Paper PO08

Generating Model Based Subgroup Analysis Using SAS® Procedures

Tracy Lin, Jie Huang
Merck & Co., Inc., Upper Gwynnedd, PA

ABSTRACT

Subgroup analysis is often carried out for clinical trials to further understand the treatment effect in some subgroups of patients. One approach is to repeat the analysis done for the whole population for the subgroup of patients of interest. Another is to generate the subgroup analysis based on a statistical model with an interaction term of subgroup and treatment. The reason the later approach is not commonly used is the complexity of interpretation of the interaction in the model. This article is an effort to provide a SAS tool to generate such an analysis with ease. Through an example, we will demonstrate the programs and results of the subgroup analysis with these two approaches.

INTRODUCTION

Subgroup analysis is often carried out for clinical trials to further understand the treatment effect in some subgroups of patients. For example, it is always required to demonstrate the consistency of the efficacy of a study agent over various subgroups of patients. On the other hand, when the efficacy of an investigational agent is less than expected, one may turn to the subgroup analysis to explore whether a subgroup of patients in the trial does better.

Typically, a subgroup analysis is done using the same analysis method for the whole study population, but based on a subgroup of patients (e.g., smokers or non-smokers). Another approach to the subgroup analysis is model based with additional interaction term of the subgroup and treatment. However, such models are not easy to interpret and translation to the subgroup analysis requires proper re-parameterization. In this article, we will present the implementation of such an analysis using SAS procedures.

EXAMPLE

We use the dataset of Low Birth Weight in Hosmer and Lemeshow (2000) as the example. The dataset is based on the research to identify possible contributing factors to low weight of new born babies. The dataset contains the following variables:

Variable	*Description*
NAME	Subject ID number
LOW	Low birth weight (1= BWT<=2500g, 0= BWT>2500g)
AGE	Age of mother in years
LWT	Mother's weight at last menstrual period
BWT	Birth weight in grams
SMOKE	Smoking status during pregnancy (1=yes, 2=no)
PTL	History of preterm labor (0=none,1=one or more)
HT	History of hypertension (1=yes, 0=no)
UI	Presence of uterine irritability (1=yes, 0=no)
FTV	Physician visits during the first trimester

The first 10 observations of the dataset are shown below:

name	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt	phy_vis
4	1	28	120	3	1	1	0	1	0	709	0
10	1	29	130	1	0	0	0	1	2	1021	1
11	1	34	187	2	1	0	1	0	0	1135	0
13	1	25	105	3	0	1	1	0	0	1330	0
15	1	25	85	3	0	0	0	1	0	1474	0
16	1	27	150	3	0	0	0	0	0	1588	0
17	1	23	97	3	0	0	0	1	1	1588	1
18	1	24	128	2	0	1	0	0	1	1701	1
19	1	24	132	3	0	0	1	0	0	1729	0
20	1	21	165	1	1	0	1	0	1	1790	1

Consider that the main question of interest is whether smoking during pregnancy is associated with babies with low birth weight. For the demonstration purposes, we only consider univariate analysis, and the subgroups of interest are history of hypertension and physician visits during the first trimester. In the following analysis, we calculate the odds ratio as the measure of difference between smoker mothers and non-smoker mothers using logistic regression. The odds ratio and its 95% confidence intervals are generated using SAS PROC LOGISTIC.

ANALYSIS BASED ON SUBGROUPS OF SUBJECTS

To perform the subgroup analysis using subgroups of patients, we need analyze the data separately for each subgroup of the subjects. The following codes were used to carry out the subgroup analysis for history of hypertension and any physician visit during first trimester. We define a new variable phy_vis as 1= at least one physician visits during the first trimester, 0= none.

```
proc sort data=lowbwt; by ht;
proc logistic data=lowbwt;
  class smoke(param=ref ref='0');
  model low(event='1')=smoke ;
  by ht;
  title "subgroup analysis by hypertension (subset)";
run;

proc sort data=lowbwt; by phy_vis;
proc logistic data=lowbwt;
  class smoke(param=ref ref='0');
  model low(event='1')=smoke ;
  by phy_vis;
  title "subgroup analysis by phyisician visits (subset)";
run;
```

MODEL BASED SUBGROUP ANALYSIS

To implement a subgroup analysis through the statistical model, we first fit the data with the logistic regression model with the variable for the treatment difference (e.g., smoke) and the variable for the subgroup of interest (e.g., ht) and the interaction term of these two variables. Then a contrast statement is constructed for the analysis of each subgroup. Below is the SAS programs to generate the subgroup analysis for history of hypertension and physician visits during the first trimester.

```
proc logistic data=lowbwt;
  class smoke(param=ref ref='0') ht(param=ref ref='0');
  model low(event='1')=smoke ht smoke*ht;
  contrast 'ht=No' smoke 1 smoke*ht 0 /estimate=exp;
  contrast 'ht=Yes' smoke 1 smoke*ht 1 /estimate=exp;
  title "model based hypertension subgroup analysis";
run;
```

```
proc logistic data=lowbwt;
  class smoke(param=ref ref='0') phy_vis(param=ref ref='0');
  model low(event='1')=smoke phy_vis smoke*phy_vis;
  contrast 'phy=No' smoke 1 smoke*phy_vis 0/estimate=exp;
  contrast 'phy=Yes' smoke 1 smoke*phy_vis 1/estimate=exp;
  title "model based physician visit subgroup analysis";
run;
```

Table 1 presents the results of the analysis. Though the results from the programs in the Section III were not presented, both approaches arrived the same results because they represent the same statistical models. In the case of univariate analysis above, the likelihood function of the statistical model with the interaction term can be easily shown to be partitioned into two distinct parts with each representing a likelihood function for each subgroup of interest.

Table 1: Results of subgroup analysis

	Smokers	Non-smokers	Odds ratio	95% CI
All	30/74 (41%)	29/115 (25%)	2.02	1.08, 3.78
History of hypertension				
Yes	3/5 (60%)	4/7 (57%)	1.13	0.11, 11.60
No	27/69 (39%)	25/108 (23%)	2.13	1.10, 4.12
Physician visits				
Yes	10/29 (34%)	13/60 (22%)	1.90	0.71, 5.08
No	20/45 (44%)	16/55 (29%)	1.95	0.85, 4.46

CONCLUSION AND DISCUSSION

This article intends to provide an alternative approach to the subgroup analysis. Using the same statistical model, both approaches came up with the same results for the subgroup analysis. Though both approaches can be implemented with relative ease, the analysis using subset of subjects requires additional step (e.g., a sort procedure or a dataset for a subgroup) prior to calling the PROC LOGISTIC. On the other hand, the model based approach fit the model with all subjects included and no additional data manipulation is necessary. Because of only one model fitting is needed in the model based approach, it may also have the advantage of less resource requirement in computer time. In addition, the use of contrast statements make the output of statistics of all subgroup of interest together and therefore, to be easily output to a dataset and be manipulated for table generation.

Statistically, though the subgroup analysis has introduced many controversy issues such as type I error control and potential selection bias, the analysis is routinely done for various purposes mainly hypothesis generating. The example shown is only for the demonstration of the approach. The approach can be easily extended to a different statistical model or SAS procedure, more covariates, or more subgroups (e.g., a combination of hypertension and physician visits). However, in those cases, conformity of the results from both approaches is not guaranteed due to either ancillary parameter or difference in the statistical models, which is out of the scope of this article.

Using model based method with contrasts provides an alternative mean of performing subgroup analysis, which offers some advantages in the ease of SAS programming and potentials for more complex statistical modeling approaches for subgroup analysis.

REFERENCES

1. Hosmer and Lemeshow (2000), Applied Logistic Regression, 2nd Edition, New York: John Wiley & Sons

ACKNOWLEDGEMENT

The Authors would like to thank Hong Qi and Richard Lowry for their strong encouragement and support in publishing this paper.

CONTACT INFORMATION

Comments and questions are greatly appreciated. Contact authors at:

Tracy Lin
Merck Research Laboratories
P.O. Box 1000
Upper Gwynedd, PA 19454-2505
(267)305-8202 (phone)
E-mail: Tracy_lin@Merck.com

Jie Huang
Merck Research Laboratories
P.O. Box 1000
Upper Gwynedd, PA 19454-2505
(267)305-5541 (phone)
E-mail: Jie_Huang4@Merck.com