Paper 2390-2018

The Function Selection Procedure

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

ABSTRACT

The function selection procedure (FSP) finds a very good transformation of a continuous predictor to use in binary logistic regression. The FSP was developed in the 1990's for applications in biostatistics. The methodology is fully presented in the book by Royston and Sauerbrei (2008). In connection with their book Royston and Sauerbrei provided a SAS® macro to implement FSP. This SAS macro has many advanced features but it is designed for the analysis of one variable at a time. A more efficient approach is needed for large scale applications in marketing and credit risk. This paper presents an alternative macro %FSP_8LR which efficiently processes multiple predictor variables (for example, 50) with minimal passing of the data. Additionally, the methodology of FSP is extended in %FSP_8LR to the cumulative logit model. This paper includes a simulation study which explores and amplifies the significance testing component of FSP which was developed in the 1990's for the binary logistic regression case.

INTRODUCTION

In *Multivariate Model-building* by Royston and Sauerbrei (2008) a class of transformations of X, called fractional polynomials (FP), is discussed.[1] The use of FP first requires that X be translated so that the values of X are positive. Then the fractional polynomials are given by:

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

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

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

FP1 produces only monotonic curves. FP2 produces curves with a variety of non-monotonic shapes. Such an example is given in Figure 1.

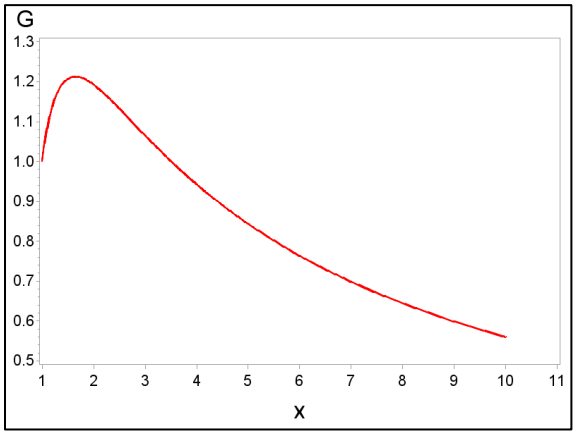

Figure 1. Graph of $G(X,-1,-1) = X^{-1} + 2\,X^{-1}\log(X)$

Hereafter, Royston and Sauerbrei (2008) will be shortened to R-S.

The function selection procedure (FSP) is a procedure that builds upon FP1 and FP2 to find a very good transformation of X for applications in binary logistic regression, ordinary least squares, and proportional

[1] X is a numeric predictor with many levels. It might measure time, distance, money, etc.

hazard modeling. R-S (p. 267) gives links to software for performing FSP including Stata, R, and SAS. The SAS version, a macro named %MFP8, was current as of 9/7/2017 but it is written in SAS version 8.[2] [3]

FOCUS OF THIS PAPER IS EXCLUSIVELY ON FSP FOR LOGISTIC REGRESSION

There are two main objectives of this paper:

- To extend FSP to the cumulative logit model.
- To present a SAS macro called %FSP_8LR that efficiently processes multiple predictor variables (e.g. 50) with minimal passing of the data.

This paper is an extension of the discussion of FSP given in Lund (2015).

FSP HAS TWO MAIN STEPS

I. Searching for Best Transformations: In the approach taken by %MFP8 and in R-S there is an exhaustive search of FP1 to find the function with maximum likelihood and a second exhaustive search of FP2 to find the function with maximum likelihood. This required the running of 44 Logistic Models in %MFP8 and, overall, PROC LOGISTIC is run 47 times by %MFP8.

In %FSP_8LR a compromised strategy is adopted which requires only 8 runs of PROC LOGISTIC. The optimal FP1 transform will always be found but it is possible that a sub-optimal FP2 transform is selected. In practice a sub-optimal result is both infrequent and not material. The details are further discussed in a later section of this paper about the operation of %FSP_8LR.

II. Performing Significance Testing: Second, significance testing is performed. The FSP significance testing follows these three steps. The test-statistic is displayed following the description of steps 1, 2, 3.

1. Perform a 4 d.f. test at the α level of the best-fitting FP2 against the null model (no predictor). If the test is not significant, drop X from consideration and stop; otherwise continue.
2. Perform a 3 d.f. test at the α level of the best-fitting FP2 against X. If the test is not significant, stop, the final model is linear X; otherwise continue.
3. Perform a 2 d.f. test at the α level of the best-fitting FP2 against the best-fitting 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 [4] as shown below:

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

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

The rationale for the degrees of freedom (4, 3, 2) in the 3-step hypothesis tests of FSP is discussed in R-S (p. 79) for binary logistic models. This rationale generalizes to the cumulative logit model.

The significance testing is further discussed in the context of the simulation study in a later section. This simulation study will explore the questions of degrees of freedom and the p-values that arises from the test-statistic for the cumulative logit model.

CUMULATIVE LOGIT MODEL

First, so that the paper can be self-contained, there is a short description of the cumulative logit model.

A Simplification for this Paper

In this paper all discussion of the cumulative logit model will assume the target has 3 levels.

[2] Meier-Hirmer, Ortseifen, and Sauerbrei (2003). Downloaded from http://portal.uni-freiburg.de/imbi/mfp Last verified 9/7/2017
[3] Meier-Hirmer, Ortseifen, and Sauerbrei (2003). Multivariable Fractional Polynomials in SAS, http://portal.uni-freiburg.de/imbi/mfp. See beschreibung.pdf in SAS downloads.
[4] The deviance is the -2 Log Likelihood value of a logistic model.

This reduces notational complexity. However, %FSP_8LR can be applied to cumulative logit models where the number of target levels is greater than or equal to 3.

Definition of the Cumulative Logit Model with the Proportional Odds (PO) Property

To define the cumulative logit model with proportional odds, the following example is given: Assume the 3 levels for the ordered target Y are A, B, C and suppose there are 2 numeric predictors X1 and X2.[5]

Let $p_{k,j}$ = probability that the k^{th} observation has the target value j = A, B or C. Let $X_{k,1}$ be the value of X1 for the k^{th} observation. Similarly, for $X_{k,2}$.

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

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

Equation 1. Response Equations for Cumulative Logit Model with ascending Target values.

Predictor X1 has the same coefficient β_{X1} in both response equations. Similarly, X2 has the same coefficient β_{X2} in both response equations.

The "cumulative logits" are the log's of the ratio of the "cumulative probability up to j" (in the ordering of the target) in the numerator to "one minus the cumulative probability up to j" in the denominator.

Formulas for the probabilities $p_{k,A}$, $p_{k,B}$, $p_{k,C}$ can be derived from the two response equations. To simplify the formulas, let T_k and U_k, for the k^{th} observation be defined by the two equations below:

T_k	$T_k = \exp(\alpha_A + \beta_{X1}*X_{k,1} + \beta_{X2}*X_{k,2})$
U_k	$U_k = \exp(\alpha_B + \beta_{X1}*X_{k,1} + \beta_{X2}*X_{k,2})$

Equation 2.

Then, after algebraic manipulation, the probability equations in Equation 3 are derived:

Response	Probability Formula
A	$p_{k,A} = 1 - 1/(1+T_k)$
B	$p_{k,B} = 1/(1+T_k) - 1/(1+U_k)$
C	$p_{k,C} = 1/(1+U_k)$

Equation 3. Cumulative Logit Model - Equations for Probabilities

The parameters for the cumulative logit model are estimated by maximizing the log likelihood equation in a manner similar to the binary case (Agresti 2010, p 58).

This cumulative logit model satisfies the following conditions for X1 (and the analogous conditions for X2):

Let "r" and "s" be two values of X1 and fix the value of X2. Using the probability formulas from Equation 3:

$$\text{Log}\left[\frac{p_{r,A}/(p_{r,B}+p_{r,C})}{p_{s,A}/(p_{s,B}+p_{s,C})}\right] = \text{Log}\,(p_{r,A} / (p_{r,B} + p_{r,C})) - \text{Log}\,(p_{s,A} / (p_{s,B} + p_{s,C})) = (r - s) * \beta_{X1}$$

$$\text{Log}\left[\frac{(p_{r,A}+p_{r,B})/p_{r,C}}{(p_{s,A}+p_{s,B})/p_{s,C}}\right] = \text{Log}\,((p_{r,A} + p_{r,B}) / p_{r,C}) - \text{Log}\,((p_{s,A} + p_{s,B}) / p_{s,C}) = (r - s) * \beta_{X1}$$

These equations display the "proportional odds" property. Specifically, the difference of cumulative logits at r and s is proportional to the difference (r - s). The proportional odds property is a by-product of the equality of the coefficients of predictors X1 and X2 across the cumulative logit response equations.[6]

[5] If a predictor X is not numeric, then the dummy variables from the coding of the levels of X appear in the right-hand-side of the response equations for j = A and j = B.

[6] A further introduction to the cumulative logit model is given by Allison (2012, Chapter 6).

DEFAULT ASSUMPTION BY PROC LOGISTIC

If the target variable in PROC LOGISTIC has more than 2 levels, PROC LOGISTIC regards the appropriate model as being the cumulative logit model with the proportional odds property.[7]

PARTIAL PROPORTIONAL ODDS (PPO)

This section defines the partial proportional odds (PPO) cumulative logit model. This is a generalization of the proportional odds model.[8]

To describe PPO, the following example is given: Assume there are 3 levels for ordered target Y: A, B, C and there are 3 numeric predictors R, S and Z.

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

In this example the PPO Model will have 6 parameters α_A α_B β_R β_S $\beta_{Z,A}$ $\beta_{Z,B}$ given in 2 equations:

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

Here, Z has different coefficients for the 2 response equations. In general, for PPO some predictors may have coefficients with values that vary across response equations.

Formulas for probabilities $p_{k,A}$, $p_{k,B}$, $p_{k,C}$ continue to be given by Equation 3 after modifications to definitions of T and U to reflect the PPO model. In unusual cases it is possible for a PPO probability to be negative.[9]

TEMPLATE FOR SIMULATIONS FOR LOGISTIC REGRESSION DATA

Equations 1, 2, and 3 (see above) provide the formulas for creating data sets for the cumulative logit PO model. Here is code that creates a data set that is generated by the FP2 model with log(X) and X^{-1}. Intercepts are set at 0 and 1. The intercept 0 corresponds to target value A and intercept 1 to target value B.

This template is used for the simulation study to be reported in a later section.

```
%LET ERROR = 0.01;
%LET SLOPE1 = 0.2;
%LET SLOPE2 = -0.5;
%LET P_Seed = 5;
%MACRO SIM(NUM);
%DO Seed = 1 %TO &NUM;
  DATA SIM_&Seed;
  do i = 1 to 8000;
     X = mod(i,16) + 1;
     rannorx = rannor(&Seed);
/* Equations 1 and 2 */
     T = exp(0 + &SLOPE1*LOG(X) + &SLOPE2*(1/X) + &ERROR*rannorx);
     U = exp(1 + &SLOPE1*LOG(X) + &SLOPE2*(1/X) + &ERROR*rannorx);
/* Equations 3 */
     PA = 1 - 1/(1 + T);
     PB = 1/(1 + T) - 1/(1 + U);
     PC = 1 - (PA + PB);
/* Assign Target Values to match model probabilities */
     R = ranuni(&P_Seed);
```

[7] Simply run: PROC LOGISTIC; MODEL Y = <X's>; where Y has more than 2 levels.
[8] See Derr (2013) for discussion of the PO and PPO models and testing of predictors for "unequalslopes".
[9] http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_logistic_examples22.htm See note at bottom of webpage for discussion. Also see slide 47 of presentation slides by Richard Williams (2008) for a discussion and references. https://www.stata.com/meeting/germany08/GSUG2008.pdf

```
      if R < PA then Y = "A";
       else if R < (PA + PB) then Y = "B";
       else Y = "C";
      output;
      end;
run;
  %END;
%MEND;
%SIM(1);
```

MACRO %FSP_8LR: PARAMETERS

The macro call is %FSP_8LR (DATASET, TARGET, INPUT, VERBOSE, ORDER);

Parameter definitions:

DATASET:	The data set containing the target and predictors
TARGET:	Target variable (character or numeric). At least two non-missing levels
INPUT:	Numeric predictors (at least one). Predictors are delimited by a space. ("-" convention is implemented ... e.g. X1 - X6). A predictor may have missing values.
VERBOSE:	YES … “YES" produces more output. Recommend not using “YES” for mass screening of predictors
ORDER:	A | D … The order for modeling the TARGET (A=ascending, D=descending). Recommend “A” for cumulative logit model to simplify interpretation of Intercepts.

The relationship between TARGET and a Predictor must be compatible with running PROC LOGISTIC.

%FSP_8LR passes the data once to translate all predictors to have a minimum value of at least 1 and then, once again, to form the 8 FP transforms for these predictors. Finally, for each predictor in INPUT, the data is passed 8 times when running PROC LOGISTIC’s. More explanation of these 8 PROC LOGISTIC’s is given in a later section.

EXAMPLE FOR %FSP_8LR WHERE TARGET HAS 3 LEVELS

The template simulation code was run with NUM=1 to produce data set SIM_1 with TARGET=Y with 3 levels and predictor X. SIM_1 generated logistic data involving transforms of X given by log(X) and X^{-1}.

INPUT has the value X, VERBOSE is YES, and ORDER is A. The macro call is:

%FSP_8LR (SIM_1, Y, X, YES, A);

The following reports are generated:

Distribution of TARGET Values
The FREQ Procedure

Y	Frequency	Cumulative Frequency
A	4496	4496
B	1742	6238
C	1762	8000

Minimum Value before Translation

Obs	var_	min_
1	X	1

Two tables below (FP1 and FP2) are printed because VERBOSE=YES. They are sorted by -2*Log(L).

The FSP1 table shows that the best FP1 transform is $X^{-0.5}$. The coefficient of $X^{-0.5}$ is -1.433. The intercept for A is 0.846 and for B it is 1.880.

Since Y has more than 2 levels, the test of proportional odds (PO) is given.[10] In this example, the p-values for the proportional odds test for best FP1 solution and best FP2 solution are 0.078 (borderline rejection) and 0.290 respectively (see table for FP2). Interpretation and actions to-be-taken related to the proportional odds test is given in the section following this section.

FP1

Obs	-2Log(L)	Trans. 1	Est. 1	Int. 1	Int. 2	ChiSq _PO	DF _PO	ProbChi Sq_PO
1	15654.34	p=-0.5	-1.433	0.846	1.880	3.112	1	0.078
2	15655.34	p=-1	-1.184	0.498	1.532	2.539	1	0.111
3	15666.16	log	0.355	-0.430	0.602	3.316	1	0.069
4	15674.34	p=-2	-1.075	0.354	1.386	1.721	1	0.190
5	15687.21	p=0.5	0.289	-0.552	0.477	2.799	1	0.094
6	15709.30	Linear	0.051	-0.181	0.846	1.934	1	0.164
7	15742.31	p=2	0.002	0.020	1.044	0.706	1	0.401
8	15762.15	p=3	0.000	0.093	1.114	0.241	1	0.623

In the FSP2 table the first two FSP2 solutions are the same. Each gives the transforms X^{-2} and log(X).

FP2

Obs	-2Log(L)	Trans. 1	Trans. 2	Est. 1	Est. 2	Int 1.	Int. 2	ChiSq _PO	DF _PO	ProbChi Sq_PO
1	15652.44	p=-2	log	-0.535	0.218	-0.115	0.919	2.479	2	0.290
2	15652.44	log	p=-2	0.218	-0.535	-0.115	0.919	2.479	2	0.290
3	15652.56	p=-1	Linear	-1.017	0.012	0.364	1.399	3.608	2	0.165
4	15652.56	Linear	p=-1	0.012	-1.017	0.364	1.399	3.608	2	0.165
5	15652.63	p=2	p=-1	0.001	-1.073	0.423	1.457	3.876	2	0.144
6	15652.69	p=0.5	p=-1	0.072	-0.962	0.252	1.286	3.402	2	0.182
7	15652.93	p=3	p=-1	0.000	-1.101	0.445	1.480	3.952	2	0.139
8	15652.94	p=-0.5	p=-2	-1.165	-0.237	0.757	1.792	2.734	2	0.255

A listing of the summary report is below (some columns are omitted). This report is produced regardless of the parameter value of VERBOSE.

The off-set is the amount added to X so that MIN(X) = 1 (if MIN(X) does not already exceed 1). Off-set = 0 for this example.

The test-statistic for FP1 v. FP2 is not significant at 0.387. Based on the 3 step significance testing, the FP1 solution is selected with transform $X^{-0.5}$.

The fitted FP2 transformations are X^{-2} and log(X), versus the simulated predictors of X^{-1} and log(X).

[10] Let S be the number of predictors in the model (S is either 1 or 2) and J be the number of levels of Target. The test statistic is a chi-square with (J-2)*S degrees of freedom. A small value of ProbChiSq_PO rejects the proportional odds assumption. However, this test may too often lead to rejection as discussed by Allison (2012 p. 168).

Significance of FP2 v Null, FP2 v Linear, FP2 v FP1
OFF-SET is added to Pred. so MIN(Pred) = 1 OR OFF-SET equals ZERO if MIN(Pred) >= 1

Off-set if PRED < 1	TEST	DEVIANCE	TEST _STAT	df	P-VALUE	MODEL	Trans. 1	Trans. 2
0	Null v. FP2	15824.49	172.05	4	0.000	Null		
0	Linear v. FP2	15709.30	56.86	3	0.000	Linear =	Linear	
0	FP1 v. FP2	15654.34	1.90	2	0.387	FP1 =	p=-0.5	
0		15652.44				FP2 =	p=-2	log

IF TEST FOR PO IS REJECTED

The test for PO might be rejected for a transform or transforms found by %FSP_8LR. If for FSP1, then the Linear or FP1 transformation has “unequalslopes” (different coefficients in the response equations). If for FSP2, then at least one of the two FSP2 transformations has unequalslopes. In either case the user should explore the usage of a PPO model.

The PPO model, including tests for unequalslopes for individual predictors in a model, is discussed by Derr (2013).

In PROC LOGISTIC a predictor is allowed to have different coefficients by entering the predictor in the UNEQUALSLOPES statement. This is shown for the predictor X in the PROC LOGISTIC code:

```
PROC LOGISTIC;
MODEL Y= X <other predictors>
/ UNEQUALSLOPES= (X <some of the other predictors>);
```

A future version of %FSP_8LR will extend to the PPO cumulative logit model. Currently being tested.

%FSP_8LR SEARCH FOR FP2 SOLUTION

Earlier in the paper it was stated that %FSP_8LR may not find the best FP2 solution. Here is the explanation: %FSP_8LR runs PROC LOGISTIC 8 times with the SELECTION options shown below. For each run there are 9 predictors listed in the MODEL statement. The 8 lists have 8 predictors in common but with a ninth that is unique to the list (in the right-most column). See Table 1.[11]

```
PROC LOGISTIC; MODEL Y = &Var<K>
/ SELECTION=FORWARD INCLUDE=1 START=1 STOP=2 SLE=1;
```

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

Table 1. Eight Sets of Predictors for %FSP_8LR

[11] There might appear to be redundancy in Table 1. 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. Suppose FORWARD selects X^{-2}. Can row#2, where X^{-2} is forced by INCLUDE=1, be simplified by checking only the newly occurring predictor X^{-2} Log(X)? The apparent reason would be that X and X^{-2} were paired by the logistic model of row#1. But this reasoning is not valid. In row#2 the predictor selected by FORWARD can be any of the remaining 8 predictors.

The INCLUDE=1 forces the selection of the left most predictor in the list. Then FORWARD with STOP=2 and SLE=1 will select one more predictor. By this method all possible FP2 pairs have a chance to be selected. But the selection of the second variable 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 the best score chi-square, not by maximizing log likelihood of the model.

E.g. Consider row#1 in Table 1.

- 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}.

Examples where %FSP_8LR produces a suboptimal FP2 solution do exist.[12] However, of all examples so far examined, the occurrence rate is not high and the severity is not material.

SIMULATION FOR 3-STEP SIGNIFICANCE TESTING

The rationale for the degrees of freedom (4, 3, 2) used in the 3-step hypothesis tests of FSP is discussed in R-S (p. 79) for the case of binary logistic models. This rationale by R-S generalizes to the cumulative logit model. But it is easier to explore the 3 step significance testing in the context of simulations.

The simulations presented here will focus on the cumulative logit model where the target has 3 levels.

This simulation study will leverage the SAS code in the template, given earlier and repeated below. In the example below there are FP2 transformations log(X) and X^{-1}.

```
%LET ERROR = <>; %LET SLOPE1 = <>; %LET SLOPE2 = <>; %LET P_Seed = <>;
%MACRO SIM(NUM);
%DO Seed = 1 %TO &NUM;
  DATA SIM_&Seed;
  do i = 1 to 8000;
    X = mod(i,16) + 1;
    rannorx = rannor(&Seed);
    T = exp(0 + &SLOPE1*LOG(X) + &SLOPE2*(1/X) + &ERROR*rannorx);
    U = exp(1 + &SLOPE1*LOG(X) + &SLOPE2*(1/X) + &ERROR*rannorx);
    PA = 1 - 1/(1 + T);
    PB = 1/(1 + T) - 1/(1 + U);
    PC = 1 - (PA + PB);
    R = ranuni(&P_Seed);
    if R < PA then Y = "A";
    else if R < (PA + PB) then Y = "B";
    else Y = "C";
    output;
    end;
  run;
%END;
%MEND;
```

DISCUSSION OF THE SIMULATION STUDY

There are too many dimensions to control in order to conduct a definitive simulation study. In the template program these dimensions are: (1) the distribution of X across its range, (2) the nature of the error term, (3) the sample size, and (4) the values of the coefficients for the FSP transformations.

In this simulation study the range of X is 1 to 16 with 500 occurrences of each level of X for a total sample of 8,000. The error term in the response equations is a standard normal but is multiplied by a

[12] See R-S p 266 Whitehall I Data. In this data set there are CIGS and ALL10 (values 0 and1). Let INPUT consist of X where X = CIGS + 1. Let TARGET = ALL10. Run %FSP_8LR. The FP2 solution is X^{-1} and X^{-1}*Log(X). This compares with the %MFP8 solution FP2 solution of X^{-2} and X^{-1}. The difference in -2*Log(L) is 10708.270 vs. 10707.827

factor = &ERROR. (Alternatively, the error term could be given by the standard logistic distribution). The number of cases in a simulation run is 300.

The transforms of X and the slope(s) (&SLOPE1 and &SLOPE2) of the transforms will change according to the needs of the simulation.

Each simulated data set is run through %FSP_8LR. The relative frequencies of the p-values for the test-statistic [13] for the 300 cases are summarized in the reports shown below. These p-values are based on degrees of freedom of 4, 3, 2 for the 3-step significance testing.

STEP1: TESTING THE NULL CASE VS. FP2

This is the most straight-forward test. For this null hypothesis the predictors have both coefficients (&SLOPE1 and &SLOPE2) set to zero. The magnitude of the error term adds one dimension to the simulation. For this simulation, the multiplier of the error term is set at 0.01. The alternative hypothesis includes all the possible FP2 selections.

For a simulation that satisfies a null hypothesis, the observed frequency of the computed p-values should correspond to expected frequencies Specifically, about 5% of the computed p-values should have values of under 0.05. Likewise, for 10% and 15%.

The macro variables for the template code are set below.

```
%LET ERROR = 0.01; %LET SLOPE1 = 0.0; %LET SLOPE2 = 0.0;
```

There were 300 cases in the simulation. The FSP test-statistic is based on 4 d.f. Table 2 shows the results of the simulation. In addition to a test-statistic with 4 degrees of freedom, Table 2 also shows alternative test-statistics computed with 2 and 3 degrees of freedom.

p-value =	5%			10%			15%		
test-stat d.f. =	2	3	4	2	3	4	2	3	4
% < p-value	11.3%	4.0%	2.0%	20.7%	10.0%	4.0%	28.7%	14.7%	6.7%

Table 2. Testing Step 1. "% < p-value" gives percent of 300 cases less than 5%, 10%, 15%

The simulation shows the 4 d.f. test-statistic, as utilized in the 3-step testing of FSP, to be overly conservative (too infrequently rejecting the null hypothesis of no relationship of Target to a function of X). The 3 d.f. test-statistic is about right. The percent of rejections of the null at 5%, 10%, and 15% p-values are respectively 4.0%, 10.0%, 14.7%. For the test-statistic using 2 d.f. there are too many rejections of the null.

STEP2: TESTING THE LINEAR CASE VS. FP2

The only predictor is X. The specification of the slope adds a new dimension to the null hypothesis. A complete simulation would involve various specifications of the slope and the error term. Here, only one specification is tested: &ERROR = 0.01 and &SLOPE1 = 0.1.

```
%LET ERROR = 0.01; %LET SLOPE1 = 0.1; %LET SLOPE2 = 0.0;
```

There were 300 cases in the simulation. The FSP test-statistic is based on 3 d.f.

p-value =	5%			10%			15%		
test-stat d.f. =	2	3	4	2	3	4	2	3	4
% < p-value	7.7%	3.3%	1.0%	14.0%	6.3%	3.3%	22.3%	10.0%	4.7%

Table 3. Testing Step 2

[13] Recall that Test-Statistic = (-2 Log Likelihood $_{\text{restricted model}}$) - (-2 Log Likelihood $_{\text{full model}}$)

The simulation shows the 2 d.f. test-statistic to be too liberal (rejecting the null too often) and 3 d.f. test-statistics to be too conservative. Their deviations from the expected rates are about equal in magnitude but different in sign. The 4 d.f. test-statistic is much too conservative.

For all 300 cases the Step 1 hypothesis of the Null Model was strongly rejected.

STEP3: TESTING FP1 VS. FP2 AND EXPECTING TO ACCEPT THE NULL HYPOTHESIS

A simulated relationship to the Target is specified by selecting an FP1 transformation for use in the template program. But which one? The selection of a specific FP1 transformation for a simulation does not really reflect the null hypothesis. The null hypothesis compares the best FP1 model to the best FP2 model where each is the best fit to the given data. But in the simulation a specific FP1 solution must be picked to generate the data.

Further complicating the choice of an FP1 transformation for the simulation is the possibility that the linear model is not rejected. Without this rejection there is no test of FP1 vs. FP2. Many FP1 transformations may look Linear over the range of X. The FP1 transformation has to exhibit enough curvature over the range of X to distinguish it from Linear. For this simulation the chosen FP1 transformation is X^2 with these macro variable values:

```
%LET ERROR = 0.01; %LET SLOPE1 = 0.02; %LET SLOPE2 = 0.0;
```

There were 300 cases in the simulation. The FSP test-statistic is based on 2 d.f.

p-value =	5%			10%			15%		
test-stat d.f. =	1	2	3	1	2	3	1	2	3
% < p-value	14.7%	5.3%	2.3%	28.0%	10.3%	4.7%	39.3%	14.7%	8.3%

Table 4. Testing Step 3

For all 300 cases the Linear hypothesis was strongly rejected and the selected FP1 transformation was X^2. The 2 d.f. test-statistic performed better than test-statistics using 1 d.f. or 3 d.f.

The simulation using X^2 was “cherry-picked” in the sense that the data generated by $0.02 * X^2$ is distinguished from data generated by X (regardless of slope of X).

In contrast, if $0.02 * X^{-1}$ were used in the simulation, a typical result (for data generated by the template program) is to accept the Null Model of no relationship between X and the Target.

STEP3: TESTING FP1 VS. FP2 AND EXPECTING TO REJECT THE NULL HYPOTHESIS

In this simulation that data set is generated by transformations log(X) and $6*X^{-1}$. The function $F(X) = \text{Intercept} + \log(X) + 6*X^{-1}$ is a non-monotonic curve with minimum at X=6. Since all FP1 solutions are monotonic, the expectation is that FP1 will be rejected in favor of FP2.

```
%LET ERROR = 0.01; %LET SLOPE1 = 1.0; %LET SLOPE2 = 6.0;
```

There were 300 cases in the simulation. The FSP test-statistic is based on 2 d.f.

p-value =	5%			10%			15%		
test-stat d.f. =	1	2	3	1	2	3	1	2	3
% < p-value	86.3%	72.3%	54.0%	90.7%	81.3%	68.3%	92.7%	86.3%	76.0%

Table 5. Testing Step 3

Now the challenge is to calculate the probability of not making a type II error since, by design, the alternative hypothesis is true. More generally, the challenge is to calculate a power curve. But the alternative hypothesis is a complex composite of FP2 transforms which defy this calculation.

It is encouraging that the null hypothesis of FP1 for the test-statistic with 2 d.f. is rejected in the simulation at a rate of 72.3% at $\alpha = 0.05$.

In all 300 cases the fitted FP1 transformation was X^{-2}. Although X^{-2} is monotonic over 1 to 16, with appropriate slope and intercept, it can track F(X) fairly closely.

The simulated data were generated by LOG(X) and $6*X^{-1}$, but in only 18 cases (6.0%) were the fitted FP2 transformations given by log(X) and X^{-1}.

CONCLUSIONS FROM THE SIMULATION

Step 1 and Step 2 can be regarded as formal hypothesis tests since the null hypothesis is fairly well defined. The degrees of freedom for these tests (4 and 3 respectively) are possibly too conservative, leading to too few rejections of the null. Importantly, accepting the null at Step 1 means abandoning this predictor all together.

The results of Step 3 should be viewed as guidelines to be evaluated by the modeler in terms of interpretability and model complexity.

In general, the simulation study supports the use of FSP 3-Step testing for the cumulative logit model with degrees of freedom 4, 3, 2 as a guideline.

But aside from the 3-Step testing, the modeler should form plots of the empirical cumulative logits and fitted cumulative logits across the range of X for each of the Linear, FP1, and FP2 solutions. A visual inspection of these plots can influence the selection of which transformation to be used in the model.

SAS MACROS DISCUSSED IN THIS PAPER

SAS macro %FSP_8LR is available from the author.

REFERENCES

Agresti, A (2010). *Analysis of Ordinal Categorical Data, 2nd Ed.*, Hoboken, NJ, John Wiley & Sons.

Allison, P.D. (2012), *Logistic Regression Using SAS: Theory and Application 2nd Ed.,* Cary, NC, SAS Institute Inc.

Derr, B. (2013). Ordinal Response Modeling with the LOGISTIC Procedure, *Proceedings of the SAS Global Forum 2013 Conference,* Cary, NC, SAS Institute Inc.

Lund, B. (2015). Selection and Transformation of Continuous Predictors for Logistic Regression, *Proceedings of the SAS Global Forum 2015 Conference*, Paper 2687-2015.

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

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

CONTACT INFORMATION

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

Bruce Lund
Magnify Analytic Solutions
Detroit MI
blund_data@mi.rr.com, blund.data@gmail.com, or blund@magnifyas.com

All code in this paper is provided by Magnify Analytic Solutions "as is" without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Recipients acknowledge and agree that Magnify Analytic Solutions shall not be liable for any damages whatsoever arising out of their use of this material. In addition, Magnify Analytic Solutions will provide no support for the materials contained herein.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.