Paper 2687-2015

Selection and Transformation of Continuous Predictors for Logistic Regression

Bruce Lund, Magnify Analytic Solutions, a Division of Marketing Associates, LLC

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) The FSP (function selection procedure) and a SAS implementation of FSP are discussed. The FSP tests all transformations from among a class of "FSP transformations" and finds the one with maximum likelihood when fitting the binary target. Royston and Sauerbrei in a 2008 book have popularized the FSP.

INTRODUCTION

The setting for this discussion is direct marketing, credit scoring, or other applications of binary logistic regression where sample sizes for model building are large 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) including finding interactions and dealing with collinearity

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]

The meaning of "continuous" for a predictor is subjective. One definition might be "when it is hard to do binning". This definition applies when many of the values of X occur only on one observation.

The use of binning of continuous predictors followed by WOE coding (often performed 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.

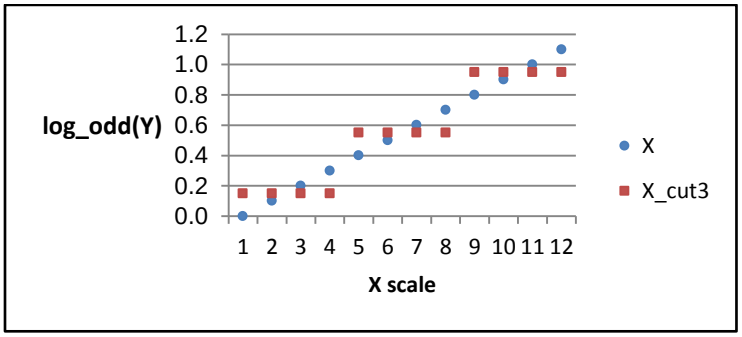

Figure 1 Hypothetical 3-cutpoint binning of a predictor X

[1] A SAS macro for binning and WOE transformations is given by Lund and Brotherton (2013).

In this paper a statistical procedure and SAS macro %LR_SCREENER 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. 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.

Once candidate predictors have passed screening by %LR_SCREENER, a final transformation for these predictors can be determined by the Function Selection Procedure (FSP) as described in Royston and Sauerbrei in *Multivariate Model-building* (2008). A SAS macro %MFP8 for FSP is provided via a web-site referenced in *Multivariate Model-building* (2008). A drawback of this SAS macro is that it processes one-predictor at the time and runs PROC LOGISTIC 47 times in doing this processing.[2] In this paper two SAS macros %FSP_8LR and %FSP_36LR are discussed which process multiple candidate predictors using the FSP in a more efficient way. FSP and its SAS implementations are discussed in Part II of this paper.

PART I: 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 hundreds of predictors in a single run of PROC TTEST. In this paper a SAS macro to implement this procedure, called %LR_SCREENER, is discussed. The macro %LR_SCREENER takes advantage of a connection between 2-group discriminant analysis, a t-test, and logistic regression. This connection is developed in Appendix A and only key results are presented below. For more discussion see Press and Wilson (1978).

A CONNECTION BETWEEN 2-GROUP DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION

Let X be a predictor of a binary target Y where Y has values 1 and 2. The logistic model relating X to Y is given by equation (A):

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

The parameters β_0 and β_1 are estimated from an (x,y) data set by maximizing the likelihood function but this estimation does not produce a closed-end solution. If it is assumed that X has a univariate normal distribution for each of the two groups {X: Y=1} and {X: Y=2} with common standard deviation σ but differing means μ_1 and μ_2, then the 2-group discriminant model also leads to equation (A). The discriminant model gives an alternative method to estimate the parameters β_0 and β_1 which does provide a closed-end solution.

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 (B):

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

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 the method of fitting is by discriminant analysis. Also, "L" is added as a subscript to "b_1", giving 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 samples, b_{1D} will be close to b_{1L}. For discussion of relative efficiency of these estimators see Efron (1975).

(2) The coefficient b_{1D} can be used to test $\beta_1 = 0$ vs. $\beta_1 \neq 0$. Under the null hypothesis of $\beta_1 = 0$, the coefficient b_{1D} is related to a t-statistic with $n_1 + n_2 - 2$ d.f. via this factorization: [3]

$$b_{1D} = t \ (1/S_p) \ \text{sqrt}(1/n_1 + 1/n_2)$$

[2] But %MFP8 is a powerful macro that also implements FSP for ordinary least squares and Cox regression.
[3] For large n_1 and n_2 the S_p is regarded as a constant.

The square of this t-statistic "$\mathbf{t^2}$" approximates a chi-square with 1 d.f.[4] The null hypothesis of $\beta_1 = 0$ is rejected if t^2 is significant. The logistic regression Wald chi-square statistic for b_{1L} and the t^2 from discriminant analysis will have roughly the same values.

USING t^2 AS A SCREENER FOR PREDICTORS FOR LOGISTIC REGRESSION

Without the normality assumptions on X, there is no guarantee that b_{1D} and b_{1L} will be close in value. But for the purpose of screening predictor variables for logistic regression the closeness of b_{1D} and b_{1L} is not a requirement. Instead, the requirement is that t^2 and the Wald chi-square from logistic regression concur in their measurement of the significance of X. Concurrence can occur without close agreement of b_{1D} and b_{1L}. Four scenarios are depicted in Table 1. The t^2 will be successful as a screener if only few examples fall into cells "B" or "C". Especially troubling would be examples in cell "C", a false negative.

Table 1 Significance by t^2 vs. Wald $\chi 2$

	Wald $\chi 2$ significant	Wald $\chi 2$ not significant
t^2 significant	A (correct)	B (false positive)
t^2 not significant	C (false negative)	D (correct)

SIMULATIONS

Three predictors X1, X2, and X3 were created. They provide three examples of the concurrence of t^2 and Wald chi-square in measuring the significance of a predictor:

EXAMPLE 1: PREDICTOR X1

```
data Example1;
do i = 1 to 4000;
   Y = (ranuni(12345) < .45) ;
   if Y = 1 then X1 = rannor(12345) + 0.1; /* normal with mean=0.1 stddev=1 */
    else if Y = 0 then X1 = rannor(12345); /* normal with mean=0 stddev=1 */
    output;
   end;
run;
```

The distributions of X1 for the two groups meet the assumptions of normality with equal standard deviations (=1). The coefficient estimates b_{1D} and b_{1L} of β_1, as well as t^2 and Wald chi-square for X1 should be nearly identical. See Figure 2.

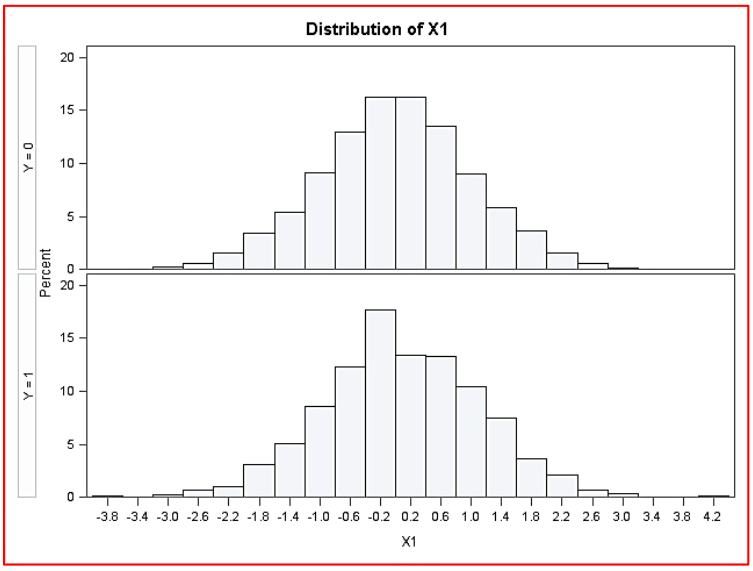

Figure 2 Histograms for Variable X1

[4] Calculation of "$\mathbf{t^2}$" by Satterthwaite (instead of pooled) variance formula gives very similar values for examples to follow in this paper. The pooled variance is used because equal population variances are assumed when deriving the formula for $\mathbf{t^2}$.

Table 2a shows coefficient estimates b_{1D} and b_{1L} to be very similar.

Table 2a Comparison of Coefficients from Discriminant and Logistic

Variable	b_{1D}	b_{1L}
X1	0.07420	0.07422

Table 2b and Table 2c show the near equality of the probabilities of the chi-squares

Table 2b Chi-Square Value for Coefficients from Discriminant Analysis

Variable	t Value	d.f.	ChiSq for b_{1D}	Prob ChiSq
X1	2.34	3998	5.467	0.0194

Table 2c Chi-Square Value from Logistic

Variable	ChiSq for b_{1L}	Prob ChiSq
X1	5.455	0.0195

The SAS code that produced b_{1D} and the t-test values is shown below but SAS code for producing the logistic results is not included.

```
data Example1;
do i = 1 to 4000;
  Y = (ranuni(12345) < .45);
  /* Normal distribution test */
  if Y = 1 then X1 = rannor(12345) + 0.1;
   else if Y = 0 then X1 = rannor(12345);
   output;
  end;
run;
ods listing;
ods output Statistics = TS;
ods output TTests = TT;
ods exclude ConfLimits;  /* suppress unneeded print */
ods exclude Equality;    /* suppress unneeded print */
ods exclude EquivLimits; /* suppress unneeded print */
ods exclude EquivTests;  /* suppress unneeded print */
proc ttest data=Example1 plots=none;
  class Y; var X1;
data TT; set TT;
  tValue = -tValue;  /* minus sign to model Y=1 as the response */
  chisq_D = tValue**2;
  label Probt = "Prob ChiSq"; /* prob. for t-value = prob. for chi-sq. */
proc print data = TT label;
var Variable tValue DF Probt chisq_D;
where method = "Pooled";
run;
data TS_2; set TS; by variable;
retain mean1 mean2 n1 n2;
keep variable n1 n2 mean1 mean2 stddev b1D;
if first.variable then do; mean1 = mean; n1 = n; end;
if first.variable + last.variable = 0 then do; mean2 = mean; n2 = n; end;
if last.variable /* this is the last row */
then
do;
  b1D = -(mean1 - mean2)/(stddev **2); /* "-" to model Y=1 as response */
  output;
  end;
proc print data = TS_2 noobs; var Variable b1D;
run;
```

EXAMPLE 2: PREDICTOR X2

```
data Example2;
do i = 1 to 500;
   X2 = ranuni(12341);
   Y = floor(X2 + ranuni(12341)); /* Y is more often equal to 1 when X2 is
large */
   output;
   end;
run;
```

The distributions of X2 for the two groups strongly fail to meet the assumptions of normality as seen from the distributional histograms in Figure 3.

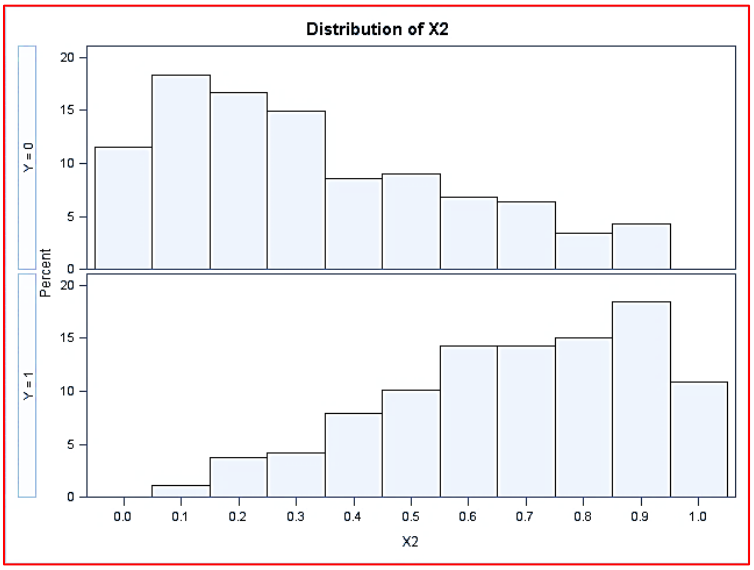

Figure 3 Histograms for Variable X2

The coefficient estimates b_{1D} and b_{1L} of β1 are not close as seen in Table 3a.

Table 3a Comparison of Coefficients from Discriminant and Logistic

Variable	b_{1D}	b_{1L}
X2	6.2743	5.3491

Table 3b and Table 3c show a divergence in value between the chi-squares, but both chi-squares are highly significant. As a screener of a predictor for a logistic model, both the t^2 from discriminant analysis and the Wald chi-square from logistic are highly significant.

Table 3b Chi-Square Value for Coefficients from Discriminant Analysis

Variable	t Value	d.f.	ChiSq for b_{1D}	Prob ChiSq
X2	16.60	498	275.397	<.0001

Table 3c Chi-Square Value from Logistic

Variable	ChiSq for b_{1L}	Prob ChiSq
X2	132.515	<.0001

EXAMPLE 3: PREDICTOR X3

```
data Example3;
do i = 1 to 1000;
   X3 = ranuni(12345);  /* Uniform on [0, 1] */
   Y = (ranuni(12345) < .2);
   output;
```

```
  end;
run;
```

Predictor X3 is uniform [0, 1] for both Y=0 or Y=1. The distributions of X3 for the two groups fail to meet the assumptions of normality as seen from the distributional histograms in Figure 4.

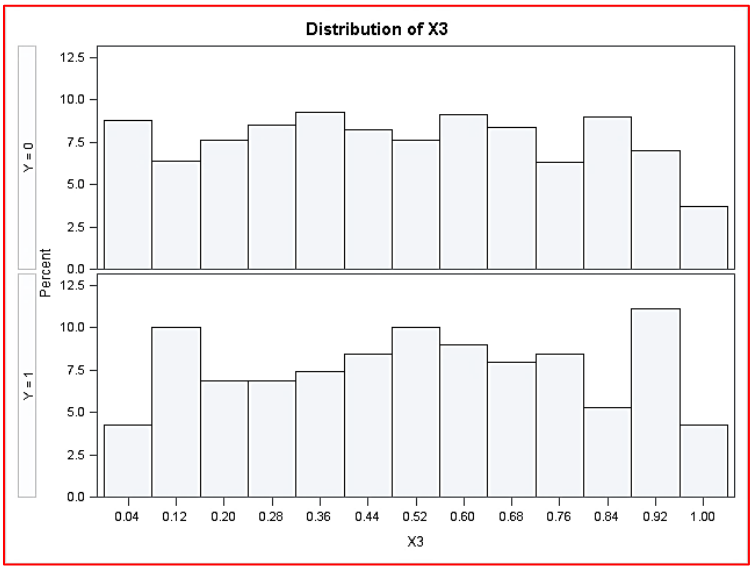

Figure 4 Histograms for Variable X3

The two coefficient estimates b_{1D} and b_{1L} of β1 are in close agreement as seen in Table 4a.

Table 4a Comparison of Coefficients from Discriminant and Logistic

Variable	b_{1D}	b_{1L}
X3	0.35467	0.35550

Table 4b and Table 4c show close agreement between the chi-squares. As a screener of X3 for a logistic model, both the t^2 from discriminant analysis and the Wald chi-square from logistic are strongly insignificant.

Table 4b Chi-Square Value for Coefficients from Discriminant Analysis

Variable	t Value	d.f.	ChiSq for b_{1D}	Prob ChiSq
X3	1.25	998	1.554	0.2129

Table 4c Chi-Square Value from Logistic

Variable	ChiSq for b_{1L}	Prob ChiSq
X3	1.552	0.2129

EXAMPLE 4: A COUNTER-EXAMPLE

For predictor X4 in the data set below the t^2 is significant but the Wald chi-square is not significant at 5%.

```
data Example4;
  Y = 0;
  do i = 1 to 50;
    X4 = rannor(12345);
    output;
    end;
  do i = 1 to 2;
    X4 = 50;
    output;
    end;
  Y = 1;
  do i = 1 to 500;
```

```
X4 = rannor(12345);
output;
end;
```

Predictor X4 is standard normal for Y=1 and X4 is also standard normal for Y=0 with the exception of two outliers at X4=50. Table 5b and Table 5c show that chi-squares from discriminant analysis and from logistic disagree. The discriminant analysis chi-square is significant at < 0.01% while logistic chi-square fails a 5% significance test with probability value of 8.90%.

Table 5a Comparison of Coefficients from Discriminant and Logistic

Variable	b_{1D}	b_{1L}
X4	-0.19914	-0.10782

Table 5b Chi-Square Value for Coefficients from Discriminant Analysis

Variable	t Value	d.f.	ChiSq for b_{1D}	Prob ChiSq
X4	-4.25	550	18.0860	<.0001

Table 5c Chi-Square Value from Logistic

Variable	ChiSq for b_{1L}	Prob ChiSq
X4	2.892	0.0890

The relationship between X4 and Y would be highly unusual in an applied setting. It is likely that a modeler would research the two observations at X4=50 and correct them or eliminate them.

CONCLUSIONS

Examples 1-3 from the section on Simulations support the conclusion that the t-test approach can differentiate between significant and insignificant predictors for a logistic model.

Support is also given by comments in *Applied Logistic Regression* by Hosmer, Lemeshow, and Sturdivant (2013 p. 91). These authors agree that if X has an approximately normal distribution for the two groups determined by values of Y, then the t-test is a good guide for screening a predictor for logistic regression.

The predictor X4 gives an example of a false positive (a "B" in Table 1) that was constructed by a trial and error process. I was unable to construct an example of a false negative (a "C" in Table 1) but, by no means, has this been ruled out. More simulations are needed to understand the risk that such examples would present to the operation of the macro %LR_SCREENER to be discussed in the next section.

%LR_SCREENER: FOR SCREENING HUNDREDS OF LOGISTIC PREDICTORS

The SAS macro %LR_SCREENER 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.[5]
   b. 10 transformations of the predictor are computed (such as LOG, X^2, and others).
3. A PROC TTEST to compute t-statistics whose squares approximate the Wald chi-squares from logistic regression.

The original X and the 10 transformations are:

- 8 monotonic: The "fractional polynomials" X^p where p belongs to S = {-2, -1, -0.5, 0, 1, 0.5, 2, 3} and where "0" denotes log(x). This list includes the original X.[6]

[5] In practice most continuous predictors in direct marketing have non-negative values domains (counts, distance, dollars, percents, time, quantity). Translation (except away from zero) is generally not needed.
[6] These transformations are motivated by the FSP (function selection procedure) of Royston and Sauerbrei (2008).

- 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.

In practice, the relationship between a predictor X and the log-odds of Y [7] is very often either monotonic or roughly quadratic (with a single maximum or minimum). Consequently, one (or more) of the 11 transformations is likely to have an approximate linear relationship to log-odds of Y, if, indeed, the predictor has predictive power.

The parameters for %LR_SCREENER are:

DATASET: The input data set name.
Y: A numeric variable with two values. The larger value is the response that is modeled.
INPUT: A list of numeric predictor variables delineated by space. The use of the dash between variables (e.g. X1 - X6) is supported. A predictor may have missing values.

EXAMPLE 5: %LR_SCREENER IS RUN ON X1 FROM EXAMPLE1.

The macro call is %LR_SCREENER (example1, Y, X1). Before the results of Table 6 are produced the predictor X1 is translated by 1 + 3.71564 so that min(X1) = 1.

Table 6 Results of %LR_SCREENER (example1, Y, X1)

Variable	Transform	b_{1D}	ChiSq for b_{1D}	Prob ChiSq
X1_p7	x**3	0.0011	6.793	0.00919
X1_p11	(x-p25)**2	0.0404	6.424	0.01130
X1_p6	x**2	0.0082	6.268	0.01233
X1_p1	linear	0.0742	5.467	0.01943
X1_p5	x**0.5	0.3019	4.945	0.02622
X1_p8	log(x)	0.2954	4.325	0.03761
X1_p4	x**-0.5	-1.0970	3.593	0.05811
X1_p3	x**-1	-0.9437	2.739	0.09798
X1_p9	(x-p50)**2	0.0306	1.958	0.16179
X1_p2	x**-2	-0.8900	0.881	0.34809
X1_p10	(x-p75)**2	-0.0095	0.331	0.56510

As shown in Table 6 the best transformations of X1 are $X1^3$, $(X1\text{-p25})^2$, and $X1^2$. These have similar chi-square values. The linear transform of X1 comes in at fourth place.

GUIDELINES FOR INTERPRETING THE RESULTS FROM %LR_SCREENER

Due to by-chance over-fitting to the data by one or more of the 11 transformations of X, the guideline threshold for significance from the "Prob ChiSq" should be set conservatively. Here is one proposal for the threshold value of the best transformation:

- "Prob Chi-Square" of the best transformation of X is less than 0.01 (1.00%)

With this guideline, X1 is a borderline case.

TRANSLATIONS AND TRANSFORMATIONS OF X

Four transformations among the 11 are unaffected by translations: X, $(X\text{-median})^2$, $(X\text{-p25})^2$, $(X\text{-p75})^2$. But could a "good" X be found that would otherwise be missed if translations were added to %LR_SCREENER to form translation-transformation combinations for the other 7?

The SAS code below creates three predictors X6, X6_1, and X6_80. A translation of 1 unit was added to X6 to form X6_1 and a translation of 80 was added to form X6_80.Then %LR_SCREENER was run on X6, X6_1, and X6_80.

EXAMPLE 6: EFFECT OF TRANSLATIONS

```
data Example6;
do i = 1 to 500;
```

[7] Here, "Y" should be taken to mean the average of Y over a small interval of X.

```
  X6 = ranuni(12341) + 1;
  Y = floor(X6 + ranuni(12341));
  X6_1 = X6 + 1;
  X6_80 = X6 + 80;
  output;
  end;

%LR_SCREENER (Example6, Y, X6 X6_1 X6_80);
```

The best transform of X6 is $X^{-0.5}$ while the best transform of X6_1 is X^{-1}. In this case the translation-transformation combination provided a very slightly better result than the transformation alone. Could a more extreme translation produce a translation-transformation that was even better? A translation of 80 units is used to create X6_80 but produces a lower chi-square of 275.88. See Table 7.

Table 7 Effect of a Translation Prior to running %LR_SCREENER

Variable	Best Transform	b_{1D}	ChiSq for b_{1D}	Prob ChiSq
X6_p4	x**-0.5	-22.270	281.24	<.0001
X6_1_p3	x**-1	-38.747	281.47	<.0001
X6_80_p2	x**-2	-1700484	275.88	<.0001

In general, the significance level of the maximum value of the chi-squares for X and its 10 transformations is not greatly affected by translations. While a translation may result in a different transform being selected (as part of a translation-transformation), the effect on significance will not be important.

CONCLUSIONS

This paper shows that a process of screening a multitude of continuous predictors by %LR_SCREENER can efficiently and effectively identify predictors to retain for further study. But how should a final transformation be found for those predictors that pass the screening? This is the topic of Part II which follows below.

PART II: 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), is presented. 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 where p is taken from S = {-2, -1, -0.5, 0, 0.5, 1, 2, 3} and 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). Such an example is given in Figure 5.

The Function Selection Procedure (FSP) is described by Royston and Sauerbrei (2008 p. 82) and a short history is given of its development. Royston and Sauerbrei (2008 p. 267) give links to software versions for performing FSP including Stata, R, and SAS.[8] The SAS version, a macro named %MFP8, was current as of 12/29/2014. It was written in SAS version 8.[9]

See notes in Appendix B which give a necessary change to the code in order to run %MFP8.

[8] Techniques for final model fitting (stepwise, best subsets, etc.) are not discussed here. But Royston & Sauerbrei (2008 chapter 6) present the "MFP Algorithm" which describes an algorithm for applying FSP to the fitting of the final model.
[9] Meier-Hirmer, Ortseifen, and Sauerbrei (2003). Downloaded from http://portal.uni-freiburg.de/imbi/mfp

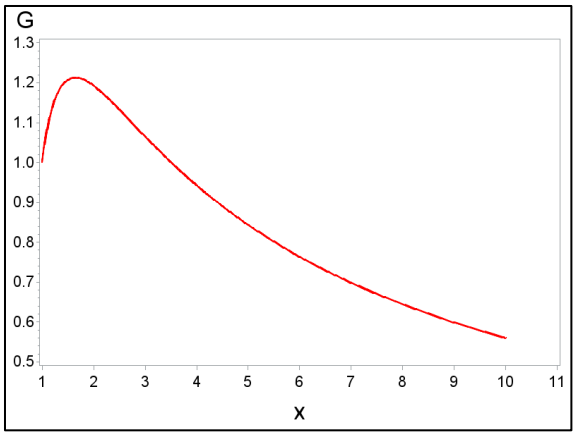

Figure 5 Graph of G(X,-1,-1) = X^{-1} + 2 X^{-1} log(X)

FSP: SEARCHING FOR BEST TRANSFORMATIONS AND SIGNIFICANCE TESTING

Searching for best transformations: First, the function within FP1 and the function within FP2 having the maximum likelihoods are found by an exhaustive search in which 44 Logistic Models are run. Overall, PROC LOGISTIC is run 47 times by %MFP8.

Performing Significance Testing: Second, significance testing is performed. The FSP significance testing follows these 3 steps.[10]

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.

The test-statistic for these three tests is the difference of deviances[11] as shown below:

Test-Statistic = (-2 Log Likelihood $_{\text{restricted model}}$) - (-2 Log Likelihood $_{\text{full model}}$)

For large samples, the Test-Statistic is approximately a chi-square.

The rationale for the degrees of freedom (4, 3, 2) used in the 3-step hypothesis tests of FSP is discussed by Royston and Sauerbrei (2008 p. 79).

EXAMPLE: RUNNING %MFP8 ON X2 FROM "EXAMPLE 2" DATA SET

The FSP macro %MFP8 was run on predictor X2 from the Example 2 data set. A translation was selected so that the minimum of the translated X2 would be 1 via this statement: X2 = X2 +1.0 - 0.001398486

Table 8 FSP Results from Running %MFP8 on Predictor X2

MFP8: Variable –X2-
Best Functions for Different Degrees m

Function	m	p1	p2	deviance	diffra2	pdiffdev	TEST:
Null	-1	.	.	691.098	208.202	0.00000	FP2 v. Null: 4 d.f.
Linear	0	.	.	489.001	6.106	0.10658	FP2 v. Linear: 3 d.f.
First Degree	1	-2	.	483.540	0.645	0.72451	FP2 v. FP1: 2 d.f.
Second Degree	2	-2	3	482.896	0.000	1.00000	

For **Step 1** of significance testing the Test-Statistic is $\chi 2$ = 691.098 - 482.896 = 208.202

[10] Meier-Hirmer, Ortseifen, and Sauerbrei (2003). Multivariable Fractional Polynomials in SAS, http://portal.uni-freiburg.de/imbi/mfp. See beschreibung.pdf in SAS downloads.
[11] The deviance is the -2 Log Likelihood value of a logistic model.

Predictor X2 passes Step 1 as seen from:

$$1- \text{Prob}(\chi2(208.202,4)) < 0.0001$$

For **Step 2** of significance testing the Test-Statistic is $\chi2 = 489.001 - 482.896 = 6.106$

Predictor X2 fails Step 2 as seen from:

$$1- \text{Prob}(\chi2(6.106,3)) = 0.10658 \text{ which is greater than } 0.05 \ (5\%)$$

On this basis FSP selects "Linear" as the transformation for X2.

VISUALIZATION OF THE LINEAR, FP1, AND FP2 SOLUTIONS

The log-odds(Y) is plotted against the Linear, FP1, and FP2 solutions at the median value of X2 across 8 equal-sized ranks of X2 values. The log-odds(Y) "open-circles" give the mean of Y within the X2 rank.

Visually, FP1 and FP2 give better fits than Linear. Since FP1 narrowly failed Step 2 at significance 10%, the FP1 solution has an appeal.

See Figure 6.

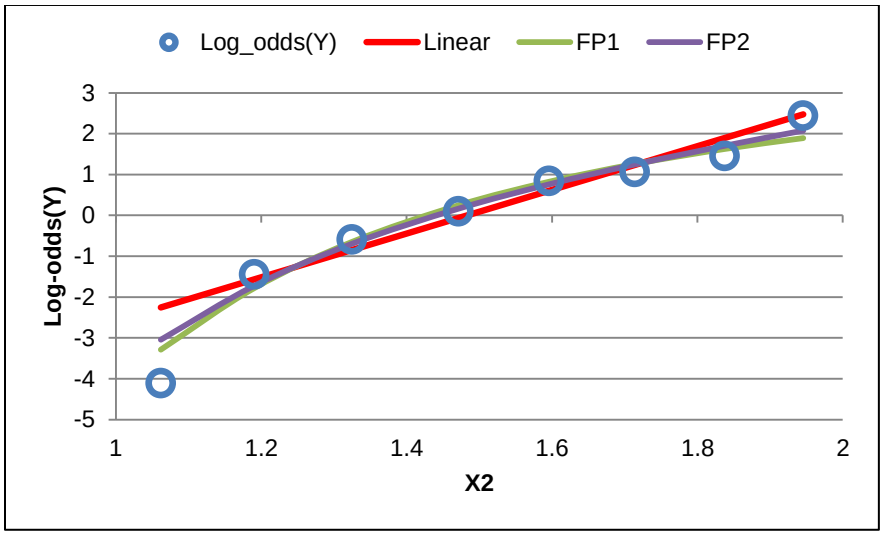

Figure 6 Linear = -7.9343 + 5.3491 X2; FP1 = 4.0901 - 8.3200 $X2^{-2}$; FP2 = 2.9666 - 6.9522 $X2^{-2}$ + 0.1284 $X2^{3}$

WHAT IF THE FP2 SOLUTION WAS SELECTED?

Recall that FP2 = 2.9666 - 6.9522 * $X2^{-2}$ + 0.1284 * $X2^{3}$. What transform is used in the fitting of the final logistic model? Choices:

1. The FP2 equation
2. $X2^{-2}$ and $X2^{3}$ as individual predictors

Discussion:

1. If the FP2 equation gives an intuitive relationship between predictor X2 and target Y, then this relationship is preserved by entering FP2 into the model.
2. Otherwise, entering $X2^{-2}$ and $X2^{3}$ into the model gives the opportunity for best fit.

TWO OTHER SAS MACROS TO IMPLEMENT THE RUNNING OF FSP

Two macros, %FSP_36LR and %FSP_8LR, are presented that implement the running of FSP. An important difference between %FSP_36LR and %FSP_8LR is the following:

- %FSP_36LR runs 36 PROC LOGISTIC's (versus 47 for %MFP8) and finds the optimal FP1 and FP2 solutions.
- %FSP_8LR runs 8 PROC LOGISTIC's and finds the optimal FP1 solution but may not find the optimal FP2 solution.

Both macros allow any number of predictors to be entered into a macro parameter. Predictors whose minimum value is less than 1 are first translated so that the new minimum value equals 1. Then the output of processing of these predictors is consolidated into a single report. See Example 7 which leads to the example of a report shown in Table 9.

EXAMPLE 7: PREPARING DATA FOR RUNNING %FSP_8LR

```
data Example2;
do i = 1 to 500;
  X2 = ranuni(12341);
  Y = floor(X2 + ranuni(12341));
  output;
  end;
data Normal;
do i = 1 to 500;
  Normal = rannor(12345);
  output;
  end;
data Example7; merge Example2 Normal;
run;

%FSP_8LR(example7, Y, X2 Normal, );
```

Table 9 FSP RESULTS FROM %FSP_8LR

PRED_ ICTOR	TEST	DEV_ IANCE	TEST_ STAT	df	P- VALUE	MODEL	var1	trans_ form1	var2	trans_ form2	inter_ cept	esti_ mate1	esti_ mate2
X2	Null v. FP2	691.098	208.202	4	0.0000	Null					0.1282		
X2	Linear v. FP2	489.001	6.106	3	0.1066	Linear =	g1	linear			-7.9343	5.3491	
X2	FP1 v. FP2	483.540	0.645	2	0.7245	FP1 =	g2	p=-2			4.0901	-8.3200	
X2		482.896				FP2 =	g2	p=-2	g7	p=3	2.9666	-6.9522	0.1284
Normal	Null v. FP2	691.098	2.805	4	0.5909	Null					0.1282		
Normal	Linear v. FP2	689.463	1.170	3	0.7601	Linear =	g1	linear			0.5918	-0.1195	
Normal	FP1 v. FP2	689.453	1.161	2	0.5597	FP1 =	g5	p=0.5			1.0086	-0.4508	
Normal		688.292				FP2 =	g2	p=-2	L2	trans_ form1*log	-0.3460	-0.0702	5.0688

Here is why %FSP_8LR can produce sub-optimal FP2 solutions: %FSP_8LR runs PROC LOGISTIC with the options shown below for each of 8 sets of predictors called &Var1 to &Var8. See Table 10.

```
PROC LOGISTIC; MODEL Y = &VarK
/ SELECTION=FORWARD INCLUDE=1 START=1 STOP=2 SLE=1;
```

Table 10 Eight Sets of Predictors for %FSP_8LR

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

By this method all possible FP2 pairs have a chance to be selected. But the selection of the second variable of a pair by FORWARD, to add to the first variable forced in by INCLUDE=1, may be sub-optimal. The reason is that the second variable is selected by best score chi-square, not by maximizing log likelihood of the model.

E.g. Consider "&Var1" (Table 10).

- First, X is forced in by INCLUDE = 1
- Now perhaps the FORWARD criterion picks X^{-2} to enter as the second variable.
- But the best log likelihood might be given by X^3.

The FSP results for variable X2 from data set Example 2 are correctly obtained by %FSP_8LR. But examples where %FSP_8LR produces a suboptimal FP2 solution do exist.[12]

A study of occurrence rate and severity of non-optimal FP2 solutions is needed. Of cases examined, the occurrence rate is not high and the severity is not material. Of course, the advantage of %FSP_8LR over %FSP_36LR is run-time speed. The run-time for %FSP_36LR is approximately 4.5 times longer.

%FSP_8LR MAY PROVIDE AN ALTERNATIVE AS A SCREENER OF PREDICTORS

When the number of predictors to be screened for potential use in logistic regression is, perhaps, a few dozen, the use of %FSP_8LR provides an alternative to %LR_SCREENER. In %FSP_8LR there is one pass of the training dataset to find the minimum value of the predictors, another pass to create the FSP transformations, and then for each predictor there are 8 runs of PROC LOGISTIC. This compares with 3 passes in total of the training data by %LR_SCREENER.[13]

CONCLUSION

This paper shows that a process of screening a multitude of continuous predictors by %LR_SCREENER 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.

Alternatively, %FSP_8LR might be run both as a screener and as a means to select the final transformations for entry into the final variable selection process for a logistic regression model.

But there are open questions:

- What is the risk of false negatives (good predictors that are screened out) when running %LR_SCREENER?
- How often and to what degree would the %FSP_8LR approach to FPS give sub-optimal FP2 solutions as measured by differences in log-likelihood?

SAS MACROS DISCUSSED IN THIS PAPER

The SAS macros %LR_SCREENER, %FSP_36LR, and %FSP_8LR continue to be under development. Contact the author for a copy of the work-in-progress.

REFERENCES

Efron, B. (1975). The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis, *Journal of the American Statistical Association*, 70, 892-898.

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

Meier-Hirmer, Ortseifen, and Sauerbrei (2003). Multivariable Fractional Polynomials in SAS, Available at http://portal.uni-freiburg.de/imbi/mfp.

Press, S. J. and Wilson, S (1978). Choosing Between Logistic Regression and Discriminant Analysis, *Journal of the American Statistical Association*, 73, 699-705.

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

[12] There might appear to be redundancy in Table 10. Specifically, consider row #1 where the variable X is forced-in. Then in the FORWARD step the next variable to be selected is the one with the greatest score chi-square. In particular, X^{-2} is one of the predictors considered in this FORWARD step. Now in row #2 X^{-2} is forced-in. Do we need to reconsider X in the FORWARD step of row #2? The answer is "yes". This is, in fact, the case for predictor X1 of Example 1. For row #1 the FORWARD step selects $X1^{-2}$. But for row #2 the FORWARD step does not select X1. Instead the FORWARD step selects $X1^2$.

[13] The same is true for %FSP_36LR except that there are 36 runs of PROC LOGISTIC for each predictor

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

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.[14]

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 random samples from populations j = 1, 2 are 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.

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

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

$$P(j \mid X=x) = P(X=x \mid j\ )\ P(j) / P(x) \quad \ldots (B)$$

Substituting (B) into (A) gives:

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

Inequality (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

Following the steps from equation (B) to equation (E) shows the odds of membership in j=1 versus membership in j=2 to be given by:

$$P(1 \mid X=x) / P(2 \mid X=x) = CF_1 / CF_2 \quad \ldots (G)$$

Logarithms are taken of equation G to give (H):

$$\text{Log } (P(1 \mid X=x) / P(2 \mid X=x)) = \text{Log}(CF_1) - \text{Log}(CF_2) \ldots (H)$$

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

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

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

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

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 log odds 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:

$$\text{Log Odds} = \beta_0 + \beta_1 X.$$

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. For a discussion of the relative efficiency of these estimators see Efron (1975).

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 $T = t\ (1/S_p)\ sqrt(1/n_1 + 1/n_2)$ satisfies $P(| T | > C) = \alpha$. This test can be changed to finding a critical C’ so that $P(| t | > C') = \alpha$ where $C' = C\ / (\ (1/S_p)\ sqrt(1/n_1 + 1/n_2)\ )$. 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.

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);
    %else /*%then*/
```

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";
```