Weight of Evidence, Dummy Variables, and Degrees of Freedom

Bruce Lund
Statistical Trainer and Consultant
blund_data@mi.rr.com

This Topic applies to Logistic Regression Models

GOALS:

1. Define Weight of Evidence (WOE) coding for “discrete” predictors
2. … As an alternative to Dummy Variables
3. Explain the “degrees of freedom” problem for WOE
4. Propose a process to assign d.f. to WOE coded predictor
5. Mention my SAS macro that implements the solution

Details in my paper on MSUG site:

“Weight of Evidence, Dummy Variables, and Degrees of Freedom”

Weight of Evidence Coding of discrete predictor D

D: Predictor
Y: Target (response)

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

D	Y = 0	Y = 1	Col % Y = 0	Col % Y = 1	WOE= Log(%Y=1/%Y=0)
D=D1	2	3	0.400	0.429	0.0690
D=D2	1	3	0.200	0.429	0.7621
D=D3	2	1	0.400	0.143	-1.0296
SUM	5	7	1.000	1.000	

If D = D3 then D_woe(D3) = -1.0296

WOE is widely used in Credit Risk Models as alternative to Dummy Variables
WOE coding supports Credit Scorecard Development

Models (A) and (B) are the Same

Let discrete predictor D have J > 2 levels.

Following 2 models are equal (i.e. produce the same probabilities):

(A) PROC LOGISTIC DESCENDING; CLASS D; MODEL Y=D;

(B) PROC LOGISTIC DESCENDING; MODEL Y=D_woe;

In Model (A) … D has J-1 degrees of freedom

Therefore,

In Model (B) … D_woe must have J-1 d.f.

If more than 1 Predictor in Model

Let D have J > 2 levels and Z be a numeric predictor

Then the following two models are NOT the same (i.e. different probabilities):

(A) PROC LOGISTIC DESCENDING; CLASS D; MODEL Y=D Z;

(B) PROC LOGISTIC DESCENDING; MODEL Y=D_woe Z;

In Model (A) … D has J-1 degrees of freedom

What about Model (B). How many d.f. to assign to D_woe?

One? … but D_woe had J-1 > 1 d.f. … before Z was considered

J-1? … but $LL_{WOE} < LL_{CLASS}$ … so, is J-1 too high?

ANSWER: Somewhere 1 ≤ d.f. ≤ J-1 … QUESTION: But Where?

Why does it matter how many d.f. to assign to D_woe?

- Consider FORWARD selection when fitting a logistic model …
- Suppose Z is already in the model
- Suppose the choice for the next predictor to enter is X or D_woe
  ➔ How to choose between X and D_woe?
- The d.f. for D_woe matters because it determines:
  - P-value of Chi-Sq to Enter … Chi-Sq probability depends on d.f.
  - AIC to Enter (AIC = -2LL + 2*(d.f.)) … depends on d.f.
- ***Normal practice*** is to assign 1 d.f. to D_woe in model fitting
- … This is bad because …
  - Favors entry of D_woe vs. X (*unfair* due to pre-coding of D_woe)
  - See paper for more discussion

Quick Detour: Nested Models and Model Comparison Test

Let the restricted model have predictors: X_1 - X_R

Let the full model have predictors: X_1 - X_R X_{R+1} - X_F

The Restricted Model is said to be "Nested" in the Full Model.

Then (for large sample) the model comparison chi-square statistic is:

$$\text{Chi-Sq}_{F\text{-}R} = -2*\text{Log}(L)_{RESTRICTED} - (-2*\text{Log}(L)_{FULL})$$

with F - R degrees of freedom

Pick your favorite **α** (e.g. 0.05) and suppose computed Chi-Sq = **t**

Restricted Model is statistically the same as the Full Model when:

$P(\text{Chi-Sq}_{F\text{-}R} > t) > \alpha$ … [= accept null hypothesis of equal models]

This is called the "Model Comparison Test"

OK … How to Assign d.f. to D_woe?

Consider two models (CLASS) and (WOE):

(CLASS) PROC LOGISTIC DESCENDING; CLASS D; MODEL Y=D Z;

(WOE) PROC LOGISTIC DESCENDING; MODEL Y=D_woe Z;

The Full Model is (CLASS) and Restricted Model is (WOE) are "nested"

"Nesting" is explained in my paper (… not the usual definition)

Let's consider (with some abuse):

$$\text{Chi-Sq}_{C\text{-}W} = -2*\text{Log}(L)_{WOE} - (-2*\text{Log}(L)_{CLASS})$$

➔ DEFINITION: The d.f. $_{C\text{-}W}$ for $\text{Chi-Sq}_{C\text{-}W}$ is *largest d.f.* which makes CLASS and WOE statistically equal (for given **α**)

The d.f. $_{C\text{-}W}$ can have a *fractional value* (e.g. 1.5)

OK … How to Assign d.f. to D_woe?

From Prior Slide:

$$\text{Chi-Sq}_{C\text{-}W} = -2*\text{Log}(L)_{WOE} - (-2*\text{Log}(L)_{CLASS})$$

DEFINITION: The d.f. $_{C\text{-}W}$ for $\text{Chi-Sq}_{C\text{-}W}$ is *largest d.f.* which makes CLASS and WOE statistically equal (for given **α**)

Let D have J levels …

then the d.f. for D_woe is given by

$$J\text{-}1 - \text{d.f.}_{C\text{-}W}$$

If $-2*\text{Log}(L)_{WOE}$ and $-2*\text{Log}(L)_{CLASS}$ are *really* close, D_woe gets J-1 d.f.

If $-2*\text{Log}(L)_{WOE}$ and $-2*\text{Log}(L)_{CLASS}$ are *far* apart, then D_woe gets 1 d.f.

How to Assign d.f. to D_woe?

Here is an example of an in-between case (using definition on prior slide)

Let **α** = 0.05 and assume D has **J** = 4 levels

Then: D_woe gets **4**-1 - **2.6** = 0.4 d.f.

… see my paper for full discussion

Model with <X> and now enter:	-2*Log(L)	t = -2*Log(L)$_{WOE}$ + 2*Log(L)$_{CLASS}$	Find max d.f. so that: P(T > t | df_{max}) > 0.05
D_woe	100.0	100-98 = 2	df_{max} = **2.6**
D	98.0	**2.6** makes CLASS and WOE statistically equal	

MACRO to Implement FORWARD with best AIC

- FORWARD selection: Select predictor with min AIC, with *adjusted* d.f. for WOE variables.
- Built upon PROC LOGISTIC or HPLOGISTIC (user option)
- Requires a run of a logistic procedure for *each* candidate variable at *each* step
  - … this is a severe problem if many X's … HPLOGISTIC can speed up the processing.
- In my paper … An example with one WOE called C_woe and 3 numeric predictors
  - C_woe entered in the 3rd step
  - *Without* d.f. adjustment, C_woe would enter in the 1st step (details not shown)

Obs	step	min AIC var	min adj AIC	best model	adj-df for min	new model df	new included var
1	1	X8	123.462		1	2	X8
2	2	X2	120.396		1	3	X8 X2
3	3	C_woe	118.577	*	6	9	X8 X2 C_woe
4	4	X10	119.175		1	10	X8 X2 C_woe X10

Weight of Evidence, Dummy Variables, and Degrees of Freedom

Bruce Lund
Statistical Trainer and Consultant
blund_data@mi.rr.com