Weight of Evidence Coded Variables for Binary and Ordinal Logistic Regression

Bruce Lund
Magnify Analytic Solutions,
Division of Marketing Associates

MSUG conference June 9, 2016

magnify
Analytic Solutions

marketing associates
creativity · technology · execution

Topics for this Talk

Focus on Logistic Regression (binary and ordinal targets):

1. Weight-of-evidence (WOE) coding of NOD (nominal/ordinal/discrete) predictors for the binary target and Information Value (IV)
2. Binning of NOD predictors before use in a Model with binary target
   a) "Optimal" binning (collapsing)
   b) Very short description of Interactive Grouping Node (SAS EM)
3. Extending WOE and IV to Ordinal Logistic Regression (… target has 3 or more levels and these levels are ordered)
4. Binning of a predictor for Ordinal Logistic Regression
5. The degrees of freedom of a Logistic Model when WOE's appear
   a) Impact on model ranking
   b) A heuristic for assigning corrected d.f. to a model

Background and Terminology … NOD

A Nominal Predictor has values that are not ordered:

- e.g. University attended (Is U. of M. > MSU ?)
- e.g. Reason for missing class (sick, overslept, boring)

An Ordinal Predictor has ordered values without an interval scale ("differences" between values don't have numeric meaning).

- e.g. Satisfaction measured as "good", "fair" "poor"

A Discrete Predictor has numeric values but only a few distinct values.

- e.g. Count of children in household

"Levels" refers to distinct values.

I'll sometimes, but not always, use "C" for a NOD. Otherwise, I'll use X.

PROC LOGISTIC

Predictors in PROC LOGISTIC can be NOD or Numeric

For NOD predictors, we want each Level(*) to enter the model as a dummy variable. Assume C has 3 levels:

PROC LOGISTIC; MODEL Y = Dum_C1 Dum_C2 <Other X>;

where Dum_C1 = (C = “C1”); Dum_C2 = (X = “C2”);

Alternatively, a CLASS statement may be used:

PROC LOGISTIC; CLASS C; MODEL Y = C <Other X>;

(*) Except one … this remaining value is determined by the others.

PROC LOGISTIC with CLASS X

```
DATA WORK;
INPUT Y X$ F;
DATALINES;
0 X1 2
1 X1 3
0 X2 1
1 X2 3
0 X3 2
1 X3 1
;
```

Model Fit Statistics	
-2 Log L	15.048
c	0.671

Maximum Likelihood Estimates			
Parameter		DF	Estimate
Intercept		1	-0.6931
X	X1	1	1.0986
X	X2	1	1.7918

```
PROC LOGISTIC DATA = work DESC;
CLASS X (PARAM = REF REF = "X3");
MODEL Y = X;
FREQ F;
```

/*PARAM=REF REF="X3" tells us how to score the MODEL if X = "X3"*/

Another way to fit this model is to transform X to weight-of-evidence.

magnify Analytic Solutions

marketing associates
creativity · technology · execution

Weight of Evidence Coding of X

X: Predictor
Y: Target (response)

Cannot have any "zeros" in Y=0 and Y=1 columns

X	Y = 0	Y = 1	Col % Y = 0	Col % Y = 1	WOE= Log(%Y=1/%Y=0)
X=X1	2	3	0.400	0.429	0.0690
X=X2	1	3	0.200	0.429	0.7621
X=X3	2	1	0.400	0.143	-1.0296
SUM	5	7	1.000	1.000	

If X = X3 then X_woe = -1.0296

magnify Analytic Solutions

X is "transformed" to X_woe

marketing associates
creativity · technology · execution

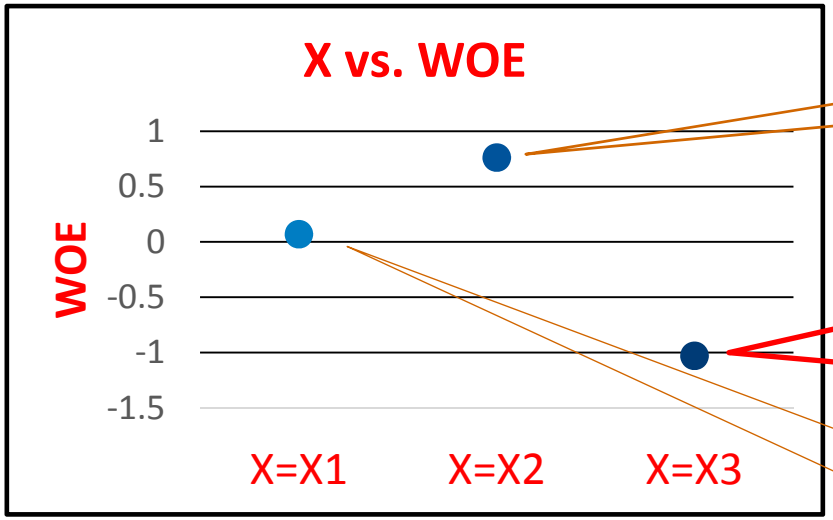
WOE -- a way to study X vs. Y

Goods and Bads

Y = 1 will be a "good" g_i = goods at Xi / total goods

Y = 0 will be a "bad" b_i = bads at Xi / total bads

X	WOE= Log(%good / %bad)
X=X1	0.069
X=X2	0.762
X=X3	-1.0296



X2: g2 > b2

X3: g3 << b3

Log(g3 / b3) = -1.0296

X3 is associated with "bads"

X1: g1 ~ b1

magnify Analytic Solutions

marketing associates creativity · technology · execution

PROC LOGISTIC with X_woe

```
DATA WORK;
INPUT Y X$ X_woe F;
DATALINES;
0 X1   0.0690  2
1 X1   0.0690  3
0 X2   0.7621  1
1 X2   0.7621  3
0 X3  -1.0296  2
1 X3  -1.0296  1
;
```

Model Fit Statistics	
-2 Log L	15.048
c	0.671

Max Likelihood Estimates	
Parameter	Estimate
Intercept	**0.3365**
X_woe	**1.0000**

```
PROC LOGISTIC DATA = work DESC;
MODEL Y = X_woe;
FREQ F;
```

α=log(G/B) and β=1

G = count of Y=1

B = count of Y=0

G = 7, B = 5 and log(7/5) = 0.3365

This characterizes WOE coding:

α=log(G/B) and β=1 if and only if X has WOE coding

Class “versus” Woe

(A) PROC LOGISTIC; CLASS C1 C2; MODEL Y = C1 C2 <other X’s>;

(B) PROC LOGISTIC; MODEL Y = C1_woe C2_woe <other X’s>;

- Log-likelihood (A) ≥ Log-likelihood (B) … better fit for (A)
  - More LL is due to dummy coefficients “reacting” to other predictors
  - The single coefficient of a WOE can only rescale the WOE values
- But probabilities from (A) and (B) are usually very similar … especially if multicollinearity is controlled (in this case LL(A) ~ LL(B))

The choice of (A) or (B) is not decisive … good models can be fit with dummies or with WOE variables.

WOE is popular in credit modeling where it naturally leads to scorecards and it is fully integrated in the Credit Scoring application in SAS® Enterprise Miner.

Topics for this Talk - #2

Focus on Logistic Regression (binary and ordinal targets):

1. Weight-of-evidence (WOE) coding of nominal/ordinal/discrete predictors for binary target. Information Value (IV), Log-likelihood, and Model c
2. **Binning predictors before use in a Model with binary target**
   a) **"Optimal" binning (collapsing)**
   b) **Very short description of Interactive Grouping Node (SAS EM)**
3. Extending WOE and IV to Ordinal Logistic Regression (… target has 3 or more levels and these levels are ordered)
4. Binning of a predictor for Ordinal Logistic Regression
5. The degrees of freedom of a Logistic Model when WOE's appear
   a) Impact on model ranking
   b) A heuristic for assigning corrected d.f. to a model

Binning (collapsing)

Regardless of (A) or (B) … NOD predictor C must be binned before modeling.
"Binning" reduces the number of levels of a NOD predictor:
e.g. Begin with 4 levels R, S, T, U and bin R and T to form R_T, S, U

- C with many levels is not parsimonious (leads to over-fitting)
- For ordered C:
  - Binning may reveal useful non-monotonic relationships of C vs C_Woe
  - Binning may be required by the business to make C_Woe monotonic
- But collapsing C reduces predictive power of C vs. target Y

  Good news: there is a Win-Win:
  - Predictive power usually decreases very little during early stages of collapsing if collapsing is done "optimally"
- How to do "optimal" binning? … next slides.

But First: Information Value (IV) of X vs. Y

X	Y = 0	Y = 1	Col % Y = 0 (A)	Col % Y = 1 (B)	Difference (B)-(A) = (C)	WOE= Log(%Y=1/%Y=0) (D)	IV term (C) * (D) = IV
X=X1	2	3	0.400	0.429	0.029	0.0690	0.0020
X=X2	1	3	0.200	0.429	0.229	0.7621	0.1742
X=X3	2	1	0.400	0.143	-0.257	-1.0296	0.2648
SUM	5	7	1.000	1.000		IV =	0.4409

IV	Interpretation
0.02	un-predictive
0.1	weak
0.2	medium
0.3	strong

VERY STRONG

Iconic table from Siddiqi
Credit Risk Scorecards

magnify Analytic Solutions

marketing associates
creativity · technology · execution

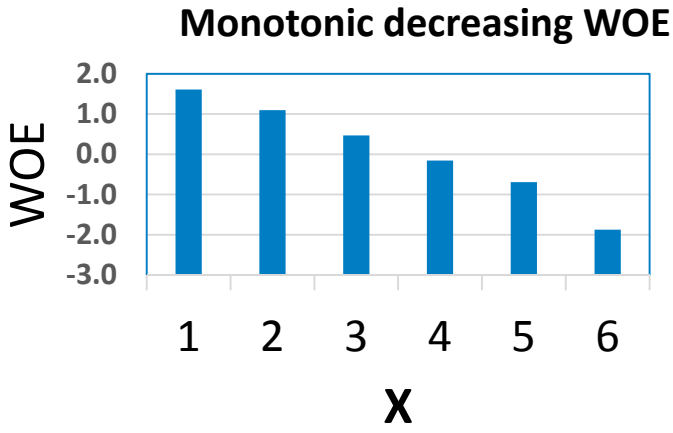
“Optimal” Binning - What is Needed

- DATA PREP:
  - No ZERO cells (a level of X where either there are no goods or no bads)
  - For Interval Predictors, use PROC HPBIN for preliminary binning
- Two MODES:
  - Adjacent (**J**) Pairs of Levels (if X is ordered) or Any (**A**) Pairs of Levels
- METHOD:
  - Find the 2 levels to collapse which maximize **IV** after the collapse vs. all other choices.
  - Repeat collapsing until a “**stopping** criterion” is satisfied
- APPROACH:
  - Collapse: A, B, C, D ➔ A_B, C, D … My approach in this talk is Collapse
  - Split (decision tree): A_B_C_D ➔ A_B, C_D

Other Methods

- Alternative collapsing criteria:
  - **LL**: Log Likelihood … MAX LL at each step … very similar to IV
  - **CS:** Chi-Sq for independence of X v. Y … MAX -log(p) at each step
    … where p is right-tail probability of Chi-Sq … very similar to IV
  - **MO:** Find "monotonic" solutions (if X ordered). This needs more explanation … ➔

A monotonic solution with 6 bins



Using base "e" for entropy:
Entropy = - LL / n

magnify Analytic Solutions

marketing associates
creativity · technology · execution

Monotonic Solutions

Suppose X has 5 levels:

1. Is the 5 bin solution monotonic? (If so, compute IV) … Continue …
2. For 4, examine the solutions having 4 bins. These are:

{1 2} {3} {4} {5} … {1} {2 3} {4} {5} … {1} {2} {3 4} {5} … {1} {2} {3} {4 5}

If any are monotonic, then find solution with best IV. Continue …

3. For 3 … same discussion

Note: Number k bin solutions when X has L levels is:

C(L-1, k-1) = combinations of L-1 taken k-1 at a time … e.g. C(5-1,4-1) = 4

$\sum C(L-1, k-1)$ across k = 2 to L …… **Total number of solutions = $2^{(L-1)} - 1$**

GOAL: For each number of BINS, there is a monotonic solution and which of these is best (as measured by IV, LL, SQ)?

magnify Analytic Solutions

marketing associates creativity · technology · execution

I have some SAS code to do "Optimal" Binning:

```
DATA EXAMPLE;
INPUT X $  Y  F;
DATALINES;
A 0 8
A 1 6
B 0 8
B 1 5
C 0 2
C 1 5
D 0 2
D 1 9
;
```

LET: Mode = **J** and Method = IV

- IV decreases at each step
- Bins become monotonic at k = 3
- Model c is "c-stat" or ROC
- Entropy = -Log Likelihood / n
- L1 to L4 give the binning history

Provide WOE code (e.g. k=3)

```
if X in ( "A","B" ) then X_woe = -0.597837001 ;
if X in ( "C" ) then X_woe = 0.6931471806 ;
if X in ( "D" ) then X_woe = 1.2809338455 ;
```

k	IV	Model c	Entropy = -LL/n	Equal Levels ChiSq	Pr > ChiSq	MONO	L1	L2	L3	L4
4	**0.623**	0.70	0.61	.	.		A	B	C	D
3	**0.618**	0.69	0.61	0.05	0.82	**YES**	A+B	C	D	
2	**0.586**	0.68	0.62	0.26	0.61	**YES**	A+B	C+D		

magnify Analytic Solutions

Illustrative SAS code will be loaded to Website

marketing associates creativity · technology · execution

Stopping Criteria for Mode = J and Method = IV

- IV falls below a threshold (e.g. < 0.1) or "dramatically" decreases
- Change in -2 * log-likelihood from step k to step k-1 is approx. a Chi-Sq (1 d.f.)
  Tests if Level "i" is equal to Level "j". Consider stopping if "significant" (unequal).

k	IV	Model c	Entropy = -LL/n	-2*Log L	Equal Levels ChiSq	Pr > ChiSq	MONO	L1	L2	L3	L4
4	0.623	0.70	0.61	55.25	.	.		A	B	C	D
3	0.618	0.69	0.61	55.31	0.05	0.82	YES	A+B	C	D	
2	0.586	0.68	0.62	55.57	0.26	0.61	YES	A+B	C+D		

Equal Levels ChiSq ≈ $\{-2*\text{Log } L_{k-1} - (-2*\text{Log } L_k)\}$ ~ ChiSq with 1 d.f.
If Prob(ChiSq) < 1 - α, then X "has not changed much". OK to collapse.

magnify Analytic Solutions

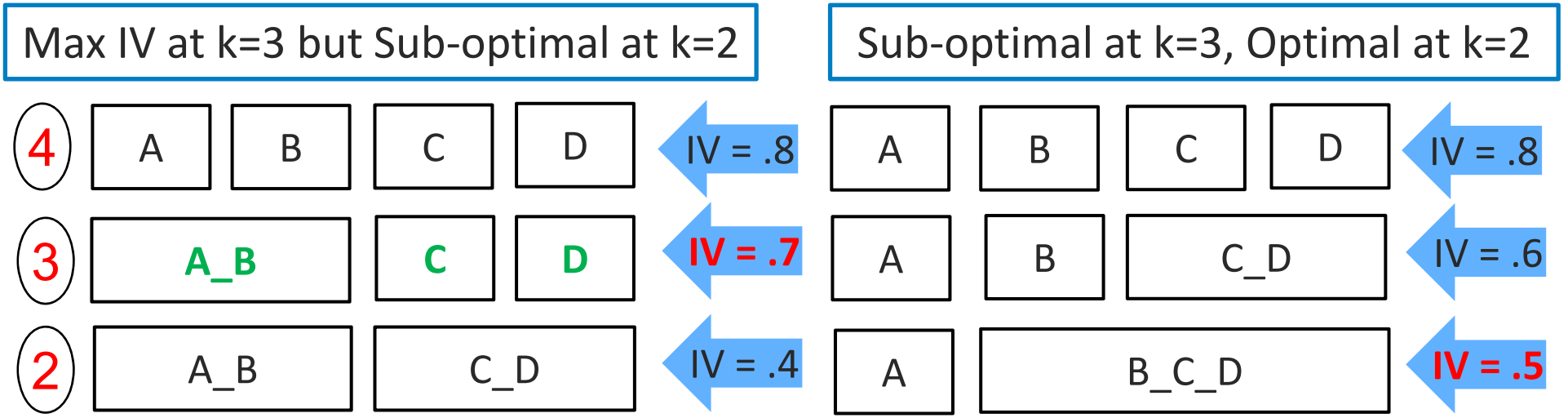
marketing associates
creativity · technology · execution

Comments

For IV and LL see Lund and Brotherton (MWSUG 2013 p. 7)

- Sequence of collapsing by IV and LL can be different.
- Worse: Each method can become sub-optimal.
  - Optimal collapse at "k" can lead to sub-optimal status at "k-1"

➢ Assume {**A_B, C, D**} is best of the 6 possible collapses at k = 3



magnify Analytic Solutions

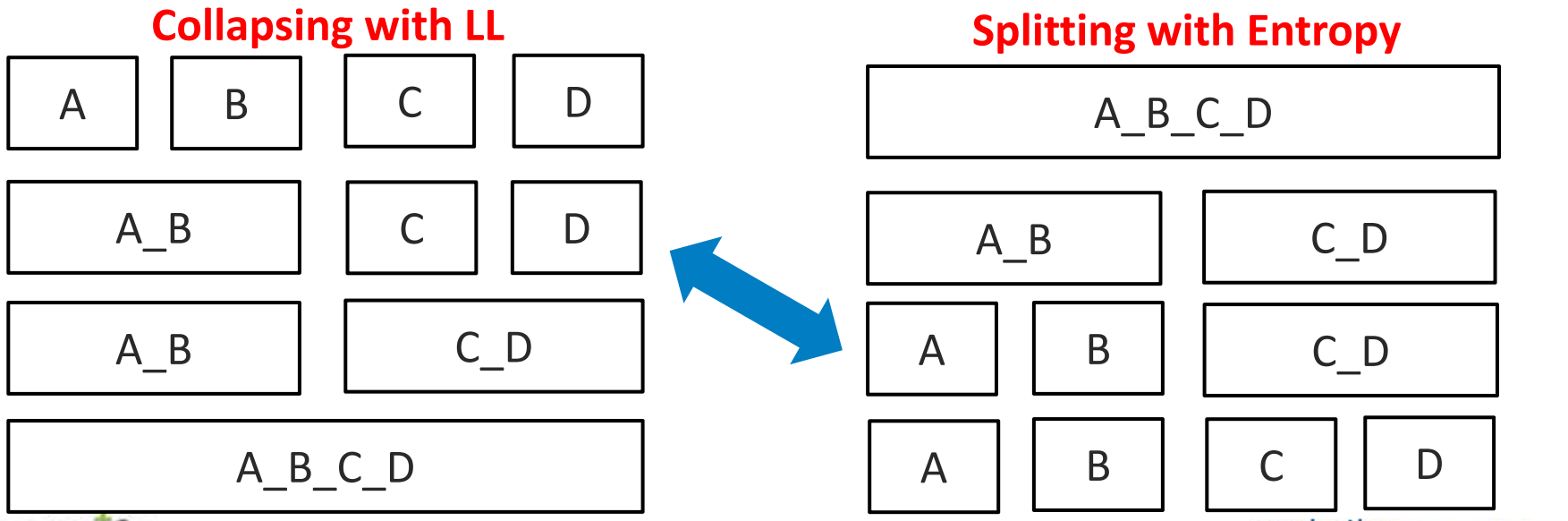
See Lund and Brotherton (MWSUG 2013 p. 6)

marketing associates creativity · technology · execution

Comments

Decision tree “splits” from bottom. Max LL “collapses” from top.

Splitting with Entropy can give different binning than does collapsing with LL (Entropy = -LL/n)



magnify Analytic Solutions

marketing associates creativity · technology · execution

See Lund and Raimi (MWSUG 2012 p. 8-9)

Interactive Grouping Node in SAS® EM

- SAS code based on the prior slides can be effective as a data exploration tool and for one-off model building for organizations having only SAS/Stat.
- For industrial-strength variable binning the Interactive Grouping Node (IGN) in SAS Enterprise Miner is the answer. IGN supports:
  - Targets can be interval or binary.
    - » In a preliminary step interval targets are converted to binary.
  - Predictors (characteristics) are interval, ordinal, or nominal
- IGN is included in the **Credit Scoring** application in SAS EM. Check with your SAS sales rep regarding licensing of this application.
- Although IGN is integrated into the Credit Scoring project flow, it can equally well be used in standalone mode in the development of other types of models.

❖ SAS Training: *Development of Credit Scoring Applications Using SAS EM*

Briefly: Interactive Grouping Node

I'll discuss binary targets with an ordinal predictor. Here is how *I think* it works:

- Among IGN's Grouping Methods for an Ordinal Predictor are
  - **Optimal** (based on a TREE using either entropy or chi-square optimization)
  - **Monotonic** (WOE is monotonic)
- **Optimal**
  - The User-provided *Max Groups* determines how many groups are in the solution and **Optimal** finds the best (tree-based) solution for *Max Groups*
- **Monotonic**
  - User-provided *Max Groups* gives upper limit for groups are in the solution.
  - **Monotonic** finds the largest number of groups ≤ *Max Groups* that gives a monotone solution. Call this *Max M*. If there are several *Max M* solutions, then best (tree-based) is found.

Very Non-Monotonic X (12 levels)

	X_NOT_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

magnify Analytic Solutions

marketing associates
creativity · technology · execution

Interactive Grouping Node for very Non-Mono-X

USER CHOICE		RESULTS			
Max Group	Method	IV	Groups found	Mono	Bins
12	Monotonic	0.316	5	YES	{1 2} {3 4} {5 6} {7 8} {9 10 11 12}
4	Monotonic	0.313	4	YES	{1 2} {3 4} {5 6} {7 8 9 10 11 12}
12	Optimal	1.113	12	NO	{1} {2} {3} {4} {5} {6} {7} {8} {9} {10} {11} {12}
4	Optimal	**0.578**	4	NO	{1 2 3 4} {5} {6} {7 8 9 10 11 12}
Versus "Max IV" collapsing:					
	Method=IV, Mode=J	**0.547**	4	NO	{1 2} {3 4} {5} {6 7 8 9 10 11 12}

IGN Optimal found a better solution for 4 groups than did "Max IV" collapsing.

magnify Analytic Solutions

marketing associates
creativity · technology · execution

Topics for this Talk - #3 and #4

Focus on Logistic Regression (binary and ordinal targets):

1. Weight-of-evidence (WOE) coding of nominal/ordinal/discrete predictors for binary target. Information Value (IV), Log-likelihood, and Model c
2. Binning predictors before use in a Model with binary target
   a) "Optimal" binning (collapsing)
   b) Very short description of Interactive Grouping Node (SAS EM)
3. **Extending WOE and IV to Ordinal Logistic Regression** (… target has 3 or more levels and these levels are ordered)
4. **Binning of a predictor for Ordinal Logistic Regression**
5. The degrees of freedom of a Logistic Model when WOE's appear
   a) Impact on model ranking
   b) A heuristic for assigning corrected d.f. to a model

Semi Warning

The following slides present a proposal for a technique for defining WOE-variables for ordinal logistic regression.

But I can't claim a practical value of the technique at this point in time.

The next step will be to apply the technique to real or realistic examples.

Cumulative Logit Proportional Odds (PO) Model

To simplify: Assume 3 levels for an ordered target Y: 1, 2, 3

Suppose there are 2 predictors X1 and X2

Let p_{ij} = prob. that i^{th} obs. has target value j = 1 to 3

PO Model has 4 parameters α_1 α_2 β_{X1} β_{X2} and is given via 2 equations:

$$\text{Log}\,(p_{i1} / (p_{i2} + p_{i3})) = \alpha_1 + \beta_{X1}*X_{i1} + \beta_{X2}*X_{i2} \quad \ldots j = 1$$

$$\text{Log}\,((p_{i1} + p_{i2}) / p_{i3}) = \alpha_2 + \beta_{X1}*X_{i1} + \beta_{X2}*X_{i2} \quad \ldots j = 2$$

Coefficients of predictors β_{X1} and β_{X2} are same in both equations.

Note: the "cumulative logits" are:

Log of ratio of cumulative probabilities

For more Explanation see:
Logistic Regression using SAS,
Paul Allison. Allison gives the motivation behind the PO model

magnify Analytic Solutions

marketing associates
creativity · technology · execution

PO Model: What are the Probabilities for Obs i ?

Let p_{ij} = prob. that i^{th} obs. has target value j = 1 to 3

PO Model has 2 equations:

$$\text{Log}\,(p_{i1} / (p_{i2} + p_{i3})) = \alpha_1 + \beta_{X1}*X_{i1} + \beta_{X2}*X_{i2} \quad \ldots j = 1$$

$$\text{Log}\,((p_{i1} + p_{i2}) / p_{i3}) = \alpha_2 + \beta_{X1}*X_{i1} + \beta_{X2}*X_{i2} \quad \ldots j = 2$$

Let $A_i = \exp(\alpha_1 + \beta_{X1}*X_{i1} + \beta_{X2}*X_{i2})$

Let $B_i = \exp(\alpha_2 + \beta_{X1}*X_{i1} + \beta_{X2}*X_{i2})$

A and B are the same except for α's

Then: $p_{i1} = 1 - 1/(1+A_i)$

$p_{i2} = 1/(1+A_i) - 1/(1+B_i)$

$p_{i3} = 1/(1+B_i)$ … messy, compared to the binary case

PO Model and PROC LOGISTIC

If a target has more than 2 Levels, then the default model in PROC LOGISTIC is PO. (Nothing more needs to be specified)

The proportional odds assumption may be too restrictive for some models. (PROC LOGISTIC provides a test statistic.)

There are alternatives to PO:

- Adjacent category logit … not discussed today
- Partial proportional odds (PPO) … to be discussed

Partial Proportional Odds (PPO) Model

Suppose there are 3 levels for an ordered target Y: 1, 2, 3

Suppose there are 3 predictors R, S, and Z

Let p_{ij} = prob. that i^{th} obs. has target value j = 1 to 3

PPO Model has parameters α_1 α_2 β_R β_S β_{Z1} β_{Z2} with 2 equations:

$$\text{Log}\,(p_{i1} / (p_{i2} + p_{i3})) = \alpha_1 + \beta_R * R_i + \beta_S * S_i + \beta_{Z\mathbf{1}} * Z_i \quad \ldots j = 1$$

$$\text{Log}\,((p_{i1} + p_{i2}) / p_{i3}) = \alpha_2 + \beta_R * R_i + \beta_S * S_i + \beta_{Z\mathbf{2}} * Z_i \quad \ldots j = 2$$

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

More Explanation, See
Ordinal Response Modeling with the LOGISTIC procedure, Bob Derr, SGF 2013

magnify Analytic Solutions

marketing associates
creativity · technology · execution

Extending WOE to Cumulative Logit Model

From binary case there are 2 **characteristics** of MODEL Y = X_woe that should be extended to the Cumulative Logit Model in order to define a WOE variable. These characteristics are:

1. Equality of CLASS variable model (A) and WOE model (B):
   A. PROC LOGISTIC; CLASS X; MODEL Y = X;
   B. PROC LOGISTIC; MODEL Y = X_woe
2. MODEL Y = X_woe: Intercept = Log(G / B) and Slope = 1

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

Extending WOE for Cumulative Logit - Step 1

If Target Y has J Levels, the need J-1 WOE's for X. One is not enough

Let Target Y have levels A, B, C … we need X_woe1, X_woe2.

Step 1 Ordinal with J = 3

DIVIDE

	Y=			Col %		Col %	
Xi	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				

1 + 2

7+10

vs. Binary, J = 2

	Y=		Col %'s	
Xi	Ai	Bi	Ai / A	Bi / B
1	2	1	0.18	0.14
2	4	3	0.36	0.43
3	4	1	0.36	0.14
4	1	2	0.09	0.29
Total	11	7		

magnify Analytic Solutions

$A = Sum(A_i)$, $B = Sum(B_i)$, $C = Sum(C_i)$

marketing associates creativity · technology · execution

Extending WOE for Cumulative Logit - Step 2

Step 2

	Y=			Col %		Col %		Ratio of Col %		Log (Ratio)	
Xi	Ai	Bi	Ci	Ai / A	(Bi+Ci) / (B+C)	(Ai+Bi) / (A+B)	Ci / C	A v. B+C	A+B v. C	X_WOE1	X_WOE2
1	**2**	**1**	**2**	**0.182**	**0.176**	**0.17**	**0.20**	**=0.182 / 0.176**	**=0.17 / 0.20**	**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**								

$A = Sum(A_i)$, $B = Sum(B_i)$, $C = Sum(C_i)$

$$X\_woe1(i) = LOG[(Ai / A) / ((Bi+Ci) / (B+C))]$$

Proportional Odds (PO) Does Not Support WOE

Try the proportional odds model:

(B) PROC LOGISTIC; MODEL Y = X_woe1 X_woe2;

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	

	Y=		
Xi	Ai	Bi	Ci
1	2	1	2
2	4	3	1
3	4	1	2
4	1	2	5

➢ And … Not equal to: (A) PROC LOGISTIC; CLASS X; MODEL Y = X;

magnify Analytic Solutions

marketing associates creativity · technology · execution

Partial Proportional Odds does have the Properties

(B) PROC LOGISTIC; MODEL Y= X_woe1 X_woe2 / unequalslopes = (X_woe1 X_woe2);

Different slopes for j = 1, j = 2

Maximum Likelihood Estimates			➢Equal to:	
Parameter	Y	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	AS REQUIRED
X_Woe1	B	-127E-12	0	
X_Woe2	A	3.2E-10	0	
X_Woe2	B	1.0000	1	AS REQUIRED

➢ Is Equal to:

(A) PROC LOGISTIC;
CLASS X;
MODEL Y = X / unequalslopes = X;

Forces us to use PPO if we want to use WOE's in Cum. Logit Model

magnify Analytic Solutions

marketing associates creativity · technology · execution

To Implement WOE in a cumulative logit model (J = 3)

- Select nominal or discrete variables that would normally be treated as CLASS and compute the WOE versions.
  - Assume C is such a predictor. Compute C_woe1 and C_woe2
- Formulate PROC LOGISTIC:

```
PROC LOGISTIC DATA = WORK;
MODEL Y = C_woe1 C_woe2 X1 X2 X3
      / unequalslopes = (C_woe1 C_woe2);
/* add other X's to unequalslopes as appropriate */
```

But, as in the binary case, the WOE variables need to be binned.

First: Compute IV for cumulative logit?

	Y=			Col %		Col %		Diff of Col %		Log (Col %)		Diff*WOE	Diff*WOE
Xi	Ai	Bi	Ci	Ai / A	(Bi+Ci) / (B+C)	(Ai+Bi) / (A+B)	Ci / C	A v B+C	A+B v C	X_Woe1	X_Woe2	IV1	IV2
1	2	1	2	0.182	0.176	0.17	0.20	**0.01**	**-0.03**	**0.03**	**-0.18**	**0.0002**	**0.0061**
2	4	3	1	0.36	0.24	0.39	0.10	0.13	0.29	0.44	1.36	0.0559	0.3923
3	4	1	2	0.36	0.18	0.28	0.20	0.19	0.08	0.72	0.33	0.1353	0.0256
4	1	2	5	0.09	0.41	0.17	0.50	-0.32	-0.33	-1.51	-1.10	0.4847	0.3662
Total	11	7	10									**0.6760**	**0.7902**

Diff of Col % = 0.182 - 0.176 = 0.01

Log of ratio of Col % = Log(0.182 / 0.176) = 0.03

if Y has J levels, then there are J-1 Information Values …

$A = Sum(A_i)$, $B = Sum(B_i)$, $C = Sum(C_i)$

magnify Analytic Solutions

marketing associates creativity · technology · execution

What are criteria for optimal binning?

if Y has J levels, there are J-1 Information Values ….

How can the multiple IV's be used?

- One idea: Total of IV's

  … e.g. TOTAL_IV = IV1 + IV2 … if Y has 3 levels.

➢ What are "good" values of TOTAL_IV ?
  - How does this depend on J?

➢ Need a chart like Siddiqi's from the binary case

➢ More work is needed

Program for binning for cumulative logit predictors

… I'm working on a pre-beta version

- TWO MODES:
  - Adjacent Pairs of Levels (if X is ordered) or Any Pairs of Levels
- METHOD:
  - Finds the 2 levels to collapse which maximize **Total IV** after the collapse vs. all the other choices.
  - Repeat this process until a "stopping criterion" is satisfied

Sample of METHOD = IV with MODE = A

	Y		
X	A	B	C
1	2	1	2
2	4	3	1
3	4	1	2
4	1	2	5
Total	11	7	10

Mode = A
(Any pairs are eligible for collapse)

step	-2_LL	IV_1	IV_2	TOTAL_IV	L1	L2	L3	L4
4	53.92	0.68	0.79	1.47	01	02	03	04
step	-2_LL	IV_1	IV_2	TOTAL_IV	L1	L2	L3	L4
3	54.27	0.62	0.76	1.39	01+03	02	04	
3	55.24	0.66	0.67	1.33	01	02+03	04	
3	55.29	0.66	0.55	1.21	01+02	03	04	
3	55.23	0.43	0.69	1.12	01+04	02	03	
3	57.45	0.09	0.51	0.61	01	02	03+04	
3	58.96	0.20	0.04	0.24	01	02+04	03	
step	-2_LL	IV_1	IV_2	TOTAL_IV	L1	L2	L3	L4
2	55.99	0.62	0.54	1.16	01_03+02	04		
2	57.54	0.08	0.50	0.58	01_03+04			
2	59.31	0.15	0.01	0.16	01_03	02+04		

magnify Analytic Solutions

marketing associates
creativity · technology · execution

Topics for this Talk - #5

Focus on Logistic Regression (binary and ordinal targets):

1. Weight-of-evidence (WOE) coding of nominal/ordinal/discrete predictors for binary target. Information Value (IV), Log-likelihood, and Model c
2. Binning predictors before use in a Model with binary target
   a) "Optimal" binning (collapsing)
   b) Very short description of Interactive Grouping Node (SAS EM)
3. Extending WOE and IV to Ordinal Logistic Regression (… target has 3 or more levels and these levels are ordered)
4. Binning of a predictor for Ordinal Logistic Regression
5. **The degrees of freedom of a Logistic Model when WOE's appear**
   a) **Impact on model ranking**
   b) **A heuristic for assigning corrected d.f. to a model**

Back to Binary

The following discussion focus on the binary target in logistic regression.

Here, this discussion has practical application.

But the discussion also applies to the ordinal target in logistic regression.

SBC (Schwarz Bayes) = -2*LL + log(n)*K (smaller is better)

Fit is measured by -2*LL while log(n)***K** adds a penalty for sample and d.f.

Schwarz-Bayes ranks models - smaller is better (*)

- Usual approach to "**K**" is to count coefficients in the model.
- With WOE there is ONE coefficient BUT there is substantial pre-modeling preparation (entire X*Y table was used).
- If WOE predictors are counted as having ONE d.f. each, then SBC may be **under-stated** (not enough penalty).

What is a better value for SBC when some predictors are WOE's?

(*) There is an extensive theory supporting SBC - See Dziak, J, Coffman, D., Lanza, S., Li, R. (2012). Sensitivity and Specificity of Information Criteria … for an overview

magnify Analytic Solutions

marketing associates creativity · technology · execution

A heuristic approach to adjusting SBC for WOE's.

This approach requires that two models be fit.

1. The model with WOE variables
   - Used for measuring predictive accuracy and scoring new observations
2. The model with WOE variables declared as CLASS variables
   - Used to obtain adjusted SBC and to rank the models

1. PROC LOGISTIC; MODEL Y = C1_woe C2_woe <other X's>; … SBC_woe

2. PROC LOGISTIC; CLASS C1_woe C2_woe;
   MODEL Y = C1_woe C2_woe <other X's>; … SBC_class

Adjusted Schwarz Bayes is: **SBC_adj** = MAX(SBC_class, SBC_woe)

magnify Analytic Solutions

NOTE: Unusual to have SBC_class < SBC_woe

marketing associates
creativity · technology · execution

A heuristic approach to adjusting SBC for WOE's.

Data set has 100 obs. with Y X1 X2 X8, C_woe … C has **7** levels

CLASS C_woe;

#	Variables in Model	SBC_ Adj	SBC_ class	-2LL_ class	DF_ class	SBC_ woe	-2LL_ woe	DF_ woe
1	X1 X2 X8 C_woe	**145.58**	**145.58**	**99.53**	**10**	**123.01**	**99.98**	**5**
2	X1 X2 X8	**131.17**				**131.17**	**112.75**	**4**

4 predictors + 1 intercept

3 predictors + 1 intercept + 6 for C_woe

- In model #1 the **-2LL's** are almost equal at **99+** … ~ same models
  - SBC_woe was under-penalized for K … (using **5** instead of ~**10**)
- #2 is selected v. #1 based on lower adjusted SBC_**adj** (=**131.17**)

magnify Analytic Solutions

marketing associates creativity · technology · execution

References

1. Paul Allison. Logistic Regression using SAS (see section on proportional odds)
2. Bob Derr. Ordinal Response Modeling with the LOGISTIC procedure, SGF 2013
3. B. Lund, D. Brotherton. Information Value Statistic, MWSUG 2013
4. B. Lund, S. Raimi. Collapsing Levels of Predictor Variables for Logistic Regression and Weight of Evidence Coding, MWSUG 2012
5. B. Lund. Finding and Evaluating Multiple Candidate Models for Logistic Regression, SGF 2016

Weight of Evidence Coded Variables for Binary and Ordinal Logistic Regression

Bruce Lund (blund@magnifyas.com)
Magnify Analytic Solutions,
Division of Marketing Associates

MSUG conference June 9, 2016

magnify
Analytic Solutions

marketing associates
creativity · technology · execution

What is the value of adjusted D.F. for WOE's?

Simply solve for **K** in the equation:

$$-2*LL_{woe} + \log(n)*K = -2*LL_{class} + \log(n)*K_{class}$$

For model #1 below, **K** = 9.9

K will be between DF_{woe} and DF_{class} … (here: 5 to 10)

#	Variables in Model	SBC_ Adj	SBC_ class	-2LL_ class	DF_ class	SBC_ woe	-2LL_ woe	DF_ woe
1	X1 X2 X8 C_woe	**145.58**	**145.58**	**99.53**	**10**	**123.01**	**99.98**	**5**

Should **K** ever be less than determined by the formula above ??

magnify Analytic Solutions

marketing associates creativity · technology · execution

Proportional Odds (PO) Model

Why “proportional odds”?

Let “r” and “s” be two values of X1 and consider log-odds ratio of 1 versus 2 + 3:

$$\text{Log}\left[\frac{p_{r1}/(p_{r2}+p_{r3})}{p_{s1}/(p_{s2}+p_{s3})}\right] = (r - s) * \beta_{X1}$$

The log-odds ratio is *proportional* to the difference (r - s)

This also holds for log-odds ratio of 1 + 2 versus 3.

Stopping Criteria

Log-odds Ratio is roughly zero, then OK to collapse.

- Assume levels "i" and "j" are selected for collapse to go to step k-1
  - If Odds G_i / B_i and Odds G_j / B_j are "close", then OK to collapse i and j
  - Measure "close" by: LOG((G_i / B_i) / (G_j / B_j)) near ZERO
  - Compute a Std Dev and 95% confidence interval
  - If C. I. contains ZERO, then OK to collapse

k	collapsing to	LO Ratio after collapse	LO Ratio Std Dev	LOminus 2SD	LOplus 2SD
4	3	0.182	0.785	-1.39	1.75
3	2	-0.588	1.145	-2.88	1.70

Finding a solution … using IV

LET IV_i = contribution to IV from the i^{th} term

LET IV_i_j = contribution to IV from collapsed i^{th} term and j^{th} term

Then always a collapsed pair has less (or equal to) IV:

$$IV\_i\_j \leq IV\_i + IV\_j$$

Criterion to collapse: Find the pair (i, j) to minimize IV_i + IV_j - IV_i_j

If **J** "adjacent", then j = i + 1. Otherwise, **A, "**any" (i, j)

LL: Same idea applies … Find the pair (i, j) to minimize the loss of LL.