Paper 317-2012
Using SAS® to Extend Logistic Regression

Dachao Liu Northwestern University Chicago

ABSTRACT

Logistic regression is widely used in analysis of categorical data especially data with variables that have binary responses. It can be used in many fields where discrete responses and a set of explanatory variables coexist. For instance, it can be used in survival analysis which often uses data with variable having values of 'success' and 'failure'. We can let logistic regression do more things for us by extending it. This paper will discuss under what circumstances we extend logistic regression by using SAS.

INTRODUCTION

Logistic regression can predict a dichotomous outcome using independent variables. This dichotomous outcome can be anything like success or failure, life or death, presence or absence, win or lose, admission or rejection, you name it. But sometimes we will see that the outcome variables are not of two levels, but of more than two levels, in other words, polychotomous. Polychotomy can be regarded as a generalization of dichotomy. We can use logistic regression to handle polychotomous outcome, which can be found in many places, like survey data. In SAS, PROC LOGISTIC is used to perform all these tasks. When sample size is small, we can use exact logistic regression. We will also see the PROC GENMOD, PROC CATMOD, PROC PROBIT used in logistic regression. Even PROC PHREG can be used to perform logistic regression. That's what I mean using SAS to extend logistic regression. In logistic regression we are trying to estimate the probability that a given subject will fall into one outcome group or the other.

DISCUSSION

Before I discuss the extensions to logistic regression, let's briefly review what is a logistic regression? Logistic regression is an extension to regression. In this world we are living, one thing, to some extent, is related to another thing. Therefore, one thing's change will lead to another thing's change. For example, the amount of carbon dioxide released in the air will lead climate change. The temperature will affect the growth of crops. The demand will affect the cost etc. We try to use regression to find relationships between variables and the causal effect of one variable upon another and to estimate the quantitative effect of the causal variables upon the variable that they influence. The value of a normally distributed dependent variable can be predicted by the values of one or more independent variables, which can be either continuous or categorical. We can use parents' height to predict a child's height and years of education to predict income. Regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

In regression analysis, a response variable Y can be predicted by a linear function of a regressor variable X. We can estimate β_0, the intercept, and β_1, the slope, in

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

In SAS, this can be performed by

```
proc reg;
  model y = x;
run;
```

If the dependent variable was not a continuous variable, but instead diagnosis of a disease (yes or not), which is dichotomous, that is, the dependent variable can take the value 1 with a probability of success p, or the value 0 with probability of failure 1-p. Then we cannot use the above procedure.

Why? This could result in negative values or values greater than 1 for dependent variable Y, which doesn't make any sense at all. Therefore, the techniques used in linear regression to estimate the regression coefficients cannot be applied to the logistic regression case. Besides that, one of the assumptions of regression is that the variance of Y is constant across values of X (homoscedasticity), which cannot be the case with a binary variable, because its variance is p(1-p). Suppose 50 percent of the people are 1s, then the variance of .25 would be its maximum value. As

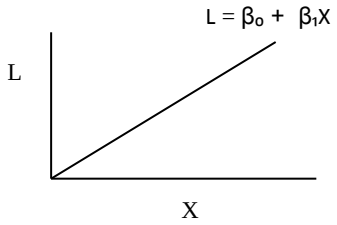
we move to more extreme values, the variance decreases, for example when p =.10, the variance is .1*.9 = .09, so as p approaches 1 or 0, the variance approaches 0. Furthermore, the significance testing of the β is based upon the assumption that errors of prediction (Y-Y') are normally distributed. Because Y now only takes the values 0 and 1, this assumption is very hard to justify. Therefore, the tests of the β are doubtful if we use linear regression with a binary dependent variable. This type of variable is called a Bernoulli variable.

An extension of regression called logistic regression comes into being. It resorts to a transformation which makes sure that value of dependent variable y is 0 (no event) and 1(event). We can also call it an event variable. How does a logistic regression work? Suppose p is the probability of the event, then the odds of the event are Odds = p / (1-p). Take the logarithms of both sides, we have Ln (Odds) = Ln (p/1-p). The left side is a logit. Ln stands for Log_e, where *e* is a mathematical expression whose numerical value is equal to 2.71828.
Suppose we use an independent variable X to predict the probability of being an event, then the following is the simple logistic model.

$$\text{Ln (Odds)} = \text{Ln}\left(p / 1\text{-}p \right) = \beta_o + \beta_1 X + e$$

This depicts a linear relationship between the natural logarithm (ln) of the odds of an event and a continuous independent variable. In other words, A logistic regression model is a liner regression equation in which the response variable is the log odds. β_0 is the intercept and β_1 is logistic regression coefficients. By means of the logistic function, we find a way of relating the outcome variable to the explanatory variables. And by means of logistic transformation of the probability p, we get rid of the probability for the sake of analysis. The left side of equation can take on any value from $-\infty$ to $+\infty$.

The linear relationship between X and Ln (Odds) can be shown in the following graph

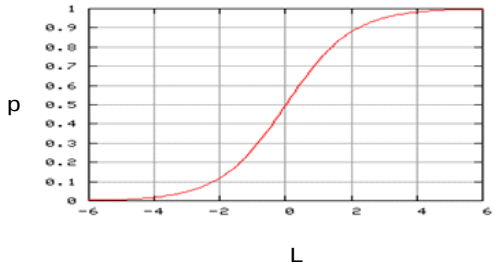


But in reality, we don't see logits in dependent variable, instead we see bunches of 1s and 0s. It's easier to use probability instead of odds to relate an independent variable to dependent variable. We can convert odds to probability by taking the logarithms of both sides.

Suppose L = Ln (Odds) = Ln (p / 1-p), then it can be shown that

$$p = e^{L} / (1 + e^{L})$$

And that is the logistic function and it has a graph as follows:



If log odds are linearly related to X, then the relation between X and p is nonlinear, and has the form of the S-shaped curve you saw in the graph. Now it is not a straight line but an upward curve. We can call L as the 'log odds' or 'logit' of p.

The simple logistic model

$$\text{Ln (Odds)} = \text{Ln}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X + e$$

can be implemented in SAS by using

```
PROC LOGISTIC;
MODEL Y=X;
RUN;
```

Usually there is more than one independent variable in the model. A typical data set for this model would look like this:

id	y	x1	x2	x3	x4
1	1	17	88	1	M
2	0	16	87	2	F
3	1	21	28	1	M
4	0	23	52	1	F
5	1	38	41	1	F
6	0	19	54	1	F
7	0	16	76	2	M
8	1	27	59	1	M
9	0	13	49	1	M
10	1	26	38	2	F

The independent variables can be continuous variables or dichotomous variables. There is no assumption about the distribution of the independent variables, like normally distributed, linearly related, and of equal variance within each group.

The above procedure is used when outcome is dichotomous. It can also be extended to the cases of polychotomous outcome. We can define obesity as a binary variable, marked by BMI >=30 and <30, we can also define it as: obese (BMI >=30) overweight (25<=BMI<30) normal weight (18<=BMI<25) . Polychotomous outcome can be ordinal (ordered categories) or nominal (unordered categories) A 5-point Likert scale is used to assess patient's opinion about a medical treatment measure in a survey. The response options are "strongly disagree", "disagree", "neutral", "agree" and "strongly agree". The researchers have reason to believe that the psychological "distances" between these points are not equal. For example, the "distance" between "strongly disagree" and "disagree" may be shorter than the distance between "disagree" and "neutral". This data can be treated as ordinal outcome. Nominal outcome is not ordered. A good example is race has values 1=White, 2=Hispanic, 3=American Indian, 4=Black, 5=Other.

Suppose we have a subset data set with polychotomous outcome and it looks like this:

id	y	x1	x2	x3	x4
1	1	17	88	1	M
2	2	16	87	2	F
3	1	21	28	1	M
4	3	23	52	1	F
5	1	38	41	1	F
6	3	19	54	1	F
7	2	16	76	2	M
8	1	27	59	1	M
9	2	13	49	1	M
10	3	26	38	2	F

Where y is outcome variable. 1= "disagree" 2 = "neutral" 3= "agree".

We can use this SAS code to run this data.

```
proc logistic;
class x3 (ref='1') x4 (ref='F') /param=ref;
model y =x1 x2 x3 x4 /rsquare;
```

```
output out=probs predprobs=(i c);
run;
```

ref allows to indicate reference category for a class variable. The quotation marks are required. The default is REF=LAST.

param specifies the parameterization method for the classification variables. A full description can be found SAS Online Documentation, procedure PROC LOGISTIC under CLASS statement. There you can find some keywords. the default is PARAM=EFFECT.

rsquare requests a generalized R^2 measure for the fitted model

predprobs=(i c) requests the predicted probability of each response level (i) and cumulative probabilities (c). In our example the individual probabilities (for each resopnse level) are: P(Y = 1, disagree), P(Y = 2, neutral), P(Y = 3, agree) and cumulative probabilities are: P(Y = 1, disagree), P(Y <= 2, disagree or neutral), and the last cumulative probability, P(Y <= 3, disagree or neutral or agree) which is equal to 1.

When sample size is large and there are a few parameters to estimate, we can use the general logistic regression mentioned above. But if sample size is very small, or data is sparse, skewed, or unbalanced with covariates and there are a lot of parameters to estimate, the test statistics may not be chi-squared, resulting an inaccurate p-value, then we have to use exact logistic regression. For exact logistic regression, you can refer to a paper 'Performing Exact Logistic Regression with the SAS system' by Robert E. Derr (P254-25)

Now I will discuss the conditional logistic regression. When we use PROC REG and PROC LOGISTIC, study subjects are independent of one another and any violation of this assumption will result in invalid statistical inference. However, many clinical and epidemiologic study designs often give the researchers correlated data. We often encounter this kind of correlated data in study designs like: prospective study, retrospective study, and cross-sectional study. A retrospective study is also called a case-control study. It compares a group of people with a disease (cases) to another group of people without the disease (controls). It is used by researchers like epidemiologists to identify and assess factors that are associated with diseases or health conditions. The researchers try to relate their prior health habits to their current disease status and compare cases and controls with respect to previous exposures to factors of interest. Different diseases have different exposures. Lung cancer may have smoking exposure. Skin cancer may have strong sunlight exposure. Depression may have specific diet exposure. In case–control studies, information about exposure is generally collected after the disease has already occurred, that's why these studies are called retrospective studies. In short, case–control studies look back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease. In case–control studies, the researcher often uses matching technique to make data analysis more efficient. In case–control studies study subjects are not independent of one another, we cannot use PROC LOGISTIC to analyze the data derived from case-control studies. An extension of logistic regression called conditional logistic regression must be used. A simple rule of thumb is use conditional logistic regression if matching has been done, and unconditional if there has been no matching. Unconditional logistic regression is the general logistic regression mentioned above.

If the likelihood function for unconditional logistic regression is $L_U = L(\beta_o \ \beta_1 \ldots \ \beta_k)$
Then the likelihood function for conditional logistic regression is

$L_C = L_C(\beta_1 \ldots \ \beta_k) = L_U / \sum(L_U)$ (Summation across all strata)

The intercept β_o is dropped out of the likelihood function for conditional logistic regression. We don't care about it (them). It is nuisance and is conditioned out of the analysis.

It happened to be that the likelihood function of the conditional logistic regression model and the partial likelihood function of the Cox proportional hazard's regression model are so similar that we can use the Cox proportional hazards model to do conditional logistic analysis using data from a matched case-control study. This is done by treating each matched set as a stratum, assuming all cases within a given matched set to have the same event time. In SAS the Cox proportional hazard's regression model is implemented by PROC PHREG.

```
proc phreg;
      model Time*Status(0)=Var;
   run;
```

Where Time can take any value larger than 0 and Status takes values of 1=event and 0=censored. When it is used in the conditional logistic regression analysis, then Status becomes Case (1=case, 0=control) and Time becomes CSCN (1=for cases, 2=for controls) All cases in the same stratum should have the same time of event, i.e. time=1. For practical reasons it is simplest to give all cases the same time regardless of stratum. All controls in the same stratum should also have the same time, but greater than that of the cases, i.e. time=2.

With that set up, now let's see an example:

Prostate cancer patients usually have a higher level of prostate-specific antigen (PSA). Therefore, PSA levels can be used to predict prostate cancer. Sample sera was frozen and stored for future analyses. We could use a case-control design to analyze the data. Here is a subset of data named ONETWOA.

ONETWOA

ID	AGE	CSCN	DX	RACE	PSA	DWDO
1799583	51	1	1	1	9	22
1799642	51	2	0	3	7	36
1799674	51	2	0	3	3	93
1799702	51	1	1	3	27	11
1799520	53	2	0	3	10	88
1799551	53	2	0	1	4	19
1799573	53	1	1	.	7	45
1799624	53	2	0	3	5	57
1799717	53	2	0	3	8	23
1799726	53	2	0	3	4	0
1799503	54	1	1	3	7	31
1799553	54	2	0	3	5	29
1799794	54	1	1	3	12	93
1799618	55	2	0	3	4	182
1799703	55	1	1	3	5	36
1799716	55	2	0	3	8	181
1799771	55	2	0	3	5	30
1799644	56	2	0	3	.	0
1799734	56	1	1	3	.	116
1799502	57	1	1	3	4	36
1799512	57	1	1	3	10	12
1799555	57	2	0	3	5	7
1799607	57	2	0	.	.	14
1794587	57	2	0	3	3	26
1794599	57	2	0	3	8	209
1794618	57	1	1	3	7	32
1794619	57	1	1	3	5	28
1794625	57	1	1	3	4	31
1794644	57	2	0	2	3	111
1794655	57	2	0	3	4	48
1794717	57	1	1	3	6	81
1794732	57	2	0	2	6	25
1799521	58	2	0	3	.	0
1799600	58	2	0	3	7	16
1799667	58	2	0	3	6	120
1799757	58	1	1	3	.	7
1799795	58	1	1	2	10	49

Since age is a risk factor for prostate cancer, we could adjust for age in the model in order to the correct risk estimates without using matched case-control technique. But if we use a matched case-control study, then we will get an even better adjustment for age. We had many cases in each matching stratum, so we have n:m matching, where number of cases and controls vary. Total number of cases and controls also varied accordingly in all strata.

Resorting to PROC PHREG, we can do conditional logistic regression on this data.

```
proc phreg data=ONETWOA;
      model CSCN*DX(0)= PSA DWDO / ties=discrete risklimits;
      strata AGE;
run;
```

Here is the part of the output from the PROC PHREG:

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	200.413	191.359
AIC	200.413	195.359
SBC	200.413	200.424

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.0544	2	0.0108
Score	2.8049	2	0.2460
Wald	6.3438	2	0.0419

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PSA	1	0.04123	0.01883	4.7959	0.0285	1.042	1.004	1.081
DWDO	1	-0.00350	0.00281	1.5539	0.2126	0.997	0.991	1.002

We see that there is a significant association between prostate cancer incidence and PSA (p = 0.0285). The relative risk or hazard ratio is 1.042 (=e to the power of 0.041). This means that if we have two men of the same age in a matched pair with one having prostate cancer and the other not, the man with the higher PSA is 1,042 times as likely to be the case than the man with the lower PSA.

The TIES=DISCRETE option is used replace the proportional hazards model by the discrete logistic model to get the conditional logistic regression. The option risklimits outputs the confidence intervals for the odds ratios. Default is 95 %.The matching variables are specified in the STRATA statement. Here we have matched on age.

We just matched age. Now if we match race, see what will happen. This is the output:

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	259.239	244.626
AIC	259.239	248.626
SBC	259.239	253.626

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.6130	2	0.0007
Score	8.3173	2	0.0156
Wald	10.7435	2	0.0046

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PSA	1	0.04503	0.01854	5.8962	0.0152	1.046	1.009	1.085
DWDO	1	-0.00629	0.00276	5.2070	0.0225	0.994	0.988	0.999

Also we see that there is a significant association between prostate cancer incidence and PSA (p = 0.0152). The relative risk or hazard ratio is 1.046 (=e to the power of 0.045). This means that if we have two men of the same race in a matched pair with one having prostate cancer and the other not, the man with the higher PSA is 1,046 times as likely to be the case than the man with the lower PSA.

Finally, let's see an example using PROC GENMOD to do logistic regression. The researchers are experimenting with a drug to see whether it can relieve the pain for patients or not. Y is outcome variable(1=pain relieved 0=not) x1 is the dose of the drug, x2 is the age of the patients, and x3 is group. Suppose the subset of data set is like this:

id	y	x1	x2	x3
1	1	17	88	1
1	0	16	88	2
2	1	21	52	1
2	1	15	52	1
3	1	30	54	2
3	0	19	54	1
4	1	16	76	2
4	1	27	76	1
5	0	13	69	1
5	1	26	69	2

This is a correlated data with each subject repeated once. We can use this SAS code to analyze the data:

```
proc genmod;
   class id;
      model y=x1 x2 x3 / dist=bin;
   repeated subject=id/ corr=unstr corrw;
run;
```

Dist=bin is specified in the model statement to conduct a logistic regression analysis. The subject=id identifies the clustering variable. This variable is listed in class statement. Unstr specifies the unstructured correlation structure. And the corrw option specifies that the final working correlation matrix be printed.
Besides PROC GENMOD, PROC CATMOD can also be used in logistic regression analysis.

Please Note: The purpose of this paper is to show briefly how to use logistic regression and its extensions in data analysis. It does not cover all aspects of the logistic regression. In particular, it does not cover data cleaning and checking, verification of assumptions, model diagnostics, multicollinearity, goodness of fit and statistical power.

CONCLUSION

Logistic regression is derived from linear regression, but the relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of p. Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous. PROC LOGISTIC is a procedure for fitting linear regression models not only for binary or dichotomous outcomes but also polychotomous outcome. SAS can extend logistic regression by using conditional regression to handle data from case-control studies by PROC PHREG procedure. SAS can also extend logistic regression to handle correlated data by PROC GENMOD procedure and PROC CATMOD procedure.

REFERENCES

Agresti, A. (2002). *Categorical data analysis (2nd Edition)*. John Wiley & Sons, New York.

Agresti, A. (2007). *An introduction to categorical data analysis (2nd Edition)*. John Wiley & Sons, New York.

Allison, Paul D. (1995). *Survival Analysis Using the SAS® System: A Practical Guide.* Cary, NC: SAS Institute Inc.

Allison Paul D. (1999). *Logistic Regression Using the SAS System*: Theory & Application. SAS Institute Inc., Cary, NC, USA.

Derr, R.E. (2000), "Performing Exact Logistic Regression with the SAS® System," *Proceedings of the 25th Annual SAS_ Users Group International Conference (SUGI 25),* 254-25.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression (2nd Edition)*. New York: Wiley.

SAS Institute Inc. (2010). "The LOGISTIC Procedure" and "The GENMOD Procedure". *SAS/STAT® 9.22 User's Guide, Second Edition*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Dachao Liu
Northwestern University
Suite 1400
680 N Lake Shore Dr.
Chicago, IL 60611
Phone (312)503-2809
Email: dachao-liu@northwestern.edu

ACKNOWLEDGEMENTS