United States
Environmental Protection
Agency

Office of Research and
Development
Washington, D.C. 20460

EPA/600/R-01/081
October 2001

Parametric and Nonparametric Logistic Regressions for Prediction of Presence/Absence of an Amphibian

Q: *What environmental factors determine the distribution of the Red-Spotted Toad in a fragmented desert landscape?*

Q: *Can logistic regression using GLM or GAM-MARS be used to address this question?*

Note: To conserve space on the cover, the title of this report has been abbreviated. The full title appears on the following page.

EPA/600/R-01/081
October 2001

Parametric and Nonparametric (MARS; Multivariate Additive Regression Splines) Logistic Regressions for Prediction of A Dichotomous Response Variable With an Example for Presence/Absence of an Amphibian*

by

Maliha S. Nash and David F. Bradford

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
Las Vegas, Nevada

* Full Title

Notice

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and performed the research described here. It has been peer reviewed by the EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Table of Contents

Page

Notice ii

List of Figures iv

List of Tables v

Acknowledgments vi

Section 1 Introduction 1
1.1 Example Data Set Used 5
1.2 Logistic Regression 5
1.2.1 Maximum Likelihood Estimator 6
1.2.2 Assumptions 6
1.2.3 Steps to Follow 7
1.2.4 Model Selection 17
1.2.5 Dependence Between Observations 17
1.2.6 Interactions 18

Section 2 MARS 19
2.1.1 Model Fitting 19
2.1.2 Final Model 21

Section 3 Conclusion 25

References 26

Appendix 1 GENMDOD Output 28

Appendix 2 SAS Statements for Standard Logistic Regression 32

Appendix 3 *MARS* Output 35

List of Figures

Figure #		Page
1	Presence (=1), absence (=0), and (–) prediction of amphibians as related with variable for latitude (UTM-N see Table 1 for variable description)	2
2	Box-plots show the presence (1) and absence (0) of toads as related to the independent variables. Independent variables are: Elevation (U), UTM-N (W; latitude), and vegetation cover (< 1 m high) over adjacent land, mean % (Q). "+" is the mean value; lines for box from top as: Maximum, Quartile 3, Quartile 2, Quartile 1 and minimum values.	13
3	Deviance (DIFDEV) values for the predicted probability of the presence and absence of toads	15
4	Chi-square (DIFCHISQ) values for the predicted probability of the presence and absence of toads	15
5	Square root of the absolute value of the deviance residual for each independent variable (x_i's) that was significant in the final model	16
6	Regression for the main effect variables in the model using MARS	24

List of Tables

Table #		Page
1	Description for Metrics Used in Analyses	2
2	SAS Output for Model Fit, Testing Global Null Hypothesis, and Association of Predicted Probability and Observed Toad Presence	10
3	Final Stepwise Logistic Regression Analysis Model (n = 122 sites)	11
4	Coefficients and Their Statistics From MARS. Model F = 14.128, p < 0.001, df = 7,114	21

Acknowledgments

We are grateful for the valuable inputs and suggestions provided by Drs. James Wickham, Sandra Catlin, Kurt Riitters, Chad Cross and Bill Bossing, which improved the comprehensiveness and clarity of this report.

Section 1
Introduction

The purpose of this report is to provide a reference manual that can be used by investigators for making informed use of logistic regression using two statistical methods (standard logistic regression and Multivariate Adaptive Regression Splines (MARS)). The details for analyses of relationships between a dependent binary response variable (e.g., presence/absence) and a set of independent variables are provided step-by-step for use by scientists who are not statisticians. Details of such statistical analyses and their assumptions are often omitted from published literature, yet such details are essential to the proper conduct of statistical analyses and interpretation of results. In this report, we use a data set for amphibian presence/absence and associated habitat variables as an example.

Relationships between a response variable and independent variable(s) are commonly quantified and described by regression models. The values of the coefficients and predictions from the fitted model are used to infer and describe patterns of relationships, the effect of the independent variables on the response, and the strength of association between the independent and response variable. All these will help to analyze and understand a phenomenon, in this case biological phenomena. The general linear model (GLM) offers a wide range of regression models where the simple regression, analysis of covariance, and ANOVA are special cases. In GLM, the functional relationships between the expected value of the response variable(s) and the independent variables are described via a link function as:

$$g(\mu) = \beta_o + \sum \beta_i x_i \quad (1)$$

where $g(.)$ is link function, μ is the expected value of the response variable, $\beta_o \& \beta_i$ are regression coefficients, and x_i's are the independent variables. In simple regression, the link function represents the mean of the response variable as:

$$g(\mu) = \mu = \beta_o + \sum \beta_i x_i \quad (2)$$

The above model represents the simplest relationship between the response and independent variables in a linear manner. The above relationship also implies that the dependent variable is continuous and random.

When the relationships between the mean response and the independent variables are not linear, a different link function can be used to describe the relationships. The loglinear model, for example, can be used where the link function is defined as:

$$\log(\mu) = \beta_o + \sum \beta_i x_i \quad (3)$$

where the $\log(\mu) = \log(\mu / (1 - \mu))$ is the "logit" link, which is the logit transformation of the probability. When the binary response variable (present/absence) is plotted against the independent variable (Figure 1), the data are on the one and zero lines. Whether the UTM-N variable (i.e., latitude) enhances or suppresses the presence of amphibians, a relationship cannot be assessed by examining Figure 1 as normally done with the continuous response variables in linear regression analysis. Therefore, the linear model (Equation 2) is not valid for count and dichotomous response variables. Instead, the response variable is related nonlinearly to the independent variable(s) via a link, such as "logit" (Equation 3; solid line in Figure 1). The latter is usually chosen as the link function in which logistic regression can be an increasing or decreasing function, the link function is differentiable, and it relates the linear predictor to the expected value of the response. The logit transformation in Equation 3 should map the binary values (0,1) to a range of $(-\infty, \infty)$ over the domain $\mathbf{x}$ $(-\infty, \infty)$.

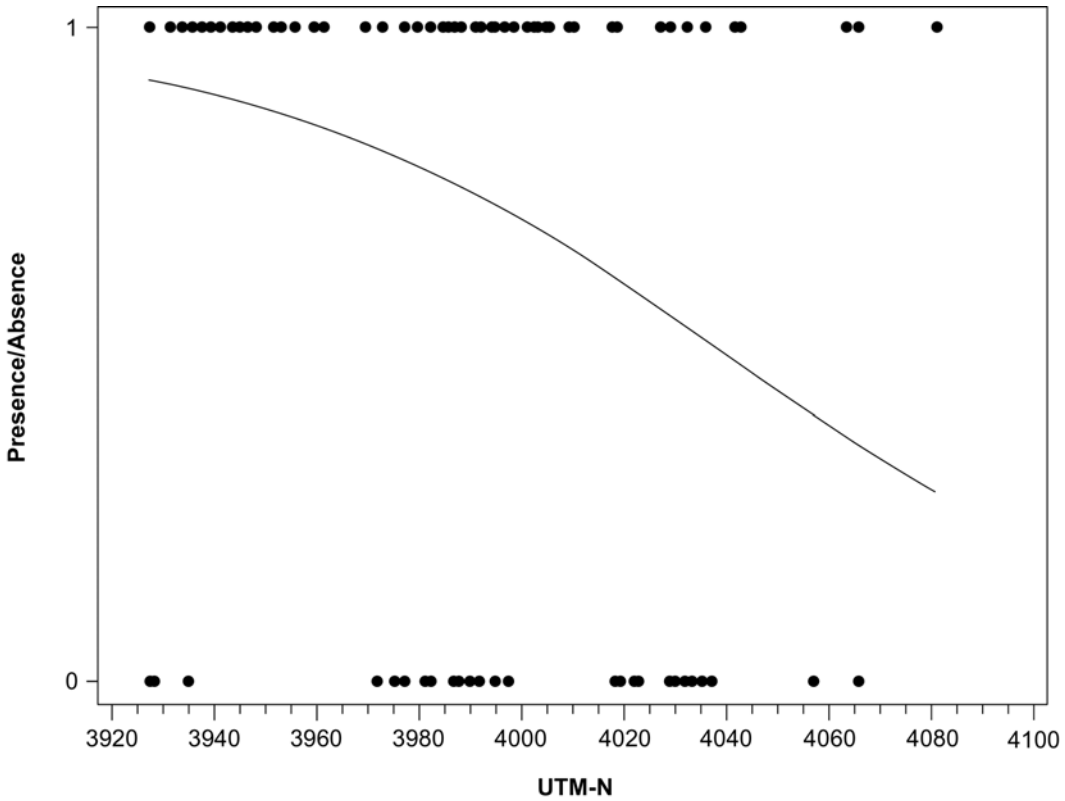

Figure 1. Presence (=1), absence (=0), and (–) prediction of amphibians as related with variable for latitude (UTM-N see Table 1 for variable description).

Table 1. Description for Metrics Used in Analyses

Variable Type/Name	*P* > P²	Description
Patch Size Metrics		
Water		
LogWPArea (A)	NS	Surface water area ($\log_{10}$ [m^2 + 10])
SiteSurf (B)	NS	Surface water linear extent (percent)

Continued.....

Table 1. *Continued*

Variable Type/Name	*P* > P²	Description
Patch Size Metrics, *Continued*		
Vegetation		
LogRiparea (C)	0.0085	Riparian zone area (log_{10} [m^2 + 10]). Zone boundaries delineated by indicator taxa defined for SITEEMER, SITENATI, and SITERIPA
SITEEMER (D)	NS	Emergent-type vegetation linear extent (percent). Indicator taxa: *Typha*, *Eleocharis*, *Scirpus*, *Mimulus*, *Anemopsis*; *Juncus* & *Carex* if in stream channel
SITENATI (E)	NS	Native riparian trees linear extent (percent). Indicator taxa: *Salix*, *Populus*, *Fraxinus*
SiteRipa (F)	0.0859	Riparian shrubs/herbs linear extent (percent). Indicator taxa: *Baccharis*, *Pluchea*, *Vitis*, *Allenrolfea*, *Equisetum*; *Juncus* or *Carex* if outside stream channel
SiteMesq (G)	NS	Phreatophytes linear extent (percent). Indicator taxa: *Prosopis*, *Chilopsis*
Patch Quality Variables		
Site Scale		
SiteOver (H)	NS	Linear extent of channel entirely overgrown (< 1 m height) with vegetation (percent)
Per_Rock (I)	0.0704	Bedrock substrate cover (percent)
AveGrain (J)	NS	Predominate substrate grain size (median of 5 categories: < 0.1, 0.1-0.5, 0.5-8, 8-30, > 30 cm)
Plot Scale		
PlotFlsu (K)	0.0142	Submerged or floating vegetation cover, including filamentous algae, mean (percent)
P_Avedep (L)	NS	Water depth, mean (cm)
P_Wetper (M)	NS	Wetted perimeter width, mean (m)
Plotsubm (N)	0.0251	Plot substrate size, for granular substrate, mean (log_{10} of cm)
PlotCanp (O)	0.0037	Vegetation cover over water, mean (percent)
PlotEmer (P)	NS	Emergent vegetation within 15 cm of point, mean (percent)
P_Cany (Q)	0.0006	Vegetation cover (< 1 m high) over adjacent land, mean (percent)
PLT_Rock (R)	0.0391	Bedrock substrate cover, mean (percent)
Water		
LogEC (S)	NS	Electrical conductivity (log_{10} of :S/cm)
PH (T)	0.0802	pH
Geographic Metrics		
Elevation (U)	0.0038	Elevation (m)
UTM-E (V)	0.0031	UTM-East coordinate (km), centered on mean (UTM - 659); approximately corresponds to longitude
UTM-N (W)	0.0006	UTM-N coordinate (km), centered on mean (UTM - 3989); approximately corresponds to latitude
Exotic Vegetation		
SiteTama (X)	0.0249	*Tamarix* spp. (exotic plant) linear extent (percent) in 400-m area
SiteTamu (Y)	0.0341	*Tamarix* spp. (exotic plant) linear extent (percent), in 40-m segments

P value represents univariate logistic regression analyses of the presence/absence of the toad with each environmental variable. NS indicates P > 0.10. Letter inside parentheses corresponds to the variable used in the SAS program for simplicity.

To relax the distributional assumption in GLM of a strictly linear relationship between response and independent variables, the general additive model (GAM) was introduced as an extension. It uses prediction via a nonparametric method. Fitting a model that accounts for local behavior may describe the behavior of the data more accurately than that described by a linear relationship. GAM uses different methods of smoothing to describe the data with little interference from the user. As described earlier, the link function in the GLM is in a linear form, whereas in GAM it is in an additive form. One of the common links in GAM is canonical and that describes the relationships between the transformed mean response and an independent variable using a nonparametric function as:

$$g\,(E\,(Y)) = g\,(u) = C_o + \mathbf{E}\;\; f_i\,(\mathrm{X}_i) \qquad (4)$$

where $g(.)$ is the link transform function, $E\,(Y)$ is the expected response, c_o is a constant (intercept), and f_i is a nonparametric function. The most commonly used link function is the canonical link (Agresti, 1996). The main difference between the GLM and GAM for binary data is that GLM-logistic regression assumes that the log odds are linearly related to the independent variables, whereas GAM assumes that the log odds are related to the sum of smooth functions of the independent variables.

A parametric method that is comparable to logistic regression analysis, but is used mainly to classify the observations into groups of populations, is known as discriminant analysis. For this data set, where the response variable is binary, discriminant analyses combine the independent variables linearly separating the data into two groups. This method requires that variance-covariance in each group is homogeneous, i.e., multivariate normal (Press and Wilson, 1978). Both discriminant and logistic regression analyses produce solutions in terms of probability of presence/absence (0-1 response as in the binary case). But the logistic model is preferable if multivariate normality is at all suspect. The logistic model requires more computational effort, but that is not of great concern. If multivariate normality holds, the coefficients of the discriminant model may have a smaller variance. In logistic regression, a maximum likelihood method for estimation is used, which does not require that independent variables be multivariate normal (see Maximum Likelihood Estimator (MLE) below).

MARS and Classification and Regression Tree (CART) are two common examples of GAM that have been used in many studies (Walker, 1990; Efron and Tibshirani, 1991; Moore et al., 1991; White and Sifneos, 1997). Both models can be used for regression modeling of binary response, but CART is more useful for classification than regression. Similar to the simple logistic regression, MARS has the backward, forward, and stepwise selections that help to choose the most related independent variables to the response. The stepwise selection is often preferred because a removed variable may have a chance to be included again.

In any GAM model, the final model is a summation of a group of functions that fit the data locally. The final model is data driven and represents closely the behavior of the data. The process of fitting and validating a model requires a large number of observations, especially when there are many independent variables relative to the number of observations (see text for MARS). For CART, for example, one needs more than 128 observations in a data set (Miller, 1994). We used MARS as a nonparametric method for logistic regression analysis, with the “stepwise” option for selecting the most significant variable, making it similar to programs for parametric logistic regression analysis.

Regression analysis and parametric generalizations make specific assumptions about the functional form of the relationship of the response to the predictions (e.g., linear); nonparametric regression, on the other hand, places minimal assumptions about the form of the relationship (e.g., Y is a smooth or additive function of the X’s). Thus, extensive computation and a decrease in theoretical framework are exchanged

for relaxation of assumptions. To decide which of the two methods to use (parametric versus nonparametric) depends on the user and the size of the data set. If the interest is for estimation and inference about the independent variables, then GLM is the preferred method. If the interest is to reveal structural behavior of the response with the independent variable, then GAM is the method of choice, especially when little is known about the nature of the relationship. Some studies have yielded similar results for parametric and nonparametric methods when applied to the same data sets (Bradford et al., 1998; Sheskin, 2000).

Below, we describe procedures for a general linear model with logistic regression analysis using SAS (SAS Institute, Cary, NC, version 8) and a general additive model (MARS) using a computer program also called *MARS* (Salford Systems, 1999, User Guide), which herein will be noted in *italics* to distinguish it from the statistical method of the same name. The *MARS* program can be easily used for continuous and binary dependent variables, handle missing values using surrogate variables, include all possible interactions, account for collinearity between independent variables, and prevent overfitting for the final model. This program has the algorithm to search for the basis function and knots and define the final optimal model with its statistics. This software offers the use of the Graphic User Interface (GUI), commands at the command prompt, and produces a classic output for the analyses.

1.1 Example Data Set Used

The main objective of conducting the regression analyses described above is to predict the presence/absence of the red-spotted toad (*Bufo punctatus*) in the northeastern Mojave Desert, USA, as a function of the surrounding environment (Bradford et al., submitted). There were 25 environmental variables that were used in this report (Table 1). These variables represent topography, patch size metrics, patch quality, exotic vegetation, and spatial direction. Although the total number of sites was 128, only 122 sites were used in the regression analysis because of missing values for some of the independent variables. Also, the data herein differed slightly from the data set and model used for biological interpretation (Bradford et al., submitted).

1.2 Logistic Regression

The response variable is dichotomous or binary (i.e., the presence or absence of toads). Our interest is in estimating the probability of the presence of toads as a function of many independent variables (see Table 1 for variable descriptions). The 25 independent variables were grouped based on habitat, scale, directional, and environmental characteristics. For simplicity hereafter, we will use alphabetic letters for the independent variables and will refer the reader to Table 1 for variable descriptions.

Equation (1) is a multiple logistic regression model that relates the proportion of presence (p(x)) as a function of independent variables (x_i's).

$$g(\mu) = Logit(p(x)) = Log\{ p(x) / (1 - p(x))\} = \beta_o + \Sigma \beta_i (x_i) \quad (5)$$

The coefficients β_o & β_i's in Equation 5 are estimated using maximum likelihood and are used to predict the probability of toad presence as a function of x_i's (Equation 6).

$$p = \exp(\beta_o + \Sigma \beta_i (x_i)) / (1 + \exp(\beta_o + \Sigma \beta_i (x_i)) \quad (6)$$

Detailed mathematical descriptions of the logistic regression analysis are given in Hosmer and Lemeshow (1989), Christensen (1997), and many others.

1.2.1 Maximum Likelihood Estimator

The Maximum Likelihood Estimator (MLE) is a method used with logistic regression analysis to estimate coefficients for the fitted model. The likelihood function for the logit model is as follows:

$$LogL = \sum \beta x_i y_i - \sum Log(1 + e^{\beta x_i}) \quad (7)$$

The coefficients βs are not linear in Equation 7. The likelihood function (Equation 7) is maximized by choosing a value of "β" in an iterative method, such as the Newton-Raphson method and the Fisher-scoring algorithm in SAS. The Fisher-scoring algorithm is the default method and Newton-Raphson is the alternative method in SAS. The former and latter methods use the observed and expected information matrix, respectively. However, when the data are binary, results are the same. Newton-Raphson is widely used by many in statistics and mathematics, and, therefore, both are explained below for the user. In the *Newton-Raphson* method, the first (U(β)) and second (I(β)) derivative of Equation 7 with respect to β is obtained and used in Equation 8 for estimation. The vector of the first derivative is called *gradients* or *score*, and the second derivative is called *Hessian*. The value of β will be estimated by the *Newton-Raphson* algorithm as:

$$\beta_{j+1} = \beta_j - I^{-1}(\beta_j) U(\beta_j) \quad (8)$$

To solve for β from Equation 8, an initial value of zero for the coefficient (β) is used to produce the first iteration estimate. This value is substituted back in the right-hand matrix (Equation 8) for the second iteration. Iterations continue until the difference between two consecutive iterations is less than or equal to a very small value, e.g., 0.001. More on maximum likelihood is given in Allison (1999, p. 36) and Hastie and Tibsherani (1990).

The *Fisher-scoring* algorithm is also known as the iteratively reweighted least squares algorithm. As mentioned earlier, the parameter estimate is the same using Newton-Raphson or Fisher-scoring, except that the covariance for the parameter estimates may not be the same. To start the estimation, a value of zero for the slopes (β) is used, and a value of logits of the observed cumulative proportions of the dependent variables is used for the intercept to produce the first iteration estimates. These values are substituted back for the second iteration. Iterations continue until the differences between two consecutive iterations are less than a very small value.

1.2.2 Assumptions

When standard regression analysis is fit to a set of data, there are basic assumptions on the errors (differences of the observed and the predicted values) that are considered normal, such as an expected mean value of zero, constant variance (homoscedasticity), no serial correlation, and no error in measurements. For multiple standard regression analysis, it is important to evaluate the collinearity between the independent variables (x_i's). There should not be "*perfect collinearity*" between the independent variables (Berry and Felman, 1985). A detailed discussion on multicollinearity is given below.

Logistic regression analyses are different from linear regression analyses because of the nature of the error. The assumption that differentiates it from linear regression is that the dependent variable is binary and has a binomial distribution with a single trial and parameter p(x) (e.g., probability of presence given x).

Logistic regression also requires that no correlation exists between observations (Hosmer and Lemeshow, 1989). If a number of observations are located close to each other and form a cluster, these observations may be dependent, and it is necessary to account for that in the model. Dependency will affect the standard error of the coefficient and make the estimate unstable. Luckily, there is a Generalized Estimation Equation (GEE) option in PROC GENMOD, SAS (Diggle et al., 1994) that accounts for the dependence between observations. The dependence between observations is run after finalizing the model and is described below.

Normality of the data is an issue of concern for many researchers. The assumption of the central limit theorem can be used when the sample size is large (> 30) and the distribution of a variable will approximate normality (Madansky, 1988). In standard linear regression analysis, however, basic assumptions need to undergo diagnostic checking on the residuals for normality or independence, for example. The link function in logistic regression allows the random component (response variable) to have a distribution other than normal. The prediction in logistic regression is done through the link function that models the mean response; the prediction is not obtained directly from the mean response as in standard linear regression. Logistic regression analysis uses the maximum likelihood estimators, which are approximately normal. Therefore, coefficients' *p*-values and confidence intervals can be estimated using the normal and chi-square distributions. Logistic regression does not require multivariate normality of the independent variables.

Linear relationships between dependent and independent variables were used in Equation 5. To linearize relationships between dependent and independent variables, log transformation of surface water area variable (A) was done as an example. Sheskin (2000) described many methods of data transformation. When nonlinear relationships exist and transformation does not linearize it, then nonparametric is the method to use, provided that the sample is large (Efron and Tibshirani, 1991).

1.2.3 Steps to Follow

Prior to running the logistic regression analysis, a few analyses are needed to understand the data better. Univariate logistic regression analysis is important to examine the strength of the relationships between the independent variables and the dependent variables. Collinearity is critical in regression analysis and needs to be studied first to exclude any collinear independent variables. Variable selection, goodness of fit, prediction of the model, and diagnostic check of the fitted model are explained in detail below.

a) Univariate Logistic Regression Analysis: Each of the variables in Table 1 was regressed with the presence/absence of toads. Univariate regression analysis reveals the strength of relationship and association of each independent variable with toad presence/absence. Table 1 gives the significance level for the slope coefficient for each of the independent variables. Eighteen of the 25 variables showed a significant relationship (p# 0.05) with the presence and absence of toads.

b) Collinearity: It is necessary to determine the magnitude of the collinearity of the independent variables. Collinearity can make the model coefficient unstable (Allison, 1999) and adversely affect the coefficient interpretation (Christensen, 1997), but it has no effect on the model prediction. A higher value of collinearity elevates the standard error of the estimated coefficient, which decreases the coefficient level of significance. In other words, a coefficient may be significant, but because of the presence of other correlated variables its significance may be diminished. Alternatively, a coefficient may be artificially significant and become insignificant when correlated variables are removed. Therefore, caution must be taken when explaining coefficients when a high degree of collinearity is present.

We examined collinearity in four ways:

1. First, we examined the simple pairwise (Pearson) correlation between the independent variables. Different cutoff values for r have been recommended as an indication of serious collinearity (e.g., *r* = 0.8, Berry and Felman, 1985; *r* = 0.9, Griffith and Amerhein, 1997). Belsley et al. (1980, p. 96) reported that *r* values < 0.7 should be considered with no fear of serious collinearity. However, Berry and Felman (1985) indicated that the cutoff value of r depends on the number of observations. When the number of observations is small (e.g., n < 30), then the recommended cutoff value of *r* is $ 0.70 and when the number of observations is > 30 then the cutoff of *r* value is $ 0.85. We used *r* $ 0.85 as the cutoff value for collinearity. If an independent variable is highly correlated with other independent variables (*r* $ 0.85), and is not highly associated with the dependent variable (presence of toads), that variable was not included in the logistic multiple regression analysis. Pairwise correlation values ranged from -0.77 to 0.77 for the independent variables that entered the final regression analysis model.

2. The second test for collinearity was using the Variance Inflation Factor (VIF) and Tolerance (VIF is the inverse of the tolerance (=1-R^2), Griffith and Amerhein, 1997) as an indication of the degree of collinearity between the independent variables. VIF represents the amount of inflation in the variance when the collinearity of a variable with others exists. A preliminary multiple regression analysis that includes all the independent variables can be used to examine VIF. A VIF value of one indicates that there is no linear relation (R^2 = zero) between the independent variables. A VIF of more than one means that R^2 is more than zero, which indicates some linear relation between the independent variables. VIF may increase to some value that makes the model unstable (i.e., "imprecise" in its prediction). A question may be asked: What is the cutoff value for VIF? Different values are given in the literature. Neter et al. (1996) and Griffith and Amerhein (1997, p. 99) indicated that a value of VIF that exceeds 10 can lead to a serious collinearity. For our analyses, we used VIF values of 10 as a cutoff for collinearity.

   Note: VIF is used later (page 13) for the diagnostic checking of the final logistic model.

3. The third test was to examine the absolute correlation between the coefficient estimates. A pairwise correlation r > 0.9 indicates that one variable has to be excluded from the regression analysis (Griffith and Amerhein, 1997). Griffith and Amerhein (1997) suggested that this test reveals a better indication of the linearity between a single variable and the linear combinations of others. The correlation between coefficient estimates can be outputted by adding an option CORRB to the Proc Logistic SAS statement. The highest r value (= 0.887) in our case was between salinity (S) and elevation (U) coefficients, and the lowest (= -0.72) was between latitude (W) and surface water area (A).

4. The fourth test was very simple and is used frequently by many researchers. It is performed by observing changes in the value and sign of a coefficient that is already in the model upon the addition of other independent variables. If a big change occurs in coefficient value or the sign is reversed, then this is an indication of collinearity. This method may be the best, though somewhat inefficient, since it is really model instability that is of concern.

Scientific knowledge is often useful in avoiding problems with collinearity. Although multicollinearity can occur with three or more variables, such that it masks its presence in any subset of two of these variables, it is most often seen with two. Hence, the first test (though without stringent cutoff values) is quite useful and simple.

c) *Variable Selection:* Stepwise selection (SAS) was used in selecting variables for the logistic multiple regression analysis. In this selection method, there are two probability values that control the variable to be entered (SLENTRY) or removed (SLSTAY) from the model. There were two important recommendations stating that the probability values for variable entry should be:

1. Higher than 0.05, giving an opportunity for an important biological variable to enter the model, and

2. Higher than that of removing the variable (Efroymson, 1960; Bendel and Afifi, 1977; Mickey and Greenland, 1989). A significance level of entry $ 0.25 has been recommended for stepwise regression analysis (Mickey and Greenland, 1989; Hosmer and Lemeshow, 1989). We used a value of 0.3 and 0.1 for SLENTRY and SLSTAY, respectively. We felt the significance level of # 0.1 was appropriate for identifying independent variables influencing the presence/absence of toads in the final model.

One key point to make is that some biological variables are important in explaining or predicting a biological phenomenon, yet their significance level may be more than 0.05 in the final model. A decision has to be made between a sound biological (or physical) model and a statistical model; therefore, it may necessitate increasing the significance level to more than 0.05.

Before finalizing the fitted modes, variables that are in the model have to be checked for VIF. A variable with high VIF should not be left in the model. A variable with high VIF would cause an unstable model, and, when deleted, would cause a major change in the coefficient estimates.

d) *Assessing Model Fit:* Proc Logistic (SAS) produces many statistics to test for predictiveness and effectiveness of the fitted model. Statistics such as Akaike Information Criterion (AIC), Shwarz Criterion (SC), and negative twice the log likelihood (-2 Log L) are given for the model with and without the independent variables (intercept only). Substantial reduction of these statistics, upon including independent variables, indicate a good predictive model. For example, the values of -2 Log L before and after including the independent variables were 144.38 and 55.58, respectively (Table 2). The difference in these two values is the maximum likelihood ratio $\mathbf{P}^2$ value (= 88.80) which was significant ($p < 0.0001$; Table 2). Analogous to a standard regression analysis, where the model F value is used to test the null hypothesis (Ho: all coefficients are equal to zero), we used Wald $\mathbf{P}^2$ in logistic regression analysis. Wald $\mathbf{P}^2$ is 21.81 ($p = 0.0027$; Table 2), rejecting the null hypothesis that states all coefficients are zero.

Table 2. SAS Output for Model Fit, Testing Global Null Hypothesis, and Association of Predicted Probability and Observed Toad Presence

Model Fit Statistics		
Criterion	**Intercept only**	**Intercept and Covariance**
AIC[a]	146.377 (=144.377 + 2*1)	71.583 (=55.583 + 2*8)
SC[b]	149.181 (=144.37 +1*Ln(122))	94.015 (=55.583 + 8*Ln(122))
-2 Log L	144.377	55.583
R-Square 0.5170	Max-rescaled R-Square 0.7453	

Testing Global Null Hypothesis: Beta = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	88.7946 (=144.377 - 55.583)	7	< 0.0001
Score	55.5509	7	< 0.0001
Wald	21.8141	7	0.0027

Association of Predicted Probabilities and Observed Responses			
Percent Concordance	95.4	Somers' D	0.909
Percent Discordant	4.5	Gamma	0.909
Percent Tied	0.0	Tau-a	0.368
Pairs	2992	c	0.954

[a] AIC = -2LogL + 2*k, where k is the number of coefficients in the model.

[b] SC = -2LogL + k*ln(n), where n is the total number of observations (SC is also known as Baysian Information Criteria).

The Wald chi-square statistic appears in logistic SAS output twice, once to test for the hypothesis that all coefficients are zero and again to test the significant level of each coefficient. A Wald statistic for the models is given in "*Testing Global Null Hypothesis,*" and for each coefficient is given in "*Analysis of Maximum Likelihood Estimates.*" For the calculations of the overall model Wald statistic, refer to Long (1997) and Rao (1973). For each coefficient, a Wald statistic is easily calculated as the squared ratio of the value of a coefficient and its standard error (Table 3).

It is important to note that when testing the best fitted model, it is safer to use a log likelihood ratio than the Wald test. When the coefficient is large, a Wald test can lead to Type II error. In general, the two statistics should not disagree substantially, and if they do, it may indicate that an asymptotic test is not appropriate. Pearson's chi-square is generally the best test of the overall model, but all fail for binary data. For nested models, the log likelihood ratio is generally preferred.

Table 3. Final Stepwise Logistic Regression Analysis Model (n = 122 sites)

Variable	Estimate ($)	Standard Error	*P*	Standardized Estimate	Odds Ratio	Variable Description
Intercept	10.9961	2.9378	0.0002			
U	-0.00858	0.00211	< 0.0001	-2.1904	0.991	Elevation
W	-0.0868	0.0211	< 0.0001	-1.7455	0.917	Latitude
S	-6.4869	1.8971	0.0006	-1.3683	0.002	Water salinity
I	0.1865	0.0635	0.0033	1.3399	1.205	Bedrock substrate cover
O	-0.0601	0.0170	0.0004	-1.0299	0.942	Vegetation over water
A	1.8175	0.7135	0.0109	0.7666	6.157	Surface water area
Q	-0.0383	0.0153	0.0125	-0.5487	0.962	Vegetation over adjacent land

Dependent variable is patch occupancy by red-spotted toad *(B. punctatus).* Variables are arranged in order of importance in influencing patch occupancy (i.e., by absolute value of a standardized estimate). **$** is a parameter estimate, and Odds Ratio is exp(**$**). P value for Wald statistic.

$$\text{standardized estimate} = \frac{\beta * std_x}{std_y}$$

Where β is the coefficient estimate in Table 3 (e.g., U = -0.00858),
std_x is the standard deviation of the independent variable (e.g., std for U = 463.159),
std_y is the standard deviation of the dependent variable (1.8138, see Allison, 1999),

$$\text{therefore the standardized estimate for U} = \frac{-\ 0.00858 * 463.159}{1.8138} = -2.1904$$

The Wald test for overall model significance, $\mathbf{P}^2 = 21.8$, df = 7, 114, P = 0.0027. Hosmer-Lemeshow goodness of fit test, $\mathbf{P}^2 = 3.91$, df = 8, P = 0.87. Percent of sites correctly classified as having present or absent populations is 95.4 percent.

R^2 is known in standard regression analysis as the coefficient of determination. In logistic regression analysis, it is known as the generalized R^2 and can be used to determine the predictive ability of the fitted model. To decide between models, choose the model with the highest R^2 value indicating a better predictive model. Allison (1999) used R^2 to identify the power of prediction of the fitted model. Logistic regression analysis R^2, known as the generalized R^2, is calculated differently than that of the standard regression analysis and should not be used to describe the percent of explained variability as in standard regression analysis.

R^2 = 1- *exp*(- *Likelihood ratio* $\mathbf{P}^2$ / *n*)
= 1- *exp*(- 88.80 /122) = 0.5170 or 52 percent (Table 2)

The generalized R^2 has an upper value which is not equal to one, as in the standard regression analysis. The logistic upper R^2 value is called "*Max rescaled R^2* " value and is determined by dividing the generalized R^2 by the upper R^2 value (see calculation below):

Upper R^2 = 1-*exp*(- *Intercept only (-2 Log L)* / *n*) =
= 1-*exp*(-144.377/122) = 0.6938 or 69 percent

Max-rescaled R^2 = 0.5170/ 0.6938 = 0.7453 or 75 percent (Table 2, Allison, 1999)

Measures of associations between observed and predicted values, also known as ordinal measures of associations, are given by SAS (Table 2). In explaining concordance, there are 7,381 pairs (122*121/2) of all possible combinations (presence/presence, absence/absence, and presence/absence). For concordance/discordance calculations, *presence/absence* pairs are considered (2,992 pairs). The model predicts the probability of presence, so if the probability of predicting presence is higher than that of absence in a pair, it is concordant. Otherwise, it is discordant. Our fitted model has a strong prediction for toad presence measured by a concordance of 95.4 percent.

There are many measures for the prediction ability of the fitted model to predict the presence of toads. These are measures of association and they are all high in values except for Tau-a (0.37; Table 2). Comparing Somers' D (0.91 percent) and R^2 (0.52 percent) values, there is a large difference between the two values, and the R^2 value is not very high. With a well-fit model in standard regression analysis, R^2 is expected to be high (e.g., > 90). Contrary to standard regression analysis, a low value of R^2 is expected in logistic regression analysis, even when applying a well-fit model to the data. Christiansen (1997, p. 128) suggested using R^2 as a *relative* measure of goodness of fit but not as a measure of the absolute goodness of fit. The value of logistic regression analysis R^2 becomes high only when the predicted probability approaches zero or one. Therefore, the value of R^2 can be used to compare two models and then to choose the one with higher value. The ability of the model to predict presence/absence can be assessed by the concordance.

The *goodness of fit* statistics is 3.905 with $p = 0.8656$, indicating a good fit (a Hosmer and Lemeshow goodness of fit test). When the probability is not significant ($p > 0.05$), a good fit is indicated. Thus, the logistic fitted model appeared to be "the right" model for presence/absence of toads.

e) *Coefficient Estimates:* Overall model $\mathbf{P}^2$ indicated that at least one coefficient is not zero. To test whether an independent variable in the model is significant, the Wald statistic was used (Table 3). All coefficients are significantly different from zero ($p < 0.013$; Table 3). An increase in surface water area and extent of bedrock substrate resulted in a higher probability for the presence of toads in the study area. Inversely, higher elevation, latitude, mean of percent vegetation cover over water, mean percent vegetation cover on the adjacent land, and electrical conductivity resulted in a lower probability of toads being present. Further interpretation of the physical relationships of these variables to the presence of toads is given in Bradford et al. (submitted).

Another statistic is the *relative importance* of the independent variables in the model, which is measured by the standardized estimate (Table 3). This statistic makes comparison possible across variables with different measurement units. Elevation is the most important environmental variable affecting the presence of toads, followed in importance by latitude (W), salinity (S), and percent bedrock substrate (I). A simple figure such as a box-plot can show the relationships between presence and absence for each of the independent variables (Figure 2). Values at quartiles 1, 2, and 3 for elevation, latitude, and vegetation cover over water over adjacent land are higher for sites where species is absent. *Odds ratio* can be defined as the ratio of odds for presence to the odds for absence. The predicted odds of toad presence for elevation are 0.991 times the odds of absence of toads. The independent variable in this study is a continuous variable, therefore, the odds ratio presents changes in toad present per one unit increase in the

independent variable. An increase of one unit (meter) in elevation decreases the predicted odds of presence of toads by 0.9 percent (1-0.991; Table 3). An increase of 100 units in elevation will change the predicted odds from 0.991 to 0.424 (exp(100*-0.00858 = 0.424). That is, with 100 more units in elevation, the predicted odds of presence are 0.424 times the odds of absence which is a decrease of 58 percent (1-0.424) in predicted odds of presence (see Hosmer and Lemeshow, 1989, pg. 63). In another example, the predicted odds of presence for bedrock substrate cover is 1.21 times the odds of absence. An increase of one unit (1 percent) in bedrock substrate cover increases the predicted odds of presence of toads by 21 percent.

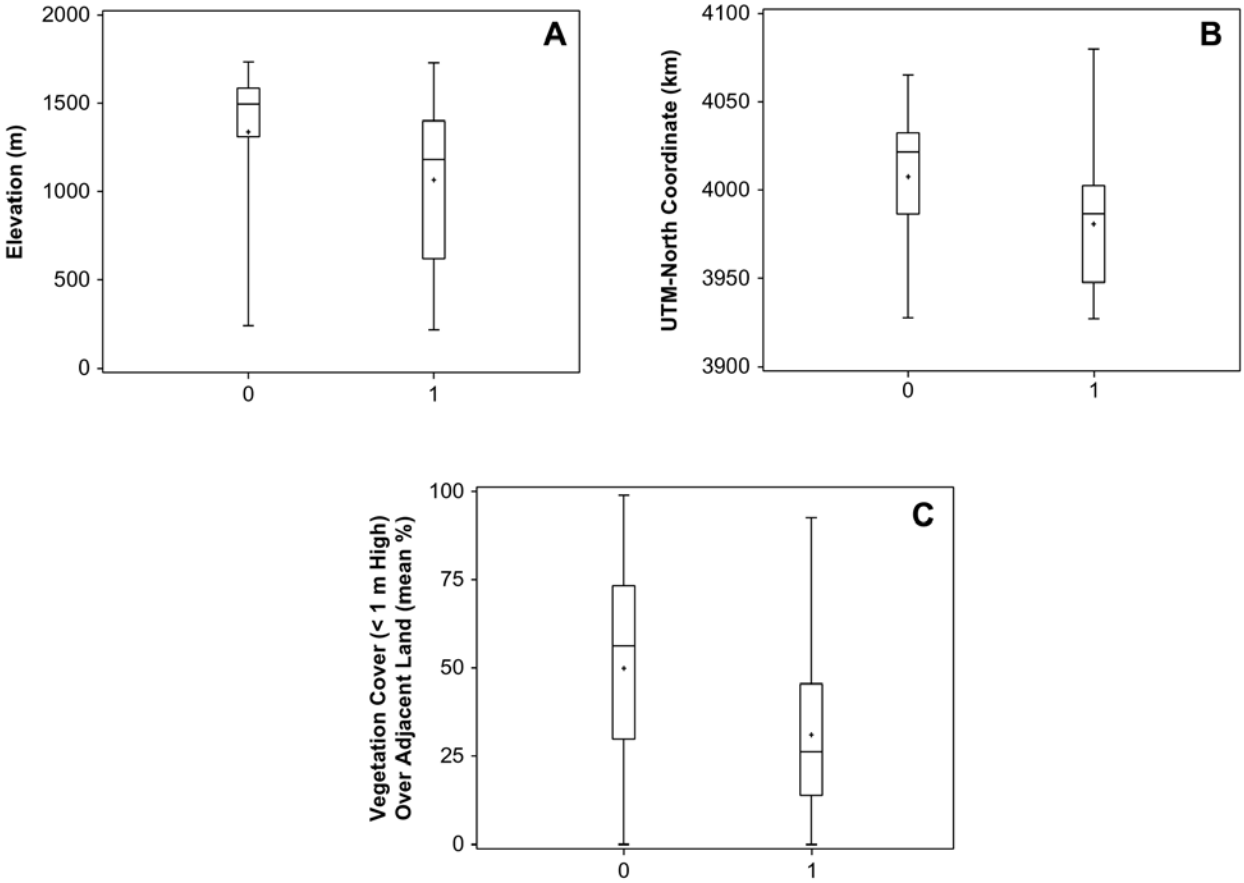

Figure 2. Box-plots show the presence (1) and absence (0) of toads as related to the independent variables. Independent variables are: Elevation (U), UTM-N (W; latitude), and vegetation cover (< 1 m high) over adjacent land, mean % (Q). "+" is the mean value; lines for box from top as: Maximum, Quartile 3, Quartile 2, Quartile 1 and minimum values.

f) Diagnostic Checking

1. VIF. Before analyzing the Logistic Regression analysis output, VIF for each of the independent variables in the final model was examined by incorporating a weighted value (variance of the binomial) into the VIF calculation to account for collinearity (Allison, 1999, p. 50). The steps below are suggested to first output the residuals into file 'O1' from the logistic regression analysis in SAS and run weighted linear regression analysis as:

   Data O2;

```
    Set O1;/*O1 is the residual from the regression equation */
    W  =  Pred*(1-Pred);
run;
Options ps  =  255 ls  =  100;
Proc Reg data  =  O2;
    Weight W;
    Model Toad  =  U W S I O A Q  / TOL VIF;
run;
```

VIF values for the independent variables in our case ranged from 1.3 to 8.0. These values are < 10, indicating no serious collinearity (Allison, 1999).

2. Deviance: A low value of the residual in linear regression analysis indicates that an observation is close to its predicted value. This is not the case in logistic regression analysis, and we may find a high residual value for normally distributed observations (Christensen, 1997). Using residuals to check for the fitted model as described in linear regression analysis is not appropriate for logistic regression analysis. SAS outputs many statistics to describe the influence of each observation on the fitted model and outliers. DFBETAS can detect an observation that causes instability for the model coefficient. DIFDEV and DIFCHISQ detect change in deviation and Pearson chi-square when an observation is deleted from the model. C and CBAR statistics are similar to that of Cook's D distance in the standard regression analysis, where a confidence interval is used to detect outliers.

   The Hat matrix diagonal is also known as leverage and is used to indicate how extreme the observation is as related to independent variables. High and low leverage are indications of a poor fit for these identified observations. A change in deviance and Pearson $\mathbf{P}^2$ with a value of more than 4 (value of 4 as the upper 95th percentile of the $^{a}\mathbf{P}^2$ distribution with one degree of freedom; $\mathbf{P}^2_{0.95} = 3.84$) needs to be considered and explained on the basis of the knowledge of these observations to make sense of the behavior. Sometimes, data like this may be deleted, but this behavior of the data should be investigated further: perhaps more data are needed. More data may result in a lower value (i.e., less than a value of 4). Plotting differences in Deviance and Pearson $\mathbf{P}^2$ are recommended to visualize the overall fitting of a covariate pattern and to look for clusters or points that are isolated from the overall pattern (Figures 3 and 4). In each figure there are two curved lines; the curve (left) that decreases with predicted probability describes the behavior of the covariates when toads are present, and the other (right) describes the pattern when they are absent. Figures 3 and 4 show that one value is more than 10 and 3 others are more than 4. These values were not removed from the data because of their physical importance to presence and absence of the toad. Hat matrix and DFBETAS are also output by SAS to show high magnitude values when an observation is deleted. More detail about this is given by Lemeshow and Hosmer (1989).

   One additional diagnostic check to make is on the variance homogeneity. This can be done by plotting the square root of the residual deviance vs. the predictor and examining the behavior of points (Figure 5). The points are scattered randomly over the domain with no clusters, indicating no heterogeneity in variance.

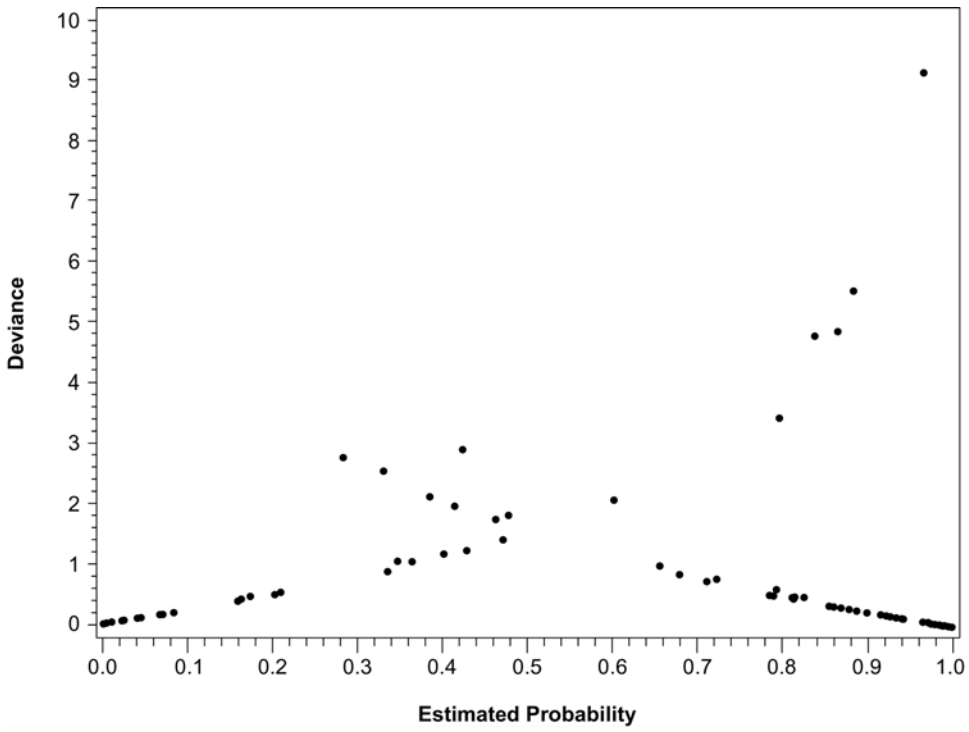

Figure 3. Deviance (DIFDEV) values for the predicted probability of the presence and absence of toads.

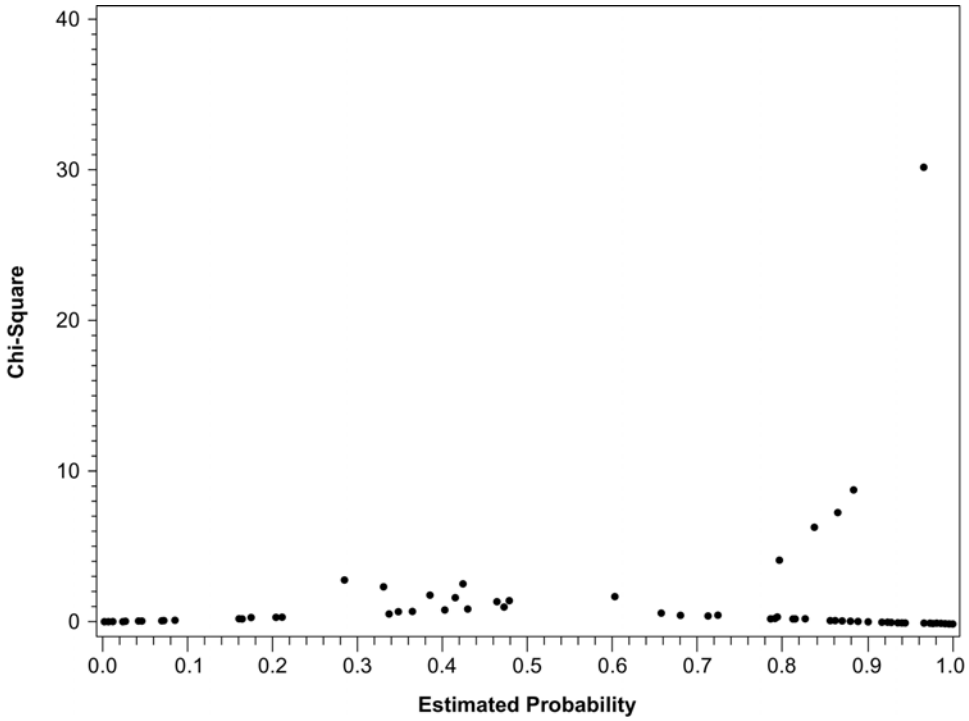

Figure 4. Chi-square (DIFCHISQ) values for the predicted probability of the presence and absence of toads.

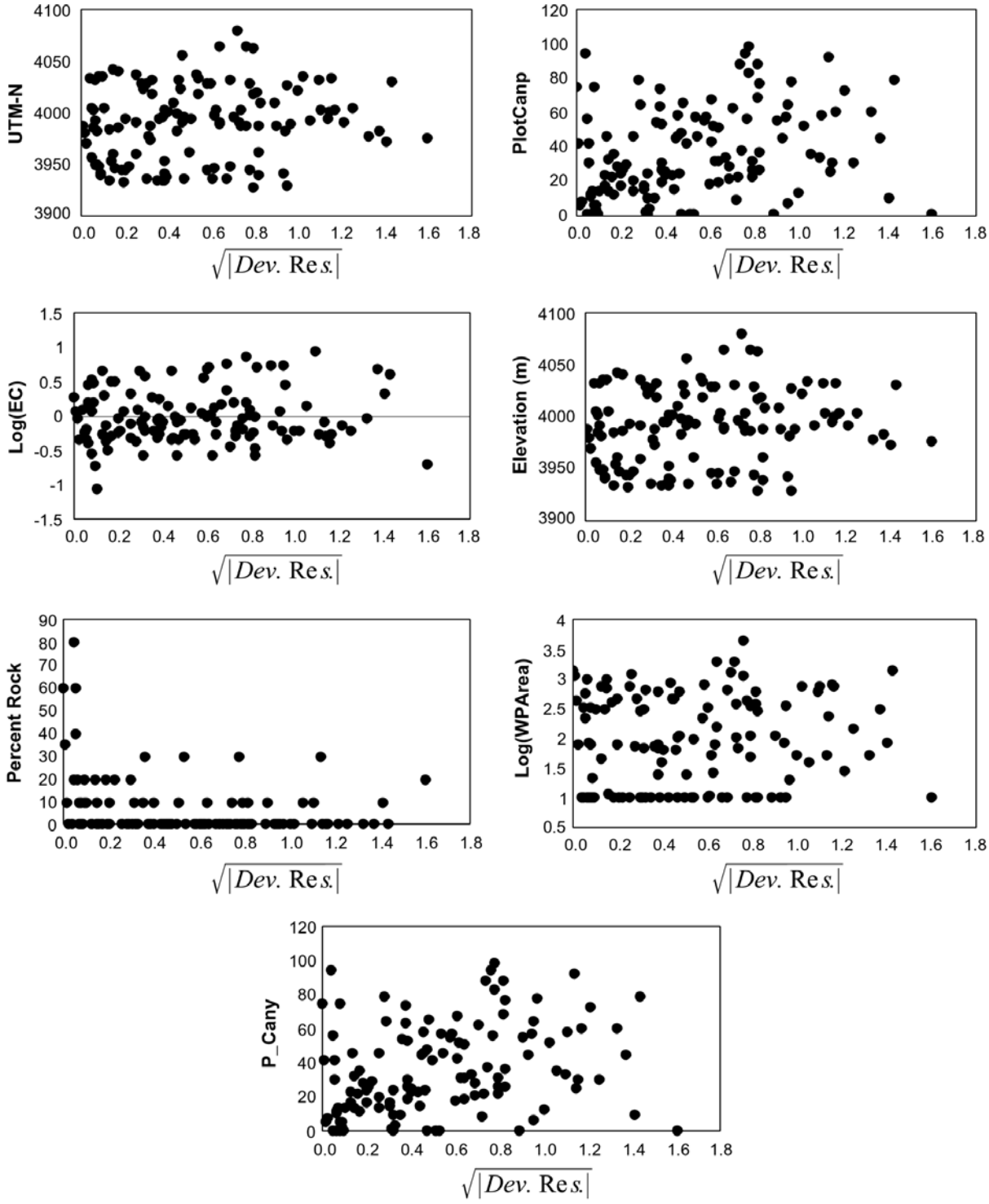

Figure 5. Square root of the absolute value of the deviance residual for each independent variable (x_i's) that was significant in the final model.

1.2.4 Model Selection

One may have many models in hand and would like to choose only one of them. A comparison has to be made between the saturated model (all independent variables are included in the model) and reduced models (model with a specific number of independent variables). Many statistics can be used in model selection such as R^2 and Wald statistics as mentioned earlier. An analog of Mallows *Cp*[1] that is used in standard linear regression (see footnote), *Cp* can be used in logistic regression to select a model. The *Cp* statistic in logistic regression is related to Akaike's criteria information (AIC). A model with the lowest *Cp* can be chosen as the best fitted model. Hosmer and Lemeshow (1989, p. 134) describe in detail the use of a modified *Cp* denoted as *Cq*. Lower *Cq* values indicate a better model to choose.

1.2.5 Dependence Between Observations

In logistic regression, it is important that observations are independent. If the dependence between observations in a group is of concern, then data may be grouped in a logical way (in our case watershed and mountain range) to examine whether sites/observations that are located near each other are dependent. The final model was run again using Proc GENMOD with the two groups/clusters (Appendix 1). Estimates of values and their probabilities were compared with that of Proc Logistic to note any change in coefficient values and their level of significance ($p > \chi^2$). If the estimate's significant levels change from significant to insignificant or vice versa, then dependence is an issue, and analysis has to be done by groups. Thus, in this situation one global logistic regression analysis would not be valid for all the observations. The following are SAS statements that use GEE through the *Repeated* option.

```
Options ps = 255 ls = 100;
Proc Genmod data = neweuc;
    Class WsGrp; /* by watershed group */
    Model Toad = A U W I O Q S / D = B;
    Repeated subject = Wsgrp / Type = exch;
run;
Options ps = 255 ls = 100;
Proc Genmod data = neweuc;
    Class Rangegrp; /* by mountain range group */
    Model Toad = A U W I O Q S / D=B;
        Repeated subject = Rangegrp / Type = exch;
run;
```

---

[1] Mallows *Cp* is a statistic that was developed by Mallows (1973) as: $Cp = \frac{RSS_{\beta}}{S^2} - (n-2\$)$

Where $RSS_{\$}$ is the residual sum of squares from regression analysis model;
S^2 is the residual mean square, an unbiased estimate of the error variance (σ^2);
$\$$ is the number of coefficients including an intercept; and
n is the number of observations.

Cp can be used to select the best fitted model. For a satisfactory model, *Cp* and $\$$ should be close in value. Normally, a "1-1" plot of *Cp* (y-axis) vs. $\$$ (number of coefficients; x-axis) is constructed to give an idea of the best fitted model (Draper and Smith, 1981, p. 300). If the model *Cp* value is above the "1-1" line, then it indicates a biased model. The best fitted models have a low *Cp* that is closer to the number of coefficients.

Results in our case indicated that estimates (coefficients) and their significant levels remain largely unchanged when either group was considered (Appendix 1). Therefore, it was concluded that there was no dependence between sites.

SAS statements for running the standard logistic regression analysis are given in Appendix 2.

1.2.6 Interactions

Interaction is referred to as the effect of the cross product of two independent variables on the mean response. For example, if the probability of presence as a function of latitude is dependent on elevation, then an interaction term (= latitude x elevation) must be included. Interaction between variables, therefore, should be considered if researchers deem it necessary to ensure that the final model is sound biologically and statistically. See Bradford et al. (submitted) for inclusion of interaction terms in the model.

Note: Researchers may find that there is a need to reduce the number of variables in the model, but this should not be done by arbitrary exclusions. An investigation of the magnitude of each coefficient may explain the importance of each variable. A coefficient's value describes the relation to the dependent variable and defines the rate of change of the dependent variable per one unit change in the independent variable. When two coefficients that are very close in value have opposite signs, the difference between them may be used in the model instead of the individual variables. One must define the statistical significance level and physical contribution of the new coefficient to decide whether the difference has better representation/contribution to the model. Then one can compare the new model R^2 with that of the original. These results may show some interesting biological phenomena.

Section 2

MARS

In the above logistic regression analysis, a linear relationship was assumed between the response variable, log(p(x)/1-p(x)), and the independent variables. A researcher may wish to explore if there are any nonlinear relationships between the response and the independent variables. That is, the response may have a piecewise behavior over the domain. In standard regression analysis, prior to fitting the model, the researcher has to specify the form of the nonlinear behavior (e.g., square). Modern nonparametric regression analysis can be used to model linear and nonlinear behavior without prior specification to the form of data behavior. To relax the assumption in GLM, the General Additive Model (GAM) was introduced. In GAM, the expected response value is a sum of smoothed functions of the independent variables. MARS and Classification and Regression Tree (CART) are two of the modern nonparametric regression analysis methods that are used in environmental studies (Walker, 1990; Moore et al., 1991; White and Sifneos, 1997; Miller, 1994). Computer algorithms, such as *MARS* (Salford Systems, 1999) and CART (Salford Systems, 1999), were developed to be used for data analysis. The selection option for variables in *MARS* (Salford Systems, 1999) algorithm and standard logistic regression (SAS) makes both methods similar in procedure for comparing results. We will use *MARS* in italics to differentiate the program from the method.

MARS, which was developed by Friedman (1991), is "*extremely promising as an automatic high dimensional smoother*" (Hastie and Tibshirani, 1990). It is data-driven more than user-driven, as in the case of simple regression analysis. For each of the variables, an algorithm is employed to determine the function of the variable (e.g., linear, cubic, etc.) by using a sequence of local nonparametric regression analysis on the data.

2.1.1 Model Fitting

In simple regression analysis, regression is done globally and may be sensitive to outliers. MARS, on the other hand, reduces the effect of outliers on the final model. It builds flexible models by fitting piecewise linear regression analysis to data to approximate the nonlinear behavior of the independent variables. Prior to building the model, we need to define and explain a few steps that need to be understood when running *MARS*.

1) Knots: When one regression line does not fit well to all the data, several regression lines (piecewise) are used to describe the overall behavior over the entire domain of the independent variable. The value of the independent variable where the slope of a line changes is called a *knot*. The *knot* defines the end of one domain and beginning of another. Between two knots, a linear (or cubic) regression line is fit to that group of data. When the slope is not changing along the entire domain, then one line fits all the data resulting in no knots.

2) Basis Functions: Basis functions are a set of functions used to reexpress the relations between dependent and independent variables. For example, basis function (BF_1) on the variable elevation is defined by MARS as:

$$BF_1 = \max(0, \text{elevation} - 219) \quad (9)$$

Data for elevation variables are grouped into two sets: the first set is assigned 0 for all elevation values that are below a threshold (e.g., c = 219 m), and the second set contains the elevation values that are more than 219 m. Elevation has no relation to the probability of presence (i.e., slope = 0) for values below the threshold of 219 m, but has a negative relationship (slope < 0) above this threshold.

3) Fitting a Model: The initial model starts with a constant only (c_o), then adds a basis function (a single or multivariate interaction term) in time to build up a comprehensive model that contains the maximum number of basis functions and their interactions, which are specified by the user (Equation 10).

$$Y = c_o + c_i * BF_i + \text{error} \quad (10)$$

where Y is the response variable, c_o is a constant and c_i is a coefficient for the basis function. Not all these basis functions contribute significantly to the model. The least contributing one is deleted by the stepwise method, and the final model contains the significant functions that contribute to the dependent variable. Analogous to CART, the initial model is overfitted, which is then pruned to give the optimum model with the lowest mean square error (MSE).

4) Fitting Logistic Regression: A modification to Equation 10, Hastie and Tibshirani (1990) introduced the logistic regression analysis using the modern additive logistic as:

$$\text{Log}[p(x)/(1-p(x))] = \$_o + \sum f_i(x_i) \quad (11)$$

where $\$_o$ is a constant, and $f_i(x_i)$ estimates local smooth functions; as mentioned earlier, this model can also contain interactions of order $ 2. Our task is to accurately predict the probability of the presence of toads given many independent variables. The same data (122 observations with 25 variables) that was used for the normal logistic regression analysis was used again in the *MARS* program.

5) Model Validation: With large sample sizes, the data are split into training (e.g., 90 percent of the data) and test sets (e.g., 10 percent of the data). The training data set is used for building the model, and the test data set is used to validate the fitted model. When the sample is not large, cross validation (CV) is the best method to use for validation. In CV, one observation is left out and smoothing is done on n-1 observations. The CV is the mean sum square of the differences between the Y_i's and their predicted values ($f^{-i}(x_i)$) where an observation is excluded:

$$CV = 1/n \sum \{Y_i - f^{-i}(x_i)\}^2 \quad (12)$$

The *MARS* outputs CV and PSE ($PSE = 1/n \sum \{Y_i - f(x_i)\}^2$) can be used to assess the final model. PSE is the mean Predictive Squared Error when all observations are included, whereas CV is a measure for n-1 observations. When CV and PSE are close in value, a minimum CV is reached and an optimal model is produced. Another measure for the optimal model is the *Generalized Cross Validation* (GCV) as a measure of mean square error. A model with minimum GCV is to be chosen. GCV is analogous to

Cp in simple regression analysis (see page 17) which is known as "*Mallows' Cp*" (see Hastie and Tibshirani (1990) for detail description of *Cp* and GCV).

2.1.2 Final Model

The model was run with 121-fold cross validation, 40 minimum numbers of observations between knots, and 15 maximum basis functions (these options are in the "Testing" option in GUI). The degrees of freedom is the number of observations between knots to be considered in smoothing and can be defined in three different ways (see Testing in *MARS* user guide GUI; Model-set-up, Model).

Different models can be applied with different options in GUI, and the model with the lowest PSE and GCV is chosen. PSE and GCV are given in *MARS* output (see Appendix 3 at the end of cross validation; *(estimated) PSE = 0.124 and GCV* = 0.136). Values of GCV and PSE are similar; therefore, we can conclude that an optimal model is reached.

The final model with all the above options yielded:

$$P(Y=1 * X) = f(U) + f(W) + f(Q) + f(S) + f(O) + f(C) + f(I).$$

The independent variables were similar to those by Proc Logistic (Table 4), except for A (surface water area) which was replaced by C (riparian zone area). Both of these metrics reflect the extent of a moist habitat. Also, the number of variables was similar for both methods. Input and output of *MARS* to this data set are given in Appendix 3.

Table 4. Coefficients and Their Statistics From MARS. Model F = 14.128, p < 0.001, df = 7,114

Variable	Estimates	t-ratio	P value	Variable description
Intercept	1.690	8.082	0.745E-12	
Basis function 1 (W)	-0.005	-5.981	0.261E-07	Latitude
Basis function 2 (U)	-0.405E-03	-4.335	0.316E-04	Elevation
Basis function 3 (O)	-0.003	-2.431	0.017	Vegetation over water
Basis function 4 (S)	-0.302	-2.902	0.004	Water salinity
Basis function 5 (C)	0.111	2.966	0.004	Riparian zone area
Basis function 6 (Q)	-0.004	-2.850	0.005	Vegetation over adjacent land
Basis function 7 (I)	0.005	2.237	0.027	Bedrock substrate cover

All 15 base variables with their contributions by GCV to the preliminary model are given in Appendix 3 ("Forward Stepwise Knot Placement"). GCV gives the amount of degradation in the model when a variable is deleted.

The importance of each variable in the model is another valuable output from *MARS*, which describes the amount of reduction in goodness of fit when any variable is removed. Latitude and elevation variables were the most important variables with 100 percent and 70 percent importance, respectively. Extent of riparian zone, salinity, and percent vegetation cover over adjacent land were similar in their importance (. 37 percent), whereas percent of vegetation cover over water and percent of bedrock

substrate were the least important (< 30 percent; see output in *Relative Variable Importance* section for these values, Appendix 3).

The Basis functions for the final model were:

	Minimum
BF1 = max (0, W - 3927.375)	3927.375
BF2 = max (0, U - 219)	219.0
BF3 = max (0, O - 0.585E-6)	0.0
BF4 = max (0, S + 1.053)	1.053
BF5 = max (0, C - 1.00)	1.00
BF6 = max (0, Q - 0.191E-5)	0.0
BF7 = max (0, I + 0.192E-6)	0.0

The above basis functions indicate the linear relationships between the dependent and independent variables (Figure 6). The basis functions all behaved linearly and started at the minimum values for each variable. In other words, there was no piecewise fitting. The final model is:

Y = 1.69 - 0.005*BF1 - 0.405E-3*BF2 - 0.003*BF3 - 0.302*BF4 + 0.111*BF5 - 0.004*BF6 + 0.005*BF7.

One important note to make here is that when different options in model specification in GUI were used, the sign for some variables was reversed in the final model. This is known as the collinearity effect between the independent variables. An option called "penalty" ("Option and Limits" Model in GUI) can be used when variables are added to reduce the potential of including collinear variables in a model.

For the final model, the "overall" F value was significant (F = 14.13, P = 0.425E-12, df = 7,114; Table 4; Appendix 3). There are three different values of R^2 which appear in the GUI and the output (Appendix 3) that can be used to describe the degree of association between the basis functions and the dependent variable, similar to that in ordinary least square regression. Therefore, these R^2 values are not appropriate for assessing the predictive power of the model when the response variable is a binary. These values, however, can be used to compare models. The description below for each R^2 is for clarification. In GUI, Naïve R^2 (final *MARS*-Model R^2), Naïve-Adjusted R^2, and GCV R^2 are given. In the output (Appendix 3), values of R^2, adjusted R^2, and the uncentered R^2 are given. An explanation of each is given below:

Naïve (GUI) and R-squared (Appendix 3): The degree of the association between the dependent variable and the basis functions. It was calculated in the same way as the R^2 for the simple ordinary least regression analysis.

Naïve-Adjusted R^2 (GUI) and Adjusted R^2 (Appendix 3): It is the same as the adjusted R^2 for the simple ordinary least regression analysis, adjusted for the number of basis functions.

GCV R^2: It is adjusted for the effective number of the parameters in the model, and it is always the lowest in value of the *MARS* R^2.

Uncentered R^2: It should not be used to test for goodness of fit. It is used in econometric diagnostic tests.

R^2 and Adjusted R^2 were 0.465 and 0.432, respectively. The R^2 values from *MARS* and simple logistic regression analysis are close to 50 percent. Values for the coefficients, Standard Error, T-Ratio and associated P-value are given in Appendix 3. All coefficients were significant ($P < 0.03$).

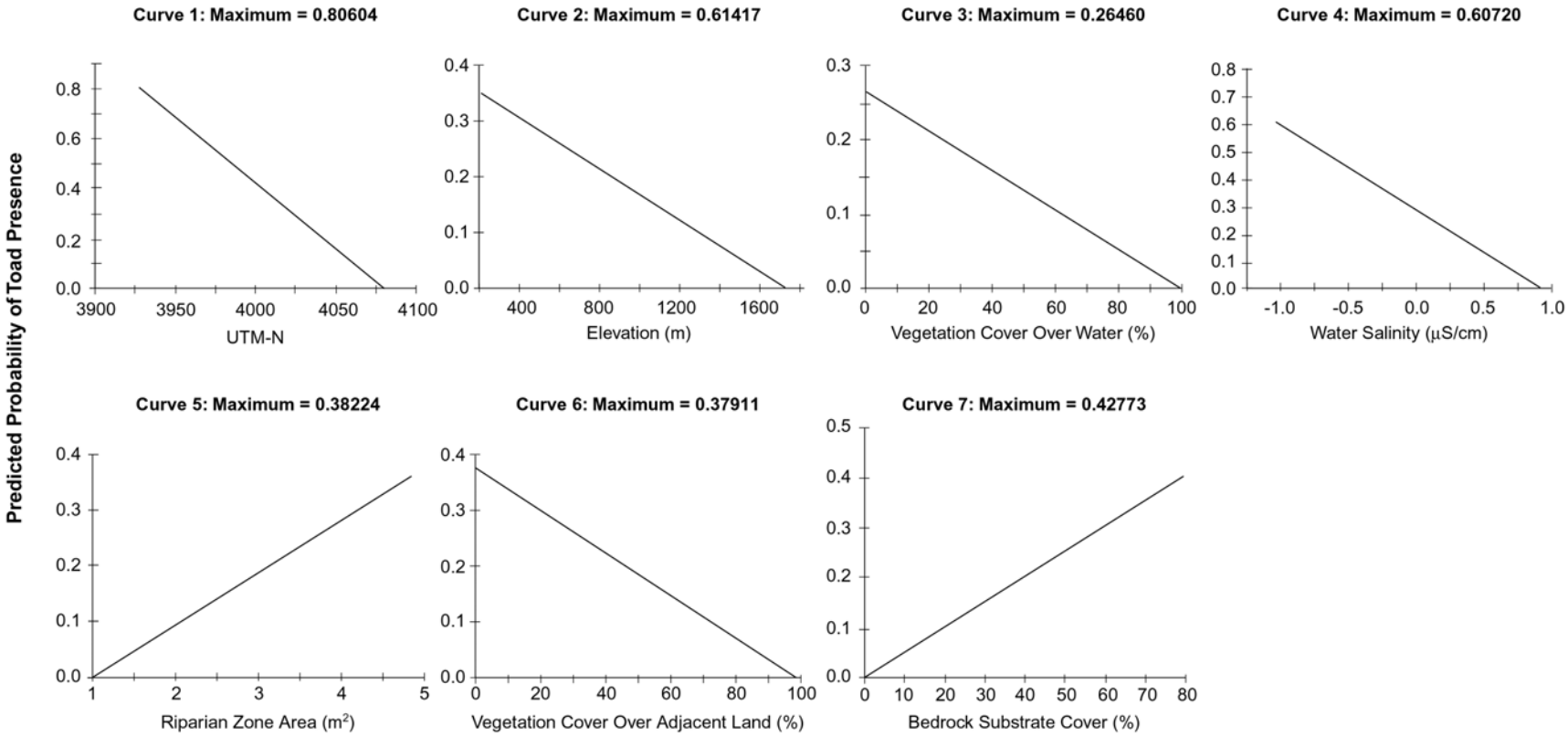

Figure 6. Regression for the main effect variables in the model using MARS.

Section 3
Conclusion

The goal of this report was to provide a reference manual to assist researchers in making informed use of logistic regression using the standard logistic regression and MARS. The details for analyses of relationships between presence/absence of an amphibian and environmental variables were exhibited in a manner to be used easily by nonstatistical users.

As noted, when comparing the standard logistic regression with another parametric method such as discriminant analysis, the former does not require multivariate normality, which often makes it more applicable to field data. Logistic regression using the nonparametric method, MARS, allows the user to fit a group of models to the data that reveal structural behavior of the data with little input from the user. Results using the standard regression (GLM) and general additive models (MARS) were similar for our example data set.

Logistic regression analysis was used to predict the probability of toad presence with respect to a number of environmental variables. Variable importance measured by the standardized estimate indicated that the geographical metrics (latitude and elevation) were the most important factors influencing toad presence, operating in a negative relationship. On the other hand, water salinity and percent bedrock substrate had lesser impacts on toad presence, with the former operating in a negative direction and the latter operating in a positive direction.

MARS is a nonparametric logistic regression analysis that is close procedurally to the simple parametric logistic regression analysis because of the variable selection through stepwise regression analysis. *MARS* and simple logistic regression analysis yielded similar models, and both indicated that latitude and elevation are the most important variables influencing toad presence. For this data set, the simple logistic regression analysis is the better method to be used because of the low number of observations (n = 122). It is recommended that if the degrees of freedom between knots were 10 and 20, the number of observations of 1,000 would be needed when using 30 variables (*MARS*; User guide, p. 34); that is, for each independent variable, 33 observations are needed. Although six independent variables were significant for both the simple logistic regression analysis and the nonparametric logistic regression analysis (*MARS*), and the R^2 values were similar, the application of MARS to this data set is still not superior to that of the parametric procedure.

On a final note, the decision to use simple logistic regression, *MARS*, or any other nonparametric method, has to be made on the basis of the suitability and interpretability of the statistical model to describe a phenomenon. If the interest is to infer the significance and importance of environmental variables in the model, then GLM logistic regression is the method to use. If the interest is to visualize and examine the structural relationship between a response and independent variable, especially when there is little or no knowledge about the data, then GAM is the method to use. A combination of the two methods may help reveal an important relationship in building the right statistical model.

References

Agresti, A. 1996. An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc. New York.

Allison, P.D. 1999. Logistic Regression Using the SAS System: Theory and application. SAS Inst. Inc., Cary, NC.

Belsley, D.A., E. Kuh, and R.E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York.

Bendel, R.B. and A.A. Afifi. 1977. Comparison of stopping rules in Forward "Stepwise" Regression. Journal of the American Statistical Association. 72:357, 46-54.

Berry, W.D. and S. Felman. 1985. Multiple Regression in Practice. Sage Publications, Beverly Hills.

Bradford, D.F., S.E. Franson, G.R. Miller, A.C. Neale, G.E. Canterbury, and D.T. Heggem. 1998. Bird species assemblages as indicators of biological integrity in Great Basin rangeland. Environmental Monitoring and Assessment. 49:1-22.

Bradford, D.F., A.C. Neale, M.S. Nash, D.W. Sada, and J.R. Jaeger. *Submitted to Ecology*. Habitat Patch Occupancy by the Red-Spotted Toad *(Bufo Punctatus)* in a Naturally Fragmented, Desert Landscape.

CART, 1999. Robust Decision-Tree Technology for data mining, predictive modeling and data preprocessing, Interface and documentation.

Christensen, R. 1997. Log-Linear Models and Logistic Regression. Springer, New York.

Diggle, P.J., K.Y. Liang, and S.L. Zeger. 1994. The analysis of longitudinal data. Oxford University Press, New York.

Draper, N.R. and H. Smith. 1981. Applied Regression Analysis, second edition, John Wiley & Sons, Inc. New York (p. 300).

Efron, B. and R. Tibshirani. 1991. Statistical data analysis in the computer age. Science 253:390-395.

Efroymson, M.A. 1960. "Multiple regression analysis" in Anthony Ralston and Herbert S. Wilf, eds., Mathematical Methods for Digital Computers, John Wiley & Sons, Inc. New York.

Friedman, J.H. 1991. Multivariate Adaptive Regression Splines (with discussion). Annals of Statistics 19, 1 (software).

Griffith, D.A. and C.G. Amerhein. 1997. Multivariate Statistical Analysis for Geographers. Prentice Hall, New Jersey.

Hastie, T.J. and R.J. Tibshirani. 1990. Generalized additive models. Chapman & Hall/CRC, New York.

Hosmer, D.W. and S. Lemeshow. 1989. Applied Logistic Regression. Wiley, New York.

Long, J.S. 1997. Regression Models for Categorical and Limited Dependent Variables. Sage Publications. London.

Mallows, C.L. 1973. Some comments on Cp. Technometrics, 15, 661-675.

Madansky, A. 1988. Prescriptions for working statisticians. Spring Verlag, New York.

Mickey, J. and S. Greenland. 1989. A study of the impact of confounder selection criteria on effect estimation. American Journal of Epidemiology, 129, 125-137.

Miller, T.W. 1994. Model selection in tree-structured regression in Proceedings of the Statistical Computing Section. ASA, p. 158-163.

Moore, D.M., B.G. Lees, and S.M. Davey. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Env. Manag. 5(1):59-71.

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. WCB McGraw-Hill, MA.

Press, S.J. and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. JASA, 73(364):699-705.

Rao, C.R. 1973. Linear statistics inferences and its application, second edition. Wiley, Inc., New York.

Salford Systems. 1999. MARS, User Guide. Cal. Stat. SoftWare, Inc., San Diego, California.

Sheskin, D.J. 2000. Handbook of Parametric and NonParametric Statistical Procedures, Second Edition. Chapman and Hall/CRD, New York.

Walker, P.A. 1990. Modeling wildlife distribution using a geographic information system: Kangaroos in relation to climate. Journal of Biogeography. 17:279-286.

White, D. and J. Sifneos. 1997. Mapping multivariate spatial relationships from regression trees by partitions of color visual variables. CSM 57th annual convention, ASPRS 63rd Annual Convention, Seattle, Washington. April 7-10, 1997. 86-95.

Appendix 1

GENMDOD Output

Dependence Between Observations
Based on Watershed Group
8:38 Thursday, April 19, 2001

The GENMOD Procedure

Model Information

Data Set	WORK.NEWEUC
Distribution	Binomial
Link Function	Logit
Dependent Variable	Toad
Observations Used	122
Probability Modeled	Pr (Toad = 1.0000)
Missing Values	6

Class Level Information

Class	Levels	Values
WSgrp	11	AMFL AMRI COLD COLO ELDO IVAN LASV MESQ PAHR PAIV VIRG

Response Profile

Ordered Level	Ordered Value	Count
1	1.0000	88
2	0.0000	34

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	A
Prm3	U
Prm4	W
Prm5	I
Prm6	O
Prm7	Q
Prm8	S

Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	114	55.5828	0.4876
Scaled Deviance	114	55.5828	0.4876
Pearson Chi-Square	114	74.4921	0.6534
Scaled Pearson X^2	114	74.4921	0.6534
Log Likelihood		-27.7914	

Algorithm Converged

Analysis of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	357.1798	85.5390	189.5266	524.8331	17.44	<.0001
A	1	1.8175	0.7135	0.4191	3.2158	6.49	0.0109
U	1	-0.0086	0.0021	-0.0127	-0.0044	16.45	<.0001
Y	1	-0.0868	0.0211	-0.1281	-0.0455	16.99	<.0001
I	1	0.1865	0.0635	0.0620	0.3109	8.62	0.0033
O	1	-0.0601	0.0170	-0.0933	-0.0268	12.54	0.0004
Q	1	-0.0383	0.0153	-0.0683	-0.0083	6.24	0.0125
S	1	-6.4872	1.8972	-10.2056	-2.7687	11.69	0.0006
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	WSgrp (11 levels)
Number of Clusters	11
Clusters with Missing Values	5
Correlation Matrix Dimension	34
Maximum Cluster Size	33
Minimum Cluster Size	1

Algorithm Converged

Analysis of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > (Z)
Intercept	354.2233	73.0343	211.0788	497.3679	4.85	<.0001
A	1.8578	0.4825	0.9122	2.8035	3.85	0.0001
U	-0.0084	0.0012	-0.0107	-0.0061	-7.22	<.0001
W	-0.0861	0.0181	-0.1216	-0.0507	-4.77	<.0001
I	0.1881	0.0857	0.0201	0.3561	2.19	0.0282
O	-0.0597	0.0083	-0.0759	-0.0434	-7.21	<.0001
Q	-0.0384	0.0236	-0.0846	0.0078	-1.63	0.1032
S	-6.3573	0.9988	-8.3149	-4.3998	-6.37	<.0001

Dependence Between Observations
Based on Mountain Range Group
8:38 Thursday, April 19, 2001

The GENMOD Procedure

Model Information

Data Set	WORK.NEWEUC
Distribution	Binomial
Link Function	Logit
Dependent Variable	Toad
Observations Used	122
Probability Modeled	Pr (Toad = 1.0000)
Missing Values	6

Class Level Information

Class	Levels	Values
Rangegrp	8	CLARK DRIVE ELDO KING MCHI RIVE SHEE SPRI

Response Profile

Ordered Level	Ordered Value	Count
1	1.0000	88
2	0.0000	34

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	A
Prm3	U
Prm4	W
Prm5	I
Prm6	O
Prm7	Q
Prm8	S

Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	114	55.5828	0.4876
Scaled Deviance	114	55.5828	0.4876
Pearson Chi-Square	114	74.4921	0.6534
Scaled Pearson X^2	114	74.4921	0.6534
Log Likelihood		-27.7914	

Algorithm Converged

Analysis of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	357.1798	85.5390	189.5266	524.8331	17.44	<.0001
A	1	1.8175	0.7135	0.4191	3.2158	6.49	0.0109
U	1	-0.0086	0.0021	-0.0127	-0.0044	16.45	<.0001
W	1	-0.0868	0.0211	-0.1281	-0.0455	16.99	<.0001
I	1	0.1865	0.0635	0.0620	0.3109	8.62	0.0033
O	1	-0.0601	0.0170	-0.0933	-0.0268	12.54	0.0004
Q	1	-0.0383	0.0153	-0.0683	-0.0083	6.24	0.0125
S	1	-6.4872	1.8972	-10.2056	-2.7687	11.69	0.0006
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	Rangegrp (8 levels)
Number of Clusters	8
Clusters with Missing Values	4
Correlation Matrix Dimension	55
Maximum Cluster Size	53
Minimum Cluster Size	4

Algorithm Converged

Analysis of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > (Z)
Intercept	367.1993	53.8241	261.7060	472.6926	6.82	<.0001
A	1.8007	0.5507	0.7215	2.8800	3.27	0.0011
U	-0.0088	0.0009	-0.0106	-0.0070	-9.60	<.0001
W	-0.0893	0.0133	-0.1154	-0.0632	-6.71	<.0001
I	0.1786	0.0210	0.1374	0.2197	8.51	<.0001
O	-0.0598	0.0089	-0.0772	-0.0425	-6.76	<.0001
Q	-0.0378	0.0239	-0.0847	0.0091	-1.58	0.1144
S	-6.2212	0.8184	-7.8252	-4.6173	-7.60	<.0001

Appendix 2

SAS Statements for Standard Logistic Regression

```
************  Comprehensive Final Model  *****************;
Options ps=255 ls=255;
Proc Corr data=neweuc;
   var A B C D E F G H I J K L M N O P Q R S T U V W X Y;
run;
Options ps=255 ls=150;
Proc Logistic data = neweuc  Descending;
   Model Toad = A B C D E F G H I J K L M N O P Q R S T U V W X Y
              / Selection = stepwise tech=newton details sle=0.30 sls=0.1
                Waldcl Waldrl Plcl Influence Iplot Lackfit Rsq CORRB ;
   output out=o1 predicted=pred difdev=dev difchisq=chi;
run;

Proc Print data=o1;
   var dev chi pred Toad;
run;

*****************************************************************;
****    Plotting  ************************************************;
goptions reset=global gunit=pct noborder
      ftext=Swissb htext=2;* horigin=0.2 in vorigin=0.2 in;

axis  length=40 order=(0 to 2000 by 500) label=(justify=c ' Elevation (m) ');
axis2 length=40 order=(0 to 1) label=(justify=c angle=-270 ' Presence/Absence');
Symbol1 v=dot       c=black h=1;
Proc gplot data=o1;
Format Toad F1.0 Elev2 F4.0;
   Plot Toad * U /
           frame
           haxis=axis
           vaxis=axis2;
        Title '   ';
        Footnote1 j=c 'Figure 1. Presence(=1) and absence(=0) of the amphibians ';
        Footnote2 j=r '  ';
        Footnote3 j=r '  ';
        Footnote4 j=r '  ';
        Footnote5 j=r 'pg  21';
run;
```

```
*************************************************************************;
*****    Plotting for Diagnostic Check  ********************************;
goptions reset=global gunit=pct noborder
       ftext=Swissb htext=2 horigin=0.2 in vorigin=0.2 in;

axis length=60;
axis2 length =60 label=(justify=c angle=-270 ' Deviance ');
Proc gplot data=o1;
    Plot Dev*Pred /
            frame
            haxis=axis
            vaxis=axis2;
            Symbol v=dot h=1.2 ;
         Title '   ';
         Footnote1 j=c 'Figure 2. Deviance (DIFDEV) values for the predicted probability for the study
area';
         Footnote2 j=r '  ';
         Footnote3 j=r '  ';
         Footnote4 j=r '  ';
         Footnote5 j=r 'pg';
run;
axis3 length=60 label=(justify=c angle=-270 ' Chi-Square');
Proc gplot data=o1;
*     format Per_min F3.0;
    Plot chi*pred /
            frame
            haxis=axis
            vaxis=axis3;
            Symbol v=dot h=1.2;
         Title '             ';

         Footnote1 j=c ' Figure 3. Chi-square (DIFCHISQ) values for the predicted probability for the study
area';
         Footnote2 j=r '  ';
         Footnote3 j=r '  ';
         Footnote4 j=r '  ';
         Footnote5 j=r 'pg';
run;
*************************************************************************;

**** Test for Collinearity for the independent variables in the final model  **********;
Data O2;
  Set O1; /* O1 is the residual from the regression equation */
  W=Pred*(1-Pred);
run;
Options ps=255 ls=100;
Proc Reg data = O2;
    Weight W;
    Model Toad = U W S I O A Q  / TOL VIF;
run;
```

```
***********    End of Test for Collinearity ***************************;

****      Test for the dependence between observations ************************;
Options ps=255 ls=100;

Title ‘ Based on Watershed groups ‘;
Proc Genmod data = neweuc ;
      Class WsGrp  ;
      Model Toad  = U W S I O A Q / D=B ;
          Repeated subject = Wsgrp / Type=exch;
run;

Title ‘ Based on Mountain Range groups ‘;
Proc Genmod data = neweuc ;
      Class Rangegrp;
      Model Toad  = U W S I O A Q / D=B ;
          Repeated subject = Rangegrp / Type=exch;
run;
```

Appendix 3

MARS Output

MARS Version 1.0.0.14

Reading Data, up to 292142 Records.

Records Read: 122
Records Kept in Learning Sample: 122

Learning Sample Statistics

(For variable description see Table 1)

Variable	Mean	SD	N	Sum
Toad	0.721	0.450	122	88.000
V	659.172	41.397	122	80418.953
W	3989.784	36.474	122	486753.597
U	1134.431	463.159	122	138400.552
H	9.180	20.076	122	1120.000
B	52.828	35.958	122	6445.000
D	35.492	35.258	122	4330.000
F	57.131	40.272	122	6970.000
E	17.500	32.367	122	2135.000
X	24.754	37.795	122	3020.000
G	29.303	35.199	122	3575.000
Y	23.402	36.592	122	2855.000
J	2.754	0.805	122	336.000
I	6.189	13.033	122	755.000
T	8.094	0.522	122	987.510
L	6.895	11.238	122	841.200
M	45.277	84.631	122	5523.800
Q	36.163	25.985	122	4411.900
N	31.311	33.763	122	3820.000

Continued....

Learning Sample Statistics, *Continued*

Variable	Mean	SD	N	Sum
K	36.221	35.146	122	4419.000
Q	23.336	31.104	122	2847.000
P	33.221	35.112	122	4053.000
R	11.484	23.568	122	1401.000
S	-0.005	0.383	122	-0.659
A	1.981	0.765	122	241.701
C	2.940	0.899	122	358.620

Ordinal Response

Variable	min	Q25	Q50	Q75	max
Toad	0	0	1	1	1

Ordinal Predictor Variables: 25

Variable	min	Q25	Q50	Q75	max
V	586.21	632.86	641.92	699.45	733.08
W	3927.38	3953.22	3990.68	4019.19	4080.23
U	219.00	620.00	1256.00	1515.00	1735.00
H	0.00	0.00	0.00	10.00	100.00
B	10.00	20.00	50.00	90.00	100.00
D	0.00	0.00	20.00	70.00	100.00
F	0.00	10.00	70.00	100.00	100.00
E	0.00	0.00	0.00	20.00	100.00
X	0.00	0.00	0.00	30.00	100.00
G	0.00	0.00	10.00	100.00	100.00
Y	0.00	0.00	0.00	30.00	100.00
J	1.00	2.20	2.90	3.10	5.00
I	0.00	0.00	0.00	10.00	80.00
T	5.55	7.80	8.10	8.40	9.37
L	0.20	1.40	3.00	9.10	92.80
M	0.20	1.00	2.60	60.40	458.00
Q	0.00	14.30	30.60	56.00	99.00
N	0.00	0.00	17.00	60.00	100.00
K	0.00	0.00	27.00	70.00	100.00

Continued....

Ordinal Predictor Variables: 25, *Continued*

Variable	min	Q25	Q50	Q75	max
Q	0.00	0.00	10.00	30.00	100.00
P	0.00	0.00	20.00	60.00	100.00
R	0.00	0.00	0.00	10.00	100.00
S	-1.05	-0.26	-0.08	0.20	0.96
A	1.00	1.00	1.91	2.65	3.66
C	1.00	2.52	3.17	3.60	4.45

121-fold cross validation used to estimate DF.
Estimated optimal DF(8) = 1.182 with (estimated) PSE = 0.124

Forward Stepwise Knot Placement

BasFn(s)	GCV	IndBsFns	EfPrms	Variable	Knot	Parent BsF
0	0.204	0.0	1.0			
1	0.185	1.0	2.8	W	3927.375	
2	0.164	2.0	4.6	U	219.000	
3	0.152	3.0	6.4	O	.584602E-06	
4	0.145	4.0	8.2	S	-1.053	
5	0.142	5.0	9.9	C	1.000	
6	0.138	6.0	11.7	Q	.191016E-05	
7	0.136	7.0	13.5	I	-.192323E-06	
8	0.139	8.0	15.3	T	5.550	
9	0.142	9.0	17.1	H	-.142210E-06	
10	0.145	10.0	18.9	A	1.000	
11	0.148	11.0	20.7	V	586.205	
12	0.152	12.0	22.5	D	-.955236E-06	
13	0.155	13.0	24.2	X	.954165E-06	
14	0.159	14.0	26.0	J	1.000	
15	0.164	15.0	27.8	F	.376499E-05	

Final Model

(After Backward Stepwise Elimination)

Basis Fun	Coefficient	Variable	Parent	Knot
0	1.69			
1	-0.005	W		3927.375
2	-0.405123E-03	U		219.00

Continued....

Final Model, *Continued*

Basis Fun	Coefficient	Variable	Parent	Knot
3	-0.003	O		0.584602E-06
4	-0.302	S		-1.053
5	0.111	C		1.000
6	-0.004	Q		0.0191016E-05
7	0.005	I		-0.192323E-06

Piecewise Linear GCV = 0.136, #efprms=13.518

ANOVA Decomposition on 7 Basis Functions

fun	std. dev.	-gcv	#bsfns	#efprms	Variable	Description
1	0.192	0.173	1	1.788	W	Latitude
2	0.187	0.153	1	1.788	U	Elevation
3	0.082	0.139	1	1.788	O	Vegetation Over Water
4	0.115	0.141	1	1.788	S	Water Salinity
5	0.099	0.142	1	1.788	C	Riparian
6	0.099	0.141	1	1.788	Q	Vegetation Over Adjacent Land
7	0.069	0.138	1	1.788	I	Bedrock Substrate Cover

Piecewise Cubic Fit on 7 Basis Functions, GCV = 0.136

Relative Variable Importance

	Variable	Importance	-gcv	Variable Description
2	W	100.000	0.173	Latitude
3	U	68.489	0.153	Elevation
25	C	39.559	0.142	Riparian
23	S	38.052	0.141	Water Salinity
17	Q	36.812	0.141	Vegetation Cover (< 1 m high) Over Adjacent Land
20	O	25.756	0.139	Vegetation Cover Over Water
13	I	19.493	0.138	Bedrock Substrate Cover
1	V	0.000	0.136	Longitude
4	H	0.000	0.136	Linear Extent of Channel with Vegetation
5	B	0.000	0.136	Surface Water Linear Extent
6	D	0.000	0.136	Emergent-type vegetation (*Typha*, *Eleocharis*, *Scirpus*, *Mimulus*, *Anemopsis*; *Juncus* & *Carex*) inside Stream Channel
7	F	0.000	0.136	Riparian Shrubs/Herbs (*Baccharis*, *Pluchea*, *Vitis*, *Allenrolfea*, *Equisetum*; *Juncus* or *Carex*) outside Stream Channel

Continued....

Relative Variable Importance, *Continued*

	Variable	Importance	-gcv	Variable Description
8	E	0.000	0.136	Native Riparian Trees (*Salix*, *Populus*, *Fraxinus*)
9	X	0.000	0.136	*Tamarix* spp. (exotic plant) in 400-m Area
10	G	0.000	0.136	Phreatophytes (*Prosopis*, *Chilopsis*)
11	Y	0.000	0.136	*Tamarix* spp. (exotic plant) in 40-m Segments
12	J	0.000	0.136	Predominate substrate grain size
14	T	0.000	0.136	pH
15	L	0.000	0.136	Water Depth
16	M	0.000	0.136	Wetted Perimeter Width
18	N	0.000	0.136	Plot Substrate Size, for Granular Substrate
19	K	0.000	0.136	Submerged or Floating Vegetation Cover
21	P	0.000	0.136	Emergent Vegetation within 15 cm of point
22	R	0.000	0.136	Bedrock Substrate Cover
24	A	0.000	0.136	Plot Surface Water Area

Basis Functions

BF1 = max (0, W - 3927.375);	Latitude
BF2 = max (0, U - 219.000);	Elevation
BF3 = max (0, O - .584602E-06);	Vegetation Over Water
BF4 = max (0, S + 1.053);	Water Salinity
BF5 = max (0, C - 1.000);	Riparian
BF6 = max(0, Q - .191016E-05);	Vegetation Over Adjacent Land
BF7 = max(0, I + .192323E-06);	Bedrock Substrate Cover

Y = 1.690 - 0.005 * BF1 - .405123E-03 * BF2 - 0.003 * BF3 - 0.302 * BF4 + 0.111 * BF5 - 0.004 * BF6 + 0.005 * BF7;

model DEPVAR = BF1 BF2 BF3 BF4 BF5 BF6 BF7;

Ordinary Least Squares Results

N:	122.000	R-Squared:	0.465
Mean Dep Var:	0.721	Adj. R-Squared	0.432

Uncentered R-squared = R-0 Squared: 0.851

Parameter	Estimate	S.E.	T-ratio	P-value
Constant	1.690	0.209	8.082	.745404E12
Basis Function 1	-0.005	.881709E-03	-5.981	260884E-07
Basis Function 2	-.405123E-03	.934568E-04	-4.335	.316045E-04
Basis Function 3	-0.003	0.001	-2.431	0.017
Basis Function 4	-0.302	0.104	-2.902	0.00
Basis Function 5	0.111	0.037	2.966	0.004
Basis Function 6	-0.004	0.001	-2.850	0.005
Basis Function 7	0.005	0.002	2.237	0.027

F-statistic	=	14.128	S.E. of Regression	=	0.339
P-value	=	.424771E-12	Residual Sum of Squares	=	13.132
[MDF, NDF]	=	[7,114]	Regression Sum of Squares	=	11.392