Paper 272-2010

Using "Recycled Predictions" for Computing Marginal Effects

Zhongmin Li and Geeta Mahendra
University of California Davis Medical Center

ABSTRACT

A conventional way to estimate the marginal effect of a risk factor is to omit the risk factor (e.g., gender, treatment) from the multivariate model and then use the omitted risk factor as class/group variable to compare observed and predicted outcomes. This can have an undesirable impact on the model since the predicted outcomes could be incorrect if the omitted variable is a significant risk factor. An alternative approach is to use the recycled prediction method to estimate and compare marginal effects without removing the risk factor from the model. While STATA (StataCorp, 2005) and SUDAAN (Research Triangle Institute, 2004) have provided sub-routines for the recycled prediction method, SAS® does not have a dedicated procedure for this purpose. This can be accomplished, however, by using simple SAS programming techniques. In this paper, we use an example to demonstrate a way to perform the recycled predictions within the SAS platform.

INTRODUCTION

Marginal effects measure the expected instantaneous change in the dependent variable as a function of a change in a specific explanatory variable while keeping all other covariates constant.

A common way to estimate the marginal effect of a specific risk factor is to omit the risk factor (e.g., gender, specific treatment) from the multivariate model, and then use the omitted risk factor as a class/group variable by which observed and predicted outcomes can be compared. This, however, can have an undesirable impact on the remaining model, because if the omitted variable is a significant risk factor, the predicted outcome for all cases could be incorrect. An alternative way is to estimate the marginal effects without removing the risk factor from the model and use the recycled prediction method. While STATA (StataCorp, 2005) and SUDAAN (Research Triangle Institute, 2004) have provided sub-routines for the recycled prediction method, SAS does not have a dedicated procedure for this purpose. However by using simple SAS programming this can be accomplished. In this paper, we use an example to demonstrate a way to perform the recycled prediction within the SAS platform.

Example: Gender Effect on Operative Mortality following Coronary Artery Bypass Surgery

The data are from the California Coronary Artery Bypass Graft (CABG) outcomes reporting program (CCORP, 2003-04) where operative mortality is reported as a binary outcome (0/1). To study the marginal effects of gender on the patients' operative mortality[1] following the CABG surgery, the first step is to read the data into SAS using the following data steps.

```
DATA cabg;
  INPUT pid $1-6 gender age
        acuity creatinine hyperTn dialysis PVD CVD diabetes … y
;
DATALINES;
000001 M 65 1 2.2 1 1 0 1 1 … 0
000002 F 72 3 2.6 0 1 0 1 0 … 1

... more lines ...

157808 F 80 4 3.1 1 1 0 1 1 … 1
;
```

[1] Operative mortality is defined as patient death occurring in the hospital after CABG surgery regardless of the length of stay, or death occurring anywhere after hospital discharge but within 30 days of the CABG surgery.

```
run;
```

The dependent variable operative mortality y is modeled as follows:

$$y_i = \pi_i + e_i,$$
$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i \quad (1)$$

where $i = 1,\ldots, I$ is the patient level indicator, and π_i is the probability of death for patient i, conditional on the risk factor x, including gender. The logit model assumes that patient level random errors e_i are independent with moments $E(e_i) = 0$ and $Var(e_i) = \sigma_e^2 = \pi_i(1-\pi_i)$. The logit model has a linear function at the logit (log odds) scale. Equation (1) implies that the probability function is

$$\pi_i = \frac{\exp(\alpha + \beta x_i)}{1+\exp(\alpha + \beta x_i)} \quad (2)$$

Using the multivariate logistic regression model, we empirically tested and found that gender difference is an important risk factor in predicting patient death following CABG surgery. However, both raw logistic regression coefficients and odds ratios are nonlinear expressions of the impact of individual covariates on the response variables. As such, they provide an imperfect picture of the impact of gender difference on probabilities of operative death. The ‘‘method of recycled predictions’’ (StataCorp, 2005), also referred to as ‘‘averaging the individual marginal effects’’ (Greene, 2002) and ‘‘predictive margins’’ (Graubard, Edward, and Korn, 1999) provides more easily interpreted statistics than raw logistic regression coefficients.

To evaluate the "average" or "overall" marginal effect, two approaches are frequently used. One approach is to compute the marginal effect at the level of the sample means. The other approach is to compute marginal effects for each observation and then calculate the sample average of these effects to obtain the overall marginal effect. Both the approaches yield similar results for large samples. For smaller samples, however, averaging the individual marginal effects is preferred (Greene, 1997, p. 876).

Marginal effects measure the instantaneous effect on the predicted probability of y due to a change in a particular explanatory variable, while keeping the other covariates fixed. We accomplish this task by using the logistic regression procedure in SAS version 9.2 along with a user-created SAS program to generate the recycled predictions. The recycled predictions method calculates the mean predicted marginal probability of death by gender, thus allowing us to compare the patients’ predicted marginal probability of death, on average, while holding constant all other model covariates except gender. We can accomplish this task with the following three steps:

Step 1: Develop risk model and output model parameters:

```
…
  PROC LOGISTIC data=cabg Descending outmodel=parms_OpDeath;
  CLASS gender (ref='M') race (ref='0: White') bmiM
        acuity dialysis diabetes
        PVD CVD cvaWhen (ref='No CVA')
        hyperTn ImmSupp ArrhyTyp (ref='None')
        CLD miWhen carshock
        chf NYHA prcIndex pciint NumDisV VDInsufM
        /PARAM=REF REF=FIRST;
  MODEL y  = gender
             age_n race bmiM
              acuity
             creatLSTPLM hyperTn dialysis PVD CVD cvaWhen
             diabetes cld Immsupp arrhyTyp
             miWhen carshock
             chf NYHA
```

```
              prcIndex pciint efpl2003m lmSten2003M
              NumDisV VDInsufM
/PARMLABEL LACKFIT RSQUARE ctable pprob=(0.05 to 0.5 by 0.05)    RIDGING=NONE;
RUN;
```

Figure 1: Logistic model fit statistics

Model Fit Statistics			
Criterion	Intercept Only	Intercept and Covariates	
AIC	10053.301	8378.175	
SC	10061.806	8752.407	
-2 Log L	10051.301	8290.175	
Responses			
Pairs	39840132	c	0.818

Partition for the Hosmer and Lemeshow Test			
Group	Total	y = 1	
		Observed	Expected
1	3651	12	11.05
2	3652	19	19.24
3	3653	17	27.14
4	3651	31	36.26
5	3652	48	47.30
6	3652	55	62.01
7	3651	85	82.60
8	3651	111	115.85
9	3651	203	182.56
10	3644	545	541.98

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.2006	8	0.4141

Figure 2: Partial display of parameter estimates output

Odds Ratio Estimates		
Effect	Point	95% Wald

	Estimate	*Confidence Limits*	
Gender: Female vs Male	**1.612**	**1.408**	**1.845**
Age	1.058	1.051	1.066
Acuity: Urgent vs Elective	1.528	1.281	1.823
Acuity: Emergent vs Elective	2.387	1.794	3.176
acuity: Salvage vs Elective	12.215	6.686	22.314
creatLSTPLM	2.685	2.036	3.541
HyperTn Yes vs No	1.064	0.893	1.268
dialysis Yes vs No	1.862	1.348	2.572
…			

Step 2: Use **inmodel=parms_OpDeath** to perform "recycled predictions". In the first run all patients were assumed to be "Male", and for the second run all patients were assumed to be "Female".

```
…
data genderM;
   length gender $8;
   set cabg(rename=(gender=sex));
gender='Male';
****Apply model to entire data;
PROC LOGISTIC inmodel=parms_OpDeath Descending;
    score data=genderM
          OUT=predGenderM(keep=pid sex age_n gender y p_1);
title "Applying the Risk Model to Entire Dataset Assuming all Patients are Male";
RUN;

data genderF;
   length gender $8;
   set cabg(rename=(gender=sex));
gender='Female';
****Apply model to entire data;
PROC LOGISTIC inmodel=parms_OpDeath Descending;
    score data=genderF
          OUT=predGenderF(keep=pid sex age_n gender y p_1);
title1 "Applying the Risk Model to Entire Dataset Assuming all Patients are Female";
RUN;
```

Please note the importance of using the RENAME statement to retain the original gender value for later comparison.

Step 3: Join the two prediction data sets and perform comparisons of marginal effects.

```
**** Join the data;
data genderM;
   set predGenderM;
p_Male=p_1;
keep pid sex age_n y p_Male;
proc sort; by pid;
data genderF;
   set predGenderF;
```

```
p_Female=p_1;
keep pid sex age_n y p_female;
proc sort; by pid;
run;
data join;
   merge genderM genderF;
   by pid;
RUN;
```

Now, we can use the t-test procedure to compare patients' observed operative mortality and use a paired comparison to test for the significant difference in predicted mortality between males and females:

```
title "Compare Observed Operative mortality: Male vs. Female";
PROC TTEST data=join;
   class sex;
   var Y;
RUN;

title "Compare predicted Operative mortality with recycle prediction method";
PROC MEANS data=join n mean std min max lclm uclm;
   var p_male p_female;
RUN;
PROC TTEST data=join;
   paired p_male*p_female;
RUN;
```

Figure 3: The difference in operative mortality by gender.

Compare Observed Operative Mortality: Male vs. Female

The TTEST Procedure

Statistics

Variable	*sex*	*N*	*Lower CL Mean*	*Mean*	*Upper CL Mean*
y	*Female*	10708	0.0421	**0.0460**	0.050
y	*Male*	29669	0.0235	**0.0253**	0.0271
y	*Diff (1-2)*		0.0169	0.0207	0.0245

T-Tests

Variable	*Method*	*Variances*	*DF*	*t Value*	*Pr > |t|*
y	*Pooled*	Equal	4.00E+04	10.65	<.0001
y	*Satterthwaite*	Unequal	1.50E+04	9.33	<.0001

Equality of Variances

Variable	*Method*	*Num DF*	*Den DF*	*F Value*	*Pr > F*
y	*Folded F*	10707	29668	1.78	<.0001

Compare Predicted Operative Mortality with Recycle Predictions Method

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum	Lower 95% CL for Mean	Upper 95% CL for Mean
p_Male	40377	0.0263239	0.048363	0.000223	0.949325	0.025852	0.026796
p_Female	40377	0.0399007	0.064223	0.00036	0.967944	0.039274	0.040527

Compare Predicted Operative Mortality with Recycled Predictions Method

The TTEST Procedure

Statistics				
Difference	N	Lower CL Mean	Mean	Upper CL Mean
p_Male - p_Female	40377	-0.014	-0.014	-0.013

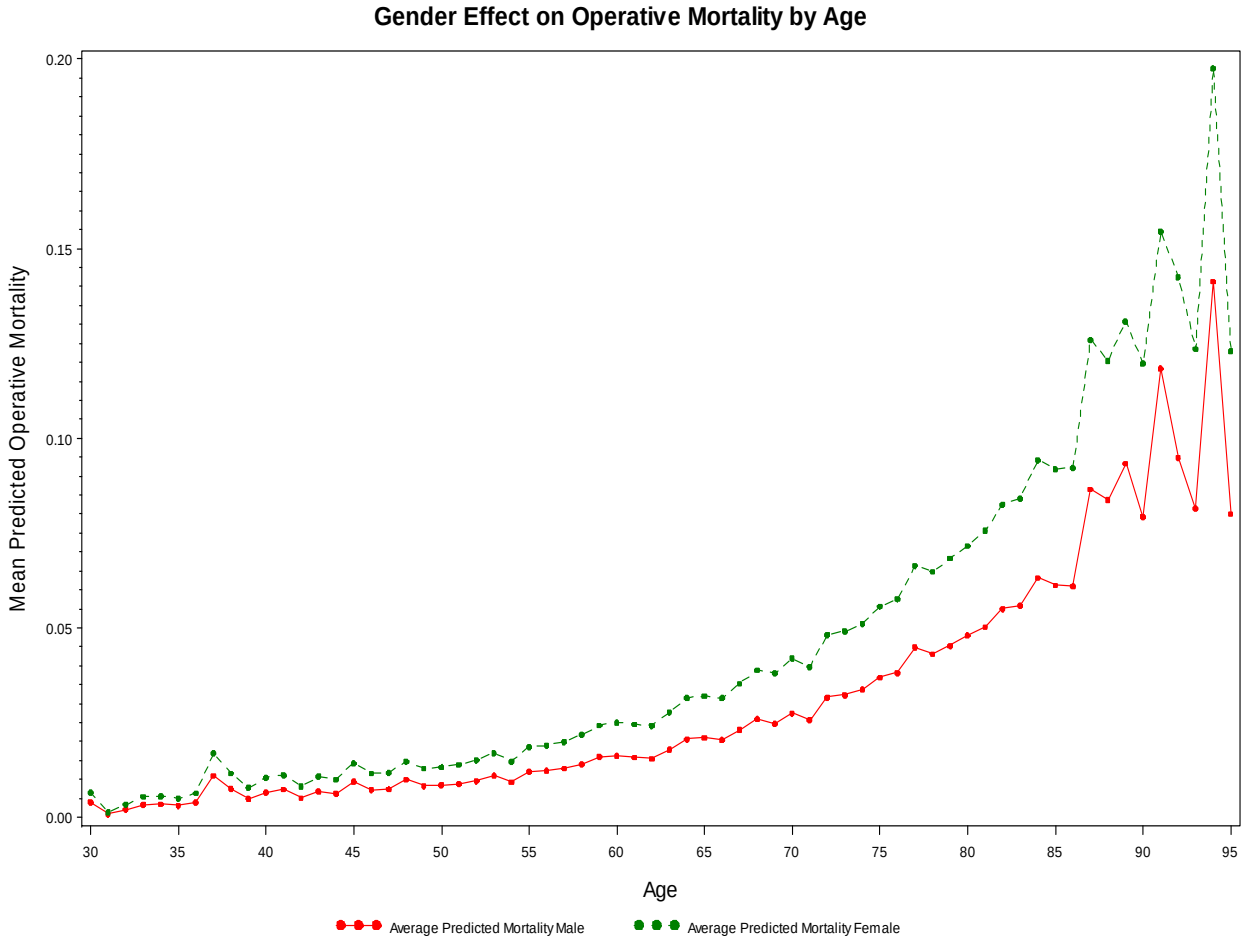
T-Tests			
Difference	DF	t Value	Pr > |t|
p_Male - p_Female	4.00E+04	-158.79	<.0001

Summary of Findings

In this example, the t-test indicates that after controlling for other risk factors, being female has a significant effect on CABG operative mortality. Notice that the ratio of the odds of predicted mortality for females (0.0399)/(1-0.0399) divided by the predicted mortality for males (0.0263)/(1-0.0263) is equal to 1.54. This result is close to the odds ratio of 1.61 from the multivariable logistic parameter estimate and confirms that the recycled method is working. Also notice that the odds ratio of the predicted mortality is less than the odds ratio of the observed mortality (i.e., (0.0460)/(1-0.0460) divided by (0.0253)/(1-0.0253) = 1.86), because women are at higher risk according to the other factors in the model (i.e., partial confounding).

Figure 4 presents the marginal effect of gender by age, on predicted operative mortality. The chart is produced by the SAS SGPLOT procedure. The chart indicates that the effect of gender on operative mortality is not evenly distributed; the gender effect is more pronounced for older patients than for younger patients.

Figure 4: Marginal effect of gender by age, on predicted operative mortality



CONCLUSION

To interpret the results of the nonlinear regression analyses, we used the method of "recycled predictions" within the SAS platform to predict marginal effects. This involves calculating predicted probabilities of the dependent variable based on the estimated model by setting key independent variables (e.g., gender in our case) to a specific value for all sample members while letting other covariates retain original values and then averaging the predictions over the entire sample. The process is repeated, using other fixed values of the key variables (Korn and Graubard, 1999). The recycled predictions method provides direct standardization and comparison of averaged predicted marginal effects.

REFERENCES

Research Triangle Institute (RTI), 2004. http://www.rti.org/sudaan/page.cfm/SUDAAN_9.

Greene, W. H. (1997), *Econometric Analysis*, Third edition, Prentice Hall, 339–350.

Graubard, B., L. Edward, and E. Korn. 1999. ''Predictive Margins with Survey Data.'' *Biometrics* 55 (2): 652–9.

Korn EL, Graubard BI. *Analysis of Health Surveys.* New York, NY: Wiley; 1999:126 –140.

Radhika N. Bukkapatnam, Khung Keong Yeo, Zhongmin Li, Ezra A. Armsterdam. Operative Mortality in

Women and Men Undergoing Coronary Artery Bypass Graft Surgery: The California Coronary Artery Bypass Graft Surgery Outcomes Reporting Program. American Journal of Cardiology (AJC), 2009; Vol. 105, Issue 3: 339-342.

ACKNOWLEDGMENTS (HEADER 1)

The authors want to thank Dr. Brian M. Paciotti at the California Office of Statewide Health Planning and Development for his paper editing and development of the SAS macro program for the applications.

CONTACT INFORMATION

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

Zhongmin Li
University of California, Davis Medical Center
4150 V Street Suite 2400
Sacramento, CA 95817
(916) 716-7736
E-mail: zmli@ucdavis.edu

Appendix A SAS Macro Code for Using “Recycled Predictions” Method

```
/************************************************************
Using “Recycled Predictions” for Computing Marginal Effects
By Zhongmin Li, Geeta Mahendra, and Brian Paciotti
Paper presented at SAS Global Forum 2010
Purpose: SAS macro allows a user to compute marginal effects for a binary
variable included in a multivariate model
************************************************************
************************************************************
DESCRIPTION OF MACRO PARAMETERS
************************************************************
DSN = Dataset Name
PARMS = Parameters from multivariate model
KEY = Unique ID or key to match observations
DEPENDENT = Dependent variable specified in model
INDEPENDENT = Independent variables separated with spaces
CLASS = Class variable of interest to calculate marginal effects
NAME_1 = User provided name for the first data element of class variable
(Name limited to 30 characters)
NAME_2 = User provided name for the second data element of class variable
(Name limited to 30 characters)
************************************************************
EXAMPLE OF MACRO CALL
************************************************************/
%MARGINAL(cabg, /* Name of original dataset */
parms_OpDeath, /* Parameter estimates from model */
medRecN, /* Key */
yom, /* Dependent variable */
age_n, /* Independent variables listed */
```

```
GENDER, /* Class variable for marginal effects */
Male, /* User-supplied name of first data element*/
Female); /* User-supplied name of second data element*/
/*************************************************************/
%MACRO MARGINAL (DSN,PARMS,KEY,DEPENDENT,INDEPENDENT,CLASS,NAME_1,NAME_2);
data OUT_1;
length &CLASS $30;
set &DSN (rename=(&CLASS = OLD_CLASS )) ;
&CLASS = "&NAME_1";
RUN;
**** Apply model to entire data;
PROC LOGISTIC inmodel= &PARMS Descending;
score data=OUT_1
OUT= PRED_1 (KEEP = &KEY &DEPENDENT &CLASS &INDEPENDENT OLD_CLASS P_1);
title "Applying the Risk Model to Entire Data Assuming Every Observation's
Class Variable Value = First Data Element of the Class Variable";
RUN;
data OUT_2;
length &CLASS $30;
set &DSN (rename=(&CLASS = OLD_CLASS )) ;
&CLASS = "&NAME_2";
RUN;
**** Apply model to entire data;
PROC LOGISTIC inmodel= &PARMS Descending;
score data=OUT_2
OUT= PRED_2 (KEEP = &KEY &DEPENDENT &CLASS &INDEPENDENT OLD_CLASS P_1);
title1 "Applying the Risk Model to Data assuming Every Observation's Class
Variable Value = Second Data Element of the Class Variable";
RUN;
**** Join the data;
data OUT_1_N;
set PRED_1;
p_&NAME_1 = p_1;
KEEP &KEY &DEPENDENT &CLASS &INDEPENDENT OLD_CLASS p_&NAME_1 ;
RUN;
proc sort data=OUT_1_N; by &KEY; RUN;
data OUT_2_N;
set PRED_2;
p_&NAME_2 = p_1;
KEEP &KEY &DEPENDENT &CLASS &INDEPENDENT OLD_CLASS p_&NAME_2 ;
RUN;
proc sort data=OUT_2_N; by &KEY; run;
data join;
merge OUT_1_N OUT_2_N;
by &KEY;
RUN;
title "Compare Observed Response Variable: Class Variable Value 1 vs. Class
Variable Value 2";
PROC TTEST data=join;
class OLD_CLASS;
var &DEPENDENT;
RUN;
title "Compare predicted Response Variable with Recycle Prediction Method";
PROC MEANS data=join n mean std min max lclm uclm;
var p_&NAME_1 p_&NAME_2 ;
```

```
RUN;
PROC TTEST data=join;
paired p_&NAME_1 * p_&NAME_2;
RUN;
%MEND MARGINAL;
```