working with these extensions to the general linear model is that there is no clear analogue to the least-squares means that can be obtained from the LSMEANS statement of PROC GLM or PROC MIXED. Those adjusted means are convenient for reporting the results of general linear models. In order to determine appropriate analogues for nonlinear and generalized linear models and to estimate their standard errors, it is first necessary to take a closer look at linear models and least-squares means.

LINEAR MODELS

The mathematics and statistics of linear regression, analysis of variance, and analysis of covariance are essentially identical; the primary differences have more to do with terminology than substance. In multiple linear regression you model a quantitative response variable as a linear combination of quantitative predictor variables. The model is linear in the sense that the relationship between the response variable and any predictor variable in the model is a straight line. It is permitted for the predictor variables to be nonlinearly related to other variables, which may or may not be included in the model. For example, the total cost of a hospital stay (the response variable) might be predicted by a linear model containing an intercept and three predictors: the length of stay, the square of the length of stay, and the cube of the length of stay. Although such a regression equation is often referred to as curvilinear, it is linear in the predictors and therefore can be considered a linear model.

In analysis of variance you model a quantitative response variable as a linear combination of effects due to qualitative factors (main effects of predictors) and combinations of factors (interactions). The possible values of a factor are called the levels of the factor and they do not need to be quantitative or even ordered. For example, the total cost of a hospital stay might be predicted by an analysis of variance model with two factors, the hospital type and the gender of the patient. The hospital type might have three levels, “public,” “private,” and “Veterans Administration.” Gender would have two levels, “male” and “female.”

Analysis of covariance in its most general form is a linear model of a quantitative response variable with both qualitative factors (as for ANOVA) and quantitative variables (as for linear regression) as predictors. This combination of qualitative and quantitative predictors is also called the “general linear model,” and some authors limit the term ANCOVA to certain special cases of the general linear model that emphasize qualitative factors and only one or sometimes two quantitative variables (covariates).

The SAS System includes many procedures for estimating linear models, each with slightly different features. For models where all or almost all of the predictors are quantitative, regression procedures such as PROC REG may be the most convenient. Qualitative predictors can be accommodated in the linear regression context by the inclusion of one or more appropriately-coded binary variables (variables that take on only two values). To represent a qualitative variable with k levels, usually “dummy-variable coding” or “effect coding” is used to construct k-1 variables that take on the value 0 or 1. For models where all of the variables are qualitative, analysis of variance procedures such as PROC ANOVA may be the most convenient. When many but not all of the predictor variables are qualitative, the more general procedures PROC GLM and PROC MIXED will almost certainly be the most convenient. The qualitative factors are specified in the CLASS statement of these procedures.

LEAST-SQUARES MEANS

In PROC GLM and PROC MIXED, the LSMEANS statement provides a convenient way to obtain least-squares means for even complicated general linear models. These least-squares means, often called adjusted means, can be thought of as the mean value that would have been obtained in a design that was balanced in a certain way (which depends on whether OBSMARGINS is specified on the LSMEANS statement). In the context of the randomized experiment with a treated group and an untreated group, the least-squares means of primary interest are usually those for the treated and untreated groups. Although the difference between these two least-squares means depends only on the specification of the model, the least-squares means themselves depend on the exact sense in which the model is assumed to be balanced. In particular, the presence of the OBSMARGINS (or OM) option on the LSMEANS statement can change the adjusted means substantially or even dramatically. For additional details on least-squares means and the OBSMARGINS option, see Potter and Pasta (1997).

For simple models without any interactions between the treatment effect and other predictors, the least-squares means for the two levels of the treatment variable with the OM option can be calculated from a model in either of two equivalent ways. One way is to set every variable except the treatment variable at its mean value and calculate the predicted value under the model as though this observation were in the treatment group, and then calculate the predicted value again as though this observation were in the untreated group. This method will be referred to as the “predicted value of the mean” approach. The precise definition of “setting a variable at its mean value” for qualitative variables has to do with the details of the OM option; it may be easiest to think of creating a set of dummy variables and taking the mean of those variables.

The other way to calculate the least-squares means is called the “mean of the predicted values” approach. In this approach, two predicted values are computed for all of the observations in the input data set, once with the treatment variable coded as though the observation were in the treated group and once as though the observation were in the untreated group. The means of those predicted values are the least-squares means for the treated group and the untreated group, respectively.

The important thing to understand is that these two methods are equivalent for the linear model, but generally are not equivalent for nonlinear models. Either method extends to nonlinear models and to the case where there are interactions between the treatment factor and other predictors. The second method, the “mean of the predicted values” method, is perhaps easier to extend. The treatment variable and all other variables (including interactions) derived from the treatment variable can be coded first as though all the observations are in the untreated group and then again as though all were in the treated group.

These methods of calculating least-squares means in the linear model do not apply to the original definition of least-squares means used in the SAS System before the introduction of the OBSMARGINS option. Without the OM option, the least-squares means assume that the observed data is spread equally over each of the possible levels of each qualitative factor. This can lead to unreasonable estimated means, as shown in Potter and Pasta (1997).

STANDARD ERRORS OF TREATMENT EFFECTS

The variance or standard error of a function of the linear model parameters can be estimated in various ways. When the function in question is a linear function, as it is for estimating treatment effects in general linear models, those methods generally reduce to formulas that are easily derived in the simple linear regression case. For more complicated models, it is convenient to write the formulas using matrix algebra. When one of the extensions to the general linear model is used, the variance can be approximated using the delta method. Rutten-van Molken et al. (1994) provide an illustration of this in the case of the logarithmic transformation.

For least-squares means for the untreated and treated groups in the linear case, the estimated variance is

$$\textbf{(1)}\quad \frac{1}{n}\; D'XVX'D\; \frac{1}{n}$$

where D is an n x 1 column vector of 1s, X is the n x k matrix of data values, and V is the estimated variance-covariance matrix of the parameters. Obtaining the variance of the difference between the two treatment groups entails calculating two predicted values for each observation, one as though it was in the untreated group and one as though it was in the treated group. The variance of the difference between the treated group “T” and the untreated group “U,” adjusted for the other variables, is

$$\textbf{(2)}\quad \frac{1}{n}\left[D_T' \,|\, -D_U'\right]\left[\begin{array}{c}X_T\\ \hline X_U\end{array}\right] V \left[X_T' \,|\, X_U'\right]\left[\begin{array}{c}D_T\\ \hline -D_U\end{array}\right]\frac{1}{n}$$

where D_T is a n x 1 column vector of 1s and $-D_U$ is an n x 1 column vector of -1s. The reason for this somewhat awkward notation will become clear when the model is generalized.

The X_T data matrix has all the observations coded as though the subjects were in the treatment group, and the X_U data matrix has all the observations coded as though the subjects were in the untreated group. The V matrix is the variance-covariance matrix of the parameters and generally can be obtained from the SAS procedure used to estimate the model.

CALCULATING STANDARD ERRORS WITH PROC IML

Given the data matrices and the variance-covariance matrix, it is not difficult to use PROC IML to calculate the standard error (square root of the variance) of the adjusted treatment effect according to the formula (2). This standard error should match the standard error reported by PROC REG for the treatment effect. We have created a macro to perform the calculations but it is too lengthy to reproduce here; it can be downloaded from www.mgcdata.com/tools. To give the reader some flavor of the necessary steps, fragments of the code for implementing this formula using REG and IML is given below.

FIG 1: CODE FOR CALCULATING THE STANDARD ERROR OF THE TREATMENT EFFECT USING PROC IML

```
* perform regression ;
proc reg covout outest=estim data=analysis;
  model &dep_var = &effect_var_list &control_var;
  output out=regout(keep=resid) residual=resid;
quit;

* obtain model “n” in a macro variable;
proc summary data=regout;
    output out=modeln (keep=modeln)
        n(pred)=modeln;
run;
data _null_;
   set modeln;
   call symput(‘modeln’,left(put(modeln,14.)));
run;

* Create a dataset concatenating both the treated
and untreated versions of the analysis dataset;
data xt;
    set analysis;
    tx=1;
run;
data xu;
    set analysis;
    tx=0;
run;
data dataplusminus;
    set xt xu;
run;

* use score to get predicted values;
proc score data=dataplusminus predict type=parms
  score=estim (drop=_NAME_)
  out=datapred (rename=(model1=lin_pred));
  var intercept &effect_var_list &control_var;
run;

* calculate derivatives;
```

```
data pred_deriv;
  set datapred;
  if &effect_var eq &plus then app_sign=1;
  else if &effect_var eq &minus then app_sign=-1;
  deriv=app_sign;
  pred=lin_pred;
run;

* extract covariance matrix;
data covonly;
  set estim;
  if _type_ eq ‘PARMS’ then delete;
  drop _model_ _type_ _name_ _depvar_ _rmse_
     &dep_var;
run;

* read in the matrices using IML and calculate
standard errors ;
proc iml;
  *create a dataset for the variance scalar;
  stderr=.;
  create stderrwk from
     stderr[colname=’stderr’];
  *get the covariance matrix;
  use covonly;
  read all into cov;
  *get standard error;
  use pred_deriv;
  read all var{Intercept &effect_var_list
     &control_var} into valuen_k;
  read all var{deriv} into derivn_1;
  valuek_n = valuen_k`;
  deriv1_n = derivn_1`;
  varpt1 = deriv1_n * valuen_k * cov * valuek_n
     * derivn_1 ;
  stderr=sqrt(varpt1)/&modeln;
  *put into external dataset;
  append from stderr;
quit;
```

GENERALIZING LEAST-SQUARES MEANS

With these formulations of the least-squares means and this way of expressing the formula for the standard error of the adjusted treatment effect, it is not difficult to express generalizations to the case of nonlinear models or generalized linear models. One analogue to the least-squares mean for the treatment groups is to calculate the predicted value under the generalized model for an observation with all variables set at their mean except the treatment variable, which is set once as though the observation were in the treated group and once as though it were in the untreated group. The other approach is to calculate predicted values for all the observations as though they were from the treated group and calculate the mean of those values and then repeat the process for the untreated group.

For the linear model, the two methods were equivalent. For the nonlinear and generalized linear models, which method is better? This is to some extent a matter of taste, but we believe that the second method, the “mean of the predicted values” provides a better analogue to the least-squares means of linear models. It provides, in a very specific sense, a mean value for each group that reflects the observed values of the predictor variables and adjusts for any imbalance between the treatment groups. It also has the advantage of calculating predicted values for plausible values of the predictor variables, rather than for some mythical hybrid observation that is, for example, 42% male and 58% female and also 27% a person with no college, 41% a person with some college, and 32% a college graduate. Another argument in support of this method is that it is closely related to the smearing estimate described below.

STANDARD ERRORS FOR NONLINEAR AND GENERALIZED LINEAR MODELS

We would like to be able to estimate standard errors for adjusted means calculated by the recommended method, the “mean of the predicted values” approach. It turns out that the estimation of the standard error of those treatment means and the standard error of the difference between treatments is not difficult. Using the delta method, the variance of a function of the parameter vector is the variance-covariance matrix of the parameters pre- and post-multiplied by the vector of derivatives of the function with respect to the parameters. This corresponds to a simple Taylor series approximation to the function. In the context of the extended models we consider here, this means that the vector D in formulas (1) and (2) needs to correspond to the derivative of the function, evaluated at the specific observation.

In the case of a logit model, we replace the vectors of +1 and –1 with vectors of the derivatives, which are +1 or –1 times exp(xβ) divided by (1+exp(xβ))$**2$.

THE SMEARING ESTIMATE FOR RETRANSFORMED DATA

For the log-transformed data, one simple estimate of the retransformed value can be obtained from exp(xβ), for which the derivative is also exp(xβ). However, Duan (1983) has proposed a simple adjustment for the bias that results from this retransformation. This estimate, called the smearing estimate, is related to the bootstrap in that it is nonparametric and does not require the underlying distribution to have any particular form.

The smearing estimate is based on the following idea. Although the retransformed data, exp(predicted), is a reasonable estimate of the median of the original distribution, it is not a very good estimate of the mean of the original distribution which, because of the long tail of the distribution, could be substantially larger. There are ways to account for this under the assumption that the original data are lognormally distributed, but the smearing estimate works well without assuming a specific distribution. To apply the smearing estimate, one simply calculates the mean value of the retransformed residuals, i.e. the mean of exp(residual), across all the observations. Although for a least-squares model the mean of the residuals will be zero (assuming a constant is included in the model), it is not necessarily the case that the mean of the retransformed residuals will be one. In fact, it will (when a log transform is appropriate), nearly always be greater than one, possibly substantially greater. The smearing estimate of predicted values is calculated by multiplying the retransformed values, exp(predicted), by the mean of the exponentiated residuals.

STANDARD ERRORS USING PROC IML

For the log-transformed and the logit models, the modification to the PROC IML code to accommodate those models is quite straightforward, as shown below.

FIG 2. STANDARD ERRORS FOR LOG-TRANSFORMED DATA

```
* get smearing estimate;
data expresid;
  set regout;
  expresid=exp(resid);
run;

proc summary data=expresid;
  var expresid;
  output out=smeared(keep=smeared)
     mean(expresid)=smeared;
run;

* calculate derivatives: note that datapred
dataset is created as is Fig 1 so the code is not
reproduced here;
data pred_deriv;
  set datapred;
  if _n_=1 then set smeared;
  if &effect_var eq &plus then app_sign=1;
  else if &effect_var eq &minus then app_sign=-1;
  exp_lin_pred=exp(lin_pred);
  deriv=app_sign*smeared*exp_lin_pred;
  pred=smeared*exp_lin_pred;
run;
```

FIG 3. STANDARD ERRORS FOR LOGIT MODEL

```
* logit version;
proc logistic descending covout outest=estim
   data=analysis;
  model &dep_var=&effect_var_list &control_var;
run;

* the dataplusminus dataset is created dataset is
created as in Fig 1 so the code is not reproduced
here;
proc score data=dataplusminus(drop=&dep_var)
   score=estim predict type=parms
   out=datapred(rename=(&dep_var=lin_pred));
var intercept &effect_var_list &control_var;
run;

* calculate derivatives;
data pred_deriv;
  set datapred;
  if &effect_var eq &plus then app_sign=1;
  else if &effect_var eq &minus then app_sign=-1;
  exp_lin_pred=exp(lin_pred);
  deriv=app_sign*exp_lin_pred/(1+exp_lin_pred)**2;
  pred=exp_lin_pred/1+exp_lin_pred);
run;
```

CONCLUSION

There are two reasonable extensions of “adjusted means” to nonlinear and generalized linear models. These two methods, the “predicted value of the mean” and the “mean of the predicted values” methods, give the same answers for the linear model but different answers for the nonlinear and generalized linear models. We recommend use of the second method, the “mean of the predicted values.” The code for calculating adjusted means by this method and associated standard errors is not difficult once the initial structure is prepared. That initial structure can be tested by comparing the results from PROC IML with the results from the REG, GLM, or MIXED procedures for linear models. The modifications needed for log-transformed and the logit models are only a few lines of code. We recommend that the smearing estimate be used when retransforming nonlinear models back to the original scale and that standard errors be routinely calculated for nonlinear and generalized linear models using the methods described here.

REFERENCES

Duan, Naihua (1983), “Smearing Estimate: A Nonparametric Retransformation Method,” Journal of the American Statistical Association, 78, 605-10.

Nelder, J.A. and Wedderburn, R.W.M. (1972), “Generalized Linear Model,” Journal of the Royal Statistical Society, Series A, 135, 761-8.

Pasta, David J., Cisternas, Miriam G., and Williamson, Cynthia L. (1998), “Estimating Standard Errors of Treatment Effects for Probit Models and for Linear Models of Log-Transformed Variables using PROC IML,” Proceedings of the 6th Annual Western Users of SAS Software Regional Users Group Conference, 211-216.

Potter, Lori and Pasta, David J. (1997), “The Sums of Squares Are All the Same – How Can the LSMEANS Be So Different?”, Proceedings of the 5th Annual Western Users of SAS Software Regional Users Group Conference, 187-92.

Rutten-van Molken, Maureen P.M.H., van Doorslaer, Eddy K.A., and van Vliet, Rene C.J.A. (1994), “Statistical Analysis of Cost Outcomes in a Randomized Controlled Clinical Trial,” Health Economics, 3, 333-45.

SAS Institute Inc. SAS/STAT Software, Changes and Enhancements through Release 6.11. Cary, NC. SAS Institute Inc. 1990.

ACKNOWLEDGMENTS

SAS and SAS/STAT software are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

We would like to thank Ed Yelin and Laura Trupin of the University of California at San Francisco for their collaboration in the development of the macro that estimates the standard errors for various levels of the class variables.

AUTHOR CONTACT

David J. Pasta, President
dmacorp@pacbell.net

DMA Corporation
2970 South Court
Palo Alto, CA 94306
(650) 213-9106 (phone)
(650) 213-9125 (fax)

Miriam G. Cisternas, Partner
miriam@mgcdata.com

MGC Data Services
5051 Millay Court
Carlsbad, CA 92008
(760) 804-5746 (phone)
www.mgcdata.com