MWSUG 2016 – Paper AA15

Weight of Evidence Coding and Binning of Predictors in Logistic Regression

Bruce Lund, Independent Consultant, Novi, MI

ABSTRACT

Weight of evidence (WOE) coding of a nominal or discrete variable is widely used when preparing predictors for usage in binary logistic regression models. When using WOE coding, an important preliminary step is binning of the levels of the predictor to achieve parsimony without giving up predictive power. These concepts of WOE and binning are extended to ordinal logistic regression in the case of the cumulative logit model. SAS® code to perform binning in the binary case and in the ordinal case is discussed. Lastly, guidelines for assignment of degrees of freedom for WOE-coded predictors within a fitted logistic model are discussed. The assignment of degrees of freedom bears on the ranking of logistic models by SBC (Schwarz Bayes). All computations in this talk are performed by using SAS® and SAS/STAT®.

INTRODUCTION

Binary logistic regression models are widely used in CRM (customer relationship management) or 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 is a count or a truncated count (e.g. number of 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.

In either case, binary or ordinal, before the WOE 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 X are discussed in the paper.

Finally, the effect of WOE coding on SBC (Schwarz Bayes criterion[3]) of a model must be considered when ranking candidate models by SBC. For example, the following two binary logistic models are equivalent (same probabilities):

```
(A)    PROC LOGISTIC; CLASS X; MODEL Y = X;
(B)    PROC LOGISTIC; MODEL Y = X_woe;
```

where X_woe is the weight of evidence transformation of X.

But Model (B) has smaller SBC than Model (A) because X_woe is counted in PROC LOGISTIC as having only 1 degree of freedom.

A discussion and recommendation for an adjustment to SBC in the case where WOE variables are included in a logistic model is given at the end of the paper.

---

[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). Unfortunately, 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] SBC = -2*Log L + log(n)*K where Log L is log likelihood, n = sample size, and K is count of coefficients in model.

TRANSFORMING 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. [4] Typically, L is 15 or less.

```
PROC LOGISTIC; CLASS C; MODEL Y = C <and other predictors>;
```

or

```
PROC LOGISTIC; MODEL Y = C_dum_<k> where k = 1 to L-1 <and other predictors>;
```

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_k. 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= 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 DESCENDING; MODEL Y = C_woe;
```

then the slope coefficient equals 1 and the intercept is the 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"	Log(g_k/b_k)	g_k - b_k	**IV Terms** (g_k - b_k) * 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

[4] "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.

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.[5]

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).[6] In logistic modeling applications it is unusual to see IV $\geq$ 0.5.

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"

There is a strong relationship of the IV of predictor C to the Log Likelihood (LL) of the model:

```
PROC LOGISTIC; CLASS C; MODEL Y = C;
```

For example, if N = 10,000 and L = 5, then a simple linear regression of IV to LL is a good model. Based on a simulation (with 500 samples) the R-square is 61%.[7]

Before a predictor is converted to WOE coding and is entered into a model, the predictor 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 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.

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.[8] 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.[9]

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

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

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

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

---

[5] See Lund and Brotherton (2013, p. 17) for a proof.
[6] See Siddiqi (2006) for the usage of WOE and IV in the preparation of predictors for credit risk models
[7] This simulation code is available from the author. See Lund and Brotherton (2013) for more discussion.
[8] Simply run: PROC LOGISTIC; MODEL Y = <X's>; where Y has more than 2 levels.
[9] 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 to j" (in the ordering of the target) in the numerator to "one minus the cumulative probability to j" in the denominator.

Formulas for the probabilities $p_{k,1}$, $p_{k,2}$, $p_{k,3}$ 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 found:

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 found by maximizing the log likelihood equation in a manner similar to the binary case.[10]

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. Using the probability formulas from Table 4:

$$\text{Log}\left[\frac{p_{rA}/(p_{rB}+p_{rC})}{p_{sA}/(p_{sB}+p_{sC})}\right] = \text{Log}\,(p_{r,A} / (p_{r,B} + p_{r,C})) - \text{Log}\,(p_{s,A} / (p_{s,B} + p_{s,C})) = (r - s) * \beta_{X1} \ldots \text{proportional odds}$$

$$\text{Log}\left[\frac{(p_{rA}+p_{rB})/p_{rC}}{(p_{sA}+p_{sB})/p_{sC}}\right] = \text{Log}\,((p_{r,A} + p_{r,B}) / p_{r,C}) - \text{Log}\,((p_{s,A} + p_{s,B}) / p_{s,C}) = (r - s) * \beta_{X1} \ldots \text{proportional odds}$$

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 for X1 is a by-product of assuming that the coefficients of predictor X1 are equal across the cumulative logit response equations.

EXTENDING WOE TO 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.

WOE TRANSFORMATIONS 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 L levels, then L-1 WOE transformations were needed.

The extension of WOE to the cumulative logit model does not require an assumption of proportional odds.

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 two tables below illustrate the steps to define the weight of evidence transformation for X. The first step is to define two sets of column percentages corresponding to the two cumulative logits.

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

Table 5. Defining WEIGHT OF EVIDENCE Predictors for Cumulative Logit Model – STEP 1

	Y=			Col %		Col %	
X=i	Ai	Bi	Ci	Ai / A	(Bi+Ci) / (B+C)	(Ai+Bi) / (A+B)	Ci / C
1	**2**	1	2	**0.182**	0.176	0.17	0.20
2	4	3	1	0.36	0.24	0.39	0.10
3	4	1	2	0.36	0.18	0.28	0.20
4	1	2	5	0.09	0.41	0.17	0.50
Total	**11**	7	10	Where A = ∑ Ai, B = ∑ Bi, C = ∑ Ci			

For the first cumulative logit the value 0.182 in column "A1 / A" is equal to 2 divided by 11. The value 0.176 in column "(B1+C1) / (B + C)" is equal to 1+2 divided by 7+10. Similarly, the columns for the second cumulative logit are computed.

Now, the second step:

Table 6. Defining WEIGHT OF EVIDENCE Predictors for Cumulative Logit Model – STEP 2

	Y=			Col %		Col %		Ratio of Col %		Log (Ratio)	
X=i	Ai	Bi	Ci	Ai / A	(Bi+Ci) / (B+C)	(Ai+Bi) / (A+B)	Ci / C	A over B+C	A+B over C	X_WOE1	X_WOE2
1	**2**	**1**	**2**	**0.182**	**0.176**	**0.167**	**0.200**	**1.034**	**0.833**	**0.03**	**-0.18**
2	4	3	1	0.36	0.24	0.39	0.10	1.55	3.89	0.44	1.36
3	4	1	2	0.36	0.18	0.28	0.20	2.06	1.39	0.72	0.33
4	1	2	5	0.09	0.41	0.17	0.50	0.22	0.33	-1.51	-1.10
Total	**11**	**7**	**10**	Where A = ∑ Ai, B = ∑ Bi, C = ∑ Ci							

The "ratio of column percentages" for the first row of the first cumulative logit is computed by 1.034 = 0.182 / 0.176. The log of this ratio gives the weight of evidence for the first row of 0.03. Likewise, the first row for the second weight of evidence is -0.18.

As equations:

X_WOE1 (X=i) = LOG [(Ai / A) / ((Bi+Ci) / (B+C))]

X_WOE2 (X=i) = LOG [((Ai+Bi) / (A+B)) / (Ci / C)]

Although X in this example is numeric, a character predictor may take the role of X.

Cumulative Logit Model with Proportional Odds Does Not Support a Generalization of WOE

Table 6 is converted to the data set EXAMPLE1 in Table 7 for the same predictor X and 3-level ordinal target Y. The use of the EXAMPLE1 will show that the cumulative logit PO model does not support the required two characteristics for a WOE predictor.

Table 7. Data Set EXAMPLE1 for Illustrations to Follow

```
DATA EXAMPLE1; Input X Y $ @@; Datalines;
1 A  2 A  3 A  4 B
1 A  2 A  3 A  4 B
1 B  2 B  3 A  4 C
1 C  2 B  3 B  4 C
1 C  2 B  3 C  4 C
2 A  2 C  3 C  4 C
2 A  3 A  4 A  4 C
;
```

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 8. Results of MODEL (II)

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	

A generalization of the PO model is needed in order to generalize the idea of weight of evidence coding. The next section describes the partial proportional odds (PPO) cumulative logit model and how weight of evidence can be generalized 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 case the PPO Model has 6 parameters α_1 α_2 β_R β_S β_{Z1} β_{Z2} given in 2 equations:

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

The coefficients of the predictors β_R and β_S are the same in the 2 equations but β_{Zj} varies with j. There are 4 β's in total.

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:

$\alpha_1 = \log\,(n_A / (n_B + n_C))$ $\alpha_2 = \log\,((n_A + n_B) / n_C)$
$\beta_{X\_woe1,1} = 1$, $\beta_{X\_woe1,2} = 0$; … (*)
$\beta_{X\_woe2,1} = 0$, $\beta_{X\_woe2,2} = 1$;

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 9.

Table 9. Results of MODEL (II)

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

Weight of evidence predictors should enter a cumulative logit model with the unequalslopes parameter in order to reproduce the 2 defining characteristics of the weight of evidence predictor from the binary case.

Comments

There are degenerate {X, Y} data sets where a cumulative logit model has no solution.[11] Setting these cases aside, I do not have a solid mathematical proof that coefficients, as given by (*), 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 for target levels A, B, and C are obtained by substitution into the equations in 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 THREE 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 have one of 3 levels of SEVERITY of backache in the “BACKACHE IN PREGNANCY” data set from Chatfield (1995, Exercise D.2). Using a statistical test called OneUp Derr shows it is reasonable to use unequalslopes for LnAGE when predicting SEVERITY.

There is a data set called BACKACHE in the Appendix with 61 observations which expands to 180 after applying a frequency variable. It has AGE and SEVERITY (and a frequency variable _FREQ_) from the BACKACHE IN PREGNANCY data set. See this data set for the discussion that follows below.

The weight of evidence transforms 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 gives the fit statistics in Table 10a and Table 10b.

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

Table 10a. 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 10b. 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?

[11] Agresti (2010 p. 64)

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 11a.

Table 11a. SEVERITY from Backache Data Predicted by WOE recoding of LnAGE 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 11b.

Table 11b. SEVERITY from Backache Data Predicted by WOE recoding of LnAGE 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 12.

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

Table 12. 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 weight of evidence predictors AGE_WOE1 and AGE_WOE2 use the 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 collapse together, at each step, so as to maximize the remaining predictive power. These measures include: (i) Information Value, (ii) Log Likelihood (equivalent to entropy), (iii) p-value from the chi-square measure of independence of X and the target. (The Interactive Grouping Node (IGN) in SAS Enterprise Miner provides the user with the choice of either (ii) or (iii) when binning predictors and IGN reports the IV for each binning solution.)

How can these binary methods be generalized to 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.

A work-in-progress macro called %CUMLOGIT_BIN is being developed to perform binning in the case of the cumulative logit model. For this macro the target has L ≥ 2 ordered values and the predictor X may be numeric or character.

Two input parameters for %CUMLOGIT_BIN are:

- MODE: The user first decides which pairs of levels of the predictor X are eligible for collapsing together. The choice is between "any pairs are eligible" or "only adjacent pairs in the ordering of X".
- METHOD: This is a criterion for selecting the pair for collapsing. The choices are TOTAL_IV and ENTROPY. 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 ENTROPY is selected.

%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 13.

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 either a 10-bin or 9-bin solution has appeal since TOTAL_IV begins to fall rapidly thereafter. These solutions give -2 * Log L values of 336.60 and 336.92 in comparison with 349.423 for LnAGE (Table 10). The correlations between AGE_woe1 and AGE_woe2 are moderate for the solutions with 10-bins and 9-bins (63% and 68%).

Table 13. 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	0.6309
11	22	336.51	0.235	0.488	0.723	380.51	450.76	0.6350
10	20	336.60	0.232	0.487	0.720	376.60	440.46	0.6343
9	18	336.92	0.229	0.484	0.713	372.92	430.39	0.6804
8	16	337.44	0.218	0.482	0.700	369.44	420.53	0.7026
7	14	339.16	0.198	0.472	0.670	367.16	411.86	0.7099
6	12	340.04	0.178	0.462	0.640	364.04	402.36	0.8082
5	10	341.54	0.144	0.461	0.604	361.54	393.47	0.8075
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	0.9996
2	4	348.01	0.049	0.382	0.430	356.01	368.78	1.0000

The selection of either the 10-bin or 9-bin WOE solution, in conjunction with all the other predictors of SEVERITY, is likely to provide an improvement in the complete Backache Model versus the usage of LnAGE with unequalslopes.

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 equal slopes, 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.

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 equal slopes. In the simulation code the slopes of X are set at 0.1 (see the 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;
run;
```

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 74.5% is near or at the borderline of being too high for both predictors to be entered into the model.

Fit statistics for the PPO model with X_woe1 and X_woe2 and for two alternative models are given in Table 15. Each of these models has a better value of -2 * Log L than the MODEL Y = X of Table 14 but at the cost of increased degrees of freedom.

But I do not know how to assign the exact degrees of freedom to the bottom two models.

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

Model	-2 Log L	MODEL DF with Intercept
PPO model with X_woe1 and X_woe2	1450.398	16
PPO model with X_woe1	1459.349	?
PO model with X_woe1	1459.683	?

High Correlation of X_woe1 and X_woe2

Conjecture: If X is a strong predictor, then the correlation of X_woe1 and X_woe2 is high.

A plausibility argument for this claim 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 while the correlation of X_woe1 and X_woe2 was high at 74.5%.

Observation: As the number of bins during the binning process for X approaches 2 the correlation of X_woe1 and X_woe2 becomes high. This is based on my empirical observations. For two bins, X_woe1 and X_woe2 are collinear.

CUMULATIVE LOGIT MODELS: MORE TO DO

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.

More to Do, Some Ideas for Further Work

- Is weight of evidence any better or worse than simply using a CLASS statement?
- When should all weight of evidence transformations of X be included in a PPO model?
  - For the case where the target has 3 levels a test can be based on the correlation between X_woe1 and X_woe2. What is a good cut-off correlation value?
    - If too highly correlated, what is the alternative? A PPO model with X_woe1 or a PO model with X_woe1? Or some other approach?
  - 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<L-1> can be used in the model.
- Is either Stepwise or Best Subsets a useful approach to deciding what WOE variables to include?
- 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? What is a low value?

BACK TO BINARY AND BINNING

Binning of predictors for binary targets has been well developed. In SAS Enterprise Miner the Interactive Grouping Node provides extensive functionality.

For users of SAS / STAT a macro %BEST_COLLAPSE for binning of binary targets was provided by Lund and Brotherton (2013). This macro was mentioned in Lund (2015). Recently the macro has been enhanced and renamed as %NOD_BIN.

%NOD_BIN addresses the binning of both an ordered or unordered predictor X. But in the case of ordered X with adjacent level binning (in the ordering of X) a complete solution is provided by a new macro called %ORDINAL_BIN.

%ORDINAL_BIN for Binning a Binary Target and an Ordered Predictor

It is assumed that X is ordered (numeric or character) and only adjacent levels of X are allowed to collapse together.

If X has L levels, then are there $2^{(L-1)}$ - 1 total solutions: $\sum_{k=2}^{L}\{\text{number of solutions with k bins}\} = 2^{(L-1)} - 1$

- If L = 3, then the solutions are: {1,2} {3}, {1} {2,3}, {1} {2} {3}, … 2^2 - 1 = 3
- If L = 20, then there are $2^{(L-1)}$ - 1 = 524,287 total solutions

%ORDINAL_BIN finds **ALL** solutions, determines which are monotonic, and computes IV.

If X has missing values, they may be included in the binning process. A WOE value is computed for missing, and IV includes the IV term for missing. The missing level may not collapse with other levels and does not affect which solutions are called monotonic.

Small-bin solutions are not reported (where bin ≤ “x”%).

WOE SAS statements are provided.

Running %ORDINAL_BIN when the number of levels of X exceeds 20 involves very long run-times. PROC HPBIN can reduce the levels of X to less than or equal to 20 before running %ORDINAL_BIN.

In the macro program the simple idea behind %ORDINAL_BIN is hidden behind multiple levels of macro statements. But the idea can be expressed through a simple example.

Assume X has 4 ordered levels 1, 2, 3, 4. There is one 4-bin solution consisting of {1}, {2}, {3}, {4}. WOE values are computed WOE(1) – WOE(4). By inspection it is determined whether these are monotonic, and the associated IV is computed. There are three 3-bin solutions: {1,2}, {3}, {4} and {1}, {2,3}, {4} and {1}, {2}, {3,4}. Again WOE’s are computed, monotonicity is determined, and IV’s are computed. Likewise, for 2-bin solutions.

The programming challenge was setting up dynamic macro DO Loops to identify all the adjacent-mode solutions.

The %ORDINAL_BIN was run on the following example of Y and X_Very_Non_Mono:

Table 15. Data Set with Target Y and X_VERY_NON_MONO

	X_VERY_NON_MONO											
Y	1	2	3	4	5	6	7	8	9	10	11	12
0	1393	60090	5083	45190	8319	48410	2689	20900	729	2920	253	2940
1	218	890	932	1035	2284	1593	1053	872	311	136	120	142
Total	1611	60980	6015	46225	10603	50003	3742	21772	1040	3056	373	3082

%ORDINAL_BIN allows the user to control the amount of output that is displayed. In the example below, the solutions for 6-bins, 5-bins, and 4-bins were obtained and the best two IV solutions were displayed and the (up to) best two monotonic solutions were displayed.

There no monotone solution with 6 bins. There is a single monotone solution with 5 bins. The best two IV solutions have much higher IV than this monotonic solution. There are multiple 4-bin monotonic solutions and the best two are displayed.

Table 16. %ORDINAL_BIN Applied to X_VERY_NON_MONO (sorted by descending IV)

BINS	Best_IV	Best_mono	Solution_num	IV	-2 * LL	L1	L2	L3	L4	L5	L6
6	*		1	0.8174	69778	1+2	3	4	5	6	7+8+9+10+11+12
6	*		2	0.7964	70128	1	2	3	4	5	6+7+8+9+10+11+12

BINS	Best_IV	Best_mono	Solution_num	IV	-2 * LL	L1	L2	L3	L4	L5
5	*		1	0.7213	70696	1+2	3	4	5	6+7+8+9+10+11+12
5	*		2	0.7124	70252	1+2+3+4	5	6	7	8+9+10+11+12
5		*****	117	0.3155	75069	1+2	3+4	5+6	7+8	9+10+11+12

BINS	Best_IV	Best_mono	Solution_num	IV	-2 * LL	L1	L2	L3	L4
4	*		1	0.5777	71875	1+2+3+4	5	6	7+8+9+10+11+12
4	*		2	0.5468	72357	1+2	3+4	5	6+7+8+9+10+11+12
4		*	26	0.3132	75095	1+2	3+4	5+6	7+8+9+10+11+12
4		*	27	0.3121	75106	1+2	3+4	5+6+7+8	9+10+11+12

WOE, DEGREES OF FREEDOM, AND SBC

A common approach to finding and ranking multiple candidate models for logistic regression is to rank the models by their Schwarz Bayes criterion (SBC) and select the top several models for further study.[12] Normally, when data are abundant, the candidate models are fit on a training data set, ranked by SBC, and the top few models are further evaluated on the validation data set.

There is, however, a problem with this approach related to the proper calculation of SBC when using WOE coded predictors.

Consider a binary target Y and the following two logistic models:

```
(A)    PROC LOGISTIC; CLASS C; MODEL Y = C <other X's>;
(B)    PROC LOGISTIC; MODEL Y = C_woe <other X's>;
```

To PROC LOGISTIC, the predictor C_woe appears to have 1 d.f. But the pre-coding of C_woe used the entire C * Y table. All the information about Y and C is reflected in the construction of C_woe and more than one degree of freedom needs to be assigned to C_woe. With only 1 d.f. the SBC for Model (B) is understated.

A Recommendation: If Ranking Models by SBC, Include WOE Predictors in a CLASS Statement.

For the purpose of ranking models by SBC I recommend computing SBC for a model by putting all WOE variables (if any) into a CLASS statement. (Once the ranking and selection of models for further study is completed, then the selected models are re-run with WOE predictors removed from the CLASS statement.)

This is theoretically correct? The answer is: No.

I can construct a contrived example where SBC for the model with WOE predictors in a CLASS statement is less than the SBC where there is no CLASS statement. That is:

The SBC for PROC LOGISTIC; CLASS C_woe1 C_woe2; MODEL Y = C_woe1 C_woe2 <other X's>;

is less than

The SBC for PROC LOGISTIC; MODEL Y = C_woe1 C_woe2 <other X's>;

In this case the "penalty" from adding degrees of freedom is offset by more "reward" from the added flexibility in fitting the dummy coefficients from the CLASS predictors.

[12] SAS Institute (2012) *Predictive Modeling Using Logistic Regression: Course Notes*. See chapter 3

But still, something has to be done. This recommendation is a heuristic that works for practical cases.

Here is some support for this recommendation:

- If X, Y, and Z are each uncorrelated with C_woe, then -2*LL for model with C_woe is equal to -2*LL for the model where C appears in CLASS C and, further, these are the same two models.[13] Therefore, the SBC for the model with C_woe must have an SBC that equals the SBC for the model with CLASS C.
- Usually, the modeler tries to reduce multicollinearity.
- Often, in the applications I've seen, -2*LL with WOE predictors is approximately equal to -2*LL when these same predictors appear in a CLASS statement.

SAS MACROS DISCUSSED IN THIS PAPER

Contact the author for copies of %NOD_BIN and %ORDINAL_BIN. However, at this time, %CUMLOGIT_BIN remains under development.

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

Lund B. (2015). Fitting and Evaluating Logistic Regression Models, *MWSUG 2015, Proceedings*, Midwest SAS Users Group, Inc., paper AA-03.

SAS Institute (2012). *Predictive Modeling Using Logistic Regression: Course Notes*, Cary, NC, SAS Institute Inc.

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

ACKNOWLEDGMENTS

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

CONTACT INFORMATION

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

Bruce Lund

blund_data@mi.rr.com, blund.data@gmail.com

[13] I do not have a mathematical proof for this claim. I am relying on examples for verification.

APPENDIX: 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: CORRELATION OF WOE PREDICTORS FOR X WHEN X IS A STRONG PREDICTOR OF TARGET Y WITH 3 LEVELS

Assume numeric predictor X is a strong predictor of target Y where “strong” will not be given an operational definition. Then the two WOE transformations of X have high correlation.

Here is a plausibility argument:

Consider the case of three levels A, B, C for target Y. If X is a strong predictor of Y, then 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 measure of the “strength” of X. But accepting that this relationship exists for a strong predictor, then for the PPO model:

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