

Screening, Transforming, and Fitting Predictors for Cumulative Logit Model

Bruce Lund, OneMagnify

ABSTRACT

The cumulative logit model is a logistic regression model where the target (or dependent variable) has 2 or more ordered levels. If only 2 levels, then the cumulative logit is the binary logistic model. Predictors for the cumulative logit model might be “NOD” (nominal, ordinal, discrete) where typically the number of levels is under 20. Alternatively, predictors might be “continuous” where the predictor is numeric and has many levels. This paper discusses methods that screen and transform both NOD and continuous predictors before the stage of model fitting. Once a collection of predictors has been screened and transformed, the paper discusses predictor variable selection methods for model fitting. One focus of this paper is determining when a predictor should be allowed to have unequal slopes. If unequal slopes are allowed, then the predictor has J-1 distinct slopes corresponding to the J values of the target variable. SAS® macros are presented which implement screening and transforming methods. Familiarity with PROC LOGISTIC is assumed.

INTRODUCTION

Binary logistic regression models are widely used in CRM (customer relationship management) and credit risk modeling. In these models it is common to consider nominal, ordinal, or discrete¹ (NOD) predictors for use in the model. Additionally, continuous predictors are often considered for these models. Generally, there are many potential predictors to consider.

Ordinal logistic regression is a generalization of the binary logistic model. In ordinal logistic regression the target has 2 or more levels and these levels have an ordering. For example, ordinal logistic regression applies to fitting a model where the target is a satisfaction rating (e.g. good, fair, poor).

The cumulative logit model is one form of the ordinal logistic model.² This paper focuses on the cumulative logit model. A special case of the cumulative logit model is the proportional odds (PO) model. The general model is called the partial proportional odds (PPO) model.

The following steps in developing a cumulative logit model are discussed in this paper:

Screening: SAS macros are presented to identify predictors that are too weak to be further considered in building the model.

Transforming: For NOD (nominal, ordinal, discrete with few levels) predictors the transformation involves binning (reducing the number of levels). A SAS macro is discussed that performs binning. In the case of a continuous predictor, a different SAS macro finds a good transformation (e.g. Log(X), X², etc.).

PO versus PPO. The use of a proportional odds versus partial proportional odds is discussed. A PPO model introduces more complexity and may, in certain cases, result in model fitting problems.

Predictor Selection Methods: Predictor selection (modeling fitting) methods are discussed. These methods select the binned or transformed predictors to include in the final model. For a PO model there are (at least) three SAS procedures that provide predictor variable selection methods, PROC LOGISTIC, PROC HPLOGISTIC, PROC HPGENSELECT.³ PROC LOGISTIC also fits a PPO model. If the input data set is suitably modified, then there is an (unconventional) approach for using PROC HPLOGISTIC and PROC HPGENSELECT for fitting a PPO model. This is discussed in Appendix B.

¹ A discrete predictor is a numeric predictor with only “few values”. Often these values are counts. The designation of “few” is subjective. It is used here to distinguish discrete from continuous (interval) predictors with “many values”.

² Allison (2012, Chapter 6) gives an introduction to proportional odds cumulative logit model. See also Agresti (2010) and Hosmer, Lemeshow, Sturdivant (2013).

³ The selection methods include FORWARD, BACKWARD, STEPWISE using significance levels. For HPLOGISTIC the predictors can be selected by best SBC, AIC, and other methods. HPGENSELECT includes selection by LASSO.

Finally, most of the methods of this paper also add value for the fitting of a binary logistic model. But the discussion of PPO models is not relevant to the binary case.

CUMULATIVE LOGIT MODEL WITH PROPORTIONAL ODDS (PO)

Consider this example: Let $p_{k,j}$ = probability that the k^{th} observation has the target value $j = A, B$ or C . Let $X_{k,1}$ be the value of predictor X_1 for the k^{th} observation. Similarly, for $X_{k,2}$.

Then this cumulative logit model has 4 parameters $\alpha_A \alpha_B \beta_{X1} \beta_{X2}$ and is given via 2 response equations:⁴

$$\begin{aligned} \text{Log} (p_{k,A} / (p_{k,B} + p_{k,C})) &= \alpha_A + \beta_{X1} * X_{k,1} + \beta_{X2} * X_{k,2} \quad \dots \text{ response equation } j = A \\ \text{Log} ((p_{k,A} + p_{k,B}) / p_{k,C}) &= \alpha_B + \beta_{X1} * X_{k,1} + \beta_{X2} * X_{k,2} \quad \dots \text{ response equation } j = B \end{aligned}$$

The coefficient β_{X1} of predictor X_1 is the same in both response equations. Similarly, for predictor X_2 .

The parameters for the cumulative logit model are estimated by maximizing the log-likelihood equation in a manner similar to the binary case (Agresti 2010, p 58).

Formulas for the probabilities $p_{k,A}, p_{k,B}, p_{k,C}$ can be derived from the two response equations. To simplify the formulas, let T_k and U_k , for the k^{th} observation be defined by the equations below:

T_k	$= \exp (\alpha_A + \beta_{X1} * X_{k,1} + \beta_{X2} * X_{k,2})$
U_k	$= \exp (\alpha_B + \beta_{X1} * X_{k,1} + \beta_{X2} * X_{k,2})$

Table 1. Supporting Notation

Then, after algebraic manipulation, the probability equations in Table 2 are derived:

Response	Probability Formula
A	$p_{k,A} = 1 - 1/(1+T_k)$
B	$p_{k,B} = 1/(1+T_k) - 1/(1+U_k)$
C	$p_{k,C} = 1/(1+U_k)$

Table 2. Cumulative Logit Model - Equations for Probabilities

This cumulative logit model satisfies the following conditions for X_1 (and the analogous conditions for X_2):

Let “r” and “s” be two values of X_1 and fix the value of X_2 . Using the probability formulas from Table 2:

$$\text{Log} \left[\frac{p_{r,A}/(p_{r,B} + p_{r,C})}{p_{s,A}/(p_{s,B} + p_{s,C})} \right] = \text{Log} (p_{r,A} / (p_{r,B} + p_{r,C})) - \text{Log} (p_{s,A} / (p_{s,B} + p_{s,C})) = (r - s) * \beta_{X1} \dots \text{proportional odds}$$

$$\text{Log} \left[\frac{(p_{r,A} + p_{r,B})/p_{r,C}}{(p_{s,A} + p_{s,B})/p_{s,C}} \right] = \text{Log} ((p_{r,A} + p_{r,B}) / p_{r,C}) - \text{Log} ((p_{s,A} + p_{s,B}) / p_{s,C}) = (r - s) * \beta_{X1} \dots \text{proportional odds}$$

These equations display the “proportional odds” property. Specifically, the difference of cumulative logits at r and s is proportional to the difference (r - s). This proportional odds property is a by-product of the equality of the coefficients of predictor X_1 across the cumulative logit response equations.

EXAMPLE OF FITTING A PO MODEL

If the target variable in PROC LOGISTIC has more than 2 levels, then default model is the cumulative logit model with the proportional odds (PO) property. No LINK statement is needed.

Let Y be an ordered target with numeric predictors X_1 and X_2 . PROC LOGISTIC statements are simply:

```
PROC LOGISTIC DATA= Test;
MODEL Y= X1 X2;
run;
```

⁴ The “cumulative logits” are the log of the ratio of the “cumulative probability up to j” (in the ordering of the target) in the numerator to “one minus the cumulative probability up to j” in the denominator. When the target has levels A, B, C, then j goes from A to B. These cumulative logits are:

$$\begin{aligned} &\text{Log} (p_{k,A} / (p_{k,B} + p_{k,C})) \text{ for the first response equation} \\ &\text{Log} ((p_{k,A} + p_{k,B}) / p_{k,C}) \text{ for the second response equation} \end{aligned}$$

CUMULATIVE LOGIT MODEL WITH PARTIAL PROPORTIONAL ODDS (PPO)

To describe PPO, the following example is given: Assume there are 3 levels for ordered target Y: A, B, C and there are 3 numeric predictors R, S and Z.

Let $p_{k,j}$ = probability that k^{th} observation has the target value $j = A, B$ or C

In this example the PPO Model will have 6 parameters $\alpha_A \alpha_B \beta_R \beta_S \beta_{Z,A} \beta_{Z,B}$ given in 2 equations:

$$\text{Log} (p_{k,A} / (p_{k,B} + p_{k,C})) = \alpha_A + \beta_R * R_k + \beta_S * S_k + \beta_{Z,A} * Z_k \quad \dots j = A$$

$$\text{Log} ((p_{k,A} + p_{k,B}) / p_{k,C}) = \alpha_B + \beta_R * R_k + \beta_S * S_k + \beta_{Z,B} * Z_k \quad \dots j = B$$

Here, Z has different coefficients for the 2 response equations. In general, for PPO some predictors may have coefficients with values that change across response equations. Predictor Z is said to have “unequal slopes”.

Formulas for probabilities $p_{k,A}, p_{k,B}, p_{k,C}$ continue to be given by Table 2 after modifications to definitions of T and U to reflect the PPO model. In unusual cases it is possible for a PPO probability to be negative.⁵

EXAMPLE OF FITTING A PPO MODEL

Suppose Y is an ordered target with 3 levels and that X1 and X2 are numeric predictors. Assume the modeler wants to allow X1 to have coefficients with different values across the response equations. The PROC LOGISTIC statements are:

```
PROC LOGISTIC DATA= Test;  
MODEL Y= X1 X2 / UNEQUALSLOPES= (X1);  
run;
```

A SIMPLIFICATION FOR THIS PAPER

All examples of the cumulative logit model in this paper will assume the target has 3 levels. However, the concepts presented in the paper extend to targets with more than 3 levels.

GOALS OF THE PAPER

Now that the cumulative logit model (PO and PPO) has been defined, the goals of the paper can be listed.

- This paper provides techniques for screening, binning, and transforming of predictors for the cumulative logit model. Additionally, SAS macros are discussed which implement the screening, binning, and transforming.
- The paper discusses tests to determine if a predictor should be allowed to have “unequal slopes”.
- After screening, binning, and transforming predictors the modeler may have many candidate predictors for potential use in a final model. This paper discusses methods for predictor selection to fit a final model. In particular, there is a discussion of SAS procedures which can fit the cumulative logit PO model and, to what extent, these procedures also extend to the fitting of a PPO model.

TEST OF THE PROPORTIONAL ODDS ASSUMPTION

PROC LOGISTIC, when fitting a cumulative logit PO model, performs a test of the proportional odds assumption. A rejection of PO implies that at least one predictor (including dummy variables for a predictor in a CLASS statement) should have unequal slopes across the response equations.

⁵ http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_logistic_examples22.htm
See note at bottom of webpage for discussion. Also see slide 47 of presentation slides by Richard Williams (2008) for a discussion and references. <https://www.stata.com/meeting/germany08/GSUG2008.pdf>

EXAMPLE

A test of the proportional odds assumption is reported by PROC LOGISTIC for this example.

```
DATA Test;
X1=1; X2=3; Y="A"; output;
X1=1; X2=3; Y="B"; output;
X1=1; X2=3; Y="C"; output;
X1=1; X2=3; Y="A"; output;
X1=2; X2=4; Y="A"; output;
X1=2; X2=3; Y="C"; output;
X1=2; X2=3; Y="C"; output;
X1=2; X2=4; Y="C"; output;
X1=2; X2=3; Y="B"; output;
X1=3; X2=3; Y="C"; output;
X1=3; X2=3; Y="A"; output;
X1=3; X2=3; Y="A"; output;
X1=3; X2=4; Y="C"; output;
X1=3; X2=4; Y="B"; output;
run;
PROC LOGISTIC DATA= Test;
MODEL Y= X1 X2;
run;
```

The proportional odds (PO) assumption test has a p-value of 0.830.⁶ This is a strong acceptance of PO. In contrast, a rejection of proportional odds would mean either X1 and/or X2 should have unequal slopes.

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
0.372	2	0.830

Table 3. PROC LOGISTIC Report of Test of Proportional Odds Assumption

SATURATED PPO CUM LOGIT MODEL WITH ONE NOD PREDICTOR

Again consider data set Test. The saturated model with single predictor X1 is the model with X1 in a CLASS statement and X1 in the UNEQUALSLOPES statement.

Since X1 has 3 levels, the CLASS statement produces 2 dummies for the 2 non-reference levels. This leads to 4 coefficients for X1 corresponding to the product of 2 dummies and 2 response equations. The results are shown in Table 4.

```
PROC LOGISTIC DATA= Test;
CLASS X1;
MODEL Y= X1
/ UNEQUALSLOPES= (X1);
```

⁶ Let S be the number of predictors in the model and J be the number of levels of Target. The test statistic is a chi-square with (J-2)*S degrees of freedom. A small value of ProbChiSq_PO rejects the proportional odds assumption. This test may too often lead to rejection as discussed by Allison (2012 p. 168).

Analysis of Maximum Likelihood Estimates							
Parameter		Y	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		A	1	-0.597	0.585	1.041	0.308
Intercept		B	1	0.366	0.577	0.402	0.526
X1	1	A	1	0.597	0.822	0.528	0.468
X1	1	B	1	0.732	0.882	0.690	0.406
X1	2	A	1	-0.789	0.871	0.820	0.365
X1	2	B	1	-0.772	0.782	0.974	0.324

Table 4. PROC LOGISTIC Report for a Saturated PPO Model

However, to fit the saturated model, PROC LOGISTIC does not need to be run. The probabilities for the target Y can be computed by using the three formulas below. These probabilities are the row percentages in the X * Y cross-tabs in Table 5.

$$p_{X1,A} = A_{X1} / (A_{X1} + B_{X1} + C_{X1})$$

$$p_{X1,B} = B_{X1} / (A_{X1} + B_{X1} + C_{X1})$$

$$p_{X1,C} = C_{X1} / (A_{X1} + B_{X1} + C_{X1})$$

where A_{X1} is the count of Y = A when X1 = 1, 2, 3 and similarly for B and C

X1	Y			Tot
	A	B	C	
1	2 .50	1 .25	1 .25	$n_{X1} = 4$
2	1 .20	1 .20	3 .60	$n_{X2} = 5$
3	2 .40	1 .20	2 .40	$n_{X3} = 5$
	$n_A = 5$	$n_B = 3$	$n_C = 6$	$n = 14$

Table 5. Probabilities from the Saturated PPO Model

The log-likelihood and likelihood ratio chi-square can be computed from these probabilities using the formulas shown below:

$$LL(\text{full}) = \text{Log-Likelihood (intercept and covariates)} = \sum_{X1=1}^3 \sum_{j=A}^C n_{X1} * \log(p_{X1,j}) = -14.185$$

$$LL(\text{restricted}) = \text{Log-Likelihood (intercept Only)} = \sum_{j=A}^C n_j * \log(n_j / n) = -14.853$$

$$\text{Likelihood Ratio Chi-Sq} = -2 * \{ LL(\text{full}) - LL(\text{restricted}) \} = 1.337 \text{ (4 d.f.)}$$

In contrast to binary logistic regression where model c (or c-statistic) is an important measure of fit, the model c for the cumulative logit is not widely discussed.⁷ Model c can be computed for the saturated model by using a DATA STEP and then a PROC FREQ to produce a Somer's D statistic. Then a simple

⁷ Definition of Model c. For each observation:

- Let Probabilities for A, B, C be p_k for $k = 1, 2, 3$
- Compute "mean score" as $Mscore = \sum_{k=1}^3 p_k * (k - 1)$.
e.g. If $p_2 = 0.4$ and $p_3 = 0.1$, then $Mscore = 0.4 + 2*0.1 = 0.6$
- IP = "Informative Pairs" of obs (r, s) where Targets $Y_r \neq Y_s$
 - If $Y_r > Y_s$ and $Mscore_r > Mscore_s$, then CONCORDANT
 - If $Y_r > Y_s$ and $Mscore_r < Mscore_s$, then DISCORDANT
 - Else TIE

$$\text{Model c} = \{ \text{CONCORDANT} + 0.5 * \text{TIE} \} / \text{IP}$$

transform gives model c. This is done by the macro %CUM_LOGIT_SCREEN_1 which is discussed after this section.

THE PURPOSE OF LOOKING AT THE SATURATED MODEL

The saturated model for X provides all the information in the predictor X about the target Y. Two ways to measure “all the information” are the significance value for likelihood ratio chi-square and the model c. These measures allow predictors to be screened and ranked. Weak predictors can be eliminated.

The availability of the simple formulas to calculate the likelihood ratio chi-square and the model c for the saturated model provide a way to efficiently compute these summary statistics for many NOD predictors with minimal passing of the data set.

If X passes this screening, the actual usage of X in a model may not be the saturated version. Instead, unequal slopes may not be needed and X may be transformed or binned.

%CUM_LOGIT_SCREEN_1 (Dataset, Target, Input,);

%CUM_LOGIT_SCREEN_1 computes the likelihood ratio chi-square (LRCS), its significance level, and the model c for the saturated model for each predictor that is entered in the INPUT parameter.⁸ Predictors in INPUT can be numeric or character. The macro is illustrated for Target Y and predictors X1 (numeric) and X2 (character) in the data set Test2.

```
DATA Test2;
  X1=1; X2="3"; Y="A"; output;
  X1=1; X2="3"; Y="B"; output;
  X1=1; X2="3"; Y="C"; output;
  X1=1; X2="3"; Y="A"; output;
  X1=2; X2="2"; Y="A"; output;
  X1=2; X2="3"; Y="C"; output;
  X1=2; X2="3"; Y="C"; output;
  X1=2; X2="2"; Y="C"; output;
  X1=2; X2="3"; Y="B"; output;
  X1=3; X2="3"; Y="C"; output;
  X1=3; X2="3"; Y="A"; output;
  X1=3; X2="3"; Y="A"; output;
  X1=3; X2="4"; Y="C"; output;
  X1=3; X2="4"; Y="B"; output;
run;

%CUM_LOGIT_SCREEN_1(Test2, Y, X1 X2, );
```

The macro call produces Table 6.

Var_name	Levels	Log Likelihood Intercept Only	Log Likelihood With Covariate	LRCS	Pr > ChiSq	Model c	ASE Model c
X1	3	-14.853	-14.185	1.337	0.85511	0.6349	0.1123
X2	3	-14.853	-13.322	3.063	0.54736	0.5556	0.0624

Table 6. Report from %CUM_LOGIT_SCREEN_1

The significance level of the LRCS can be used to rank the predictors. Similarly, Model c can rank the predictors. More work is needed in order to give a cut-off for a “bad” Model c. But the rankings, along with the modeler’s subject matter expertise, can screen out obviously weak predictors.

⁸ The macro passes the full data set only once. Then a second pass is made of the output from a PROC SUMMARY.

The significance level is sample size dependent. For example, if a data set is exactly duplicated (two copies), then the LRCS is doubled but has the same degrees of freedom. It is not really practicable to look for an absolute cut-off for all applications.

The right-most macro parameter position (a space in this example) is called SORT, It designates the sort sequence of Table 6. The choices are “LRCS” (for significance), “Model_c”, or “space”. If space, then the order in INPUT is used.

Additionally, a second macro is described below which also bears on the screening question.

%CUM_LOGIT_SCREEN_2 (Dataset, Target, N_Input, C_Input, IV_ADJ);

Weight of evidence (WOE) is often used to transform a NOD predictor in a binary logistic model. WOE can also be defined each “binary split” of the target in a cumulative logit model.

In the case where the target has levels A, B, C there are two splits:

A vs. B,C and A,B vs. C

These two splits lead to two WOE predictors. The first is for the binary pair A vs. B,C and the second is for the binary pair A,B vs. C. If the target has J levels, then there are J-1 WOE's

A numeric X is monotonic for a “split” when the associated WOE is monotonic vs. X. The example in Table 7 shows a predictor X and target Y. WOE1_X is monotonic while WOE2_X is not. The information values, IV1 and IV2, for the associated binary splits are given in the right two columns.

X	Y=A	Y=B	Y=C	Binary: A vs. BC WOE1_X	Binary AB vs. C WOE2_X	Binary: A vs. BC IV1	Binary AB vs. C IV2
1	4	1	1	0.811	0.734	0.225	0.159
2	3	1	3	-0.170	-0.588	0.012	0.157
3	1	2	1	-0.981	0.223	0.204	0.011
				Monotonic	NOT Monotonic	IV1= 0.441	IV2= 0.327

Table 7. WOE and IV for Cumulative Logit with Target Y and Predictor X

Macro %CUM_LOGIT_SCREEN_2 looks at the relationship of a predictor (numeric or character) to the binary splits of the target. The macro parameters are shown:

- Dataset:** Data set containing the Target and Predictors
- Target:** At least 2 levels (missing ignored)
- N_Input:** Numeric Predictors (space delimited, missing ignored)
- C_Input:** Character Predictors (space delimited, missing ignored)
- IV_ADJ:** YES, adds 0.1 to a zero cell to allow IV calculation

```
%CUM_LOGIT_SCREEN_2 (Test2, Y, X1, X2, );
```

Var Name	Spit Point	Levels	Nominal	Monotonic	C_Stat	Model c	IV
X1	A - B	3	NO	NO	0.533	0.644	0.312
X1	B - C	3	NO	NO	0.552	0.656	0.347
X2	A - B	3	YES	n/a	n/a	0.633	0.564
X2	B - C	3	YES	n/a	n/a	0.542	0.034

Table 8. Report from %CUM_LOGIT_SCREEN_2

In Table 8 the Split-Point column shows where the target was split to form the cumulative logit.

The macro produces familiar and readily interpretable statistics which apply to the binary split. The IV column shows if the predictor provides a useful information value with respect to the associated split.

Binary IV guidelines are given below ⁹

IV Range	Interpretation
IV < 0.02	“Not Predictive”
IV in [0.02 to 0.1)	“Weak”
IV in [0.1 to 0.3)	“Medium”
IV ≥ 0.3	“Strong”

Table 9. Practical Guide for Interpreting IV

The c-stat (computed only for numeric variables) and model c (computed for the saturated binary model) provide additional measures of association for the predictor and the binary splits.¹⁰

For this very small example, X1 is strongly associated with both binary splits of the cum logits but this association is not close to being monotonic.¹¹ X2 is strongly associated with the first binary split A vs BC and very weakly associated with the second.

One strategy may be to insist that the predictor provide a “medium” or “strong” IV for at least one split in order to pass the screening.

Example for CUM_LOGIT_SCREEN 1 and 2

The data set CLS_Example is used to illustrate the two Cum_Logit_Screen macros. The target is Y with 3 levels and there are 6 predictors X1-X5 and C5. By construction, X5 has unequal slopes.

```
%LET ERROR = 0.01;
%LET SLOPE1 = 0.01;
%LET SLOPE2 = 0.05;
%LET SLOPE3 = 0.10;
%LET SLOPE4 = 0.20;
%LET SLOPE5 = 0.99;
%LET P_Seed = 5;

DATA CLS_Example2;
Do i = 1 to 8000;
  X1 = floor(12*ranuni(2)) - 1.5;
  X2 = floor(2*ranuni(2)) - .5;
  X3 = floor(2*ranuni(2)) - .5;
  X4 = floor(2*ranuni(2)) - .5;
  X5 = floor(2*ranuni(2)) - .5;
  C1 = put(floor(4*ranuni(2)), z2.);
  C1_all = &SLOPE1*(C1='00') + &SLOPE2*(C1='01') + &SLOPE3*(C1='02');
  rannorx = rannor(1);
/* See Table 1 */
  T = exp(0 + C1_all + &SLOPE1*X1 + &SLOPE2*X2 + &SLOPE3*X3 + &SLOPE4*X4 +
    &SLOPE5*X5 + &ERROR*rannorx);
  U = exp(1 + C1_all + &SLOPE1*X1 + &SLOPE2*X2 + &SLOPE3*X3 + &SLOPE4*X4 +
    &SLOPE5*X5 + &ERROR*rannorx);
/* See Table 2 */
  PA = 1 - 1/(1 + T);
  PB = 1/(1 + T) - 1/(1 + U);
  PC = 1 - (PA + PB);
```

⁹ Siddiqi (2006, p.81)

¹⁰ For the binary logistic, see Lund and Brotherton (2013) for an extensive study of model c.

¹¹ X is monotonic for a split if and only if c-stat = model c. Otherwise, c-stat < model c.

```

/* Assign Target Values to match model probabilities */
R = ranuni(&P_Seed);
if R < PA then Y = "A";
else if R < (PA + PB) then Y = "B";
else Y = "C";
OUTPUT;
end;
run;

```

```
%CUM_LOGIT_SCREEN_1(CLS_Example2, Y, X1 X2 X3 X4 X5 C1, LRCS);
```

The predictors are sorted by Pr > ChiSq (parameter=LRCS). The best ranked predictor is X5. The Model c for X5 is 0.573 (best among the six). Predictor X5 would be retained for further analysis. A question to consider for X5 would be whether unequal slopes are required.

Obs	Var_Name	Levels	Log_L_ Intercept	Log_ Likelihood	LRCS	df	Pr>ChiSq (Num)	Pr > ChiSq	MODEL c	MODEL c _ASE
1	X5	2	-8209.1	-7822.8	772.64	2	0.0000	<.0001	0.573	0.0048
2	X4	2	-8209.1	-8198.4	21.42	2	0.0000	<.0001	0.522	0.0048
3	X1	12	-8209.1	-8190.8	36.63	22	0.0260	0.0260	0.523	0.0055
4	X2	2	-8209.1	-8205.7	6.82	2	0.0331	0.0331	0.510	0.0048
5	C1	4	-8209.1	-8205.1	8.00	6	0.2383	0.2383	0.513	0.0054
6	X3	2	-8209.1	-8207.7	2.68	2	0.2613	0.2613	0.508	0.0048

Table 10: Cum_Logit_Screen_1 applied to CLS_Example

```
%CUM_LOGIT_SCREEN_2(CLS_Example2, Y, X1 X2 X3 X4 X5, C1, YES);
```

Predictor is X5 is very strong for split A v. BC but weak for AB v. C. Strength on (at least) one split is reason to keep X5. The difference in strength of X5 for the 2 binary splits suggests that X5 could have unequal slopes.

Obs	Split_Point	Var_Name	Levels	Nominal	Mono tonic	C_Stat	Model c	IV (Info Value)
1	A - B	C1	4	YES	N/A	n/a	0.515	0.003
2	B - C	C1	4	YES	N/A	n/a	0.516	0.003
3	A - B	X1	12	NO		0.518	0.524	0.011
4	B - C	X1	12	NO		0.517	0.532	0.014
5	A - B	X2	2	NO	YES	0.514	0.514	0.003
6	B - C	X2	2	NO	YES	0.506	0.506	0.001
7	A - B	X3	2	NO	YES	0.509	0.509	0.001
8	B - C	X3	2	NO	YES	0.507	0.507	0.001
9	A - B	X4	2	NO	YES	0.525	0.525	0.010
10	B - C	X4	2	NO	YES	0.523	0.523	0.009
11	A - B	X5	2	NO	YES	0.621	0.621	0.240
12	B - C	X5	2	NO	YES	0.500	0.500	0.000

Table 11: Cum_Logit_Screen_2 applied to CLS_Example

BINNING OF NOD PREDICTORS

Binning is the process of reducing the number of levels of a NOD predictor to achieve parsimony while preserving, as much as possible, the predictive power of the predictor. In addition to parsimony, the purpose of binning is to establish a relationship between the bins and empirical cumulative logits that satisfies business expectations and common sense.

Binning of NOD predictors for the cumulative logit model (PO and PPO) is covered in Lund (2017b). Background regarding binary-target binning is given in Lund (2017a).

A macro %CUMLOGIT_BIN is presented in Lund (2017b) for binning for a cumulative logit model. For this macro the predictor X may be integer (with values 0 to 99) or be character. The target variable is character or numeric with at least 2 levels. Observations with missing values are ignored.

One of the macro parameters is MODE. MODE defines the pairs of levels of predictor X that are eligible for collapsing together. Choices are:

- "A": Any pairs are eligible.
- "J": Only pairs of levels that are adjacent in the ordering of X (this includes character X)

Whatever the value of MODE, a criterion is needed for deciding which levels of X to collapse together. One criterion is to select levels to collapse that minimize, for that the step, the "Total_IV". Total_IV is the sum of the IV's that come from the binary splits of the cum logit model. If the target has J levels, there are J-1 terms to sum for Total_IV. Another criterion is to minimize $-2 * \log$ likelihood for the saturated model.

The macro parameter METHOD is used to specify the criterion to be used for collapsing. Choices are:

- "IV" for TOTAL_IV. Here, the two levels of the predictor which give the greatest "IV" after collapsing (versus all other choices) are the levels which are collapsed at that step.
- "LL" for $-2 * \log(L)$ computed for the saturated model ($-2 * \log$ likelihood). Here, the two levels of the predictor which give the smallest " $-2 * \log(L)$ " after collapsing (versus all other choices) are the levels which are collapsed at that step.

%CUMLOGIT_BIN

The parameters for %CUMLOGIT_BIN are given below:

DATASET: Data set to be processed

TARGET: Target variable with numeric or character values.

TARGET must have at least 2 levels. Missing values are ignored.

X: Predictor variable.

If numeric, then X must have integer values from 0 to 99. There is no restriction on character values. If the character values represent an ordered predictor, then care must be taken to assign values that give the intended ordering. X must have at least 2 levels. Missing values are ignored.

W: A frequency variable if present in DATASET. Otherwise enter 1. Space is not permitted as entry.

METHOD: IV or LL

MODE: A or J

ONE_ITER: YES | <other>.

YES | <other> restricts reports to only statistics for bins before any collapsing. Priority over MIN_BIN

MIN_BIN:

INTEGER > 1 | space. Integer value restricts the processing to bin solutions where the number of BINs is greater or equal to the INTEGER. If <space>, then all bin solutions are processed.

VERBOSE:

YES | <other>. The value YES significantly increases the volume of printed output. The use of YES is not recommended for the initial run of the macro. Once the number of bins in the binning solution is determined by the model, then VERBOSE=YES is run to obtain the SAS coding for weight of evidence and classification level coding. The initial run, without VERBOSE, will create only Summary

Reports. By viewing the Summary Reports the modeler can usually decide on the number of bins for the final binning solution.

EXAMPLE

Consider data set called BACKACHE in Table 12. It gives the age of women who were pregnant and the severity of backache they experienced.¹² Severity has three levels: 1, 2, and 3 with 1 being least severe.

Age Group	Severity		
	1	2	3
15to19	10	5	2
20to22	20	12	2
23to24	19	8	3
25to26	15	13	4
27to28	8	7	2
29to30	6	7	3
31to32	4	5	3
33to35	5	1	4
36andUP	6	2	4
Total	93	60	27

Table 12. Data Set BACKACHE in Pregnancy

The macro call is given below. ANY pairs of levels can be collapsed and the method is IV.

```
%CUMLOGIT_BIN (Backache, Severity, Age_Group, _freq_, IV, A, NO, , NO);
```

Since VERBOSE=NO, only the summary report is printed.

Table 13 shows the columns of the summary report. Under the column “Collapsed” the levels that were collapsed together at that step have a “+” between. Then Minus2_LL (-2*Log Likelihood) and Total_IV are given, followed by IV1 and IV2. Corr_woe_1_2 gives the correlation between the two WOE’s.

A good solution appears to be the 5 bin solution because there is a large drop-off in Total_IV when going to the 4 bin solution (0.598 down to 0.561). Guidelines (like those in Table 12 for binary targets) are, however, needed to interpret TOTAL_IV. An alternative approach may be to apply the binary guidelines to the individual IVs from the cumulative logit model.

Step	Collapsed	minus2_LL	Total_IV	IV_1	IV_2	Corr_woe_1_2
9	None	339.5	0.614	0.138	0.476	0.581
8	25to26+27to28	339.5	0.613	0.138	0.476	0.581
7	15to19+23to24	339.6	0.609	0.135	0.474	0.578
6	33to35+36UP	339.9	0.605	0.135	0.470	0.579
5	29to30+31to32	340.0	0.598	0.134	0.464	0.578
4	15to19_23to24+20to22	341.0	0.561	0.133	0.429	0.638
3	25to26_27to28+29to30_31to32	342.4	0.493	0.111	0.381	0.607
2	25to26_27to28_29to30_31to32+33to35_36andUP	350.4	0.324	0.103	0.221	1

Table 13. Step by Step Results from %CUMLOGIT_BIN with MODE=A

¹² AGE_GROUP of pregnant women who had one of 3 levels of SEVERITY (1 = none or very little pain, 2 = troublesome pain, and 3 = severe pain) of backache from the “BACKACHE IN PREGNANCY” data set in Chatfield (1995, Exercise D.2).

The five-bin solution is shown in Table 14. The bins are not monotonic. Of course, this was not required by the parameter choice of MODE=A.

BIN1	BIN2	BIN3	BIN4	BIN5
15to19_23to24	20to22	25to26_27to28	29to30_31to32	33to35_36andUP

Table 14. Five-Bin Solution from %CUMLOGIT_BIN with MODE=A

If dummy variables for the 5 bins are entered into a model, there is a question of whether unequal slopes should be allowed. One test, which gives an all or none result, is the model comparison test shown below.

```

DATA BACKACHE_5; SET BACKACHE;
  if Age_group in ( "15to19","23to24" ) then Age_group_bin = 1;
  if Age_group in ( "20to22" ) then Age_group_bin = 2;
  if Age_group in ( "25to26","27to28" ) then Age_group_bin = 3;
  if Age_group in ( "29to30","31to32" ) then Age_group_bin = 4;
  if Age_group in ( "33to35","36andUP" ) then Age_group_bin = 5;
PROC LOGISTIC DATA= BACKACHE_5;
CLASS AGE_GROUP;
MODEL SEVERITY= AGE_GROUP;
PROC LOGISTIC DATA= BACKACHE_5;
CLASS AGE_GROUP;
MODEL SEVERITY= AGE_GROUP / UNEQUALSLOPES= (AGE_GROUP);
run;

```

The Model Comparison Test shows that unequal slopes give a significantly different model:

Likelihood Ratio without unequal slopes = 7.553 with 4 d.f.

Likelihood Ratio with unequal slopes = 17.086 with 8 d.f.

Model Comparison Test ChiSq (4 d.f.) = 17.086 - 7.553 = 9.533, with Pr > ChiSq = 0.049

The 5 bin solution could also be implemented by using the two WOE transformations, again with unequal slopes. The correlation between WOE1_X and WOE2_X for 5 bins is only 57.8% (Table 13) which indicates that both WOE's could be entered into the model. See Lund (2017b) for more discussion.

TRANSFORMATION OF CONTINUOUS PREDICTORS FOR CUM LOGIT MODEL

Some modelers may decide to bin a continuous predictor. Some form of fine classing is needed to create preliminary bins. Then a binning algorithm such as %CUMLOGIT_BIN would be applied. However, binning creates arbitrary cut-points (discontinuities) which could distort identification of a true trend. Binning is especially inappropriate for small sample data sets since reproducible bins are unlikely to be found during binning. The alternative to binning is to find a good transformation for a numeric predictor X.

In *Multivariate Model-building* by Royston and Sauerbrei (2008) (R-S) discuss the Function Selection Procedure (FSP).¹³ FSP finds a very good transformation of X in the case of the binary logistic model.

In this paper FSP is extended to finding transformations of X for the cumulative logit PO model. The mechanics of FSP extend without change from the binary case. See Lund (2018) for more detail on FSP and the extension to the cumulative logit model. Later in the paper an extension to the cumulative logit PPO model is discussed.

FSP HAS TWO PRELIMINARY STEPS

First, the predictor X must be positive and, if needed, a translation of X is applied. In fact, for numerical stability, it is better to translate X (if needed) so that the minimum of X is at least 1.

¹³ The development of FSP is due to P. Royston P. and W. Sauerbrei and other researchers in bio-statistics during the 1990's. See Royston and Sauerbrei (2008) for much more detail on this topic.

Next a class of transformations of X , called fractional polynomials (FP), is defined. These fractional polynomials are given by:

X^p where p is taken from $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ and where " X^0 " denotes $\log(X)$

FP1 refers to the collection of linear functions formed by the selection of one X^p . That is,

$$g(X,p) = \beta_0 + \beta_1 X^p$$

FP2 refers to the collection of linear functions formed by selection of two X^p as shown below:

$$G(X,p_1,p_2) = \beta_0 + \beta_1 X^{p_1} + \beta_2 X^{p_2} \quad p_1 \neq p_2$$

However, there is also the special case where the same fractional polynomial is selected twice:

$$G(X,p_1,p_1) = \beta_0 + \beta_1 X^{p_1} + \beta_2 X^{p_1} \log(X) \quad p_1 = p_2$$

FP1 produces only monotonic curves. FP2 produces curves with a variety of non-monotonic shapes. Such an example is given in Figure 1.

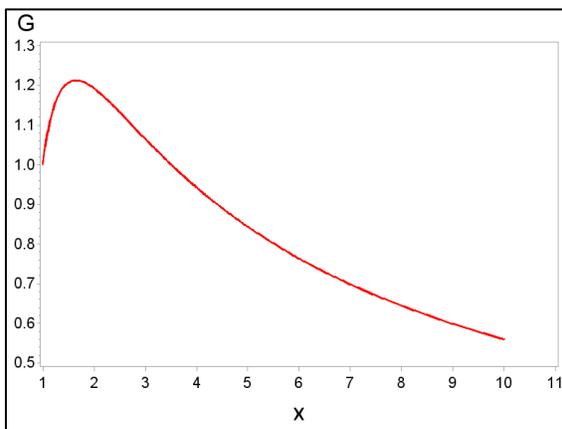


Figure 1. Graph of $G(X,-1,-1) = X^{-1} + 2 X^{-1} \log(X)$

FSP HAS TWO MAIN STEPS

I. Searching for Best Transformations: There is an exhaustive search of FP1 to find the function with maximum likelihood in a cumulative logit regression. This is called the FSP1 solution. Then, a second exhaustive search of FP2 is conducted to find the function with maximum likelihood within this collection. This is called the FSP2 solution.

II. Performing Significance Testing: Second, significance testing is performed. The FSP significance testing follows these three steps. The test-statistic is displayed following the description of steps 1, 2, 3.

1. Perform a 4 d.f. test at the α level of the FSP2 solution against the null model (no predictor). If the test is not significant, drop X from consideration and stop; otherwise continue.
2. Perform a 3 d.f. test at the α level of the FSP2 solution against X . If the test is not significant, stop, the final model is linear X ; otherwise continue.
3. Perform a 2 d.f. test at the α level of the FSP2 solution against the FSP1 solution. If the test is significant, the final model is the FSP2;¹⁴ otherwise the FSP1 is the final model.

The test-statistics for these three tests are the difference of -2 Log Likelihood's, as shown below:

$$\text{Test-Statistic} = (-2 \text{ Log Likelihood}_{\text{restricted model}}) - (-2 \text{ Log Likelihood}_{\text{full model}})$$

For large samples, the Test-Statistic is approximately a chi-square.

¹⁴ If FSP2 solution is selected, then, normally, the two transforms are entered separately into the model. However, if the linear combination has a special appeal, then the FSP2 solution can be entered.

The rationale for the degrees of freedom (4, 3, 2) in the 3-step hypothesis tests of FSP is discussed in R-S (p. 79) for binary logistic models.¹⁵ This rationale generalizes to the cumulative logit model. See Lund (2018) for discussion.¹⁶

IMPLEMENTATION OF FSP BY %FSP_8LR

The macro call is %FSP_8LR (DATASET, TARGET, INPUT, VERBOSE, ORDER);

Parameter definitions:

DATASET: The data set containing the target and predictors

TARGET: Target variable (character or numeric). At least two non-missing levels

INPUT: Numeric predictors (at least one). Predictors are delimited by a space. ("- " convention is implemented ... e.g. X1 - X6).

VERBOSE: YES ... "YES" produces more output.

ORDER: A | D ... The order for modeling the TARGET (A=ascending, D=descending).

%FSP_8LR passes the data once to translate all predictors to have a minimum value of at least 1 and then, once again, to form the 8 fractional polynomial transforms for these predictors.

Next, for each predictor in INPUT, there are 8 PROC LOGISTIC's that are run. For each of the 8 PROC LOGISTIC runs, there are 9 predictors listed in the MODEL statement. The 8 MODEL statements have 8 predictors in common and a ninth that is unique to the statement. See Table 15 in the right-most column.

	COMMON								UNIQUE
%LET Var1=	X	X ⁻²	X ⁻¹	X ⁻⁵	X ^{.5}	X ²	X ³	Log(X)	X Log(X)
%LET Var2=	X ⁻²	X	X ⁻¹	X ⁻⁵	X ^{.5}	X ²	X ³	Log(X)	X ⁻² Log(X)
%LET Var3=	X ⁻¹	X	X ⁻²	X ⁻⁵	X ^{.5}	X ²	X ³	Log(X)	X ⁻¹ Log(X)
%LET Var4=	X ⁻⁵	X	X ⁻²	X ⁻¹	X ^{.5}	X ²	X ³	Log(X)	X ⁻⁵ Log(X)
%LET Var5=	X ^{.5}	X	X ⁻²	X ⁻¹	X ⁻⁵	X ²	X ³	Log(X)	X ^{.5} Log(X)
%LET Var6=	X ²	X	X ⁻²	X ⁻¹	X ⁻⁵	X ^{.5}	X ³	Log(X)	X ² Log(X)
%LET Var7=	X ³	X	X ⁻²	X ⁻¹	X ⁻⁵	X ^{.5}	X ²	Log(X)	X ³ Log(X)
%LET Var8=	Log(X)	X	X ⁻²	X ⁻¹	X ⁻⁵	X ^{.5}	X ²	X ³	Log(X) Log(X)

Table 15. Eight Sets of Predictors for %FSP_8LR

This PROC LOGISTIC is executed for each predictor in INPUT for the 8 MODEL statements of Table 15.

```
PROC LOGISTIC; MODEL Y= &Var<K>
/ SELECTION=FORWARD INCLUDE=1 START=1 STOP=2 SLE=1;
```

The INCLUDE=1 forces the selection of the left most predictor in the list. Then FORWARD with STOP=2 and SLE=1 will select one more predictor. By this method all possible FP2 pairs have a chance to be selected. But the selection of the second variable by FORWARD, to add to the first variable forced in by INCLUDE=1, may be sub-optimal. The reason is that the second variable is selected by the best score chi-square, not by maximizing log likelihood of the model.¹⁷

¹⁵ Loosely, for the first test ("Reject X") there are four parameters to be selection for the FSP2 solution. These are the two exponents (transforms) and the two coefficients. No parameters are required for the null model. The difference is 4 degrees of freedom. Similar logic applies to steps 2 and 3 which have test-statistics of 3 and 2 degrees of freedom.

¹⁶ In Lund (2018) a simulation study is reported which extends the use of 4, 3, 2 d.f. to the cumulative logit PO model. So far, no simulations have been conducted to justify the d.f. that are suggested in the later section for the cumulative logit PPO model.

¹⁷ There might appear to be redundancy in Table 15. Specifically, consider row #1 where the variable X is forced-in. Then in the FORWARD step the next variable to be selected is the one with the greatest score chi-square. Suppose FORWARD selects X⁻². Can row#2, where X⁻² is forced by INCLUDE=1, be simplified by checking only the newly occurring predictor X⁻² Log(X)? The apparent reason would be that X and X⁻² were paired by the logistic model of

E.g. Consider row#1 in Table 15.

- First, X is forced in by INCLUDE=1
- Now perhaps the FORWARD criterion picks X^{-2} to enter as the second variable.
- But the best log likelihood might be given by X^3 .

Examples where %FSP_8LR produces a suboptimal FP2 solution do exist.¹⁸ However, of all examples so far examined, the occurrence rate is not high and the severity is not material.

EXAMPLE OF %FSP_8LR

The code below creates a data set for use in %FSP_8LR.¹⁹ The target Y has three levels. The predictor X has levels 1 through 8, each with a frequency of 500. An "FP2" function is used to generate the values of Y. This FP2 function is: $0.2 \cdot \text{LOG}(X) - 0.5 \cdot (1/X) + 0.01 \cdot \text{error-term}$. The random error-term is drawn from a standard normal and then multiplied by 0.01.

```
%LET ERROR = 0.01;
%LET SLOPE1 = 0.2;
%LET SLOPE2 = -0.5;
%LET P_Seed = 5;
%MACRO SIM(NUM);
%DO Seed = 1 %TO &NUM;
  DATA SIM_&Seed;
  do i = 1 to 8000;
    X = mod(i,16) + 1;
    rannorx = rannor(&Seed);
/* Table 1 */
    T = exp(0 + &SLOPE1*LOG(X) + &SLOPE2*(1/X) + &ERROR*rannorx);
    U = exp(1 + &SLOPE1*LOG(X) + &SLOPE2*(1/X) + &ERROR*rannorx);
/* Table 2 */
    PA = 1 - 1/(1 + T);
    PB = 1/(1 + T) - 1/(1 + U);
    PC = 1 - (PA + PB);
/* Assign Target Values to match model probabilities */
    R = ranuni(&P_Seed);
    if R < PA then Y = "A";
    else if R < (PA + PB) then Y = "B";
    else Y = "C";
    output;
  end;
end;
run;
%END;
%MEND;
%SIM(1);
```

The macro call is: %FSP_8LR (SIM_1, Y, X, NO, A); The summary report is given in Table 16.

Based on three-step testing the recommended transformation is the FSP1 solution $f(X) = X^{-0.5}$. The first two steps ("eliminate X" and "Use Linear") are both strongly rejected with p-values of essentially 0. The FSP1 solution is accepted with p-value of 0.387.

row#1. But this reasoning is not valid. In row#2 the predictor selected by FORWARD can be any of the remaining 8 predictors.

¹⁸ See R-S (2008) p 266 Whitehall I Data. In this data set there are CIGS and ALL10 (values 0 and 1). Let INPUT consist of X where $X = \text{CIGS} + 1$. Let TARGET = ALL10. Run %FSP_8LR. The FP2 solution is X^{-1} and $X^{-1} \cdot \text{Log}(X)$. This compares with the R-S solution of X^{-2} and X^{-1} . The difference in $-2 \cdot \text{Log}(L)$ is 10708.270 vs. 10707.827

¹⁹ This macro program is used in Lund (2018) for a simulation study to support the "d.f. of 4, 3, 2" for FSP significance testing in the case of the cumulative logit PO model.

Since Y has more than 2 levels, the test of proportional odds (PO) is given. In this example, the p-value for the proportional odds test for best FP1 solution is **0.078** (borderline rejection).

Given the borderline rejection the modeler might choose to add unequalslopes to $X^{-0.5}$.

Test Step	-2*Log(L)	Test Stat	d.f.	P-Value	Trans1	Trans2	ChiSq _PO	d.f. _PO	Prob _PO
Eliminate X	15824.5	172.05	4	0					
Use Linear	15709.3	56.86	3	0			1.934	1	0.164
Use FP1	15654.3	1.90	2	0.387	p=-0.5		3.112	1	0.078
Use FP2	15652.4				p=-2	log	2.479	2	0.290

Table 16. Report from %FSP_8LR for Target Y and Predictor X

BETA VERSION: %FSP_8LR_PPO

A second macro %FSP_8LR_PPO implements FSP where UNEQUALSLOPES is specified for each transformation. Since additional parameters are required for PPO, the new 3-step testing involves degrees of freedom as given by these formulas: (Target_Levels is the number of levels of Target.)

- FSP2 vs Null: d.f. = 2*(Target_Levels - 1) + 2
- FSP2 vs Linear: d.f. = 2*(Target_Levels - 1)
- FSP2 vs FSP1: d.f. = (Target_Levels - 1) + 1

However, these degrees of freedoms have not been tested via simulations as was done for the cumulative logit PO model in Lund (2018).

The macro call is: %FSP_8LR_PPO (SIM_1, Y, X, NO,);

The parameters are the same except that 5th parameter "ORDER" is now a dummy parameter with no meaning. (Since coefficients are not reported by %FSP_8LR_PPO, there is no reason to specify ORDER.)

Again, the FSP1 Solution $f(X) = X^{-0.5}$ is the final transformation selected by FSP. The degrees of freedom for the three-step testing are 6, 4, 3. There is little to gain by using the PPO solution in this case.

Test Step	-2*Log(L)	Test Stat	d.f.	P-Value	Trans1	Trans2
Eliminate X	15824.5	175.53	6	0	Null	
Use Linear	15707.4	58.40	4	0		
Use FP1	15651.3	2.37	3	0.499	p=-0.5	
Use FP2	15649.0				p=-1	p=2

Table 17. Report from %FSP_8LR_PPO for Target Y and Predictor X

A PPO cum logit model may have difficulty in converging and some probabilities from a PPO may be negative. This behavior and the greater model complexity of a PPO model argues against the automatic generalization to PPO.

MODEL FITTING METHODS

After binning and transforming the predictors, modeler will probably need to use a variable selection method in order to select the predictors for the final model.

The cumulative logit PO model is fit by three procedures that provide predictor variable selection methods: LOGISTIC, HPLOGISTIC, HPGENSELECT.

PROC LOGISTIC supports FORWARD, BACKWARD, STEPWISE selection where the predictor variables are selected according to user-specific significance levels (also the SCORE method is supported). In

2013 (SAS 9.3) the UNEQUALSLOPES statement was added to PROC LOGISTIC to enable the fitting of PPO models. UNEQUALSLOPES allowed the designated predictors to have different coefficients across the response equations. If UNEQUALSLOPES appears, then all predictor SELECTION methods are available except SELECTION = SCORE.

Both HPLOGISTIC and HPGENSELECT can fit the cumulative logit PO model, and all advanced selection methods that are available for the binary logistic are also available here. These methods include selection by SBC, AIC, VALIDATE, LASSO (HPGENSELECT). But HPLOGISTIC and HPGENSELECT do not fit the PPO model (but see the discussion of this issue near the end of the paper).

EXAMPLE: PREDICTOR SELECTION BY PROC LOGISTIC WITH UNEQUALSLOPES

The predictor selection methods of PROC LOGISTIC with UNEQUALSLOPES can be used to “decide” if a designated predictor should have unequalslopes. Following Derr (2013), the input data set to PROC LOGISTIC requires that “duplicate” predictors be added for the predictors that are thought to be candidates for unequalslopes. In the DATA step creating RANDOM2 a duplicate of X1 is created.

```
DATA RANDOM;
Do ID = 1 to 5000;
  random = ranuni(1);
  If random < .5 then Y = "A";
  else if random < .8 then Y = "B";
  else Y = "C";
  X1 = floor(ranuni(1)*5) * random;
  X2 = rannor(1) * random;
  X3 = ranuni(10) * random;
  X4 = X3*ranuni(10);
  output;
end;
run;
DATA RANDOM2; SET RANDOM;
  X1_Duplicate = X1;
run;
```

The predictor X1_Duplicate appears in UNEQUALSLOPES. If X1_Duplicate enters the model by FORWARD selection, then X1 will have unequalslopes in the model. The significance level to enter (SLE) is set at 0.15 in the FORWARD selection given below.

```
PROC LOGISTIC DATA= RANDOM2;
MODEL Y= X1 X2 X3 X4 X1_Duplicate
/ UNEQUALSLOPES= X1_Duplicate SELECTION= FORWARD SLE= 0.15;
run;
```

X1_Duplicate enters, and therefore, in the model, X1 has unequalslopes. In this model X2 and X4 do not enter.

Analysis of Maximum Likelihood Estimates						
Parameter	Y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	A	1	2.2503	0.0641	1230.6562	<.0001
Intercept	B	1	3.9821	0.0900	1956.0357	<.0001
X3		1	-5.1705	0.1607	1035.6637	<.0001
X1_Duplicate	A	1	-1.1617	0.0478	589.8959	<.0001
X1_Duplicate	B	1	-0.6735	0.0388	301.7948	<.0001

Table 18. Coefficients for PPO Model by PROC LOGISTIC with UNEQUALSLOPES

A disadvantage of PROC LOGISTIC with UNEQUALSLOPES is that the only criterion for predictor selection is “significance level”. The many criticisms of FORWARD, BACKWARD, STEPWISE selection by “p-values” will therefore apply.

PPO MODELS BY HPLOGISTIC OR HPGENSELECT?

Can HPLOGISTIC or HPGENSELECT somehow be “tricked” to fit a cumulative logit PPO Model? The answer is “maybe”, at least for the simple case where only one or second predictors are designated for unequalslopes. Such a method is given in Appendix B. More investigation is needed.

CONCLUSIONS

This paper describes methods for the process of Screening, Transforming, and Fitting Predictors for Cumulative Logit Model:

1. **Screening:** Two macros were provided to screen NOD predictors. The FSP screens continuous predictors.
2. **Binning and Transforming:** A binning macro was provided for NOD predictors. FSP selects a transformation for continuous predictors. The use of unequalslopes must be resolved.
3. **Predictor Selection:** The PO model can be fit by LOGISTIC, HPLOGISTIC, and HPGENSELECT with the same selection methods as used for a binary model. PROC LOGISTIC with UNEQUALSLOPES can fit the PPO model using any selection method except SCORE.

As discussed in Appendix B: For PPO models with only one or two predictors designated to have unequalslopes, it may be possible to use HPLOGISTIC or HPGENSELECT to fit a model. More investigation is needed. Then the advanced predictor selection methods of HPLOGISTIC or HPGENSELECT would be used to find candidate models. Lastly, these *candidate* models would be refit by PROC LOGISTIC with UNEQUALSLOPES to produce the final model.

Indianapolis, IN, MWSUG 2018 v10

REFERENCES

- Agresti, A (2010). *Analysis of Ordinal Categorical Data, 2nd Ed.*, Hoboken, NJ, John Wiley & Sons.
- Allison, P.D. (2012), *Logistic Regression Using SAS: Theory and Application 2nd Ed.*, Cary, NC, SAS Institute Inc.
- Chatfield, C. (1995). *Problem Solving: A Statistician’s Guide, 2nd Ed.*, Boca Raton, FL: Chapman & Hall/CRC.
- Derr, B. (2013). Ordinal Response Modeling with the LOGISTIC Procedure, *Proceedings of the SAS Global Forum 2013 Conference*, Cary, NC, SAS Institute Inc.
- Hosmer D., Lemeshow S., and Sturdivant R. (2013). *Applied Logistic Regression, 3rd Ed.*, New York, John Wiley & Sons.
- Lund B (2017a). SAS® Macros for Binning Predictors with a Binary Target, *Proceedings of the SAS Global Forum 2017 Conference*, Cary, NC, SAS Institute Inc., paper 969.
- Lund, B. (2017b). Binning of Predictors for the Cumulative Logit Model, SESUG 2017, *Proceedings*, Southeast SAS Users Group, paper SD-69.
- Lund, B. (2018). The Function Selection Procedure, *Proceedings of the SAS Global Forum 2018 Conference*, Cary, NC, SAS Institute Inc., paper 2390.
- Lund B. and Brotherton D. (2013). Information Value Statistic, *MWSUG 2013, Proceedings*, Midwest SAS Users Group, paper AA-14.
- Royston P. and Sauerbrei W. (2008). *Multivariate Model-building*, John Wiley & Sons, West Sussex, UK.
- Siddiqi, N. (2006). *Credit Risk Scorecards*, Hoboken, NJ, John Wiley & Sons, Inc.
- Stokes, M., Davis, C., and Koch, G. (2000), *Categorical Data Analysis Using the SAS System, 2nd Edition*, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Bruce Lund
OneMagnify, Detroit, MI
blund_data@mi.rr.com, blund.data@gmail.com, blund@onemagnify.com

All code in this paper is provided by OneMagnify "as is" without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Recipients acknowledge and agree that OneMagnify shall not be liable for any damages whatsoever arising out of their use of this material. In addition, OneMagnify will provide no support for the materials contained herein.

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.

APPENDIX A

EXAMPLE WOE AND INFORMATION VALUE (IV) FOR PREDICTOR X FOR TARGET Y

Table 19 shows the calculations for WOE and IV for a binary target.

X	Y = 0 "B _k "	Y = 1 "G _k "	Col % Y=0 "b _k "	Col % Y=1 "g _k "	Log(g _k /b _k) = X_woe	g _k - b _k	IV Terms (g _k - b _k) * Log(g _k /b _k)
X1	2	1	25.0%	12.5%	-0.69315	-0.125	0.08664
X2	1	1	12.5%	12.5%	0.00000	0	0.00000
X3	5	6	62.5%	75.0%	0.18232	0.125	0.02279
SUM	8	8	100%	100%		IV =	0.10943

Table 19. Binary Target Example of WOE and IV for Binary-Target Y and Predictor X

APPENDIX B

THE DATA RECODING TRICK

The code below shows how observations from a data set for a cumulative logit model can be disaggregated to become observations for a binary logistic. The data set "RANDOM" was created earlier in the MODEL FITTING METHODS section of the paper.

The cumulative logit target Y has 3 levels (A, B, C). In data set "RECODE" there are two outputs for every input observation. The first output statement corresponds to the first cumulative logit split and the second output statement to the second cumulative logit split. The split number (0 or 1) is assigned to SPLIT. A new binary variable TARGET is created where TARGET = 0 if the current Y value corresponds to the current SPLIT, else TARGET = 1.

```
DATA RECODE; Set RANDOM;
  Do; if Y="A" then TARGET=0; else TARGET=1; SPLIT=0; output;
  end;
  Do; if Y="A" or Y="B" then TARGET=0; else TARGET=1; SPLIT=1; output;
  end;
```

PROC GENMOD can fit a Cumulative Logit PPO model if the input data is recoded as above. But this recoding is not sufficient to fit the PPO model. PROC GENMOD has a REPEATED statement that is needed in conjunction with the recoding.²⁰

²⁰ Stokes, Davis, Koch (2000) Categorical Data Analysis, 2nd ed. P. 533

PROC GENMOD

Suppose the modeler wants predictor X1 to have unequal slopes. This is achieved in GENMOD by:

```
PROC GENMOD DATA= RECODE; CLASS ID SPLIT;
MODEL TARGET= SPLIT X1 X2 X3 X4 X1*SPLIT / LINK= logit DIST= bin type3;
REPEATED subject= ID / type= unstructured;
run;
```

The unequal slopes for X1 is accomplished through the inclusion of X1 and the interaction of X1 with SPLIT in the MODEL statement. The coefficient of X1 for the first response equation is the sum of the coefficient of X1 plus the coefficient associated with X1*SPLIT(0). For the second response function it is coefficient of X1 plus the coefficient associated with X1*SPLIT(1)

The statement "REPEATED subject=ID / type =unstructured" corrects for correlation of TARGET values within the two observations of an ID.²¹

PROC LOGISTIC WITH UNEQUALSLOPES, THE GOLD STANDARD

The "gold standard" for fitting the PPO model to Data Set RANDOM with X1 having unequal slopes is PROC LOGISTIC with UNEQUALSLOPES = (X1).

```
PROC LOGISTIC DATA= RANDOM; /* Model #1 */
MODEL Y= X1 X2 X3 X4 / UNEQUALSLOPES= (X1);
run;
```

PROC HPLOGISTIC AND PROC HPGENSELECT WITH DATA CODING TRICK

The data coding trick that was employed for PROC GENMOD can also be used to enable HPLOGISTIC and HPGENSELECT to approximate the fitting of a PPO method, thereby making their advanced predictor SELECTION methods available for model fitting. However, neither procedure has a statement like the REPEATED statement of GENMOD.

But how different will be the results from those of Model #1 (i.e. PROC LOGISTIC above)?

```
PROC HPLOGISTIC DATA= RECODE; /* Model #2 */
CLASS SPLIT /PARAM = ref;
MODEL TARGET= SPLIT X1 X2 X3 X4 X1*SPLIT ;
run;
PROC HPGENSELECT DATA= RECODE; /* Model #3 */
CLASS SPLIT /PARAM = ref;
MODEL TARGET= SPLIT X1 X2 X3 X4 X1*SPLIT / DISTRIBUTION = BINARY;;
run;
```

COEFFICIENTS FROM THE THREE MODELS

In Models #2, #3 some coefficients must be added together in order to have their values be comparable to Model #1. After these adjustments the 3 models have very similar coefficients as shown in Table 20.

²¹ Discussion of REPEATED:

http://documentation.sas.com/?docsetId=statug&docsetVersion=14.3&docsetTarget=statug_genmod_syntax26.htm&locale=en

	LOGISTIC with unequalslopes RANDOM Data MODEL #1	HPLOGISTIC RECODE Data MODEL #2	HPGENSELECT RECODE Data MODEL #3
Intercept A	2.2504	2.2051	2.2051
Intercept B	3.9821	3.9326	3.9326
X1 A	-1.1611	-1.1240	-1.1240
X1 B	-0.6728	-0.6415	-0.6415
X2	0.0332	0.0308	0.0308
X3	-5.1511	-5.2855	-5.2855
X4	-0.0427	-0.0646	-0.0646

Table 20. Coefficients from Fitting PPO Model to Data Set Random or Recode

A comparison of probabilities was made for PROC HPLOGISTIC (Model #2) and PROC LOGISTIC with UNEQUALSLOPES (Model #1). In Table 21 the “Diff A” column is the difference in the probability of A for the same ID’s from Model #1 and Model #2. The average value of the probability of A in Model #1 and in Model #2 is also shown in Table 21. Similar descriptions apply to “Diff B” and “Diff C”.

As shown in Table 21 these probabilities are similar. There are moderately important differences in probabilities for A, B, C for high and low quantiles. The lack of a REPEATED statement in Model #2 (and Model #3) has not greatly degraded the approximation to Model #1.

	Diff A	Diff B	Diff C
Level	Quantile	Quantile	Quantile
100% Max	0.0303	0.0273	0.0203
99%	0.0281	0.0226	0.0164
95%	0.0217	0.0152	0.0117
90%	0.0171	0.0103	0.0084
75% Q3	0.0110	0.0032	0.0019
50% Median	0.0058	-0.0033	-0.0083
25% Q1	0.0029	-0.0058	-0.0154
10%	-0.0004	-0.0089	-0.0243
5%	-0.0029	-0.0108	-0.0306
1%	-0.0081	-0.0133	-0.0387
0% Min	-0.0120	-0.0145	-0.0417
	Prob A	Prob B	Prob C
Mean Prob (Model #1)	0.495	0.312	0.193
Mean Prob (Model #2)	0.488	0.313	0.199

Table 21. Difference in Probabilities for PPO Fit by LOGISTIC vs. HPLOGISTIC with coding trick

STRATEGY FOR PREDICTOR SELECTION FOR PPO MODELS

The hope is that HPLOGISTIC or HPGENSELECT provides a good approximation to the ideal PPO model. With this belief, then it is attractive to use the advanced predictor selection methods of HPLOGISTIC or HPGENSELECT to find good candidate models.

HPLOGISTIC or HPGENSELECT would identify the predictors for candidate models as well as which of these predictors would have unequalslopes. Once a good model or models were identified, then a final fit could be implemented by running PROC LOGISTIC with UNEQUALSLOPES.

EXAMPLES OF SELECTION OF PREDICTORS BY HPLOGISTIC WITH FORWARD AND AIC

HPLOGISTIC runs FORWARD with selection of predictors and choosing of the final model by AIC.

```
PROC HPLOGISTIC DATA= RECODE;
CLASS SPLIT /PARAM = ref;
MODEL TARGET= SPLIT X1 X2 X3 X4 X1*SPLIT / INCLUDE=1;
SELECTION METHOD= FORWARD (SELECT=AIC CHOOSE=AIC STOP=NONE) DETAILS=ALL;
```

Here is the selection summary. SPLIT is forced in (INCLUDE=1) to ensure that two intercepts are created. If X1*SPLIT is entered (as it is), then unequalslopes are selected for X1. The alternative would be to enter X1 with equalslopes or not enter X1 at all. The lowest AIC is achieved at Step 3 and therefore X2 and X4 are not entered.

Selection Summary			
Step	Effect Entered	Number Effects In	AIC
0	Intercept	1	
	SPLIT	2	11928
1	X3	3	9460.96
2	X1	4	8329.04
3	X1*SPLIT	5	8267.28*
4	X2	6	8268.06
5	X4	7	8270.01

Table 22. FORWARD Selections by HPLOGISTIC with SELECT=AIC CHOOSE=AIC

Now a final model would be fit by PROC LOGISTIC with the predictors identified by HPLOGISTIC.

```
PROC LOGISTIC DATA= RANDOM;
MODEL Y= X1 X3 /* X2 X4*/ / UNEQUALSLOPES= (X1);
run;
```