PharmaSUG 2017 - Paper HA02

Multinomial Logistic Regression Models with SAS® PROC SURVEYLOGISTIC

Marina Komaroff, Noven Pharmaceuticals, New York, NY

ABSTRACT

Proportional odds logistic regressions are popular models to analyze data from the complex population survey design that includes strata, clusters, and weights. However, when the proportional odds assumption is violated (p-value < .05 for chi-square statistic), the use of multinomial logistic regression models for survey designs becomes challenging.

This paper provides guidance in using multinomial logistic regression models to estimate and correctly interpret the relationships between predictor and multiple levels of nominal outcome with and without interaction term. The author developed a SAS MACRO utilizing PROC SYRVEYLOGISTIC that will help researchers to conduct statistical analyses. The U.S. National Health and Nutrition Examination Survey (NHANES) is a probability sample of the US population. These data sets were used in the examples of multinomial logistic regression modeling techniques.

Statistical analysis was conducted using the SAS System for Windows (release 9.3; SAS Institute Inc., Cary, N.C.) The author is convinced that this paper will be useful to SAS-friendly researchers who analyze the complex population survey data with multinomial logistic regression models.

INTRODUCTION

Multinomial logistic regressions model log odds of the nominal outcome variable as a linear combination of the predictors. A multivariate method for multinomial outcome variable compares one for each pair of outcomes. For example, if the outcome variable has three categories then two models are tested with multinomial regression comparing simultaneously the second and third level versus the first (reference). The ratio of the probability of one outcome category over the probability of the reference category is often referred to as relative risk or odds, and regression coefficients are relative risk ratios or odds ratios for a unit change in the predictor variable.

The complexity increases when multinomial models are applied to data from population survey designs. The recent updates in PROC SURVEYLOGISTIC made the use of multinomial logistic regressions more inviting, but left users with challenging interpretations of the results.

This paper concentrates on use and interpretation of the results from multinomial logistic regression models utilizing PROC SURVEYLOGISTIC. The user-friendly SAS MACRO written by the author can easily be applied for analysis of different research questions.

DESCRIPTION

DATABASE

Eight cross-sectional (NHANES 2-year cycles) data sets were concatenated to examine relationships between predictor and multinomial outcome. Eight time points (NHANES cycles from 1999-2000 through 2013-2014) were used to determine if the relationships (odds ratios) have changed over 16-year time period.

METHOD

The analysis was conducted by multinomial logistic regression models across all surveys that accommodated the complex multistage sample survey design utilizing appropriate sampling weights following NHANES Analytic and Reporting guidance.[1,2] The models utilized the NOMCAR option in PROC SURVEYLOGISTIC to treat missing values in the variance computation as not missing completely at random for Taylor series variance estimation.

In the models, a set of k levels of outcome variable are modeled as generalized logits that contrast each level with the reference: $GLogit_i$ { Pr [Outcome i] } = Log {Pr[Outcome i] / Pr[Outcome j] }, for Outcome i=1, 2, ..,k where j=1, 2,…, k-1, j < i estimating probability of each level versus the reference ("2 vs. 1", "3 vs. 1", etc. if 1-ref.). For the predictor variable (for example, gender), the coefficient β_{gender} = Log (odds ratio) = Log {$odds_{female}$ / $odds_{male}$}= Log[$odds_{female}$] – Log[$odds_{male}$]. This odds ratio estimates the relationship between predictor and outcome. Particularly, the odds ratio for gender estimates the ratio between odds of advanced versus early level of outcome (for example, "3 vs. 1") for females and the same odds for males.

The interaction term between predictor and time (eight NHANES cycles) can be tested. A significant interaction indicates that the relationship (odds ratio) between predictor and outcome has changed over time. In the model without interaction term, the Odds Ratio (OR) greater than 1 indicates that probability of advance versus earlier level of outcome (reference) is higher among females versus males keeping the covariates at the constant level. With the interaction term, the Logit equals to $\beta_{gender} + \beta_{gender*time,}$ where $\beta_{gender*time}$ is the coefficient of the interaction. In other words, with increase of one unit of time (NHANES cycle), the Log (OR) is adding the value of Time*$\beta_{gender*time;}$ which is the same as OR changing exponentially by multiplying exp(β_{gender}) to the (exp($\beta_{gender*time}$))^{time}, where time = 1, 2,..8. If $\beta_{gender*time}$ is close to zero which is the same as exp($\beta_{gender*time}$) is close to one, then no significant change in the relationship is observed over the years.

APPLICATION

Objectives

The objective was to estimate the association between Gender (female vs. male) and BMI categories (normal, overweight, and obese), and how the associations have changed from 1999 to 2014 for the US American adults (18 years or older).

Outcome Variable - Three levels of BMI: normal (<25 kg/m^2), overweight (25 - < 30 kg/m^2), and obese (≥30 kg/m^2).

Predictor Variable - Gender: Females versus Males (ref.).

Covariates – Age, and Race groups (NHANES[2]: 1='Non-Hispanic White', 2 = 'Non-Hispanic Black', 3 = 'Mexican American', 4 = 'Other').

EXAMPLE № 1: Evaluate the association between Gender and BMI categories and examine if the associations have changed from 1999 to 2014.

EXAMPLE № 2: Evaluate the association between Gender and BMI categories and examine if the associations have changed from 1999 to 2014 adjusting for Age and Race.

MACRO TO PERFORM MULTINOMIAL LOGISTIC REGRESSION MODELS

Parameters for Multinomial Logistic Regression Model

```
%MNLRM(   sds=A,              /* The name of SAS data set              */
          inp=gender,         /* Predictor Variable                    */
          refinp=%str(Male) , /* Reference group for Predictor         */
          outp=BMIGRP,        /* Outcome Variable                      */
          cov=%str(race),     /* Covariates: names of the variables    */
          domain=sel,         /* Domain used for selected population   */
          domainx=%str( ),    /* Population from Domain to exclude     */
          num=1  );           /* Model Number                          */
```

```
%MACRO MNLRM(sds= A,     inp=gender, refinp=%str(Male), outp=BMIGRP,
             cov=%str(), domain=sel, domainx=%str( ),   num=1  );
```

```
*RUN MODEL ***;
proc surveylogistic data=&SDS NOMCAR;
format cycle cyclef. age agef. gender genderf.
  race racef.;                /* prepare for possible covariates */
  strata sdmvstra;
  cluster sdmvpsu;
  class &inp(ref="&refinp") &cov/param=glm ;
  domain &domain ;
  model &outp(descending)  = &inp|cycle  &cov  /link=glogit expb ;
  weight mec16yr;                /* recalculate weight to combine surveys */
  %DO i=1 %To 8;
  lsmeans &inp /at cycle=&i e ilink oddsratio cl diff;
  %END;
  ods output lsmeans=LSM&num  Diffs=D&num Type3=T3A&num
             ParameterEstimates = PE&num  DomainSummary=DS&num ;
  store Model&num;
run;
quit;
Title "Table: Odds Ratios ";
proc sort data=D&num
out=D1&num(keep=effect cycle probZ &outp DOMAIN OddsRatio LOWEROR UPPEROR);
by domain &outp cycle  ; where (domain ne "&domainx");
run;

*PUT ODDS RATIOS FROM EACH TIME POINT INTO THE ONE LAST RECORD ***;
data D2&num(rename=(effect=variable));
 set D1&num;
 by domain &outp cycle;
 format response $1.;
 retain OR1-OR8 LowerOR1-LowerOR8 UpperOR1-UpperOR8 .;
 response=put(&outp, 1.);
 if first.&outp then do;
    %DO j=1 %TO 8; OR&j=.; UpperOR&j=.; LowerOR&j=.; %END;
 end;
 %DO cycle=1 %TO 8;
  if cycle=&cycle then do;
   OR&cycle=OddsRatio; UpperOR&cycle=UpperOR; LowerOR&cycle=LowerOR;
  end;
 %END;
 if last.&outp then output;
run;

*GENERATE OUTPUT ***;
filename outf&num "C:\&study\out\&pname._Model&num..rtf";
ODS RTF file=outf&num;

title "Output 1: Type 3 Analysis of Effects";
proc print data=T3A&num; where domain ne "&domainx";
run;
title "Parameter Estimates";
proc print data=PE&num;
where (domain ne "&domainx") and (ClassVal0 ne "&refinp");
run;
title "Output 2: Odds Ratios with 95% Confidence Intervals (CI)";
proc print data=D2&num; by domain ;
var variable &outp OR1-OR8 LowerOR1- LowerOR8 UpperOR1-UpperOR8 ;
run;
```

```
** FOREST PLOTS *********;
data DD&num;
 set D&num;
 reference=1;
 col1=1;
 response=&outp;
 ef1=put(cycle, cyclef.); ef2=put(cycle, cyclef.);
 if (domain ne "&domainx")  and (response > reference) and (oddsratio ne .);
 label col1="Reference "
       response="&outp" ;
run;

ODS graphics on;
proc sort data=DD&num; by domain response cycle; run;
title "Output 3: Forest Plots";
proc sgplot data=DD&num UNIFORM=all ;
 by domain response ;
 format response respf.;
 keylegend /title="";
    scatter x=oddsratio y=ef1 / xerrorlower=lowerOR xerrorupper=upperOR
    markerattrs=(symbol=DiamondFilled size=8);
    refline 1 / NAME= "BMI" axis=x;
     xaxis label="OR and 95% CI " min=0;
     yaxis label="Female vs. Male ";
run;

* REGRESSION ***;
proc sort data=LSM&num out=LSM&num._1; where domain ne "";
by domain &outp;
run;
title "Output 4: Analysis of Covariates" ;
proc GLM data=LSM&num._1;
 by domain &outp;
 format &inp &inp.f. cycle cyclef.;
 class  &inp (ref="&refinp");
  model estimate= cycle|&inp /solution;
 output out=r&num p=pred&num L95=lower&num U95=upper&num;
run;
quit;

ODS graphics off;
ODS RTF close;
%MEND;
```

EXAMPLE № 1

```
%MNLRM(sds=a4, inp=gender, refinp=%str(Male) , outp=BMIGRP, cov=%str(),
domain=sel, domainx=%str( ), num=1 )
```

Output 1: Type 3 Analysis of Effects

Variable	DF	WaldChiSq	P-value
Gender	2	72.2829	<.0001
NHANES cycle	2	36.5854	<.0001
NHANES cycle*Gender	2	3.5658	0.1681

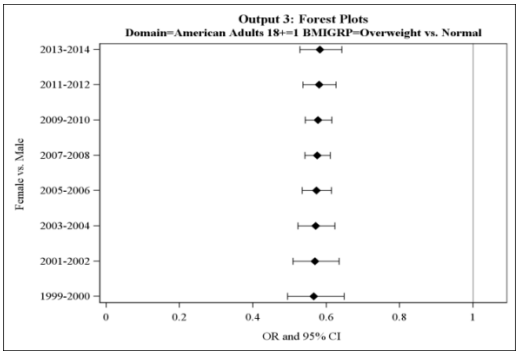
Output 2: Odds Ratios with 95% Confidence Intervals (CI) for Females compared to Males*

BMI Group	1999-2000 OR (95% CI)	2001-2002 OR (95% CI)	2003-2004 OR (95% CI)	2005-2006 OR (95% CI)	2007-2008 OR (95% CI)	2009-2010 OR (95% CI)	2011-2012 OR (95% CI)	2013-2014 OR (95% CI)	p-value
A. Overweight vs Normal	**0.57 (0.49 -0.65)**	**0.57 (0.51-0.64)**	**0.57 (0.52-0.62)**	**0.57 (0.53-0.61)**	**0.58 (0.54-0.61)**	**0.58 (0.54-0.62)**	**0.58 (0.54-0.63)**	**0.58 (0.53-0.64)**	0.7828
B. Obese vs Normal	0.96 (0.86 -1.07)	0.94 (0.86-1.03)	0.92 (0.86-1.00)	**0.91 (0.85-0.97)**	**0.89 (0.84-0.95)**	**0.88 (0.82-0.94)**	**0.86 (0.79-0.94)**	**0.84 (0.76-0.94)**	0.1727

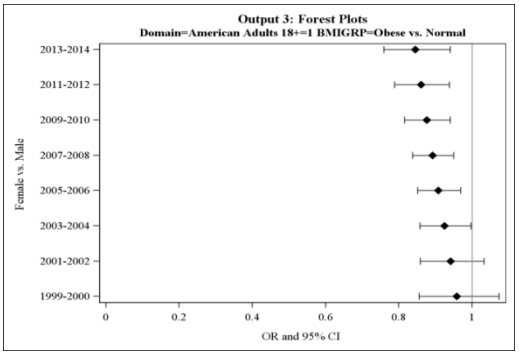
*Significant associations are presented in bold

Output 3: Forest Plots

A. BMI Groups: Overweight versus Normal

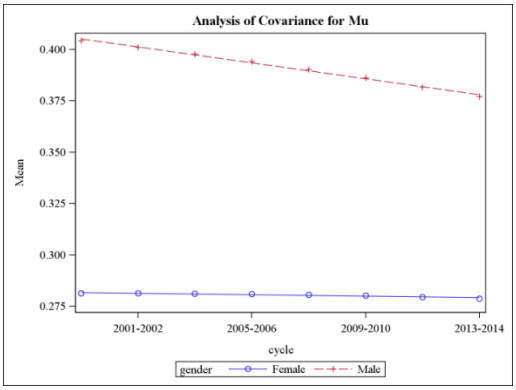


B. BMI Groups: Obese versus Normal

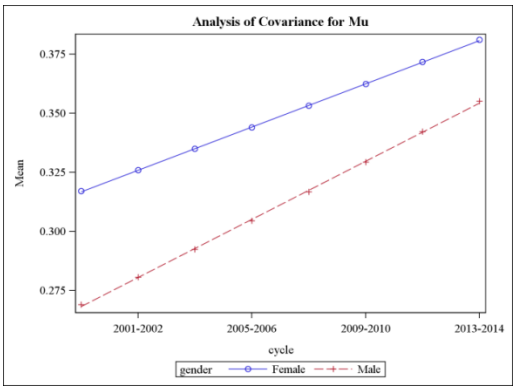


Output 4: Analysis of Covariates

A. BMI Groups: Overweight versus Normal



B. BMI Groups: Obese versus Normal



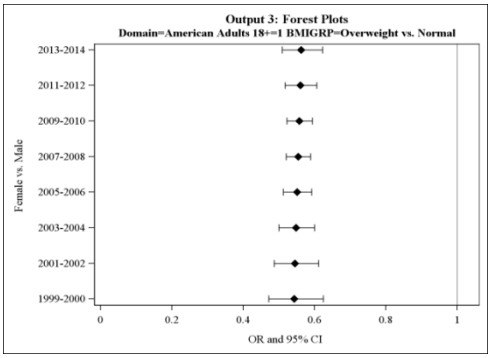
CONCLUSIONS FOR EXAMPLE #1:

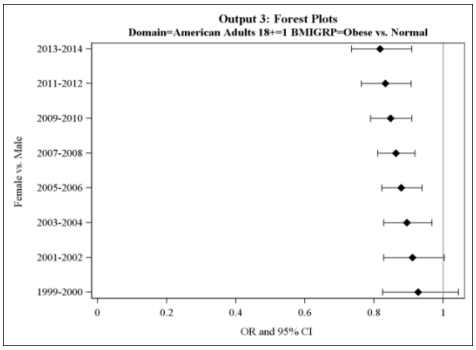
1. Type 3 analysis of effects demonstrated that gender and time are significant predictors ($p < .05$) for BMI. The interaction term was not significant ($p > .05$) indicating these relationships have not changed over time. (Output 1)
2. The chance of being in "overweight vs normal" BMI category is significantly lower for Females compared to Males across all years from 1999 through 2014, but the relationships (odds ratios) have not changed over time. (Output 2.A and Output 3.A)
3. The chance of being in "obese versus normal" BMI category is significantly lower for Females compared to Males starting from 2005 and through the 2014. The relationships (odds ratios) have not changed over the years. (Output 2.B and Output 3.B)
4. Predicted probability for being "overweight vs normal" BMI category was stable for Females over the years, but slightly decreased for Males as can be seen in Output 4.A.
5. Predicted probability of "obese vs normal" BMI category increased faster for Males as represented by steeper slope for Males versus Females in Output 4.B.

EXAMPLE № 2

```
%MNLRM(sds=a4, inp=gender, refinp=%str(Male) , outp=BMIGRP, cov=%str(race
age), domain=sel, domainx=%str( ), num=2  );
```

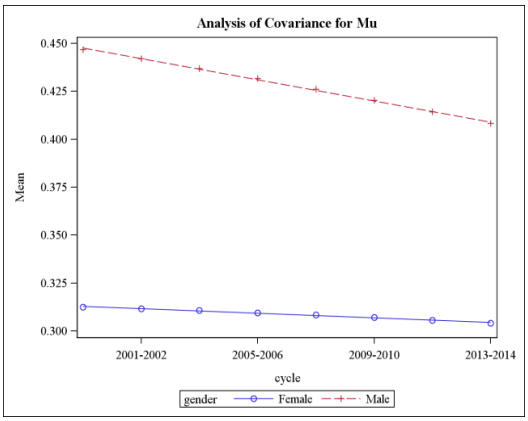
Output 5: Forest Plots



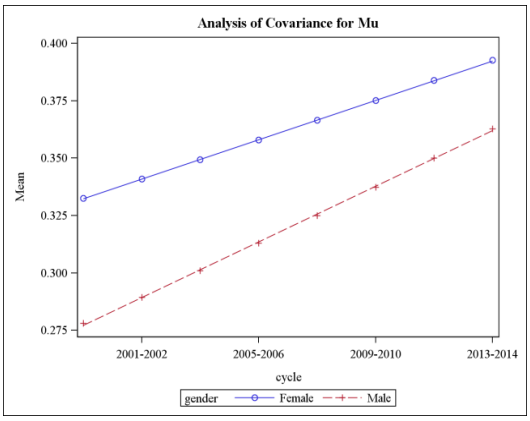


Output 6: Analysis of Covariates

A. BMI Groups: Overweight versus Normal



B. BMI Groups: Obese versus Normal



CONCLUSIONS FOR EXAMPLE #2:

Adjustment for Age and Race confirmed and demonstrated slightly stronger association between Gender and BMI. The results also supported the conclusion that these associations have not changed from 1999 to 2014. (Output 5 and Output 6)

CONCLUSION

Multinomial Logistic Regression Models are statistical analysis technique applicable to population survey designs. The MACRO in this paper was developed with use of SAS PROC SURVEYLOGISTIC to examine the associations between predictor and multi-level outcome. This MACRO can be easily adapted for different research questions to address not just associations, but also if and how they changed over time.

REFERENCES

[1] Centers for Disease Control and Prevention, National Center for Health Statistics Division of Health and Nutrition Examination Surveys: National Health and Nutrition Examination Survey: Analytic Guidelines, 1999-2010, and 2011-2012; September 30, 2013; http://www.cdc.gov/nchs /nhanes/

[2] National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES) manuals (1999-2000, 2001-2002, 2003-2004, 2005-2006, 2007-2008, 2009–2010, 2011–2012, and 2013-2014). Available from: http://www.cdc.gov/ nchs/nhanes/nhanes [accessed June 2013].

CONTACT INFORMATION

Marina Komaroff
Director – Biometrics
Product Development/Clinical Opertations/Regulatory
Noven Pharmaceuticals, Inc.
Empire State Building
350 Fifth Avenue, 37th Floor
New York, NY 10118
Tel: 212 299 4202
Email: Mkomaroff@noven.com