Paper 785-2017

Weight of Evidence Coding for the Cumulative Logit Model

Bruce Lund, Magnify Analytic Solutions, Detroit MI, Wilmington DE, Charlotte NC

ABSTRACT

Weight of evidence (WOE) coding of a nominal or discrete variable X is widely used when preparing predictors for usage in binary logistic regression models. The concept of WOE is extended to ordinal logistic regression for the case of the cumulative logit model. If the target (dependent) variable has J levels, then J-1 WOE variables are needed to recode X. The appropriate setting for implementing WOE coding is the cumulative logit model with partial proportional odds. As in the binary case it is important to bin X to achieve parsimony before the WOE coding. SAS® code to perform this binning is discussed. An example is given that shows the implementation of WOE coding and binning for a cumulative logit model with target variable having three levels.

INTRODUCTION

Binary logistic regression models are widely used in CRM (customer relationship management) and credit risk modeling. In these models it is common to use weight of evidence (WOE) coding of a nominal, ordinal, or discrete[1] (NOD) variable when preparing predictors for use in a logistic model.

Ordinal logistic regression refers to logistic models where the target has more than 2 values and these values have an ordering. For example, ordinal logistic regression is applied when fitting a model to a target which is a satisfaction rating (e.g. good, fair, poor). Here, the scale is inherently non-interval. But in other cases the target could be a count or a truncated count (e.g. children in household: 0, 1, 2, 3+).

The cumulative logit model is one formulation of the ordinal logistic model.[2] In this paper the idea of WOE coding of a NOD predictor is extended to the cumulative logit model. Examples are given where WOE coding of a predictor is used in the fitting of a cumulative logit model. The comparative benefits of WOE coding and dummy-variable coding are discussed.

In either case, binary or ordinal, before the WOE or dummy variable coding it is important that the predictor be “binned”. 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. SAS macros for “optimal” binning of NOD predictors are discussed in the paper.

TRANSFORMING A PREDICTOR BY WOE FOR BINARY LOGISTIC REGRESSION

A NOD predictor C (character or numeric) with L levels can be entered into a binary logistic regression model with a CLASS statement or as a collection of dummy variables. [3] Typically, L is 15 or less.

```
PROC LOGISTIC; CLASS C; MODEL Y = C <and other predictors>;
    or
PROC LOGISTIC; MODEL Y = C_dum_k <and other predictors>;  where k = 1 to L-1
```

These two models produce exactly the same probabilities.

An alternative to CLASS / DUMMY coding of C is the weight of evidence (WOE) transformation of C. It is notationally convenient to use G_k to refer to counts of Y = 1 and B_k to refer to counts of Y = 0 when C = C_k. Let $G = \sum_k G_k$. Then g_k is defined as $g_k = G_k / G$. Similarly, for $B = \sum_k B_k$ and $b_k = B_k / B$.

[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] An introduction to the cumulative logit model is given by Allison (2012, Chapter 6). See also Agresti (2010) and Hosmer, Lemeshow, Sturdivant (2013). These references do not discuss in any detail a generalization of cumulative logit called partial proportional odds (PPO). The PPO model will appear later in this paper.

[3] “CLASS C;” creates a coefficient in the model for each of L-1 of the L levels. The modeler’s choice of “reference level coding” determines how the L^{th} level enters into the calculation of the model scores. See SAS/STAT(R) 14.1 User's Guide (2015), LOGISTIC procedure, CLASS statement.

For the predictor C and target Y of Table 1 the weight of evidence transformation of C is given by the right-most column in the table.

Table 1. Weight of Evidence Transformation for Binary Logistic Regression

C	Y = 0 "B_k"	Y = 1 "G_k"	Col % Y=0 "b_k"	Col % Y=1 "g_k"	WOE= $\mathrm{Log}(g_k/b_k)$
C1	2	1	0.250	0.125	-0.69315
C2	1	1	0.125	0.125	0.00000
C3	5	6	0.625	0.750	0.18232

The formula for the transformation is: If C = "C_k" then C_woe = log (g_k / b_k) for k = 1 to L where g_k, b_k > 0. WOE coding is preceded by binning of the levels of predictor C, a topic to be discussed in a later section.

A Property of a Logistic Model with a Single Weight of Evidence Predictor

When a single weight of evidence variable C_woe appears in the logistic model:

```
PROC LOGISTIC DATA = <> DESCENDING; MODEL Y = C_woe;
```

then the slope coefficient equals 1 and the intercept is log (G/B). This property of a WOE predictor is verified by substituting the solution α = log (G/B) and β = 1 into the maximum likelihood equations to show that a solution has been found. This solution is the global maximum since the log likelihood function has a unique extreme point and this point is a maximum (ignoring the degenerate cases given by data sets having quasi-complete and complete separation). See Albert and Anderson (1984, Theorem 3).

Information Value of C for Target Y

An often-used measure of the predictive power of predictor C is Information Value (IV). It measures predictive power without regard to an ordering of a predictor. The right-most column of Table 2 gives the terms that are summed to obtain the IV. The range of IV is the non-negative numbers.

Table 2. Information Value Example for Binary Logistic Regression

C	Y = 0 "B_k"	Y = 1 "G_k"	Col % Y=0 "b_k"	Col % Y=1 "g_k"	$\mathrm{Log}(g_k/b_k)$	g_k - b_k	**IV Terms** (g_k - b_k) * $\mathrm{Log}(g_k/b_k)$
C1	2	1	0.250	0.125	-0.69315	-0.125	0.08664
C2	1	1	0.125	0.125	0.00000	0	0.00000
C3	5	6	0.625	0.750	0.18232	0.125	0.02279
SUM	8	8				**IV =**	**0.10943**

IV can be computed for any predictor provided none of the g_k or b_k is zero. As a formula, IV is given by:

$$\mathbf{IV = \sum_{k=1}^{L} (g_k - b_k) * \log (g_k / b_k)}$$

where L $\geq$ 2 and where g_k and b_k > 0 for all k = 1, …, L

Note: If two levels of C are collapsed (binned together), the new value of IV is less than or equal to the old value. The new IV value is equal to the old IV value if and only if the ratios g_r / b_r and g_s / b_s are equal for levels C_r and C_s that were collapsed together.[4]

Predictive Power of IV for Binary Logistic Regression

Guidelines for interpretation of values of the IV of a predictor in an applied setting are given below. These guidelines come from Siddiqi (2006, p.81). In logistic modeling applications it is unusual to see IV $\geq$ 0.5.

[4] See Lund and Brotherton (2013, p. 17) for a proof.

Table 3. Practical Guide to Interpreting IV

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 $\geq$ 0.3	“Strong”

Before the final coding of WOE variables, these variables should undergo a binning process to reduce the number of levels in order to achieve parsimony but while maintaining predictive power to the fullest extent possible. This important topic is discussed in a later section.

THE DEFINING CHARACTERISTICS OF WEIGHT OF EVIDENCE CODING

The next step is to explore the extension of weight of evidence coding and information value to the case of ordinal logistic regression and, in particular, to the cumulative logit model.

There are two defining characteristics of the weight of evidence coding, X_woe, of a predictor X when the target is binary and X_woe is the single predictor in a logistic model. These are:

1. Equality of Model (I) and Model (II):

   (I) **PROC LOGISTIC** DESCENDING; CLASS X; MODEL Y = X;

   (II) **PROC LOGISTIC** DESCENDING; MODEL Y = X_woe;

2. The values of the coefficients for Model (II): Intercept = Log (G / B) and Slope = 1

GOAL: Find a definition of WOE to extend to the cumulative logit model so that the appropriate generalizations of (1) and (2) are true.

CUMULATIVE LOGIT MODEL

If the target variable in PROC LOGISTIC has more than 2 levels, PROC LOGISTIC regards the appropriate model as being the cumulative logit model with the proportional odds property.[5] An explanation of the cumulative logit model and of the proportional odds property is given in this section.

A Simplification for this Paper

In this paper all discussion of the cumulative logit model will assume the target has 3 levels. This reduces notational complexity. The concept of weight of evidence for the cumulative logit model does not depend on having only 3 levels. But the assumption of 3 levels does provide crucial simplifications when applying the weight of evidence approach to examples of fitting cumulative logit models, as will be seen later in the paper.

Definition of the Cumulative Logit Model with the Proportional Odds (PO) Property

To define the cumulative logit model with PO, the following example is given: Assume the 3 levels for the ordered target Y are A, B, C and suppose there are 2 numeric predictors X1 and X2.[6]

Let $p_{k,j}$ = probability that the kth observation has the target value j = A, B or C. Let $X_{k,1}$ be the value of X1 for the kth observation. Similarly, for $X_{k,2}$.

Then this cumulative logit model has 4 parameters α_A α_B β_{X1} β_{X2} and is given via 2 response equations:

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

The coefficients β_{X1} and β_{X2} of predictors X1 and X2 are the same in both response equations.

[5] Simply run: PROC LOGISTIC; MODEL Y = <X’s>; where Y has more than 2 levels.

[6] If a predictor X is not numeric, then the dummy variables from the coding of the levels of X appear in the right-hand-side of the response equations for j = A and j = B.

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.

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:

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

Then, after algebraic manipulation, these probability equations are derived:

Table 4. Cumulative Logit Model - Equations for Probabilities

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	$p_{k,C} = 1/(1+U_k)$

The parameters for the cumulative logit model are estimated by maximizing the log likelihood equation in a manner similar to the binary case.[7]

This cumulative logit model satisfies the following conditions for X1 (and the analogous conditions for X2):

Let “r” and “s” be two values of X1 and fix the value of X2. Using the probability formulas from Table 4:

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

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

These equations display the “proportional odds” property. Specifically, the difference of cumulative logits at r and s is proportional to the difference (r - s). The proportional odds property is a by-product of the equality of the coefficients of predictors X1 and X2 across the cumulative logit response equations.

WOE AND INFORMATION VALUE FOR THE CUMULATIVE LOGIT MODEL

After trial and error, when trying to define an extension of weight of evidence coding of X for the cumulative logit model, I realized that if Y had J levels, then J-1 WOE transformations were needed. WOE can be defined for the cumulative logit model without assuming proportional odds (i.e. without assuming equal predictor coefficients across response equations).

Consider an ordinal target Y with levels A, B, C and predictor X with levels 1, 2, 3, 4. Here, Y has 3 levels and, therefore, 2 weight of evidence transformations are formed.

The formulas for the two weight of evidence variables X_woe1 and X_woe2 are given below.

$X\_woe1\ (X=i) = \text{LOG}\ [(A_i / n_A) / ((B_i + C_i) / (n_B + n_C))]$
$X\_woe2\ (X=i) = \text{LOG}\ [((A_i + B_i) / (n_A + n_B)) / (Ci / n_C)]$

where A_i = count of A’s for X = i, similarly for B_i and C_i, and $n_A = \sum A_i$, $n_B = \sum B_i$, $n_C = \sum C_i$

The (X, Y) values from Table 5 may be substituted into the formulas for X_woe1 and X_woe2 to verify the values in the right two columns of Table 5.

Table 5. Example Data Set with Target Y and Predictor X and Weight of Evidence Coding of X

	Y=				
X=i	Ai	Bi	Ci	X_woe1	X_woe2
1	2	1	2	0.03	-0.18
2	4	3	1	0.44	1.36
3	4	1	2	0.72	0.33
4	1	2	5	-1.51	-1.10
Total	11	7	10		

[7] See Agresti (2010, p 58).

Although X in this example is numeric, any NOD predictor may take the role of X.

There is no mystery associated with X_woe1 and X_woe2. In the case of X_woe1 this is simply the binary weight of evidence transform for the binary target A vs {B and C}. Similarly, X_woe2 is the binary weight of evidence transform for the binary target {A and B} vs C.

There is also the natural extension of Information Value to the cumulative logit model. Information value is computed for the binary target A vs {B and C}. This is called IV1. Similarly, for {A and B} vs C there is IV2. One approach to defining a single information value for predictor X is to sum IV1 and IV2 and call this TOTAL_IV.

The reader may check for the data of Table 5 that IV1 = 0.6760 and IV2 = 0.7902.

Cumulative Logit Model with PO Does Not Support a Generalization of WOE

Data set EXAMPLE1 is created from the counts of Table 5. The use of EXAMPLE1 will show that the cumulative logit PO model does not support the required two characteristics for a WOE predictor. To show the failure of the WOE definitions in the cumulative logit PO case, the Models (I) and (II) are considered:

```
(I)  PROC LOGISTIC DATA = EXAMPLE1; CLASS X; MODEL Y = X;
(II) PROC LOGISTIC DATA = EXAMPLE1; MODEL Y = X_woe1 X_woe2;
```

The reader may verify the Models (I) and (II) do not produce the same probabilities. In addition, the coefficients of Model (II) do not have the required values.

Table 6. Results of MODEL (II) for the PO Model

Maximum Likelihood Estimates			Not Equal to:	
Parameter		Estimate		
Intercept	A	-0.4870	≠ -0.4353	=Log(A/(B+C))
Intercept	B	0.7067	≠ 0.5878	=Log((A+B)/C)
X_Woe1		0.6368	≠1	
X_Woe2		0.2869	≠1	

The next section describes the partial proportional odds (PPO) cumulative logit model and how weight of evidence does naturally extend to this setting.

Partial Proportional Odds (PPO) Cumulative Logit Model

To describe the PPO cumulative logit model, the following simple example is given: Assume there are 3 levels for the ordered target Y: A, B, C and suppose 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 α_A α_B β_R β_S $\beta_{Z,A}$ $\beta_{Z,B}$ given in 2 equations:

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

The coefficients of the predictors β_R and β_S are the same in the 2 equations but $\beta_{Z,j}$ varies with j. In general, for PPO some predictors may have coefficients with values that vary across response equations.

The formulas for the probabilities $p_{k,A}$, $p_{k,B}$, $p_{k,C}$ continue to be given by Table 4 after modifications to the definitions of T and U to reflect the PPO model.

Weight of Evidence in the Setting of PPO Cumulative Logit Model

Models (I) and (II) are modified to allow the coefficients of the predictors to depend on the cumulative logit response function. This is accomplished by adding the UNEQUALSLOPES statement.

```
(I)  PROC LOGISTIC DATA = EXAMPLE1; CLASS X;
      MODEL Y = X / unequalslopes = (X);
(II) PROC LOGISTIC DATA = EXAMPLE1;
      MODEL Y = X_woe1 X_woe2 / unequalslopes = (X_woe1 X_woe2);
```

For data set EXAMPLE1, Models (I) and (II) are the same model (produce the same probabilities). Model (II) produces coefficients which generalize WOE coefficients from the binary case. Formulas for these coefficients are shown below:

$$\alpha_A = \log (n_A / (n_B + n_C)) \quad \alpha_B = \log ((n_A + n_B) / n_C)$$
$$\beta_{X\_woe1,A} = 1,\ \beta_{X\_woe1,B} = 0; \qquad \ldots \text{ (Eq. A)}$$
$$\beta_{X\_woe2,A} = 0,\ \beta_{X\_woe2,B} = 1;$$

where n_A is count of Y = A, n_B is count of Y = B, n_C is count of Y = C

The regression results from running Model (II) are given in Table 7.

Table 7. Results of MODEL (II) for the PPO Model with WOE Predictors

Maximum Likelihood Estimates			Equal to:	
Parameter		Estimate		
Intercept	A	-0.4353	-0.4353	=Log(A/(B+C))
Intercept	B	0.5878	0.5878	=Log((A+B)/C)
X_Woe1	A	1.0000	1	
X_Woe1	B	-127E-12	0	
X_Woe2	A	3.2E-10	0	
X_Woe2	B	1.0000	1	

Conclusion Regarding the Usage of Weight of Evidence Predictors

In order to reproduce the two defining characteristics of the weight of evidence predictor from the binary case, the weight of evidence predictors should enter a cumulative logit model within the unequalslopes parameter.

Comments

There are degenerate {X, Y} data sets where a cumulative logit model has no solution.[8] Setting these cases aside, I do not have a solid mathematical proof that coefficients, as given by Eq. A, always produce the maximum likelihood solution for Model (II) or that Model (I) and Model (II) are always equivalent. I am relying on verification by examples.

Using the parameter values found for Model (II) the probabilities when X = r for target levels A, B, and C are obtained by substitution into the equations of Table 4.

$$p_{r,A} = A_r / (A_r + B_r + C_r)$$
$$p_{r,B} = B_r / (A_r + B_r + C_r)$$
$$p_{r,C} = C_r / (A_r + B_r + C_r)$$

where A_r is the count of Y = A when X = r, etc.

EXAMPLE: BACKACHE DATA, LOG OF AGE, AND SEVERITY WITH 3 LEVELS

A paper by Bob Derr (2013) at the 2013 SAS Global Forum discussed the cumulative logit PO and PPO models. In the paper Derr studied the log transform of the AGE (called LnAGE) of pregnant women who had one of 3 levels of SEVERITY[9] of backache from the "BACKACHE IN PREGNANCY" data set from Chatfield (1995, Exercise D.2). In the Appendix there is a data set called BACKACHE with 61 observations which expands to 180 after applying a frequency variable. It has AGE and SEVERITY (and frequency variable _FREQ_) from the BACKACHE IN PREGNANCY data set. See this data set for the discussion that follows below.

Using a statistical test called OneUp, Derr shows it is reasonable to use unequalslopes for LnAGE when predicting SEVERITY in a logistic model.[10]

[8] Agresti (2010 p. 64)
[9] 1 = none or very little pain, 2 = troublesome pain, and 3 = severe pain
[10] P-value of 0.06 to reject equal slopes. See the Appendix for explanation and SAS code to conduct the test.

The weight of evidence transformations of AGE will be used in a PPO model for SEVERITY and will be compared with the results of running a cumulative logit model for LnAGE with unequalslopes.

The logistic model for SEVERITY with unequalslopes for LnAGE has the fit statistics in Table 8a and Table 8b.

```
PROC LOGISTIC DATA = Backache;
MODEL SEVERITY = LnAGE / unequalslopes = LnAGE;
Freq _freq_;
run;
```

Table 8a. SEVERITY from Backache Data Predicted by LnAGE with Unequalslopes

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	361.104	357.423
SC	367.490	370.194
-2 Log L	357.104	349.423

Table 8b. SEVERITY from Backache Data Predicted by LnAGE with Unequalslopes

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	7.6819	2	0.0215
Score	7.5053	2	0.0235
Wald	7.3415	2	0.0255

Replacing LnAGE by Weight of Evidence

What improvement in fit might be achieved by replacing LnAGE with AGE_woe1 and AGE_woe2? This is explored next.

The AGE * SEVERITY cells have zero counts when AGE < 19, AGE = 22, and AGE > 32. To eliminate these zero cells, AGE levels were collapsed as shown. AGE had 13 levels after this preliminary binning.

```
DATA Backache2; Set Backache;
if AGE < 19 then AGE = 19;
if AGE = 22 then AGE = 23;
if AGE > 32 then AGE = 32;
```

Next, AGE_woe1 and AGE_woe2 were computed. Before entering AGE_woe1 and AGE_woe2 into the MODEL their correlation should be checked. The correlation of AGE_woe1 and AGE_woe2 was found to be 58.9% which is suitably low to support the use of both predictors in a model.

Now the PPO model, shown below, was run;

```
PROC LOGISTIC DATA = Backache2;
MODEL SEVERITY = AGE_woe1 AGE_woe2 / unequalslopes = (AGE_woe1 AGE_woe2);
Freq _freq_;
run;
```

The fit was improved, as measured by -2 * Log L, from 349.423 to 336.378 as seen in Table 9a.

Table 9a. SEVERITY from Backache Data Predicted by WOE coding of AGE with Unequalslopes

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	361.104	348.378
SC	367.490	367.536
-2 Log L	357.104	336.378

Penalized Measures of Fit Instead of Log-Likelihood

But the measures AIC and SC (Schwarz Bayes criterion) of parsimonious fit of 348.378 and 367.536 are not correctly computed when weight of evidence predictors appear in a model. The weight of evidence predictors should count for a total of 24 degrees of freedom and not the 4 counted by PROC LOGISTIC, as shown in the Testing Global Null Hypothesis report, Table 9b.

Table 9b. SEVERITY from Backache Data Predicted by WOE coding of AGE with Unequalslopes

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	20.7264	4	0.0004
Score	22.0324	4	0.0002
Wald	20.2476	4	0.0004

The penalized measures of fit, AIC and SC should be recomputed to match the Model Fit Statistics for the equivalent model with a CLASS statement for AGE shown below in Table 10.

```
PROC LOGISTIC DATA = Backache2;
CLASS AGE;
MODEL SEVERITY = AGE / unequalslopes = (AGE);
Freq _freq_;
run;
```

Table 10. Model Fit Statistics with Adjusted Degrees of Freedom

Model Fit Statistics (adjusted)		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	361.104	388.378
SC	367.490	471.395
-2 Log L	357.104	336.378

The adjusted SC of 471.395 is much higher than the SC of 370.194 from the PPO model with LnAGE. Similarly, the adjusted AIC of 388.378 is much higher than the 357.423 from the PPO model with LnAGE.

BINNING PREDICTORS FOR CUMULATIVE LOGIT MODELS

The predictors AGE_woe1 and AGE_woe2 used all 13 levels of AGE. Perhaps these 13 levels could be binned to a smaller number to achieve parsimony and still retain most of the predictive power?

For logistic models with binary targets there are methods to decide which levels of the predictor to bin together, at each step, so as to maximize the remaining predictive power. These methods involve binning so as to optimize: (i) Information Value, (ii) Log Likelihood (equivalent to entropy), or (iii) p-value from the chi-square measure of independence of X and the target. How could these binary methods be generalized for binning decisions for the cumulative logit model?

For the cumulative logit model, the use of Information Value for binning is complicated because each weight of evidence predictor has its own IV. One approach for binning decisions is to compute TOTAL_IV by simply summing the individual IV's. This approach is adopted in this paper.

A macro called %CUMLOGIT_BIN performs binning for the cumulative logit model. For this macro the target has J ≥ 2 ordered values and the predictor X may be integer (with values 0 to 99) or character.

Two of the parameters for %CUMLOGIT_BIN are:

- MODE: Defines the pairs of levels of predictor X that are eligible for collapsing together. The choice is (1) "any pairs are eligible" or (2) "only pairs of levels that are adjacent in the ordering of X".
- METHOD: Defines the rule for selecting a pair for collapsing. The choices are TOTAL_IV and -2*LOG(L). For TOTAL_IV the two levels of the predictor which give the greatest TOTAL_IV after collapsing (versus all other choices) are the levels which are collapsed at that step. A similar description applies if -2*LOG(L) is selected for minimizing. (Note: -2*LOG(L) is equivalent to entropy.)

Once selected, MODE and METHOD are applied at each step in the binning process.

%CUMLOGIT_BIN APPLIED TO AGE AND SEVERITY FROM BACKACHE

TOTAL_IV and adjacent-only collapsing were selected for %CUMLOGIT_BIN and applied to AGE from the Backache data set. There were 13 levels for AGE after the initial zero-cell consolidation.

The summary results of the binning are shown in Table 11.

The AIC and SC columns have been adjusted for degrees of freedom for weight of evidence. If AIC and SC are not a concern for predictor variable preparation before modeling, then the 9-bin solution has appeal since TOTAL_IV begins to fall rapidly thereafter. This solution gives -2 * Log L = 336.92 in comparison with 349.423 for LnAGE (Table 8a). The correlation between AGE_woe1 and AGE_woe2 is moderate at 61%.

Table 11. Binning of AGE vs. SEVERITY from BACKACHE DATA. MODE = ADJACENT, Method = TOTAL_IV

BINS	MODEL DF With Intercept	-2_LL	IV_1	IV_2	Total_IV	Adj. AIC	Adj SC	Correlation of AGE_woe1 and AGE_woe2
13	26	336.38	0.237	0.489	0.726	388.38	471.39	0.5886
12	24	336.46	0.236	0.489	0.725	384.46	461.09	0.5924
11	22	336.51	0.235	0.488	0.723	380.51	450.76	0.5916
10	20	336.60	0.232	0.487	0.720	376.60	440.46	0.5922
9	18	336.92	0.229	0.484	0.713	372.92	430.39	0.6119
8	16	337.44	0.218	0.482	0.700	369.44	420.53	0.6156
7	14	339.16	0.198	0.472	0.670	367.16	411.86	0.7002
6	12	340.04	0.178	0.462	0.640	364.04	402.36	0.6893
5	10	341.54	0.144	0.461	0.604	361.54	393.47	0.7336
4	8	344.50	0.121	0.443	0.564	360.50	386.04	0.8827
3	6	345.34	0.108	0.409	0.517	357.34	376.50	0.8606
2	4	348.01	0.049	0.382	0.430	356.01	368.78	1.0000

Selection of the 9-bin WOE solution with unequalslopes, in conjunction with other predictors of SEVERITY, is likely to provide an improvement in the full Backache Model versus using LnAGE with unequalslopes.

AN ALTERNATIVE TO %CUMLOGIT_BIN

Now let us assume that the medical expectation is that severity of backache should increase with age. In this case, LOG(P1/(P2+P3)) should be decreasing (more SEVERITY with AGE). Although the general trend of AGE_woe1 is decreasing for the 9-bin solution (Table 12), it is not monotonically decreasing. The same observations apply to AGE_woe2. Small sample size might be the problem. Nonetheless, to address this issue we could revise our approach and look for monotonic solutions.

Table 12. Weight of Evidence for Age with 9 Bins

Age	AGE_woe1	AGE_woe2
19 & under	0.290	0.280
20	-0.354	0.057
21	1.032	0.973
22-23	0.020	0.617
24	0.626	1.099
25	0.270	-0.125
26-29	-0.341	0.320
30-31	-0.878	-0.531
32 & over	-0.200	-1.041

But %CUMLOGIT_BIN maximizes IV and, in doing so, is likely to bypass good monotonic solutions.

A new approach is provided by binary binning of severity 1 vs. severity {2, 3} and / or binary binning of severity {1, 2} vs. severity 3 where the binning process is able to find the highest IV solutions subject to these solutions being monotonic. Such a binning process is given in Lund (2017). This paper presents a macro called %ORDINAL_BIN. %ORDINAL_BIN is applied to ordinal predictors and finds the best IV (or Log Likelihood) binning solution subject to the solution being monotonic.[11]

First, a choice has to be made regarding the properties of the solutions that are found. Here are the choices: (1) Both AGE_woe1 and AGE_woe2 are monotonic, (2) Only AGE_woe1 has to be monotonic, or (3) Only AGE_woe2 has to be monotonic.

Let's assume that choice (2) is the most important requirement. For each k (number of bins), the monotonic solutions for AGE_woe1 will be found by %ORDINAL_BIN. These solutions will be merged to the corresponding solutions for AGE_woe2 (i.e. those with the same binning). The IV for AGE_woe1 and the IV for AGE_woe2 are then added to give Total_IV. The best solution for k bins is the one with greatest Total_IV.

There are no monotonic solutions for AGE_woe1 where k > 7 and there is just one for k = 7. The TOTAL_IV for AGE_woe1 and the matching AGE_woe2 is only 0.476. This compares to 0.670 for the 7-bin solution of Table 11. The correlation between the WOE's is 86.8%.

But allowing AGE_woe1 to be monotonic with 6 bins leads to a much better solution. Now TOTAL_IV is 0.564 and correlation is 66.3%.

Table 13. Six Bin Solution for Age with Greatest Total_IV Subject to AGE_woe1 being monotonic

Age	AGE_woe1	AGE_woe2
19 - 21	0.444	0.463
22 - 24	0.278	0.804
25	0.270	-0.125
26 - 28	-0.339	0.376
29	-0.354	0.057
30 and Up	-0.395	-0.898

The discussion above shows this process to be rather ad hoc. Software development is needed for an easy-to-use process which finds optimal binning solutions while being subject to the monotonicity conditions (1), (2), or (3). A prototype process is available from the author.

If X is ordered and the user requires either or both of X_woe1 and X_woe2 to be monotonic, then using %ORDINAL_BIN is the best approach.[12] If X is nominal (unordered), then %CUMLOGIT_BIN is required.

%ORDINAL_BIN can be used to bin ordinal predictors for the cumulative logit model for targets with more than 3 levels. But the extension beyond 3 brings increased complexity.

PREDICTORS WITH EQUAL SLOPES

For the cumulative logit model example of AGE and SEVERITY the predictor LnAGE was judged to have unequal slopes according to the OneUp test. When using 13 bins for AGE the weight of evidence variables, AGE_woe1 and AGE_woe2, were only moderately correlated.

What about the case of "equal slopes"? If a target Y has three levels and a predictor X has nearly equal slopes, then can X_woe1 and X_woe2 still be used to replace X? The answer is "Yes" unless X_woe1 and X_woe2 are too highly correlated.

[11] %ORDINAL_BIN finds only binning solutions where the bins are ordered with respect to the ordering of X. A monotonic solution has monotonic event rates (or odds) with respect to the ordering of X.

[12] In fact, if the user wants only adjacent level binning (for ordered X), %ORDINAL_BIN can be used to bin 1 vs 2_3 and 1_2 vs. 3 without regard to searching for monotonic solutions. Solutions with same bins from 1 vs 1_2 and 1_2 vs 3 are merged and the best TOTAL_IV solution for k-bins for each k would be considered by the modeler. Although more computationally intense, this approach will give better results than would %CUMLOGIT_BIN.

The DATA Step creates data for a cumulative logit model where the target has 3 levels, the predictor X has 8 levels, and X has nearly equal slopes in the two response equations. In the simulation code the coefficients of X are set at 0.1 (See the code statements for T and U).

```
DATA EQUAL_SLOPES;
do i = 1 to 800;
   X = mod(i,8) + 1;
   T = exp(0 + 0.1*X + 0.01*rannor(1));
   U = exp(1 + 0.1*X + 0.01*rannor(3));
   PA = 1 - 1/(1 + T);
   PB = 1/(1 + T) - 1/(1 + U);
   PC = 1 - (PA + PB);
   R = ranuni(5);
   if R < PA then Y = "A";
   else if R < (PA + PB) then Y = "B";
   else Y = "C";
   output;
   end;
run;
```

The OneUp test for X has a p-value of 0.56 and the null hypothesis of equal slopes is accepted. The results for the cumulative logit PO model for X with target Y are shown in Table 14. The fit is given by -2 * Log L = 1463.462 and the estimated slope for X is 0.1012 with Pr > ChiSq = 0.0012.

```
PROC LOGISTIC DATA = EQUAL_SLOPES;
MODEL Y = X;
```

Table 14. The Cumulative Logit PO Model for X and Target Y

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1477.909	1469.462
SC	1487.279	1483.516
-2 Log L	1473.909	1463.462

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Std. Error	Wald Chi-Sq	Pr > ChiSq
Intercept	A	1	0.0632	0.1549	0.1666	0.6831
Intercept	B	1	0.9798	0.1603	37.3757	<.0001
X		1	0.1012	0.0314	10.4179	0.0012

%CUMLOGIT_BIN was run on X from the data set EQUAL_SLOPES to form weight of evidence predictors X_woe1 and X_woe2 before any binning (X still has 8 levels). The correlation of X_woe1 and X_woe2 at 78.5% is at the borderline of being too high for both predictors to be entered into the model.

Fit statistics for the PPO model (unequalslopes) with X_woe1 and X_woe2 and for three alternative models are given in Table 15. Each of these models has a better value of -2 * Log L than MODEL Y = X of Table 14 but at the cost of increased degrees of freedom. Model A has 16 degrees of freedom and Model D has 8, but I do not know how to determine the degrees of freedom for models B and C. [13]

Table 15. Weight of Evidence Models for X and Target Y

Model	-2 Log L	MODEL DF with Intercept
A. PPO model with X_woe1 and X_woe2	1450.398	16
B. PPO model with X_woe1	1459.349	?
C. PO model with X_woe1	1459.683	?
D. PO model with CLASS X	1459.317	8

[13] I think Model D will always have similar but smaller -2*Log L than Model C or Model B. This suggests that 8 d.f. is an appropriate conservative assignment to Model C and Model B.

High Correlation of X_woe1 and X_woe2

In the event of high correlation, I think the best approach is to use only X_woe1 (or X_woe2) and to specify equal slopes. Alternatively, X can be included as a class variable with equal slopes.[14] There seems to be little benefit from specifying unequal slopes but I do not have a mathematical demonstration. Here are situations where strong correlation is likely to arise:

- Conjecture: If X is a strong predictor, then the correlation of X_woe1 and X_woe2 is high. A plausibility argument is given in the Appendix. In this plausibility argument, the meaning of "strong" is left vague. The preceding example supports this conjecture since X had a strongly significance chi-square with p-value of 0.0012 and the correlation of X_woe1 and X_woe2 was high at 78.5%.
- If there is a monotonic relationship between X and both of X_woe1 and X_woe2, then, almost necessarily, there is a strong correlation of X_woe1 and X_woe2.
- Based on empirical observation, as number of bins during the binning process for X decreases, the correlation of X_woe1 and X_woe2 increases. For two bins, X_woe1 and X_woe2 are collinear.

WOE FOR CUMULATIVE LOGIT: WHAT WE KNOW AND MORE TO DISCOVER

What We Know about the Case Where the Target has Three Levels

In the case of a target with 3 levels and predictor X, the usage of X_woe1 and X_woe2 in place of X in a PPO model is very likely to provide more predictive power than X or some transform of X.

The process of binning X can help to achieve parsimony while maintaining predictive power.

The measurement of correlation between X_woe1 and X_woe2, as the binning process proceeds, can signal when these predictors are too highly correlated for both to be entered into the model.

WOE Coding versus CLASS / Dummy Variable Coding for Cumulative Logit Model

For a NOD predictor X, it is clearly easier to put X in a CLASS statement than it is to code WOE variables when fitting a cumulative logit model. A macro program or tedious hand coding is needed to form the WOE transformation.

In the one variable case the use of CLASS X with UNEQUALSLOPES = (X) provides the same fit as WOE variables for X with UNEQUALSLOPES. However, the use of CLASS X in the presence of other predictors will give a somewhat better fit (measured by log likelihood) than using WOE predictors.

But a drawback to using CLASS statements is the proliferation of coefficients. If the target has 3 levels and X has 10 levels, then 18 coefficients are introduced by CLASS X with UNEQUALSLOPES while only 4 are added by WOE coding. The use of CLASS and UNEQUALSLOPES becomes awkward if not impractical for a model with 10 or 20 NOD predictors.

Finally, the modeler might find a pattern in X_woe1 and X_woe2 that is consistent with expectations. For example, it might be that both X_woe1 and X_woe2 increase as X is increases. After fitting the model, the product of the coefficient and a WOE variable will still be monotonic versus X. But when CLASS X (with unequalslopes) is used, the coefficients for the response equations may not have a monotonic relationship to X.

As in the case of binary logistic models, however, either WOE or CLASS variable may be used successfully in model fitting.

Some Ideas for Further Work

- When should all weight of evidence transformations of X be included in a cumulative logit model?
  - For the case where the target has 3 levels a test can be based on the correlation between X_woe1 and X_woe2. But what is a good cut-off correlation value?

[14] Model with CLASS X is not equivalent to model with X_woe1. Model with CLASS X has greater log likelihood.

- For the case where the target has more than 3 levels a correlation technique is needed to decide which of X_woe1 to X_woe<J-1> should be used in the model.

- In the case of J > 3, is FORWARD or STEPWISE a useful approach to deciding what WOE variables to include? [15]
- When Binning: Is TOTAL_IV a good measure of the association of X to the target? What is a good value of TOTAL_IV and is there a parallel to the Table 3 in this paper taken from Siddiqi's book?
- Does a low IV for X_woe<k> indicate that X_woe<k> should not be included in the model? The Siddiqi guidelines seem to apply here since IV<k> is the IV of a binary target versus X.

SAS MACRO DISCUSSED IN THIS PAPER

Contact the author for an experimental beta version of %CUMLOGIT_BIN and for SAS code that applies %ORDINAL_BIN to binning of ordered predictors for the cumulative logit model.

REFERENCES

Albert, A and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, 71, 1, pp. 1-10.

Allison, P.D. (2012), *Logistic Regression Using SAS: Theory and Application 2nd Ed.*, Cary, NC, SAS Institute Inc.

Agresti, A (2010). *Analysis of Ordinal Categorical Data, 2nd Ed.*, Hoboken, NJ, John Wiley & Sons.

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 (2017). SAS® Macros for Binning Predictors with a Binary Target, *Proceedings of the SAS Global Forum 2017 Conference,* Cary, NC, SAS Institute Inc.

Lund B. and Brotherton D. (2013). Information Value Statistic, *MWSUG 2013, Proceedings*, Midwest SAS Users Group, Inc., paper AA-14.

Siddiqi, N. (2006). *Credit Risk Scorecards*, Hoboken, NJ, John Wiley & Sons, Inc.

ACKNOWLEDGMENTS

All SAS programming for this paper was done by using SAS University Edition.
Dave Brotherton of Magnify Analytic Solutions of Detroit provided helpful insights and suggestions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author: blund@magnifyas.com, blund_data@mi.rr.com, or blund.data@gmail.com

All SAS code in this paper is provided by Marketing Associates, LLC "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 Marketing Associates, LLC shall not be liable for any damages whatsoever arising out of their use of this material. In addition, Marketing Associates, LLC will provide no support for the materials contained herein.

[15] SELECTION = SCORE with UNEQUALSLOPES is not supported in PROC LOGISTIC. The CLASS statement is not supported with SELECTION = SCORE

APPENDIX: DATA SET BACKACHE

Data Set BACKACHE
(A change to raw data has been made: if SEVERITY = 0 then SEVERITY = 1)

Obs	SEVERITY	AGE	_FREQ_
1	1	16	1
2	1	18	4
3	1	19	5
4	1	20	3
5	1	21	12
6	1	22	5
7	1	23	7
8	1	24	12
9	1	25	7
10	1	26	8
11	1	27	4
12	1	28	4
13	1	29	3
14	1	30	3
15	1	31	1
16	1	32	3
17	1	33	2
18	1	34	2
19	1	35	1
20	1	37	1
21	1	39	1
22	1	42	4
23	2	17	1
24	2	18	3
25	2	19	1
26	2	20	3
27	2	21	3
28	2	22	6
29	2	23	3
30	2	24	5
31	2	25	3
32	2	26	10
33	2	27	3
34	2	28	4
35	2	29	3
36	2	30	4
37	2	31	2
38	2	32	3
39	2	35	1
40	2	36	1
41	2	37	1
42	3	15	1
43	3	19	1
44	3	20	1
45	3	21	1
46	3	23	2
47	3	24	1
48	3	25	2
49	3	26	2
50	3	27	1
51	3	28	1
52	3	29	1
53	3	30	2
54	3	31	1
55	3	32	2
56	3	33	1
57	3	34	1
58	3	35	2
59	3	36	1
60	3	38	1
61	3	39	2

APPENDIX: ONEUP TEST (FOLLOWING B. DERR (2013))

In the DATA STEP a second copy of LnAge called LnAgeB is created. In PROC LOGISTIC the predictors LnAge and LnAgeB are entered by FORWARD with LnAge being forced in by the INCLUDE = 1 and LnAgeB being entered in “step 1” (by use of SLE = .999).

The log likelihood chi-squares for Step 0 (Include = 0) and for Step 1 (from FORWARD) are captured in the ODS OUTPUT GLOBALTESTS. The difference of these chi-squares gives a chi-square test statistic for the significance of a model with unequalslopes for LnAge versus the model with equal slopes.

This test statistic and its p-value are computed in the final DATA STEP. The p-value is 0.06. This result is small enough to favor the use of unequalslopes for the predictor LnAge.

```
DATA Backache; SET Backache;
LnAge = log(AGE);
LnAgeB = LnAge;
run;
ods output globaltests = globaltests;
PROC LOGISTIC DATA = Backache;
MODEL Severity(descending)=lnAge lnAgeB / unequalslopes = lnAgeB
Selection = forward sle = .999 include = 1 stop = 2;
Freq _freq_;
run;
DATA OneUp_LRT; SET globaltests end = eof;
retain ChiSq_0 ChiSq_1 DF_0 DF_1;
if Step = 0 & Test = "Likelihood Ratio" then do;
   ChiSq_0 = ChiSq;
   DF_0 = DF;
   end;
if Step = 1 & Test = "Likelihood Ratio" then do;
   ChiSq_1 = ChiSq;
   DF_1 = DF;
   end;
if eof then do;
   OneUp_LRT = 1 - PROBCHI(ChiSq_1 - ChiSq_0, DF_1 - DF_0);
   output;
   end;
run;
PROC PRINT DATA = OneUp_LRT; Var OneUp_LRT;
title "OneUp Likelihood Ratio Test";
run;
```

APPENDIX: CORRELATION OF WOE PREDICTORS FOR 3 LEVEL TARGET

Consider the case of three levels A, B, C for target Y. If X is a *strong* predictor of Y, then we'll assume that the probabilities of A, B, C can be approximated by the empirical probabilities as shown:

Prob (Y = A | X = x) = $p_{x,A} \sim A_x / (A_x + B_x + C_x)$ … and likewise for B and C

where A_x gives the count of occurrences of A when X = x and similarly for B_x and C_x

I do not have a way to quantify this approximating relationship in terms of some well-defined measure of the strength of X. But accepting that this relationship exists for a strong predictor, then for the PPO model for values r and s of X:

$\text{Log}\ [p_{r,A}/(p_{r,B} + p_{r,C})] - \text{Log}\ [p_{s,A}/(p_{s,B} + p_{s,C})] = (r - s) * \beta_{X,1}$ ... from response equation 1
$\text{Log}\ [(p_{r,A} + p_{r,B})/p_{r,C}] - \text{Log}\ [(p_{s,A} + p_{s,B})/p_{s,C}] = (r - s) * \beta_{X,2}$ ... from response equation 2

and via substitution of the approximations for $p_{x,A}$, $p_{x,B}$, $p_{x,C}$:

X_woe1(X=r) - X_woe1(X=s) = Log [Ar / (Br + Cr)] - Log [As / (Bs + Cs)] ~
$\text{Log}\ [p_{r,A}/(p_{r,B} + p_{r,C})] - \text{Log}\ [p_{s,A}/(p_{s,B} + p_{s,C})] = (r - s) * \beta_{X,1}$

X_woe2(X=r) - X_woe2(X=s) = Log [(Ar + Br) / Cr] - Log [(As + Bs) / Cs] ~
$\text{Log}\ [(p_{r,A} + p_{r,B})/p_{r,C}] - \text{Log}\ [(p_{s,A} + p_{s,B})/p_{s,C}] = (r - s) * \beta_{X,2}$

Fixing the value of s, these equations above imply that X_woe1(X=r) and X_woe2(X=r) are approximately collinear as a function of r.