

Paper 427-2012

Beyond Binary Outcomes:**PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables**

Eric Elkin, ICON Late Phase & Outcomes Research, San Francisco, CA, USA

ABSTRACT

The most familiar reason to use PROC LOGISTIC is to model binary (yes/no, 1/0) categorical outcome variables. However, PROC LOGISTIC can handle the case where the dependent variable has more than two categories. PROC LOGISTIC uses a cumulative logit function if it detects more than two levels of the dependent variable, which is appropriate for ordinal (ordered) dependent variables with 3 or more levels. A generalized logit function for the LINK= option is available to analyze nominal (un-ordered) categorical variables with 3+ levels (i.e., multinomial logistic regression). Detailed examples will be given, emphasizing procedure syntax, data structure, interpretation of statistical output, and ODS output data sets.

INTRODUCTION

PROC LOGISTIC is one of the tools in SAS[®] for multivariate modeling of categorical outcome variables (the CATMOD procedure, among others, can also be used). Binary (or dichotomous) response variables are the most familiar categorical variables to model using logistic regression. These are often yes/no variables coded as 0=no and 1=yes. However, there are situations when the categorical outcome variable has more than two levels.

These polychotomous variables may be either ordinal or nominal. Ordinal variables have some intrinsic order to them, such as a variable with responses coded as 1=low, 2=intermediate, 3=high. Nominal variables have no intrinsic order to them – they are simply labels for different categories, such as a variable with responses coded as 1=surgery, 2=medication, 3=radiation.

In order to use the more familiar binary logistic regression, we might consider collapsing a dependent variable with three or more categories into two categories. Another idea might be to model each pairwise comparison (low vs. intermediate, low vs. high, intermediate vs. high). However, neither of these options is a satisfying solution. In the first case, we might lose information depending upon how the variable is collapsed into two groups. In the second case, it would be unclear how to interpret the results – not to mention the additional work of three models instead of one.

Instead, this paper will demonstrate how to analyze these variables in one model retaining the polychotomous nature of the dependent variable. Depending on whether the multi-level outcomes variable is ordinal or nominal will determine procedure syntax and interpretation of results.

DATA EXAMPLE

Procedure syntax and interpretation of results will be provided through specific examples. The data for these examples comes from sample data of men with prostate cancer. The ordinal dependent variable will be stage at time of diagnosis, coded as: 1="Early", 2="Middle", 3="Late". The nominal dependent variable will be type of treatment, coded as: 10=Radical prostatectomy (surgery), 30=Radioactive seed implant (brachytherapy), 40=External beam radiation therapy, 60=Hormonal therapy, 80=Watchful waiting (surveillance). Independent variables considered for this example are age at diagnosis (1="<60", 2="60-69", 3="70+"); results of prostate-specific antigen (PSA) test at diagnosis (1="≤20 ng/ml", 2=">20 ng/ml"); and history of previous cancer (0="no", 1="yes").

REVIEW OF PROC LOGISTIC

A quick review of PROC LOGISTIC syntax may be helpful. We will use "history of cancer" as a binary outcome for this example to see how independent categorical variables are specified using the CLASS statement, as well as the logistic regression model specification using the MODEL statement:

```
proc logistic data=temp01;
  class age_cat (ref="<60") psa_cat (ref="<=20") / param=ref;
  model cancer = age_cat psa_cat;
  format cancer yesno. age_cat agecat. psa_cat psacat.;
  title3 "Binary outcome";
run;
```

The dependent (outcome) variable goes on the left side of the equal sign in the model statement, and the independent (predictor) variables on the right side. FORMAT and TITLE statements are optional.

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

The REF=*refcat* option after each variable in the CLASS statement allows us to control which category is used as the reference category in the design matrix. In this example, I've used formats for these variables, so the format label is used to specify *refcat*. Be careful: the text in *refcat* must match the format label exactly, including capitalization! The PARAM=ref option on the CLASS statement tells the procedure to use reference coding for the model design matrix.

In the example above, the default is to model the first level of the outcome variable, which in this case is 0="no". If we want the procedure to model the risk of having a previous cancer (1="yes") there are two ways to do that. One way is to use the DESCENDING option on the PROC LOGISTIC statement:

```
proc logistic data=temp01 descending;
  class age_cat (ref="<60") psa_cat (ref="<=20") / param=ref;
  model cancer = age_cat psa_cat;
  format cancer yesno. age_cat agecat. psa_cat psacat.;
  title3 "Binary outcome";
run;
```

Alternatively, we can also include a REF=*refcat* specification after the outcome variable in the MODEL statement:

```
proc logistic data=temp01;
  class age_cat (ref="<60") psa_cat (ref="<=20") / param=ref;
  model cancer (ref="No") = age_cat psa_cat;
  format cancer yesno. age_cat agecat. psa_cat psacat.;
  title3 "Binary outcome";
run;
```

ORDINAL DEPENDENT VARIABLE

Traditional binomial logistic regression uses the binary logit function for statistical analysis (based on the binomial distribution). When the dependent variable is a multi-level ordinal variable, the cumulative logit is appropriate.

When PROC LOGISTIC encounters a model with a dependent variable that has more than two categories, it automatically uses the cumulative logit to perform the analysis. Be careful: make sure that the dependent variable is ordinal and not nominal!

There is essentially no additional syntax to specify. In order to keep the ordering of the coded values (1, 2, 3) for the dependent variables, I would suggest *not* formatting the dependent variable. Otherwise the design matrix will be based upon the formatted alphanumeric order (early-late-middle), rather than the true order (early-middle-late).

```
proc logistic data=temp01 ;
  class age_cat (ref="<60") psa_cat (ref="<=20") cancer (ref="No") / param=ref;
  model tstage = age_cat psa_cat cancer ;
  format age_cat agecat. psa_cat psacat. cancer yesno. ;
  title3 "Ordinal outcome (cumulative logit)";
run;
```

Much of the output is the same as binary logistic (response profile of dependent variables, model fit statistics, testing global null hypothesis, type III analysis of effects, etc.).

The odds ratio estimates also look the same but are interpreted somewhat differently than binary logistic. In the case of the cumulative logit, odds ratios are interpreted as the association between that variable and being in a lower level of the dependent variable. For example, patients 70 or older are less likely (0.8 times less) to be diagnosed at an earlier stage of disease when compared to men younger than 60 years old.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age_cat 60-69 vs <60	0.950	0.831	1.086
age_cat 70+ vs <60	0.792	0.691	0.908
psa_cat >20 vs <=20	0.163	0.136	0.196
CANCER Yes vs No	1.059	0.891	1.258

Another difference is seen in the analysis of maximum likelihood estimates. There are two intercepts. As with binary logistic regression, we can ignore the meaning of the intercepts. However, I will explain why there are two later.

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 1	1	-0.0654	0.0546	1.4308	0.2316
Intercept 2	1	3.0996	0.0776	1595.5517	<.0001
age_cat 60-69	1	-0.0510	0.0682	0.5578	0.4551
age_cat 70+	1	-0.2328	0.0698	11.1325	0.0008
psa_cat >20	1	-1.8114	0.0937	373.6556	<.0001
CANCER Yes	1	0.0570	0.0879	0.4213	0.5163

As with the binary logistic regression example, the procedure defaults to modeling the first level of the outcome variable. In the previous example, this means modeling the likelihood of being diagnosed with an earlier stage of disease (toward level 1="Early"). If we want to interpret results the other way – the likelihood of being diagnosed with later stage disease (toward level 3="Late") – then we can use the DESCENDING option on the PROC LOGISTIC statement.

```
proc logistic data=temp01 descending;
  class age_cat (ref="<60") psa_cat (ref="<=20") cancer (ref="No") / param=ref;
  model tstage = age_cat psa_cat cancer ;
  format age_cat agecat. psa_cat psacat. cancer yesno. ;
  title3 "Ordinal outcome (cumulative logit)";
run;
```

With the following results for the odds ratios:

Effect	Point Estimate	95% Wald Confidence Limits	
age_cat 60-69 vs <60	1.052	0.921	1.203
age_cat 70+ vs <60	1.262	1.101	1.447
psa_cat >20 vs <=20	6.119	5.092	7.353
CANCER Yes vs No	0.945	0.795	1.122

The direction of the odds ratios is now reversed. For example, patients aged 70 or older are 1.3 times *more* likely to be diagnosed with a *later* stage of disease than patients younger than 60.

For more insight into what the cumulative logit model is doing, consider what would have happened if we had followed the idea of collapsing the 3-level variable into two categories. We might have grouped the variable as "Early" vs. "Middle/Late", and seen the following results for the maximum likelihood estimates:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0670	0.0559	1.4380	0.2305
age_cat 60-69	1	-0.0781	0.0702	1.2386	0.2657
age_cat 70+	1	-0.2756	0.0723	14.5349	0.0001
psa_cat >20	1	-1.0524	0.1021	106.1478	<.0001
CANCER Yes	1	0.0322	0.0915	0.1241	0.7246

We could also have grouped the variable as "Early/ Middle" vs. "Late":

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.2773	0.1320	616.7605	<.0001
age_cat 60-69	1	0.1292	0.1601	0.6506	0.4199
age_cat 70+	1	-0.0366	0.1561	0.0549	0.8148
psa_cat >20	1	-2.5978	0.1175	488.9915	<.0001
CANCER Yes	1	0.2425	0.2046	1.4056	0.2358

Comparing these two outputs to the maximum likelihood estimates shown earlier for the model with the 3-level ordinal outcome, provides clues as to what the cumulative logit model does. The intercepts from the two models are very close to the two intercepts shown in the cumulative logit results. Each of the parameter estimates from the cumulative logit model lies between the two estimates from the two binary models.

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

The cumulative logit model allows the intercepts to differ but restricts the coefficients from the two binary models to be the same (basically, weighted averages). Any differences between the coefficients in the two binary estimates are assumed to be due to random error.

This leads to the other difference in the output from the cumulative logit. Near the top of the output is the "score test for the proportional odds assumption," which tests the validity of the ordinal model restricting to one coefficient. From the original cumulative logit model output:

```

          Score Test for the Proportional Odds Assumption
Chi-Square    DF      Pr > ChiSq
151.0851      4        <.0001

```

The null hypothesis is that the assumption is valid. In this case, the low p-value would reject the null hypothesis, so the assumption may not be valid. However, the *SAS/STAT@User's Guide* cautions that the test may reject more often than it should, especially if the sample size is large, as in this example, or if there are many independent variables in the model.

In sum, the cumulative logit model assumes that it does not matter how we dichotomize the ordinal variable, the effects will be fairly similar across the ordering of that variable. This is why the results are interpreted somewhat differently for these models.

NOMINAL DEPENDENT VARIABLE

When the dependent variable is nominal with three or more categories, then the generalized logit function is appropriate. Beginning with version 8.2, the generalized logit became available as an option in PROC LOGISTIC (prior to this release PROC CATMOD could be used). Based on the multinomial distribution, this type of analysis is also called "multinomial logistic regression".

To specify the generalized logit function, use the LINK=glogit option on the model statement:

```

proc logistic data=temp01 ;
  class age_cat (ref="<60") psa_cat (ref="<=20") cancer (ref="No") / param=ref;
  model deftx1 (ref="RP") = age_cat psa_cat cancer / link=glogit;
  format deftx1 tx. age_cat agecat. psa_cat psacat. cancer yesno.;
  title3 "Nominal outcome (generalized logit)";
run;

```

In this case, I recommend specifying which category to use as the reference for the dependent variable (in this case "RP" or radical prostatectomy surgery). The default is for the procedure to use the first ordered category. When a format is used, that will be the category with the format label that comes first alphabetically.

The output for this model will look like the following:

```

          Analysis of Maximum Likelihood Estimates

Parameter          deftx1    DF      Estimate      Standard      Wald
                   deftx1    DF      Estimate      Error         Chi-Square    Pr > ChiSq
Intercept          BT         1      -2.2785      0.1004         514.5663     <.0001
Intercept          EBRT        1      -3.2945      0.1533         461.9548     <.0001
Intercept          HT         1      -2.9824      0.1250         569.4310     <.0001
Intercept          WW         1      -4.2822      0.2606         269.9769     <.0001
age_cat 60-69      BT         1       0.8536      0.1169          53.3243     <.0001
age_cat 60-69      EBRT        1       1.1623      0.1706          46.4282     <.0001
age_cat 60-69      HT         1       0.8793      0.1410          38.8869     <.0001
age_cat 60-69      WW         1       1.2733      0.2878          19.5673     <.0001
age_cat 70+       BT         1       2.5113      0.1248          404.7992     <.0001
age_cat 70+       EBRT        1       3.3565      0.1701          389.2146     <.0001
age_cat 70+       HT         1       3.4165      0.1413          584.6879     <.0001
age_cat 70+       WW         1       4.0723      0.2733          222.0353     <.0001
psa_cat >20        BT         1       0.1100      0.1953           0.3173     0.5732
psa_cat >20        EBRT        1       1.5902      0.1545          106.0009     <.0001
psa_cat >20        HT         1       2.5037      0.1310          365.1683     <.0001
psa_cat >20        WW         1       0.2070      0.2815           0.5407     0.4621
CANCER Yes        BT         1       0.0357      0.1373           0.0677     0.7947
CANCER Yes        EBRT        1      -0.2144      0.1675           1.6371     0.2007
CANCER Yes        HT         1       0.1398      0.1389           1.0134     0.3141
CANCER Yes        WW         1      -0.00547     0.1940           0.0008     0.9775

```

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

Odds Ratio Estimates				
Effect	deftxl	Point Estimate	95% Wald Confidence Limits	
age_cat 60-69 vs <60	BT	2.348	1.867	2.953
age_cat 60-69 vs <60	EBRT	3.197	2.289	4.467
age_cat 60-69 vs <60	HT	2.409	1.827	3.176
age_cat 60-69 vs <60	WW	3.573	2.032	6.281
age_cat 70+ vs <60	BT	12.321	9.647	15.735
age_cat 70+ vs <60	EBRT	28.690	20.555	40.045
age_cat 70+ vs <60	HT	30.461	23.093	40.180
age_cat 70+ vs <60	WW	58.689	34.351	100.273
psa_cat >20 vs <=20	BT	1.116	0.761	1.637
psa_cat >20 vs <=20	EBRT	4.905	3.624	6.639
psa_cat >20 vs <=20	HT	12.227	9.458	15.807
psa_cat >20 vs <=20	WW	1.230	0.708	2.135
CANCER Yes vs No	BT	1.036	0.792	1.356
CANCER Yes vs No	EBRT	0.807	0.581	1.121
CANCER Yes vs No	HT	1.150	0.876	1.510
CANCER Yes vs No	WW	0.995	0.680	1.455

The output here shows the effect of each independent variable on each category of the dependent variable compared to the reference category (RP in this case). For example, men with higher PSA test values (>20) are more likely to receive external beam radiation (OR=4.9) and hormones (OR=12.2) rather than surgery when compared to men with lower PSA values. With a binary outcome, the comparison to the dependent variable reference group still occurs but is less apparent since there are only two categories.

Although the reference category we chose for the dependent variable is RP surgery, it is possible to compare other treatments directly. For example, what if we want to know the effect of PSA level on use of external beam radiation versus brachytherapy radiation? We can subtract the coefficients for BT vs. RP and EBRT vs. RP to get a coefficient for BT vs. EBRT. In this case, $1.59 - 0.11 = 1.48$. Exponentiating 1.48 gives an odds ratio of 4.39. A high PSA increases a patient's likelihood of getting external radiation rather than seed radiation by more than 4 times. This result is similar to what we would find from a bivariate model with only EBRT and BT patients.

It is somewhat more complicated to work with output and interpret the results as the number of nominal categories increases, unlike the ordinal situation where there is one set of parameter estimates and odds ratios for each independent variable.

I recommend outputting the odds ratio results using the ODS facility.

```
ods output oddsratios=nominalOR;
proc logistic data=temp01 ;
  class age_cat (ref="<60") psa_cat (ref="<=20") cancer (ref="No") / param=ref;
  model deftx1 (ref="RP") = age_cat psa_cat cancer /link=glogit;
  format deftx1 tx. age_cat agecat. psa_cat psacat. cancer yesno.;
  title3 "Nominal outcome (generalized logit)";
run;
ods output close;
```

If we print this new data set, we see that it is structured as multiple observations per independent parameter:

Obs	Effect	Response	RatioEst	LowerCL	UpperCL
1	age_cat 60-69 vs <60	BT	2.348	1.867	2.953
2	age_cat 60-69 vs <60	EBRT	3.197	2.289	4.467
3	age_cat 60-69 vs <60	HT	2.409	1.827	3.176
4	age_cat 60-69 vs <60	WW	3.573	2.032	6.281
5	age_cat 70+ vs <60	BT	12.321	9.647	15.735
6	age_cat 70+ vs <60	EBRT	28.690	20.555	40.045
7	age_cat 70+ vs <60	HT	30.461	23.093	40.180
8	age_cat 70+ vs <60	WW	58.689	34.351	100.273
9	psa_cat >20 vs <=20	BT	1.116	0.761	1.637
10	psa_cat >20 vs <=20	EBRT	4.905	3.624	6.639
11	psa_cat >20 vs <=20	HT	12.227	9.458	15.807
12	psa_cat >20 vs <=20	WW	1.230	0.708	2.135
13	CANCER Yes vs No	BT	1.036	0.792	1.356
14	CANCER Yes vs No	EBRT	0.807	0.581	1.121
15	CANCER Yes vs No	HT	1.150	0.876	1.510
16	CANCER Yes vs No	WW	0.995	0.680	1.455

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

This is a classic case of converting multiple observations per unit into one observation per unit. This can be accomplished using a RETAIN statement. The one observation per unit data set can then be displayed easily using PROC PRINT to any of the ODS output destinations (ODS HTML is used here).

```

    *re-format output from multinomial logistic to make presentation table;

proc sort data=nominalOR;
by effect response;
run;

data temp01 (drop= response j oddsratioest lowercl uppercl);
    set nominalOR;

    by effect response;

    retain BT_OR BT_LCI BT_UCI
           EBRT_OR EBRT_LCI EBRT_UCI
           HT_OR HT_LCI HT_UCI
           WW_OR WW_LCI WW_UCI
           ;

    array headervars(12) BT_OR BT_LCI BT_UCI
                        EBRT_OR EBRT_LCI EBRT_UCI
                        HT_OR HT_LCI HT_UCI
                        WW_OR WW_LCI WW_UCI
                        ;

    if first.effect then do j=1 to 12;
        headervars{j}=.;
    end;

    if response="BT" then do;
        BT_OR=oddsratioest;
        BT_LCI=lowercl;
        BT_UCI=uppercl;
    end;

    if response="EBRT" then do;
        EBRT_OR=oddsratioest;
        EBRT_LCI=lowercl;
        EBRT_UCI=uppercl;
    end;

    if response="HT" then do;
        HT_OR=oddsratioest;
        HT_LCI=lowercl;
        HT_UCI=uppercl;
    end;

    if response="WW" then do;
        WW_OR=oddsratioest;
        WW_LCI=lowercl;
        WW_UCI=uppercl;
    end;

    label Effect="Independent Variables"
          BT_OR="BT vs. RP/Odds Ratio"
          BT_LCI="BT vs. RP/Lower CI"
          BT_UCI="BT vs. RP/Upper CI"

          EBRT_OR="EB vs. RP/Odds Ratio"
          EBRT_LCI="EB vs. RP/Lower CI"
          EBRT_UCI="EB vs. RP/Upper CI"

          HT_OR="HT vs. RP/Odds Ratio"

```

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

```

HT_LCI="HT vs. RP/Lower CI"
HT_UCI="HT vs. RP/Upper CI"

WW_OR="WW vs. RP/Odds Ratio"
WW_LCI="WW vs. RP/Lower CI"
WW_UCI="WW vs. RP/Upper CI"
;

if last.effect then output;
run;

ods html file="nominal print example";
proc print data=temp01 split="/" width=uniform;
format BT_OR BT_LCI BT_UCI EBRT_OR EBRT_LCI EBRT_UCI
HT_OR HT_LCI HT_UCI WW_OR WW_LCI WW_UCI oddsr8.1;
run;
ods html close;

```

We will then get a concise and organized table that can be used as the basis for reporting:

Independent Variables	BT vs. RP Odds Ratio	BT vs. RP Lower CI	BT vs. RP Upper CI	EB vs. RP Odds Ratio	EB vs. RP Lower CI	EB vs. RP Upper CI	Etc.
CANCER Yes vs No	1.0	0.8	1.4	0.8	0.6	1.1	
age_cat 60-69 vs <60	2.3	1.9	3.0	3.2	2.3	4.5	
age_cat 70+ vs <60	12.3	9.6	15.7	28.7	20.6	40.0	
psa_cat >20 vs <=20	1.1	0.8	1.6	4.9	3.6	6.6	

Table 1. Sample Presentation of Results for Multinomial Logistic Regression

CONCLUSION

Using the methods explained in this paper, there is no need to reduce dependent categorical variables with more than two categories or levels into a binary variable simply to fit the constraints of the more familiar bivariate logistic regression. For ordinal variables, the cumulative logit model is appropriate as it will retain the ordered nature of the variable. Be aware of the directionality of the variable as this will affect interpretation of results. For nominal variables, use the LINK=glogit option on the MODEL statement to specify the generalized logit function. Finally, specify which category of the dependent variable to use as the referent in the MODEL statement with the REF=refcat option.

REFERENCES

The following resources were invaluable in putting together this presentation:

- Allison PD (1999). *Logistic Regression Using the SAS® System: Theory and Application*, Cary, NC: SAS Institute Inc.
- Stokes ME, Davis CS, Koch GG (1995). *Categorical Data Analysis Using the SAS® System*, Cary, NC: SAS Institute Inc.
- Pritchard ML, Pasta DJ (2004). "Head of the CLASS: Impress Your Colleagues with a Superior Understanding of the CLASS Statement in PROC LOGISTIC." *Proceedings of the 29th Annual SAS Users Group International Conference*, paper 194-29.

Beyond Binary Outcomes: PROC LOGISTIC to Model Ordinal and Nominal Dependent Variables, continued

CONTACT INFORMATION

Your comments and questions are welcomed. Contact the author at:

Eric Elkin
ICON Late Phase & Outcomes Research
188 The Embarcadero, Suite 200
San Francisco, CA 94105
Phone: 415-371-2153
Fax: 415-856-0840
E-mail: eric.elkin@iconplc.com
Web: www.iconplc.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.