MWSUG 2016 - Paper AA18

To be two or not be two, that is a LOGISTIC question

Robert G. Downer, Grand Valley State University, Allendale, MI

ABSTRACT

A binary response is very common in logistic regression modeling. The binary outcome could be the only possible construction of the response but it also could be the result of collapsing of additional response categories. Potential advantages of a binary response include easier interpretation of odds ratios and a single fitted model. Some information will be sacrificed through collapsing but what about other implications? Consequences such as model simplicity and prediction performance are explored through the investigation of data involving an immigration program. Two detailed PROC LOGISTIC examples give relevant syntax and output for a baseline multinomial logit model and a standard binary logistic model. Utilizing standard SAS Stat® procedures for exploratory analysis is shown to be very practical for understanding the modeling. Some familiarity with logistic regression would be helpful for understanding this paper.

INTRODUCTION

This paper is a data driven investigation of collapsing response categories in logistic regression modeling. More specifically, it compares the modeling of a binary logistic regression model to a nominal multi-category logistic model for the same data set. It is meant to provide some insight into some of the decision making associated with collapsing multiple categories to two responses while illustrating relevant features, code, and output of PROC LOGISTIC. In the two detailed examples given, it also illustrates the application of other procedures which might be useful in understanding fitted models. The prediction performance simulation of the second example gives a noteworthy result which may be practically relevant to modelers who might consider collapsing multiple nominal categories.

MODELING BACKGROUND

BINARY LOGISTIC MODEL

As described in Downer (2013) and applied statistics references such as Agresti (2007), a standard logistic regression model with two response categories expresses the log odds of presence versus absence p/(1-p) as a linear function of the predictor variables. The logistic regression model for predictors $X_1 \ldots X_k$ is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + \ldots \beta_k x_k$$

The estimated coefficients $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ can be interpreted on the log-odds or odds scale. Indicator variables are coded for categorical predictors and (in the case of 0,1 predictor coding), exponentiation of the estimated coefficient represents the odds of the response at the given level of the categorical

variable versus the baseline category. For continuous predictors, exponentiation of the estimated coefficient $\hat{\beta}_i$ represents the estimated odds of the response for a unit change in the predictor X_i

The fitted probability $\hat{p}$ can obtained for each observation from a generated output file and the plot of a fitted logistic curve as a function of a continuous predictor can be obtained through a variety of ODS graphics options that have generally been available in PROC LOGISTIC since SAS/STAT 9.1 (SAS/STAT 9.4 was utilized for this work). From a given binary logistic fit, the model can be used with a new observed set of predictors to predict success or failure and hence the regression is being utilized as a classifier for future observations.

GENERALIZED LOGISTIC REGRESSION MODEL

The modeling set-up changes with multiple categories in the response. Assuming a nominal ordering to a response with K categories, then there will be (K-1) models fit by PROC LOGISTIC as a generalized logit model. Ordinal models such as the cumulative logistic model will not be discussed in this paper. It typically makes sense to consider a meaningful baseline nominal category for suitable estimation or predictive interpretation

Following the notation of Agresti (2007), and assuming we label category J as the baseline then the baseline logit model with a single predictor x has the form:

$$\log\left(\frac{\pi_i}{\pi_J}\right) = \alpha_i + \beta_i x$$

Category J typically has the most meaning as the first or last category and J is actually category 1 in the examples of this paper. The left-hand side is the log-odds that the response is classified into category I category as opposed to the baseline category J. If there are only 2 categories, we are in the binary logit model described in the previous section.

So if K=3 and the first category is the baseline, then there will be 3-1=2 logit models fit as:

$$\log\left(\frac{\pi_3}{\pi_1}\right) = \alpha_3 + \beta_3 x \quad \text{and} \quad \log\left(\frac{\pi_2}{\pi_1}\right) = \alpha_2 + \beta_2 x$$

There is a separate intercept and slope for each log-odds (a separate model for 3 vs 1 and 2 vs. 1). This is the basic form of the generalized logit models to be discussed in the two examples to follow. For each of the two models there will be a coefficient fit for the continuous predictor age and C-1 coefficients for a factor predictor variable with C levels (eg. marital status will have 1 coefficient for its main effect in each of the two models).

For a given observation (i.e. a set of predictors), there will be an estimated individual probability from the generalized logit model for each of the k-1 categories. In a generated output file from PROC LOGISTIC, these will be stored in the automatically generated variables _IP_1 through _IP_k.

For the same data set, a comparison of a binary logistic model and a multinomial logit model will be simpler if interaction terms are not significant. One can simply interpret the estimates with respect to odds in the manner described in the previous section. It is much more obvious where differences in the modeling are occurring and exploratory analysis may reveal the reason(s) more explicitly. In Example A, the interaction term is not significant

In a binary logistic model, an interaction between a continuous predictor and categorical predictor will graphically correspond to a comparison of C-1 S-shaped logistic curves where C is the number of levels of the categorical predictor. If the continuous predictor is age and the categorical predictor is gender, for example, the interaction term will represent a differing slope in the possible logistic S curves. If the interaction is significant and the corresponding estimated coefficient is positive (with males coded as1), a change in age of 1 year will result in a significant increase in the odds of the response for males as compared to females. A significant interaction suggests at least some difference from the baseline predictor category to another predictor category as the second variable changes. The **Type 3 analysis of effects** in the LOGISTIC output will be a reasonable initial indicator of interaction significance while the **Analysis of Maximum Likelihood Estimates and Odds Ratio Estimates** will be best for overall understanding. The interpretation of significant interaction terms for a generalized logit fit will be similar for the K-1 models generated..

APPLICATION DATA SET

The data set utilized in this paper is a subset of public data from the New Immigrant Survey (NIS). Versions of the data set are available via registration through the Office of Population Research (OPR). The study and survey involved new legal immigrants to the United States. It involved an initial response upon immigration and a follow-up interview. The goals and description of the study can be found at http://nis.princeton.edu/project.html. Research papers and goals focusing on immigration can be found in Guillemena et al (2006), (2014). One of the goals of the survey was investigating the living conditions of legal immigrants. Observations included in the survey (and those exclusively included in this analysis) are immigrants admitted to the USA under the diversity immigrant visa program (https://travel.state.gov/content/visas/en/immigrate/diversity-visa/entry.html) For investigating the SAS applications and statistical goals of this paper, a real data set with a multi category response was of interest. The housing categorization for these immigrants in the USA satisfied this response criterion for modeling and was viewed as nominal. The mix of continuous and categorical predictors was also desirable.

The only variables from the data set illustrated within this paper are: housing: (3= own or buying a home, 2 = renting, 1=free residence or other), age (continuous in years), marital status (1-married, 0 otherwise), adjustee (1=visa status changed after entering the USA, 0 otherwise), americas (1=migrated from north, central or south America, 0 otherwise). The multi-category housing response appears as pydwell in the examples. For the binary logistic model, the response y has original housing categories 2 and 3 combined into a binary response (paying for housing) and appears as the variable pybin in the examples.

There were 8559 total possible observations available for consideration after deletion of missing housing information

EXAMPLE A: TWO PREDICTORS, SMALL DATA SET

To illustrate estimation in the two modeling strategies, a subset of the immigration data of n=100 was chosen. It was decided that a smaller data set would be more likely to show a meaningful relative magnitude to the impact of each observation in terms of the effect of collapsing response categories and the effect on a final model. The small data focus of this example also provides contrast to working with the entire data set (Example B of the next section).

Age and adjustee and their interaction were selected as predictors for this example. The interaction was insignificant in both types of modeling and removed. With ODS graphics previously invoked, the following code was used for the modeling of the binary response pybin:

```
proc logistic data = Ex1 descending plots = effect;
class  adjustee/ param = glm descending;
model pybin = age adjustee ;
run;
```

DESCENDING on the PROC LOGISTIC line ensures modeling will involve the probability of a 1 response and the options in the CLASS statement ensure a (0,1) indicator set-up for the absence/presence of the adjustee characteristic. Options such as the REF= option are other candidates to achieve the same purpose. The PLOTS = EFFECT option on the first line generates the logistic S-curves for the 2 models (see Figure 1).

As can be seen in Output 1, age and adjustee are both significant. Older immigrants in both adjustee groups are less likely to be paying for housing when the follow-up responses on living conditions were obtained. After accounting for age ,an adjustee still had increased odds of paying for housing (either renting or owning, either 2 or 3 as the response value). The curves have the same steepness due to the fact that interaction has not been included in the model.

Output 1 (Binary Model fit, Example A)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.5273	0.6462	0.6658	0.4145
age		1	-0.0346	0.0157	4.8307	0.0280
adjustee	**1**	1	0.8459	0.4282	3.9015	0.0482
adjustee	**0**	0	0	.	.	.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.966	0.937	0.996
adjustee 1 vs 0	2.330	1.007	5.393

After accounting for age, an adjustee still had increased odds of paying for housing (either renting or owning, 2 or 3 combined as the response variable pybin). The curves in Figure 1 below have the same steepness due to the fact that interaction has not been included in the model.

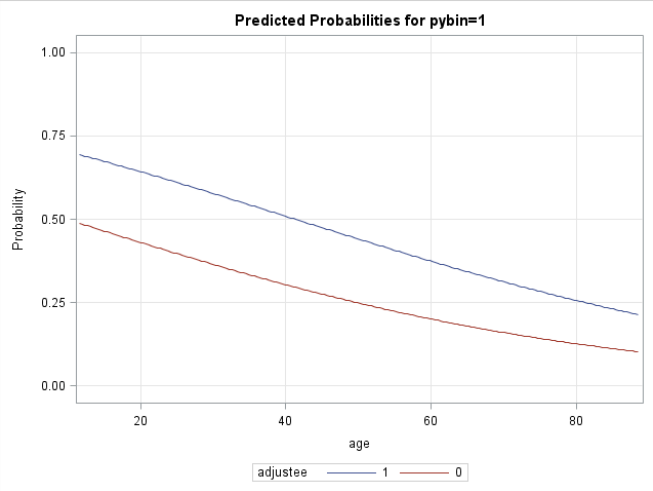

Figure 2 Logistic curves from EFFECTPLOT statement in Binary Model of Example A

The following code generates Output 2 and was used to fit the generalized logistic model to the small data set of 100 observations

```
proc logistic descending ;
class  adjustee/ param = glm descending;
model pydwell = age adjustee / link = glogit;
run
```

Output 2 (Multinomial Model fit, Example A)

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	2	3.7267	0.1552
adjustee	2	6.1309	0.0466

Analysis of Maximum Likelihood Estimates

Parameter		pydwell	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		**3**	1	-1.8032	1.4606	1.5241	0.2170
Intercept		**2**	1	0.0647	0.6660	0.0094	0.9226
age		**3**	1	-0.0488	0.0330	2.1896	0.1389
age		**2**	1	-0.0237	0.0158	2.2392	0.1346
adjustee	**1**	**3**	1	2.5823	1.1098	5.4140	0.0200
adjustee	**1**	**2**	1	0.5473	0.4501	1.4787	0.2240
adjustee	**0**	**3**	0	0	.	.	.
adjustee	**0**	**2**	0	0	.	.	.

Odds Ratio Estimates

Effect	pydwell	Point Estimate	95% Wald Confidence Limits	
Age	**3**	0.952	0.893	1.016
Age	**2**	0.977	0.947	1.007
adjustee 1 vs 0	**3**	13.227	1.502	116.446
adjustee 1 vs 0	**2**	1.729	0.715	4.176

Age is not significant after adjusting for adjustee in the generalized logit model. Why does this difference occur ? We would appear to have a more noteworthy result for the binary model. In general, there will be much less information available for modeling in the smaller data set and sparseness will be evident in combinations of the multi-category response with categorical variables. Continuous predictors will also need to be well represented across each of the response categories. Interactions between the predictors will be even more difficult to detect with less information at predictor combinations across the multi-category response levels. Hence, collapsing to two categories could definitely have some benefit for smaller data sets but understanding through exploratory analysis might be appropriate

Descriptive analysis revealed that the age distribution has a median of 36. To investigate the significance of both predictors in the simpler binary model, a three-way table was produced using PROC FREQ (using less than median age of 36 as the third variable). For the 49 individuals less than median age, 13/21 adjustees (62 percent) were paying for housing. In contrast, in the older group (age greater or equal in age than the median), only 12/28 adjustees (43 percent) were paying for housing. These fractions differ enough to be detected as significant within the estimation of the binary logit model. In the fitting of the multinomial model, there's not enough information in the age distribution (for such a small data set) to detect a possible differing odds of renting as compared to the 'other' category. Histograms of the age distribution across combinations of adjustee and the 3 response categories were generated by the following application of PROC SGPANEL

```
proc sgpanel;
panelby pydwell adjustee /columns = 2 ;
histogram age;
run;
```

As can be seen in the generated display (Figure 2 below), there is little to be gained in modeling the multi-category response by separating out the rent category with both age and adjustee considered (and hence another model comparing renters to 'other' will be redundant). For this small data set, the age distribution of renters (pydwell =2) does not differ much from the age distribution of the 'other' nonpaying housing group (pydwell=1). There is contrast in the age distributions the between pydwell =1 and pydwell =3. However, the differing age distribution for owning a home (pydwell =3) now also directly corresponds to adjustee versus non-adjustee. Hence, for this small data set, investigating the modeling through exploratory analysis has shown that there is little to be gained by having both age and adjustee in this multi-category response model. It may make sense to recommend a binary model and use both predictors in this small data context.

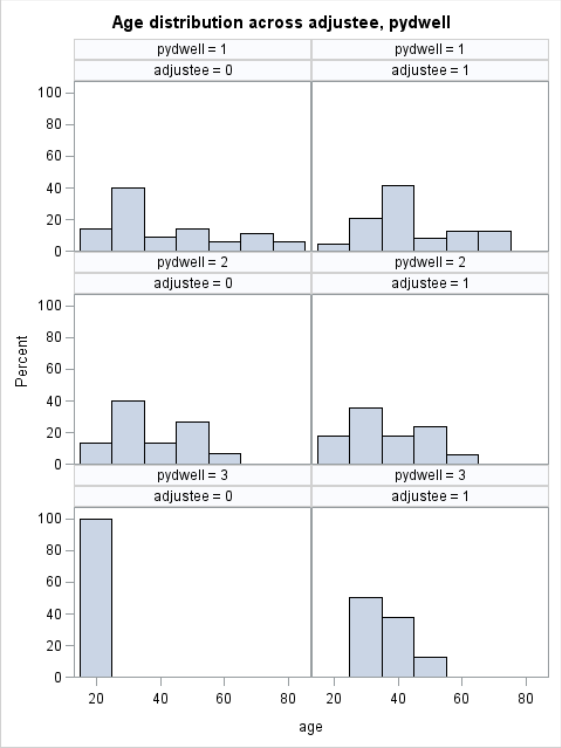

Figure 2 SGPANEL display investigating multinomial fit of Example A.

EXAMPLE B: EVALUATING PREDICTION PERFORMANCE, FULL DATA SET

Exploratory analysis was done with most of the predictor variables and some variables were highlighted for investigating the ability of models to classify new observations as either the binary or multi-category housing response. To investigate the prediction performance comparison adequately, a model with variables leading to only a fair (but not strong) concordance index (area under the ROC curve) for a binary model was deemed to be an appropriate setting for the desired goal. This focus would allow any improvement in prediction performance through a multinomial fit to be more easily identified and

quantified. The full data set and a binary response logistic model were utilized with the following variables and their 2 way interactions: age, marital status, americas, and adjustee (c= 0.617).

The final binary logit model had the 4 main effect variables as well as the interactions americas*age, adjustee*age and adjustee*marital status. The final multinomial model included the main effects and interactions adjustee*marital status, adjustee*Americas, Americas*adjustee and Americas*marital status. So, for example, (for the interaction held in common), the effect of a visa adjustment after arrival depended on marital status. This makes sense since relationships involving immigrants can often lead to a change in status at some point after arrival.

Investigating and exploring prediction performance of the binary and multinomial model for the same training and test data sets was of interest. As a result, a Monte Carlo re-sampling simulation was conducted. A random sample of 100 observations was held out of the data set for prediction with 1000 replications. The response for these 100 was left missing after keeping a copy of the true response. Modeling was based on the other 8459 observations for each replication of the simulation. The binary model and the generalized logit model were each then used to predict the response in each replicate. The simulation process was conducted through the use of PROC SURVEYSELECT. The REP = option allows the sampling to be repeated and indexed by the REPLICATE variable of the output data set. The OUTALL option allows one to keep track of the test set (newly created variable has SELECTED = 1) of the 100 predicted test set observations for each replicate.

The following code performed the generalized logit on each of the 1000 data sets generated by PROC SURVEYSELECT (with actual response copied and set to missing for selected = 1) SURVEYSELECT (with actual response copied and set to missing for selected = 1) .

```
proc logistic descending noprint;
class americas marstat adjustee /param=glm descending;
model pydwell = americas marstat age adjustee
            americas*marstat americas*adjustee adjustee*age
            marstat*adjustee /link = glogit ;
by replicate;
output out = simgl predprobs = individual ;
run;
```

.

In the output data set simgl, the individual predicted probabilities of the output data set are automatically named _IP_1, _IP_2 and _IP_3 as default by PROC LOGISTIC. For this data set and response configuration, the respective predicted probabilities for (1) ‘Other housing’, (2) ‘Renting’ and (3), ‘Own home’ at the time of the survey. Categories (2) and (3) have been combined for the binary response model (with the positive response having a meaning as paying for housing). In the output data set for the generalized logit model, there is also an _INTO_ variable automatically created which contains the category of the maximum of the estimated probabilities _IP_1, _IP_2 and _IP_3. For the binary response, an estimated probability greater than 0.5 (in the output file) was predicted to be a success (paying for housing).

To evaluate prediction performance, (absolute) correct performance was actually predicting the correct category for each of the 100 observations in the test set (and this process was repeated 1000 times). Since chance probability for the generalized logit model would be an estimated probability of 1/3 in each

category, the generalized logit model was at a natural disadvantage. Performance was also evaluated in which the multinomial model would be used to estimate separate category probabilities but collapsing would occur at the estimation stage.

Very interesting results were obtained after an application of PROC MEANS to the simulation performance of 1000 replicates for each of the two modeling strategies. In Output 3 below, we can see the percent correct (mean.617, median 0.62) obtained by the binary logit model was higher than the generalized logit model (.506) as expected since there are only 2 categories. As expected, the percent correct for the binary logit model is very close to the concordance index (0.618) for the full data set for that model. However the .5 median obtained by the multinomial logit model across the 3 categories is more above chance (one-third) than the percent above chance correctness by the binary logit model.

Even more interesting results pertain to the nature of the errors for the multinomial model. If a binary categorization could indeed be acceptable for prediction, then initially using the 3 category response model would appear to reap benefits (at least for this model and data set). If we would have classified a correct prediction as either predicting category 2 or 3 based on the sum of the estimated probabilities, then we would gain an additional 15 percent correct (pctc23)using the multinomial model as compared to the binary model. A fraction of 0.26 would be classified as paying for dwelling when _IP_ 2 + _IP_3 was higher than _IP_1 when _IP_ had the highest individual probability. The correct category in these instances was indeed 2 or 3 but _IP_ 1 was the highest so 1 was chosen as the predicted category. Since the correctly classified _IP_3's were already 0.506 based on _IP_3 being the highest estimated probability (and the correct category was 3), we'd ultimately have on average 0.767 correct if binary prediction was done on collapsing estimated probabilities after the generalized logit model has been run. We had 0.617 correct on average based on collapsing prior to applying the model and using an estimated probability of 0.5 as the classifier. This noteworthy result suggests that post-fit collapsing to two categories from a fit of a multinomial model could be very beneficial if a binary classification is acceptable

Output 3 (Simulation Prediction Evaulation, Example B)

Overall Simulation Summary, Multinomial Model using all 3 categories

The MEANS Procedure

Variable	Mean	Median	Std Dev	Minimum	Maximum
pctcor	0.5058000	0.5000000	0.0445884	0.4200000	0.6000000
pcterr	0.4838000	0.4900000	0.0442115	0.3900000	0.5600000
pctc23	0.2626000	0.2700000	0.0403965	0.1600000	0.3700000

Overall Simulation Summary, Binary Model (Paying for Housing versus 'Other')

The MEANS Procedure

Variable	Mean	Median	Std Dev	Minimum	Maximum
pctcorr	0.6168000	0.6200000	0.0514460	0.4500000	0.7400000
Pcterr	0.3770000	0.3700000	0.0493736	0.2600000	0.5200000

CONCLUSION

This paper demonstrates some aspects of logistic regression modeling for both a binary response and a multi-category nominal response. As well as illustrating features of PROC LOGISTIC, other SAS procedures were utilized to further understand the model fitting in Example A. In Example B, a simulation evaluation of prediction performance showed that collapsing to two categories only after a multinomial fit had been performed could provide potential improvement in prediction accuracy over a binary logistic fit. The application data set was used in order to investigate the SAS and statistical methodology. It is recognized that there are limitations to making general modeling strategy decisions based on this one data but the results provide interesting suggestions for decision making in situations involving a multi-category response.

REFERENCES

Agresti, A. (2007) *An Introduction to Categorical Data Analysis*, Second Edition, Wiley, New York

Downer, R. G. (2013), Improved Interaction Interpretation: Application of the EFFECTPLOT statement and other useful features in PROC LOGISTIC, MWUG 2013, Proceedings of the Midwest SAS Users Group Meeting, Inc., Paper AA-08

Guillermina, J. (2006) Douglas S. Massey, Mark R. Rosenzweig and James P. Smith. "The New Immigrant Survey 2003 Round 1 (NIS-2003-1) Public Release Data." Funded by NIH HD33843, NSF, USCIS, ASPE & Pew. http://nis.princeton.edu.

Guillermina, J (2014) Douglas S. Massey, Mark R. Rosenzweig and James P. Smith. "The New Immigrant Survey 2003 Round 2 (NIS-2003-2) Public Release Data." Funded by NIH HD33843, NSF, USCIS, ASPE & Pew. http://nis.princeton.edu.

CONTACT INFORMATION

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

Robert G. Downer
Biostatistics Director & Professor
Department of Statistics, Grand Valley State University
Allendale, Michigan 49401
downerr@gvsu.edu