Paper AA-09-2014

Selection and Transformation of Continuous Predictors for Logistic Regression

Bruce Lund, Magnify Analytic Solutions
A Division of Marketing Associates, Detroit, MI

ABSTRACT

This paper discusses the selection and transformation of continuous predictor variables for the fitting of binary logistic models. The paper has two parts: (1) A procedure and associated SAS® macro is presented which can screen hundreds of predictor variables and 10 transformations of these variables to determine their predictive power for a logistic regression. The SAS macro passes the training data set twice to prepare the transformations and one more time through PROC TTEST. (2) SAS macros are discussed which conduct the FSP (function selection procedure) for logistic regression by P. Royston and W. Sauerbrei.

INTRODUCTION

The setting for this discussion is direct marketing or credit scoring or other applications of binary logistic regression where sample sizes for model building are large (perhaps exceeding 1000 observations for each target value) and the emphasis is on building predictive models to be used for scoring future observations. In this setting:

The preparation of predictors for binary logistic regression includes the following phases:

- Screening predictors to detect predictive power
- Transforming the predictors to maximize the predictive power
- Other phases (not discussed) include finding interaction of predictors and elimination of collinear predictors

Predictors fall into three broad categories:

(1) Nominal and ordinal
(2) Counts
(3) Continuous

An effective and widely used transformation for nominal and ordinal predictors is weight-of-evidence (WOE) coding. WOE is also often applied to count predictors. Optimal binning before WOE transformation is a key requirement.[1] [2]

But sometimes the routine use of binning of continuous predictors followed by WOE coding in direct marketing or credit scoring may over-fit and complicate these predictive models. Specifically, when a functional form can be accurately identified for a continuous predictor, the application of binning and WOE coding will lose predictive power.

FIGURE 1 provides an illustration. In this loose hypothetical case, the relationship between predictor X and log-odds of Y is linear. But the approximation to log-odds(Y) by X_cut3 creates three abrupt jumps in the prediction of log-odds(Y). These jumps are not related to underlying behavior of X and Y and create prediction error.

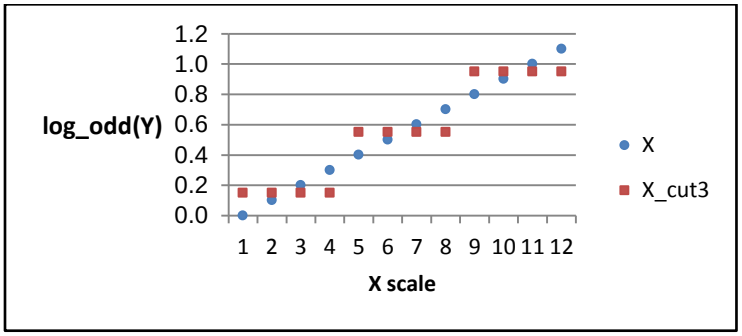

FIGURE 1: Hypothetical 3-cutpoint binning of a predictor X

THIS PAPER HAS TWO MAIN SECTIONS

(1) In the first section a statistical procedure and SAS macro for this procedure are given for screening hundreds of continuous predictors for logistic regression. The goal is to identify predictor candidates that merit further study.

[1] A SAS macro for binning and WOE transformations is given by Lund and Brotherton (2013) and Lund (2013). An updated version (version 8f) of the macro is available from the author.
[2] The binning (using cut-points) of a predictor in bio-statistics is criticized by Royston and Sauerbrei (2008 p. 58)

This is done by measuring the predictive power of the original predictor and 10 transformations of the predictor. The entire procedure requires 3 passes of the original data set regardless of the number of predictors.

(2) In the second section there is a discussion of the FSP (function selection procedure) of Royston and Sauerbrei (2008).[3] The book by Royston and Sauerbrei, *Multivariable Model-building*, (2008) includes a link to a SAS implementation of FSP (see p. 267). FSP can be applied to ordinary regression, logistic regression, and Cox regression. Our focus is only on logistic regression. We run the SAS implementation of FSP on a test case and then present an alternative version of a SAS macro for FSP for logistic models. This alternative macro is relatively simple and includes a helpful and necessary pre-processing DATA step.

SCREENING CONTINUOUS PREDICTORS FOR PREDICTIVE POWER

Given dozens or hundreds of candidate continuous predictors, the "screening problem" is to test each predictor as well as a collection of transformations of the predictor for, at least, minimal predictive power in order to justify further investigation. If the number of candidate predictors is only a few, a brute force approach of fitting the predictor and transformations of the predictor by PROC LOGISTIC provides a simple, direct solution. However, each PROC LOGISTIC requires a pass of the training data set.

Instead, an alternative procedure is described which screens dozens or hundreds of predictors in a single run of PROC TTEST. In this paper this procedure is given as a SAS macro %LOGIT_CONTINUOUS. The use of this macro is recommended wherever the number of candidate predictors is more than, perhaps, five. %LOGIT_CONTINUOUS takes advantage of a connection between 2-group discriminant analysis, a t-test, and logistic regression. This connection is fully developed in Appendix A and only the result keys are presented below.

THE CONNECTION BETWEEN 2-GROUP DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

Let X be a predictor for a binary target Y. The values of Y identify two groups (j = 1, 2) of observations. It is assumed that X has a univariate normal distribution for each group with common standard deviation σ but differing means μ_1 and μ_2. Then, just as in logistic regression, the 2-group discriminant analysis leads to the equation:

$$\text{Log } (P(Y=1 \mid X=x) / P(Y=2 \mid X=x)) = \beta_0 + \beta_1 x$$

The key results are:

(1) When fitting the 2-group discriminant model to a sample, the coefficient β_1 is estimated by b_{1D}, where b_{1D} is found by substituting sample statistics from the two groups, $\bar{x}_1$, $\bar{x}_2$, and S_p^2 as shown in equation (A):

$$b_{1D} = (\bar{x}_1 - \bar{x}_2) / S_p^2 \ldots \text{(A)}$$

Here, the pooled variance S_p^2 estimates σ^2 where S_j^2 are the sample variances for samples j = 1, 2 and

$$S_p^2 = \{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\}/(n_1 - 1 + n_2 - 1)$$

The "D" is added as a subscript to "b_1" to indicate that the method of fitting is by discriminant analysis. Additionally, "L" will be added as a subscript to "b_1"to give b_{1L} to indicate when the estimation of β_1 is by logistic regression maximum likelihood.

Both b_{1D} and b_{1L} are consistent estimators of β_1. For large enough samples, b_{1D} will be close to b_{1L}

(2) In order to test that $\beta_1 = 0$ vs. $\beta_1 \neq 0$ the discriminant analysis coefficient b_{1D} can be regarded as a t-statistic with $n_1 + n_2 - 2$ d.f. via this factorization: $b_{1D} = t\ (1/S_p)\ \text{sqrt}(1/n_1 + 1/n_2)$. The square of this t-statistic is essentially a chi-square with 1 d.f. and it will be close to the Wald chi-square statistic for b_{1L} from logistic regression.

COMPARISON OF DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION WITHOUT NORMALITY

Without the normality assumptions on X, there is not a theory that guarantees that b_{1D} and b_{1L} will be close in value.

Press and Wilson (1978 p. 705) state that b_{1D} and b_{1L} will give similar results except if the response (Y=1) or non-response (Y=0) is heavily concentrated over a sub-region of the domain of X. Hosmer, Lemeshow, and Sturdivant (2013, p. 21-22) note that when X is binary, then b_{1D} will over-estimate β_1.

The observations of Press and Wilson and Hosmer, Lemeshow, and Sturdivant rule out the usage of b_{1D} when X is a binary variable, especially when the response is concentrated over one of the two values of X. (Consider variables X3 and X4 in the section on simulation which directly follows.)

[3] See Royston and Sauerbrei (2008 p. 82) for the FSP and references to their earlier papers.

DESCRIPTION OF SIMULATION OF ONE-VARIABLE DISCRIMINANT AND LOGISTIC MODELS

To test the robustness of the approximate equality of b_{1D} and b_{1L} and their associated chi-square statistics, the distributions of six predictors X1 to X6 were simulated. There were 4,000 observations with 2,232 observations where binary target Y = 1 and 1,768 observations where Y = 2.

X1: Distribution of X1 when Y=1 is N(0, 1).[4] Distribution of X1 when Y=2 is N(0.1, 1). The distributions for the two groups meet the assumptions of normality with equal variances as given for equation (A). The values of b_{1D} and b_{1L} and their chi-squares should be nearly identical.

X2: Distribution of X2 when Y=1 is N(0, 1). Distribution of X2 when Y=2 is N(0.1, 1.2). The distributions for the two groups meet the assumptions of normality but without equal variances.

X3: Distribution of X3 is binary with P(Y=2|X3=0) = 0.358 and P(Y=2|X3=1) = 0.544. The distribution of Y strongly depends on X3. There are 2,088 observations (52.2%) where X3 = 0.

X4: Distribution of X4 is binary with P(Y=2|X4=0) = 0.459 and P(Y=2|X4=1) = 0.426. The distribution of Y is much less dependent on X4. There are 1,949 observations (48.7%) where X4 = 0.

X5: Distribution of X5 has 5 values having a quadratic relationship to P(Y = 2| X5) as shown in the table.

TABLE 1: Distribution of Rate of Y=2 for X5

X5	N Obs	P(Y=2 | X5)
1	835	0.380
2	764	0.424
3	799	0.454
4	797	0.504
5	805	0.450

X6: Distribution of X6 is uniform on 0 to 1 for both Y = 1 and Y = 2. Both b_{1D} and b_{1L} should be insignificant.

The simulation code is given below.

```
data example;
do i = 1 to 4000;
   Y = (ranuni(12345) < .45) + 1;  /* Y=2 are ~ 45% */
   /* Normal distribution tests */
   if Y = 2 then X1 = rannor(12345) + 0.1;
    else if Y = 1 then X1 = rannor(12345);
   if Y = 2 then X2 = 1.2*rannor(12345) + 0.1;
    else if Y = 1 then X2 = rannor(12345);
   /* binary variable tests */
   if Y = 2 then X3 = (ranuni(12345) > .40);
    else if Y = 1 then X3 = (ranuni(12345) > .60);
   if Y = 2 then X4 = (ranuni(12345) > .50);
    else if Y = 1 then X4 = (ranuni(12345) > .50);
   /* discrete variable test */
   r = ranuni(12345);
   if Y = 2 then do;
      if r < .20 then X5 = 1;
       else if r < .40 then X5 = 2;
       else if r < .60 then X5 = 3;
       else if r < .80 then X5 = 4;
       else X5 = 5;
      end;
   r = ranuni(12345);
   if Y = 1 then do;
      if r < .23 then X5 = 1;
       else if r < .42 then X5 = 2;
       else if r < .61 then X5 = 3;
       else if r < .80 then X5 = 4;
       else X5 = 5;
      end;
   /* equal distributions test */
   r = ranuni(12345);
```

[4] Normal with mean 0 and standard deviation 1

```
   if Y = 2 then X6 = r;
   r = ranuni(12345);
   if Y = 1 then X6 = r;

   output;
   end;
```

The code below computes the values for b_{1D} (coefficient for X from the discriminant model) for X1 – X6. First, preliminary calculations are performed by PROC TTEST. An ODS OUTPUT data set for PROC TTEST is the ODS data set STATISTICS. It is saved as TS in the code below. TS is further processed by a DATA step to compute b_{1D}. Here, Y = 2 is being modeled as the "response" versus Y = 1. The results are printed in TABLE 2.

Note, no discriminant analysis procedure appears in this code,

```
ods listing;
ods output Statistics = TS;
ods exclude ConfLimits;
ods exclude Equality;
ods exclude EquivLimits;
ods exclude EquivTests;
ods exclude TTests;

proc ttest data=example plots=none;
   class Y;
   var X1 X2 X3 X4 X5 X6;

data TS_2; set TS;
retain mean1 mean2 n1 n2;
keep variable n1 n2 mean1 mean2 sp b1D;
if class = "1" then do; mean1 = mean; n1 = n; end;
if class = "2" then do; mean2 = mean; n2 = n; end;
if class = "Diff (1-2)"
then
do;
   sp = stddev;
   b1D = -(mean1 - mean2)/(sp**2); /* minus sign to model Y=2 as the response */
   output;
   end;
proc print data = TS_2 noobs; var Variable b1D;
run;
```

The coefficients from PROC LOGISTIC for the same six models are compared to the results of discriminant analysis. The subscript "L" indicates the coefficient was fitted by logistic regression. SAS code for computing b_{1L} is not included.

TABLE 2: Comparison of Coefficients from Discriminant and Logistic

Variable	b_{1D}	b_{1L}
X1	0.089	0.089
X2	0.111	0.111
X3	0.828	0.805
X4	-0.136	-0.136
X5	0.090	0.090
X6	0.160	0.160

Although only X1 satisfied the assumptions for equation (A), the discriminant analysis coefficient b_{1D} is close to the coefficient b_{1L} from logistic regression for 5 of the 6 variables. Only for X3 is there a noticeable difference. Here, as predicted by Hosmer, Lemeshow, and Sturdivant, the value of b_D exceeds b_{1L} in absolute value. Of more importance is whether the square of the t-statistic associated with b_{1D} is a good approximation to the Wald chi-square for the logistic coefficient b_{1L}. The code below computes the square of the t-statistic which underlies b_{1D}.

```
ods listing;
ods output TTests = TT;
proc ttest data = example plots = none;
   class Y;
   var X1 X2 X3 X4 X5 X6;

data TT; set TT;
```

```
tValue = -tValue; /* minus sign to model Y=2 as the response */
chisq_D = tValue**2;
label Probt = "Prob ChiSq";

proc print data = TT label;
var Variable tValue DF Probt chisq_D;
where method = "Pooled";
```

The square of the t-statistic is a chi-square with 1 degree of freedom since a t-statistic with a large number of degrees of freedom (3,998 in this example) is essentially a standard normal and the square of a standard normal is distributed as a chi-square with 1 degree of freedom.

Chi-square values and significance levels from PROC TTEST are given by TABLE 3.

TABLE 3: Chi-Square Values for Coefficients from Discriminant Analysis

Variable	t Value	d.f.	ChiSq for b_{1D}	Prob ChiSq
X1	2.80	3998	7.823	0.0052
X2	3.77	3998	14.243	0.0002
X3	12.73	3998	162.138	<.0001
X4	-2.14	3998	4.568	0.0326
X5	4.03	3998	16.222	<.0001
X6	1.47	3998	2.148	0.1428

TABLE 4 gives the chi-squares from logistic. With the exception of X3 the chi-square values and significance from PROC LOGISTIC are very close to the values from PROC TTEST and, in particular, are close enough to function as a screener for predictor variable candidates. Even for X3 both the chi-squares are large and indicate the significance of X3 as a predictor. (SAS code for computing logistic results is not included.)

TABLE 4: Chi-Square Values from Logistic

Variable	ChiSq for b_{1L}	Prob ChiSq
X1	7.796	0.0052
X2	14.145	0.0002
X3	153.718	<.0001
X4	4.563	0.0327
X5	16.127	<.0001
X6	2.146	0.1430

CONCLUSIONS

The risk in using the t-statistic as a screener is that a good predictor for logistic regression would be rejected. A less serious risk is that non-productive predictors are retained for the next step of modeling. Neither of these outcomes is indicated by this simulation nor by other tests of predictors from actual data (not shown here). Still, more extensive simulation testing of the t-test screener is warranted.

The X5 variable with only 5 values had very similar t-statistic and logistic results. However, binning and WOE coding of X5 remains the preferred approach rather than applying a continuous transformation to X5.

MACRO FOR SCREENING HUNDREDS OF LOGISTIC MODEL PREDICTORS

The SAS macro %LOGIT_CONTINUOUS can screen hundreds of numeric predictors for logistic regression as well as 10 transformations of these predictors using the chi-square which is computed from a t-statistic. This t-statistic is mathematically derived from the coefficient of 2-group discriminant analysis as explained in Appendix A. The 10 transformations include 7 monotonic transformations and 3 quadratic transformations. Three passes of the data set are required for this computation.

1. PROC MEANS to determine the minimum value of each predictor
2. A DATA STEP:
   a. Predictors with minimum < 1 are shifted to have minimum value of 1
   b. 10 transformations of the predictor are computed (such as LOG, X^2, and others)
3. A PROC TTEST to compute t-statistics

The original X and the 10 transformations are:

- 8 monotonic: X^p where p is taken from S = {-2, -1, -0.5, 0, 1, 0.5, 2, 3} where "0" denotes log(x). This list includes the original X. [5]

[5] These transformations anticipate the discussion of the second section of the paper of the FSP by Royston and Sauerbrei. A future enhancement to %LOGIT_CONTINUOUS will be automatic scaling of X^2 X^3 and the quadratic transforms to reduce extreme values.

- 3 quadratic: $(X\text{-median})^2$, $(X\text{-p25})^2$, $(X\text{-p75})^2$ where median, p25, and p75 are respectively the 50^{th}, 25^{th}, and 75^{th} percentiles for X.

The parameters for %LOGIT_CONTINUOUS are:

DATASET: The input data set name.
Y: A numeric binary target variable where the larger value is the "positive" response.
INPUT: A list of numeric predictor variables delineated by space. A predictor may have missing values.

EXAMPLE:

%LOGIT_CONTINUOUS is run on X1 and X6 from the data set "example". The macro call is:

```
%LOGIT_CONTINUOUS(example, Y, X1 X6);
```

As shown in the TABLE 5 the best transformations of X1 are $X1^3$, $X1^2$ and X1. These have very similar chi-square values. Support for the cubic and square transformations is shown by inspection of TABLE 6 where X1 is ranked into 8 equal groups and the rates of Y=2 within each group are given.

The final choice of a transformation of X1 is addressed by the FSP (function selection procedure) of Royston and Sauerbrei. The FSP is discussed in the second major section of this paper. The FSP might or might not find a stronger transformation of X1 than simply $X1^3$ or FSP might reject X1 altogether.

For X6 all the chi-squares are insignificant, as expected.

TABLE 5: Results of %LOGIT_CONTINUOUS(example, Y, X1 X6)

Variable	Transform	b_{1D}	ChiSq for b_{1D}	Prob ChiSq
X1_p7	X**3	0.001	8.466	0.004
X1_p6	X**2	0.010	8.382	0.004
X1_p1	linear	0.089	7.823	0.005
X1_p5	X**0.5	0.369	7.342	0.007
X1_p8	log(X)	0.368	6.727	0.010
X1_p11	(X-p25)**2	0.042	6.568	0.010
X1_p4	X**-0.5	-1.410	5.990	0.014
X1_p3	X**-1	-1.281	5.158	0.023
X1_p2	X**-2	-1.688	3.380	0.066
X1_p10	(X-p75)**2	-0.021	1.581	0.209
X1_p9	(X-p50)**2	0.022	0.947	0.331
X6_p10	(X-p75)**2	-0.354	3.561	0.059
X6_p2	X**-2	-0.290	3.537	0.060
X6_p3	X**-1	-0.391	3.010	0.083
X6_p4	X**-0.5	-0.633	2.769	0.096
X6_p8	log(X)	0.254	2.545	0.111
X6_p5	X**0.5	0.405	2.338	0.126
X6_p1	linear	0.160	2.148	0.143
X6_p9	(X-p50)**2	-0.576	1.871	0.172
X6_p6	X**2	0.049	1.821	0.177
X6_p7	X**3	0.019	1.558	0.212
X6_p11	(X-p25)**2	0.137	0.492	0.483

TABLE 6: Distribution of Rate of Y=2 for Octile Ranks of X1

Octile ranks: X1	N Obs	Rate of Y=2
0	500	0.402
1	500	0.426
2	500	0.430
3	500	0.440
4	500	0.448
5	500	0.434
6	500	0.472
7	500	0.484

GUIDELINES FOR INTERPRETING THE RESULTS FROM %LOGIT_CONTINUOUS

In TABLE 5 the "Prob ChiSq" value should be viewed as a guide rather than a firm standard for significance. Due to by-chance over-fitting to the data by one or more of the 11 functions of X, the thresholds for the "Prob ChiSq" should be set at 0.01 (1%) and/or the "ChiSq for b_{1D}" might be set at 10.0 or more. By either criterion X6 is rejected while X1 will be further analyzed through the FSP.

THE FUNCTION SELECTION PROCEDURE (FSP) FOR A CONTINUOUS PREDICTOR X

BACKGROUND

In *Multivariate Model-building* by Royston and Sauerbrei (2008) a class of transformations of X called fractional polynomials (FP) are given. The fractional polynomial transformations first require that X be translated so that the values of X are positive. Then the fractional polynomial transformations of X are given by:

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

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

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

FP2 refers to the collection of functions formed by selection of two X^p. That is,

$$G(X,p_1,p_2) = \beta_0 + \beta_1 X^{p1} + \beta_2 X^{p2} \qquad p_1 \neq p_2$$
$$G(X,p_1,p_1) = \beta_0 + \beta_1 X^{p1} + \beta_2 X^{p1} \log(X) \qquad p_1 = p_2$$

FP2 produces curves with a variety of non-monotonic shapes as shown by Royston and Sauerbrei (2008 p. 76).

The Function Selection Procedure (FSP) is described by Royston and Sauerbrei (2008 p. 82) and a short history is given of its development. The following 3 steps for FSP are taken from documentation of their SAS implementation of FSP.[6] [7]

1. Perform a 4 d.f. test at the α level of the best-fitting second-degree FP (i.e. FP2) against the null model. If the test is not significant, drop X and stop, otherwise continue.
2. Perform a 3 d.f. test at the α level of the best-fitting second-degree FP against a straight line. If the test is not significant, stop (the final model is a straight line), otherwise continue.
3. Perform a 2 d.f. test at the α level of the best-fitting second-degree FP against the best-fitting first-degree FP (i.e. FP1). If the test is significant, the final model is the FP2, otherwise the FP1 is the final model.

A rationale for the degrees of freedom and hypothesis testing used in FSP is given by Royston and Sauerbrei (2008 p. 79). The FSP involves testing a total of 8 FP1 and 28 FP2 models.

Royston and Sauerbrei (2008 p. 267) give links to software versions for performing FSP including Stata, R, and SAS. I downloaded the SAS version **%mfp8** on 8/7/2014. It was written in SAS version 8. See notes in Appendix B which discuss a necessary change to the code before the running of %mfp8.

EXAMPLE: RUNNING %MFP8 ON X1 FROM "EXAMPLE" DATA SET

The FSP macro %mfp8 was run on predictor X1 from the "example" simulation data set. Due to transformations, such as LOG, that are only defined for positive values, a translation of X1 is required so that the translated X1 is positive.

Before using %mfp8 the user has the responsibility to translate X1 so that X1 is positive. I selected a translation so that the minimum of the translated X1 would be 1. This requires 2 passes of the data set: (1) Find the minimum of X1 across the "example" data set. This minimum equals -3.70327; (2) Apply the translation X1 = X1 + 1 + 3.70327 = X1 + 4.70327.

OBSERVATIONS ABOUT %MFP8

Before mentioning some minor issues with %mfp8, it must be acknowledged that this macro has very many powerful features and should be considered for usage by any predictive modeler.

- PROC LOGISTIC was run 47 times by %mfp8 . However, the run time was only a few seconds for running %mfp8 on predictor X1 from the "example" data set. This data set has 4,000 observations.
- The selection of the amount of translation (shifting to right) of X1 can affect the results of the FSP. If X1 were to be translated by, for example, 100 units, then a different FP2 solution is found.[8]
- My choice of translation was designed to make the minimum of X1, after translation, to equal 1. This minimum value of X avoids, at least on the left end, extreme values of the predictors. I don't know if there is a choice of translation that would optimize the log likelihood for the best FP1 or FP2 solutions.[9]

[6] Meier-Hirmer, Ortseifen, and Sauerbrei (2003). Multivariable Fractional Polynomials in SAS, http://portal.uni-freiburg.de/imbi/mfp, beschreibung.pdf in SAS downloads.
[7] There is a second algorithm called the "sequential model selection" in beschreibung.pdf which is not discussed in this paper.
[8] The FP2 solution in this case is p1 = -2 and p2 = -1. Deviance = 5482.654. Compare this to TABLE 7.
[9] Royston and Sauerbrei (2008). See p. 79 for a discussion of translation and p. 84 for a discussion of scaling.

RESULTS FROM RUNNING %MFP8 ON X1

Results are given in TABLE 7. The "Deviance" in TABLE 7 is another term for -2 Log Likelihood of the fitted model. Findings include:

- The best FP1 function is $X1^3$, in accord with the results of screening by %LOGIT_CONTINUOUS.
- The degrees of freedom used in the FSP are pivotal. The difference in deviance between FP2 and the null model is "Diffra2" = 5491.232 - 5482.564 = 8.668.
  - When significance is computed for X1 by FSP using 4 d.f., then pdiffdev = 6.995%, as given in TABLE 7.
  - If α is set at 10%, then the FSP stops at step 2 and the final model is linear, $\beta_0 + \beta_1$ X1.
  - If, on the other hand, α is set at 5%, then X1 is rejected altogether. This rejection at 5% occurs despite some promising results regarding X1 from running %LOGIT_CONTINUOUS.
- In contrast, when significance is computed by %LOGIT_CONTINUOUS for $X1^3$ (the best transformation of X1) using 1 d.f. the significance is 0.364%.[10] This achieved significance level meets our suggested guideline of 1% but does not exceed the Wald chi-square threshold of 10. (See TABLE 5.)
- We might have anticipated that the FP1 solution for X1, if not an FP2 solution, could be selected for the logistic model. Instead, the conservative FSP degrees-of-freedom adjustment has dropped X1 at the 5% level from further consideration.

TABLE 7: FSP Results from MFP8 for Predictor X1

MFP8: Variable -X1-
Best Functions for Different Degrees m

Function	m	p1	p2	deviance	diffra2	pdiffdev	TEST:
Null	-1	.	.	5491.232	8.668	0.06995	FP2 v. Null: 4 d.f.
Linear	0	.	.	5483.413	0.849	0.83763	FP2 v. Linear: 3 d.f.
First Degree FP1	1	3	.	5482.802	0.238	0.88783	FP2 v. FP1: 2 d.f.
Second Degree FP2	2	0.5	1	5482.564			

%FSP_LR - A SAS MACRO FOR FSP FOR LOGISTIC REGRESSION

Macro %FSP_LR runs PROC LOGISTIC 36 times and will find the optimal FSP solution.[11] The predictor X is pre-processed by %FSP_LR so that if X has minimum less than 1, it is translated to have minimum of 1. The three-step FSP tests are performed. The code is simpler (but does less) than %mfp8.

SUMMARY

This paper shows that a process of screening a multitude of continuous predictors by %LOGIT_CONTINUOUS can efficiently and effectively identify predictors to retain for further study. The FSP can then be focused on the surviving candidate predictors for either the selection of a final transformation or final elimination of the variable.

SAS MACROS DISCUSSED IN THIS PAPER

The SAS macros %LOGIT_CONTINUOUS and %FSP_LR would require about 4 pages each to list in this paper. Instead, these macros will be provided by the author upon request.

REFERENCES

Hosmer D., Lemeshow S., and Sturdivant R. (2013). *Applied Logistic Regression, 3rd Ed.,* John Wiley & Sons, New York.

Huberty C. and Olejnik S. (2006), *Applied MANOVA and Discriminant Analysis, 2nd Ed.*, John Wiley & Sons, Hoboken, N.J.

Lund B. (2013). Preparing Interaction Variables for Logistic Regression. *SCSUG2013, Proceedings*, South Central SAS Users Group, Inc.

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

Press S. J. and Wilson S. (1978). Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association*, **73**, pp. 699-705.

Royston P. and Sauerbrei W. (2008). *Multivariate Model-building,* John Wiley & Sons, Ltd, West Sussex, England.

[10] Chi-square to enter the model in FORWARD or STEPWISE selection.

[11] See Appendix C for a discussion of attempts to reduce the number of PROC LOGISTIC runs that are needed to implement FSP.

CONTACT INFORMATION

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

Bruce Lund
Magnify Analytic Solutions, A Division of Marketing Associates, LLC
777 Woodward Ave, Suite 500,
Detroit, MI, 48226
blund@marketingassociates.com
blund_data@mi.rr.com

APPENDIX A: TWO-GROUP DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

In discriminant analysis it is assumed there are "p" populations G1, …, Gp. The members of these p populations are characterized by predictor variables X_1 to X_K. The purpose of discriminant analysis is to develop for each population a linear combination of the predictors called the classification function (CF). Then an observation (x_1 …, x_K) is assigned by the CF's to the population which has the largest CF value.

The development and application of the CF's follows two steps:
(i) The "p" CF's are fitted using observations from G1, …, Gp (i.e. observations where population membership is known).
(ii) The CF's are applied to observations of X_1 to X_K where population membership is not known. An observation is assigned to the population with the largest CF value.

To support statistic inference, it is assumed the X_1 … X_K follow a multivariate normal distribution.[12]

THE CASE WHERE p = 2, K = 1, AND $\sigma_1 = \sigma_2$

The simplest case of discriminant analysis is when there are 2 populations, one predictor X, and the distributions of X for the two populations have a univariate normal distribution with common standard deviation σ but differing means μ_1 and μ_2.

The formula for the distribution of X for population j = 1 or 2 is given by:

$$P(X=x \mid j) = (2\pi\sigma^2)^{-1/2} \exp(-0.5 ((x - \mu_j) / \sigma)^2) \quad \text{… where } \mu_j \text{ is the mean for X for population j.}$$

DEVELOPMENT OF THE CLASSIFICATION FUNCTIONS

Suppose a random sample for populations j = 1, 2 is taken and n_1 is the size from population 1 and n_2 is the size from population 2. The base-rate population probability of j is denoted by P(j). These probabilities are estimated by $q_j = n_j/(n_1 + n_2)$.

The probability that an observation with value X = x belongs to population j = 1, 2 is the conditional probability expressed by:

$$P(j \mid X=x) \text{ for } j = 1, 2.$$

These are the probabilities which are needed to classify an observation x into a population. The classification rule is to assign x to j=1 if:

$$P(1 \mid X=x) > P(2 \mid X=x) \quad \text{… (A)}$$

Otherwise assign x to j=2.

The P(j | X=x) probabilities can be calculated from the P(X=x | j) distributions using Bayes theorem as shown:

$$P(j \mid X=x) = P(X=x \mid j) P(j) / P(x) \quad \text{… (B)}$$

[12] See Huberty and Olejnik (2006, chapter 13) for discussion.

Substituting (B) into (A) gives:

$$P(X=x \mid 1)\, P(1) / P(x) > P(X=x \mid 2)\, P(2) / P(x) \quad \ldots (C)$$

The P(x) will cancel when forming the classification rule (D) and there is no need to evaluate them. Equation (C) simplifies by cancelling the P(x) as well as the common factors in the normal distributions and then taking logarithms to produce:

$$-0.5 * ((x - \mu_1)/\sigma)^2 + \log(q_1) > -0.5 * ((x - \mu_2)/\sigma)^2 + \log(q_2) \ldots (D)$$

The classification function CF_j for j = 1, 2 is:

$$CF_j = -0.5 * ((x - \mu_j)/\sigma)^2 + \log(q_j) \ldots (E)$$

ESTIMATING CF FROM THE SAMPLES

The samples from populations 1 and 2 are used to estimate the parameters μ_j and σ. The sample means $\bar{x}_j$ estimate μ_j. The pooled variance S_p^2 estimates σ^2 where S_j^2 is the sample variance for sample j and

$$S_p^2 = \{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\}/(n_1 - 1 + n_2 - 1)$$

The sample CF is used in the sample based classification rule:

$$-0.5 * ((x - \bar{x}_1)/ S_p)^2 + \log(q_1) > -0.5 * ((x - \bar{x}_2)/ S_p)^2 + \log(q_2) \ldots (F)$$

THE CLASSIFICATION FUNCTIONS AND LOG-ODDS-RATIO

Returning to equation (A), the odds ratio of membership in j=1 versus j=2 is given by:

$$P(1 \mid X=x) / P(2 \mid X=x) = (P(X=x \mid 1)\, P(1) / P(x)) / (P(X=x \mid 2)\, P(2) / P(x)) \quad \ldots (G)$$

Logarithms are taken of equation G, and P(x) is cancelled to give equation (H):

$$\text{Log } (P(1 \mid X=x) / P(2 \mid X=x)) = \text{Log}(P(X=x \mid 1)\, P(1)) - \text{Log}(/ P(X=x \mid 2)\, P(2)) \ldots (H)$$

Using equation (D), equation (H) becomes:

$$\text{Log Odds-Ratio} = -0.5(-2x(\mu_1 - \mu_2) + \mu_1^2 - \mu_2^2) / \sigma^2 + \log(q1/q2) \ldots (I)$$

Equation (I) shows that the Log-Odds-Ratio from 2-group discriminant analysis is a linear function of x

$$\text{Log Odds-Ratio} = \beta_0 + \beta_1 x \quad \ldots (J)$$

where $\beta_0 = -(\mu_1^2 - \mu_2^2) / 2\sigma^2 + \log(q1/q2)$ and $\beta_1 = (\mu_1 - \mu_2) / \sigma^2$

When fitting the 2-group discriminant analysis model to the sample, the estimates b_{0D}, b_{1D} for β_0, β_1 are found by replacing μ_1, μ_2, σ with $\bar{x}_1$, $\bar{x}_2$, and S_p in equation (J)

$$b_{0D} = -(\bar{x}_1^2 + \bar{x}_2^2) / 2S_p^2 + \log(q1/q2) \text{ and } b_{1D} = (\bar{x}_1 - \bar{x}_2) / S_p^2 \ldots (K)$$

The "D" is added as a subscript to indicate that the method of fitting is by discriminant analysis.

If we make the assumption that $(1/S_p)$ is a constant, then, as shown by (L), the expression for b_{1D} is a linear transformation of a t-statistic with $n_1 + n_2 - 2$ d.f. The term K is the constant $(\mu_1 - \mu_2) / S_p^2$.

$$b_{1D} = \{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\} / S_p^2 + (\mu_1 - \mu_2) / S_p^2 = t\ (1/S_p)\ \text{sqrt}(1/n_1 + 1/n_2) + K \ldots (L)$$

The appropriateness of the assumption of constant S_p is justified by having large sample sizes n_1 and n_2.

CONNECTION WITH LOGISTIC REGRESSION

The logistic regression model with predictor x also is formulated as:

$$\text{Log Odds-Ratio} = \beta_0 + \beta_1 x.$$

In the case of logistic regression the parameter estimates b_{0L}, b_{1L} are fitted by maximum likelihood.

The "L" is added as a subscript to indicate that the method of fitting is Logistic Regression with maximum likelihood.

CONNECTING b_{1D} AND b_{1L}

Under the assumptions of Appendix A, b_{1D} is essentially an unbiased estimator of β_1. The actual expectation is given by $E(b_{1D}) = ((n_1 + n_2 - 2)/ (n_1 + n_2 - 4))\, \beta_1$. Meanwhile, b_{1L} is asymptotically an unbiased estimator of β_1.

The Wald chi-square for b_{1L} gives the significance for rejecting the null that $\beta_1 = 0$.

For discriminant analysis, the null hypothesis that $\beta_1 = 0$ makes $(\mu_1 - \mu_2) = 0$. The hypothesis H0: $\beta_1 = 0$ vs. H1: $\beta_1 \neq 0$ at significance "α" is tested by finding a critical value C so that the test statistic $T = t\ (1/S_p)\ \text{sqrt}(1/n_1 + 1/n_2)$ satisfies $P(|T| > C) = \alpha$. This test can be changed to finding a critical C' so that $P(|t| > C') = \alpha$ where $C' = C\ /\ ((1/S_p)\ \text{sqrt}(1/n_1 + 1/n_2))$. But then C' is the t-statistic value $t_{\alpha/2}$.

For large $n_1 + n_2$ the square of the t-statistic is essentially a chi-square with 1 d.f. The test above can be rephrased in terms of a chi-square.

Finally, Hosmer, Lemeshow, and Sturdivant (2013, p 91) give the opinion that in cases where the predictor X has approximately a normal distribution for both of the groups determined by values of Y, the t-test, discussed above, is a good guide for screening a predictor for logistic regression.

APPENDIX B: NECESSARY CHANGES TO %MFP8

A. The "%then" must be removed in the sub-macro "fpmodels"

```
%macro fpmodels(model,y,x,pref,base,m,stvars);

   %local i j k;

    %if %upcase(&model)=N %then
      %do;
         %let procname=REG;
         %let procopt=SSE OUTSEB;
         %let modelpar=;
         %let modelopt=;
       %end;
    %else %if %upcase(&model)=L %then
       %do;
         %let procname=LOGISTIC;
         %let procopt=DESCENDING COVOUT;
         %let modelpar=;
         %let modelopt=;
       %end;
      %else /*%then*/
       %do;
         %let procname=PHREG;
         %let procopt=COVOUT;
         %let modelpar=*&censvar(&censval);
         %let modelopt= / Ties=&ties;
       %end;
```

B. For purposes of running in Windows %mfp8 must be modified by replacing "/" with "\" as shown in the statements below:

```
%include "&MacPath.\boxtid.sas";
%include "&MacPath.\xtop.sas";
%include "&MacPath.\xvars.sas";
%include "&MacPath.\fpmodels.sas";
%include "&MacPath.\datasave.sas";
%include "&MacPath.\exlabb.sas";
%include "&MacPath.\exinc.sas";
%include "&MacPath.\labs.sas";
%include "&MacPath.\brename.sas";
%include "&MacPath.\funcfm.sas";
```

Appendix C: ATTEMPTS TO REDUCE THE NUMBER OF PROC LOGISTIC RUNS IN FSP

A. The number of PROC LOGISTIC's is reduced to 8 by running the code below.

PROC LOGISTIC; MODEL Y = Var<k> / SELECTION = FORWARD INCLUDE=**1** START=**1** STOP=**2** SLE=**1**;

The Var<k>'s are defined by the table:

Var1=	X^{-2}	X^{-1}	$X^{-.5}$	$X^{.5}$	X	X^2	X^3	Log(X)	X^{-2} Log(X)
Var2=		X^{-1}	$X^{-.5}$	$X^{.5}$	X	X^2	X^3	Log(X)	X^{-1} Log(X)
Var3=			$X^{-.5}$	$X^{.5}$	X	X^2	X^3	Log(X)	$X^{-.5}$ Log(X)
Var4=				$X^{.5}$	X	X^2	X^3	Log(X)	$X^{.5}$ Log(X)
Var5=					X	X^2	X^3	Log(X)	X Log(X)
Var6=						X^2	X^3	Log(X)	X^2 Log(X)
Var7=							X^3	Log(X)	X^3 Log(X)
Var8=								Log(X)	Log(X) Log(X)

All possible FP2 pairs have a chance to be selected but the selection of the second variable in a pair, to accompany the first variable, forced in by INCLUDE=1, is based on best Wald chi-square. Examples can be found that show that this selection may not find the FP2 pair with the best log likelihood.

B. The number of PROC LOGISTIC's might be reduced by running PROC LOGISTIC with SELECTION = SCORE. For example, one would hope that the code below finds the best FP1 based on the Score Statistic (not Log Likelihood):

PROC LOGISTIC; MODEL Y = X^{-2} X^{-1} $X^{-.5}$ $X^{.5}$ X X^2 X^3 Log(X) / SELCTION=SCORE BEST=1 START=1 STOP=1;

This fails because PROC LOGISTIC will detect a near collinear relationship among X^{-2} X^{-1} $X^{-.5}$ $X^{.5}$ X X^2 X^3 Log(X) and will remove one or more variables from its processing.