Research Papers Graduate School

2011

Plots and Prediction Intervals for Generalized Additive Models

Joshua E. Powers
Southern Illinois University Carbondale, jpowers@siu.edu

PLOTS AND PREDICTION INTERVALS FOR

GENERALIZED ADDITIVE MODELS

by

Joshua Powers

Bachelor of Science in Mathematics, Southeast Missouri State University, 2009

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
March, 2011

RESEARCH PAPER APPROVAL

PLOTS AND PREDICTION INTERVALS FOR

GENERALIZED ADDITIVE MODELS

By

Joshua Powers

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

Dr. David Olive, Chair

Dr. Sakthivel Jeyaratnam

Dr. Randy Hughes

Graduate School
Southern Illinois University Carbondale
April 7, 2011

ACKNOWLEDGMENTS

I would like to thank Dr. Olive for his invaluable assistance and insights leading to the writing of this paper. Many of the examples and definitions have been quoted directly from his book. My sincere thanks also goes to the members of my graduate committee, Dr. Jeyratnam and Dr. Hughes for their patience and time dedicated to read and discuss this paper. I would also like to thank Linda Gibson for all of her help during the writing of this paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In a *generalized additive model* (GAM), Y is conditionally independent of the predictors $\boldsymbol{x}$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some functions S_j. Plots for generalized linear models (GLM) using the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ can be extended to generalized additive models by replacing the ESP by the estimated additive predictor $EAP = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$. The response plot and transformation plots are examples. Since a GLM is a special case of a GAM, a plot of EAP versus ESP is useful for checking goodness of fit of the GLM.

The prediction intervals are for a future response Y_f given a vector $\boldsymbol{x}_f$ of predictors when the regression model has the form $Y_i = m(\boldsymbol{x}_i) + e_i$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are iid. The techniques perform well for moderate sample sizes as well as asymptotically.

This research paper gives information on presenting plots and asymptotically optimal prediction intervals for generalized additive models (GAM). In particular for the binomial, negative binomial, and Poisson models.

Chapter 1 gives information on the generalized linear model (GLM). It will give binomial, Poisson, and negative binomial models regarding the GLM. Then it will give information on generalized additive models (GAM), including the binomial, Poisson, and negative binomial models.

Chapter 2 introduces plots used to visualize the data involved in generalized additive models. It will also give several figures of such plots.

Chapter 3 deals with finding prediction intervals for the GAM, $Y_i = m(\boldsymbol{x}_i) + e_i$. It will give information as well as the results from a simulation used to find the prediction intervals.

CHAPTER 1

GENERALIZED LINEAR MODELS AND GENERALIZED ADDITIVE MODELS

1.1 INTRODUCTIONS TO GENERALIZED LINEAR MODELS

Following Olive [19, ch. 13], generalized linear models are a class of parametric regression models that include logistic regression and loglinear Poisson regression. Assume that there is a response variable Y and a $k \times 1$ vector of nontrivial predictors $\mathbf{x_i}$. Before we define a generalized linear model, the definition of a one parameter exponential family is needed. Let $f(y)$ be a probability density function (pdf) if Y is a continuous random variable and let $f(y)$ be a probability mass function (pmf) if Y is a discrete random variable. Assume that the *support of the distribution* of Y is $\mathcal{Y}$ and that the *parameter space* of θ is Θ.

Definition. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)exp[w(\theta)t(y)] \tag{1.1}$$

where $k(\theta) \geq 0$ and $h(y) \geq 0$. The functions h, k, t, and w are real valued functions.

It is crucial that in the definition, k and w do not depend on y and that h and t do not depend on θ. Note that the parameterization is not unique since, for example w could be multiplied be a nonzero constant m if t is divided by m. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $k(\theta)$ and $g(y)$ are positive, so another parameterization is

$$f(y|\theta) = exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{1.2}$$

where $S(y) = log(g(y)), d(\theta) = log(k(\theta))$, and the support $\mathcal{Y}$ does not depend on θ. Here the indicator function $I_{\mathcal{Y}}(y) = 1$ if $y \in \mathcal{Y}$ and $I_{\mathcal{Y}}(y) = 0$, otherwise.

Definition. Assume that the data is $(Y_i, \mathbf{x_i})$ for $i = 1, \ldots, n$. An important type of **generalized linear model (GLM)** for the data states that the $Y_1, \ldots, Y_n$ are independent random variables from a 1-parameter exponential family with pdf or pmf

$$f(y_i|\theta(\mathbf{x_i})) = k(\theta(\mathbf{x_i}))h(y_i) \exp\left[\frac{c(\theta(\mathbf{x_i}))}{a(\phi)} y_i\right] \tag{1.3}$$

Here ϕ is a known constant (often a dispersion parameter), $a(\cdot)$ is a known function, and $\theta(\mathbf{x_i}) = \eta(\alpha + \boldsymbol{\beta}^T \mathbf{x_i})$. Let $E(Y_i) \equiv E(Y_i|\mathbf{x_i}) = \mu(\mathbf{x_i})$. The GLM also states that $g(\mu(\mathbf{x_i})) = \alpha + \boldsymbol{\beta}^{\mathbf{T}} \mathbf{x_i}$ where the **link function** g is a differentiable monotone function. Then the **canonical link function** uses the function c given in (1.3), so $g(\mu(\mathbf{x_i})) \equiv c(\mu(\mathbf{x_i})) = \alpha + \boldsymbol{\beta}^T \mathbf{x_i}$, and the quantity $\alpha + \boldsymbol{\beta}^T \mathbf{x_i}$ is called the **linear predictor** and the **sufficient predictor** (SP).

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\mathbf{x_i}) = g^{-1}(\alpha + \boldsymbol{\beta}^T \mathbf{x_i}). \tag{1.4}$$

Also notice that the Y_i follow a 1-parameter exponential family where

$$t(y_i) = y_i \text{ and } w(\theta) = \frac{c(\theta)}{a(\phi)},$$

and notice that the value of the parameter $\theta(\mathbf{x_i}) = \eta(\alpha + \boldsymbol{\beta}^{\mathbf{T}} \mathbf{x_i})$ depends on the value of $\mathbf{x_i}$. Since the model depends on $\mathbf{x}$ only through the linear predictor $\alpha + \boldsymbol{\beta}^T \mathbf{x}$, a GLM is a 1D regression model: Y depends on $\mathbf{x_i}$ only through $\boldsymbol{\beta}^T \mathbf{x_i}$. Thus the linear predictor is also sufficient predictor.

1.2 EXAMPLES OF GENERALIZED LINEAR MODELS

In many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labeled as a 1 or a "success," while the non-occurrence of the category that is counted is labeled as a 0 or a "failure." For example, a "success"="occurrence" could be a person who died as a result from having cancer in a study. For a binary response variable, a binary regression model is often appropriate. This model is a special case of the binomial regression model with $m_i \equiv 1$.

Definition. The **binomial regression model** states that $Y_1, \ldots, Y_n$ are independent random variables with

$$Y_i \sim binomial(m_i, \rho(\mathbf{x_i})).$$

If the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x_i}$, then the most used binomial regression models are such that $Y_1, \ldots, Y_n$ are independent random variables with

$$Y_i \sim binomial(m_i, \rho(\alpha + \boldsymbol{\beta}^T \mathbf{x_i})),$$

or

$$Y_i | SP_i \sim binomial(m_i, \rho(SP_i)) \tag{1.5}$$

where the logistic regression model uses $\rho(SP) = \frac{e^{SP}}{1+e^{SP}}$.

If the response variable Y is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and Y_i is the number of a specified type of animal found in the subregion.

Definition. The **Poisson regression model** states that $Y_1, \ldots, Y_n$ are independent random variables with

$$Y_i \sim Poisson(\mu(\mathbf{x_i})).$$

The **loglinear Poisson regression models** is the special case where

$$\mu(\mathbf{x_i}) = \exp(\alpha + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x_i}). \tag{1.6}$$

If the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\mathbf{x_i}$, and $Y_1, \ldots, Y_n$ are independent random variables we have the Poisson model

$$Y_i \sim Poisson(\exp(\alpha + \boldsymbol{\beta}^T\mathbf{x_i})),$$

or

$$Y_i|SP_i \sim Poisson(\exp(SP_i)). \tag{1.7}$$

Some notation is needed for the negative binomial regression model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of Y is

$$P(Y = y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)}\left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa}\left(1 - \frac{\kappa}{\mu+\kappa}\right)^{y}$$

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\mathrm{V}(Y) = \mu + \mu^2/\kappa$. If $\tau = 1/\kappa$, then as $\tau \to 0$ the negative binomial distribution converges to the Poisson(μ) distribution.

Definition. The **negative binomial regression (NBR) regression model** states that $Y_1, \ldots, Y_n$ are independent random variables where

$$Y_i \sim NB(\mu(\mathbf{x_i}), \kappa).$$

with $\mu(\mathbf{x_i}) = \exp(\alpha + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x_i})$.

Now the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\mathbf{x_i}$, and $Y_1, \ldots, Y_n$ are independent random variables and we have the NBR model

$$Y_i \sim NB(\exp(\alpha + \boldsymbol{\beta}^T\mathbf{x_i}), \kappa),$$

or

$$Y_i|SP_i \sim NB(\exp(SP_i), \kappa). \tag{1.8}$$

1.3 GENERALIZED ADDITIVE MODELS AND EXAMPLES

Following Olive [21], *regression* is the study of the conditional distribution $Y|\boldsymbol{x}$ of the scalar response Y given the predictors $\boldsymbol{x}$. In a *1D regression model*, Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination of the predictors, called the *sufficient predictor* $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. See Cook and Weisberg [10, pp. 414-415].

In a *generalized additive model* (GAM), Y is conditionally independent of $\boldsymbol{x}$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some functions S_j. See Hastie and Tibshirani [13], Wood [29] and Zuur, Ieno, Walker, Saveliev and Smith [30]. Note that a 1D regression model is a special case of a GAM where $S_j(x_j) = x_j\beta_j$. The following examples are important.

1) The *multiple linear regression* model

$$Y|SP = SP + e \tag{1.9}$$

has GAM analog

$$Y|AP = AP + e. \tag{1.10}$$

2) For the binomial *logistic regression* model, $Y_1, ..., Y_n$ are independent with

$$Y_i|SP_i \sim \mathrm{binomial(m_i, \rho(SP_i))}. \tag{1.11}$$

The GAM analog is

$$Y|AP_i \sim \mathrm{binomial(m_i, \rho(AP_i))}. \tag{1.12}$$

The binary model is a special case with $m_i \equiv 1$.

3) For the *Poisson regression* model, $Y_1, ..., Y_n$ are independent random variables with

$$Y|SP \sim \mathrm{Poisson(exp(SP))}. \tag{1.13}$$

The GAM analog is

$$Y|AP \sim \mathrm{Poisson(exp(AP))}. \tag{1.14}$$

4) For the *negative binomial regression model*, $Y_1, ..., Y_n$ are independent random variables with

$$Y|SP \sim \text{NB}(\exp(\text{SP}), \kappa). \tag{1.15}$$

The GAM analog is

$$Y|AP \sim \text{NB}(\exp(\text{AP}), \kappa). \tag{1.16}$$

For a GLM, the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ while for a GAM, the estimated additive predictor $EAP = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$. It is well known that the residual plot of ESP or EAP versus the residuals (on the vertical axis) is useful for checking the model, but there are several other plots using the ESP that can be generalized to a GAM by replacing the ESP by the EAP.

Chapter 2 considers the response plot, plots for response transformations and additional plots such as the plot of the EAP versus the ESP.

CHAPTER 2

PLOTS FOR GENERALIZED ADDITIVE MODELS

This chapter follows Olive [21] closely.

2.1 RESPONSE PLOTS

Response plots are used to visualize 1D regression models in the background of the data. See Brillinger [3], Chambers, Cleveland, Kleiner and Tukey [6, p. 280], Cook and Weisberg [9],[10, ch. 18], and Olive and Hawkins [24]. For 1D regression, a response plot is the plot of the ESP versus the response Y with the estimated model conditional mean function and a scatterplot smoother often added as visual aids. Note that the response plot is used to visualize $Y|SP$ while a residual plot of the ESP versus the residual is used to visualize $e|SP$. For a GAM, these two plots replace the ESP by the EAP. Assume that the ESP or EAP takes on many values.

Suppose the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. For models (1.9) and (1.10) the estimated mean function is the identity line with unit slope and zero intercept. If the sample size n is large, then the plotted points should scatter about the identity line and the residual $= 0$ line in an evenly populated band for the response and residual plots, with no other pattern. For model (1.9), the two plots often look good if $n > 5p$. For the GAM, often much larger n is needed.

If $Z_i = Y_i/m_i$, then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the binomial regression model can be visualized with a response plot of the ESP versus Z_i with the estimated mean function of the Z_i

$$\hat{E}(Z|SP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. Cook and Weisberg [10] add a lowess curve to the plot.

Alternatively, divide the ESP into J slices with approximately the same number of cases in each slice. Then compute $\hat{\rho}_s = \sum_s Y_i / \sum_s m_i$ where the sum is over the cases in slice s. Then plot the resulting step function. For binary data the step function is simply the sample proportion in each slice.

The binomial GAM response plot is a plot of EAP versus Z_i with

$$\hat{E}(Z|AP) = \frac{\exp(EAP)}{1 + \exp(EAP)}$$

added as a visual aid. Lowess or the step function will also be added to the plot. For both the GAM and the GLM, the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(AP)$ or $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated mean) closely, then the logistic mean function is a reasonable approximation to the data. For the GLM, this plot is a graphical approximation of the logistic regression goodness of fit tests described in Hosmer and Lemeshow [14, pp. 147-151].

For Poisson regression, the response plot is a plot of ESP versus Y with $\hat{E}(Y|SP) = \exp(ESP)$ and lowess added as visual aids. The Poisson GAM response plot is a plot of EAP versus Y with $\hat{E}(Y|AP) = \exp(EAP)$ and lowess added as visual aids. For both the GAM and the GLM, the lowess curve should be close to the exponential curve, except possibly for the largest values of the ESP or EAP in the upper right corner of the plot. Here, lowess often underestimates the exponential curve because lowess downweights the largest Y values too much. Similar plots can be made for a negative binomial regression or GAM.

2.2 PLOTS FOR RESPONSE TRANSFORMATIONS

The applicability of the multiple linear regression model (1.9) or GAM (1.10) can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter λ_o, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i \tag{2.1}$$

where $E(Y_i|\boldsymbol{x}_i) = SP_i$ or $E(Y_i|\boldsymbol{x}_i) = AP_i$. If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow model (1.9) or (1.10) with p predictors. The $p \times 1$ vector $\boldsymbol{\beta}$ or the p functions S_j depend on λ_o, the p predictors x_j are assumed to be measured with negligible error, and the zero mean constant variance errors e_i are assumed to be iid from a unimodal distribution that is not highly skewed.

Next, two important response transformation models are given. Assume that *all* of the values of the "response" Z_i are *positive.* A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

The *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{2.2}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Often $Z_i^{(1)}$ is replaced by Z_i for $\lambda = 1$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as Λ_L. This family is a special case of the response transformations considered by Tukey [26].

A graphical method for response transformations computes the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$. If the plotted points follow the identity line for λ^*, then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. After selecting the transformation, the usual checks should be made. In particular, the transformation plot for the selected transformation is the response plot, and a residual plot should also be made. This technique is simple and can be used for regression methods with additive errors: $Y = t_{\lambda_o}(Z) = m(\boldsymbol{x}) + e$ where $m(\boldsymbol{x}) = E(Y|\boldsymbol{x})$. Olive [23] suggested the method for linear models including experimental design models.

Each transformation plot is a "response plot" for the seven values of $W_\lambda = t_\lambda(Z)$, and the method chooses the "best response plot" where the model (1.9) or (1.10) seems "most reasonable." If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding "residual plots" of $\hat{W}_\lambda$ versus $W_\lambda - \hat{W}_\lambda$ look reasonable. According to Mosteller and Tukey [18, p. 91], the values of λ in decreasing order of importance are $1, 0, 1/2, -1$ and $1/3$. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good. Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of λ_o by adding $\hat{\lambda}$ to Λ_L. For linear models, Box and Cox [2] is widely used.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge (e.g. in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey [26] showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid Λ_L. Useful powers are $\pm 1/4, \pm 2/3, \pm 2,$ and ± 3. Powers from numerical methods can also be added.

2.3 ADDITIONAL PLOTS

2.3.1 A Plot for Variable Selection

Variable selection is the search for a subset of variables that can be deleted without important loss of information. Olive and Hawkins [24] make an EE plot of ESP(I) versus ESP where ESP(I) is for a submodel I and ESP is for the full model. If model I is good, then the plotted points will follow the identity line with correlation near one.

Next we show that this result will hold for the plot of EAP(I) versus EAP. Assume that there exists a subset S of predictor variables such that if $\boldsymbol{x}_S$ is in the model, then none of the other predictors is needed in the model. Write E for these ('extraneous') variables not in S, partitioning $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$. Then

$$AP = \alpha + \sum_{j=1}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) + \sum_{k \in E} S_k(x_k) = \alpha + \sum_{j \in S} S_j(x_j). \qquad (2.3)$$

The extraneous terms that can be eliminated given that the subset S is in the model have $S_k(x_k) = 0$ for $k \in E$.

Now suppose that I is a candidate subset of predictors and that $S \subseteq I$. Then

$$AP = \alpha + \sum_{j=1}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) = \alpha + \sum_{k \in I} S_k(x_k) = AP(I),$$

(if I includes predictors from E, these will have $S_k(x_k) = 0$). For any subset I that includes all relevant predictors, the correlation $\text{corr}(\text{AP}, \text{AP(I)}) = 1$. Hence if the full model and submodel are reasonable and EAP and EAP(I) are good estimators of AP and AP(I), then the plotted points in the EE plot of EAP(I) versus EAP will follow the identity line with high correlation.

2.3.2 Plots for Checking the GLM

A plot of the estimated additive predictor $EAP = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$ versus the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ should be useful for checking the goodness of fit of the GLM since the GLM is a special case of the corresponding generalized additive model. The plotted points should follow the identity line with very high correlation if the GLM and GAM are roughly equivalent. If the correlation is not very high and the GAM has a nonlinear $\hat{S}_j(x_j)$, add x_j^2 and possibly x_j^3 to the GLM and remake the EAP versus ESP plot.

As another example, take a candidate GLM and fit the corresponding GAM. Since the GAM software can choose $S_j(x_j)$ to be general or linear $S_j(x_j) = x_j\beta_j$, choose all S_j to be linear except for S_k for $k = 1, ..., p$. Use the GAM software to check the shape of S_k for linearity. These p plots could be used to check the linearity of the x_j in the GLM, and the plots may be a competitor of the CERES plots described in Cook and Weisberg [10, ch. 16, p. 519].

2.3.3 A Plot for Checking Overdispersion

Conditional mean and variance functions are needed to study overdispersion. For binomial regression, the conditional mean function $E(Y_i|SP_i) = m_i\rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i\rho(SP_i)(1-\rho(SP_i))$. For the binomial GAM, the conditional mean function $E(Y_i|AP_i) = m_i\rho(AP_i)$ and the conditional variance function $V(Y_i|AP_i) = m_i\rho(AP_i)(1-\rho(AP_i))$. For Poisson regression, $V(Y|SP) = E(Y|SP) = \exp(SP)$. For the Poisson GAM, $V(Y|AP) = E(Y|AP) = \exp(AP)$. For negative binomial regression, $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right).$$

For the negative binomial GAM, $E(Y|AP) = \exp(AP)$ and

$$V(Y|AP) = \exp(AP)\left(1 + \frac{\exp(AP)}{\kappa}\right).$$

Overdispersion occurs when $V(Y|\boldsymbol{x})$ is larger than the model conditional variance function. Overdispersion can occur even if the model conditional mean function $E(Y|SP)$ or $E(Y|AP)$ is a good approximation to the data. For example, for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. Similarly, for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, it turns out that $V(Y|\boldsymbol{x}) > \exp(SP)$. See Cameron and Trivedi [5, p. 64].

To check for overdispersion in parametric models, we suggest using the *OD plot* of the estimated model variance $\hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. This plot has been used by Winkelmann [28, p. 110] for the Poisson regression model where $\hat{V}(Y|SP) = \hat{E}(Y|SP) = \exp(ESP)$. For binomial and Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi [5], Collett [8, ch. 6], and Winkelmann [28]. For a GAM, use the OD plot of the estimated model variance $\hat{V}(Y|AP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|AP)]^2$.

For the Poisson GAM, $\hat{V}(Y|AP) = \hat{E}(Y|AP) = \exp(EAP)$. For binomial regression, $\hat{E}(Y_i|SP_i) = m_i\rho(ESP_i)$ and $\hat{V}(Y_i|SP_i) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$. For the binomial GAM, $\hat{E}(Y_i|AP_i) = m_i\rho(EAP_i)$ and $\hat{V}(Y_i|AP_i) = m_i\rho(EAP_i)(1 - \rho(EAP_i))$. For negative binomial regression, $\hat{E}(Y|SP) = \exp(ESP)$ and

$$\hat{V}(Y|SP) = \exp(ESP)\left(1 + \frac{\exp(ESP)}{\hat{\kappa}}\right) = \exp(ESP) + \hat{\tau}\exp(2\ ESP).$$

For the negative binomial GAM, $\hat{E}(Y|AP) = \exp(EAP)$ and

$$\hat{V}(Y|AP) = \exp(EAP)\left(1 + \frac{\exp(EAP)}{\hat{\kappa}}\right) = \exp(EAP) + \hat{\tau}\exp(2\ EAP).$$

For generalized linear models, numerical summaries are also available. The deviance G^2 and Pearson goodness of fit statistic X^2 are used to assess the goodness of fit of the Poisson regression model much as R^2 is used for multiple linear regression. For Poisson regression (and binomial regression if the counts are neither too small nor too large), both G^2 and X^2 are approximately chi-square with $n-p-1$ degrees of freedom. Since a χ^2_d random variable has mean d and standard deviation $\sqrt{2d}$, the 98th percentile of the χ^2_d distribution is approximately $d + 3\sqrt{d} \approx d + 2.121\sqrt{2d}$. If G^2 or $X^2 > (n-p-1) + 3\sqrt{n-p-1}$, then overdispersion may be present.

For Poisson regression, Winkelmann [28, p. 110] suggested that the plotted points in the OD plot should scatter about the identity line and that the OLS line should be approximately equal to the identity line if the Poisson regression model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use.

First, recall that a normal approximation is good for the Poisson distribution if the count Y is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line. Similar remarks apply to negative binomial regression and also to binomial regression if the counts are neither too big nor too small.

Second, the evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 5 to 10 times that of the horizontal axis. (The scale of the vertical axis tends to depend on the few cases with the largest $\hat{V}(Y|SP)$, and $P[(Y - \hat{E}(Y|SP))^2 > 10\hat{V}(Y|SP)]$ can be approximated with a normal approximation or Chebyshev's inequality.) There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

Hence the identity line and slope 4 line are added to the OD plot as visual aids, and one should check whether the scale of the vertical axis is more than 10 times that of the horizontal. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot. For the Poisson, negative binomial and binomial GAM models, replace SP by AP.

2.3.4 Plots for the Poisson GLM and GAM

For the Poisson models, judging the mean function from the response plot may be rather difficult for large counts for two reasons. First, the mean function is curved. Secondly, for real and simulated Poisson regression data, it was observed that scatterplot smoothers such as lowess tend to underestimate the mean function for large ESP or EAP.

The basic idea of the following two plots for Poisson regression is to transform the data towards a linear model, then make the response plot and residual plot for the transformed data based on weighted least squares (WLS).

The *weighted forward response plot* is a plot of $\sqrt{Z_i}ESP$ versus $\sqrt{Z_i}\log(Z_i)$ where $Z_i = Y_i$ if $Y_i > 0$, and $Z_i = 0.5$ if $Y_i = 0$. The *weighted residual plot* is a plot of $\sqrt{Z_i}ESP$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}ESP$. The WLS residuals are often highly correlated with the deviance residuals. When the counts Y_i are small, the WLS residuals can not be expected to be approximately normal. Often the larger counts are fit better than the smaller counts and hence the residual plots have a "left opening megaphone" shape. This fact makes residual plots for Poisson regression rather hard to use, but cases with large WLS residuals may not be fit very well by the model. Both the weighted forward response and residual plots perform better for simulated Poisson regression data with many large counts than for data where all of the counts are less than 10.

To motivate the above two plots, recall that the minimum chi–square estimator $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ for Poisson regression is found from the WLS regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the OLS regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$. Then the plot of the "fitted values" $\sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\beta}}_M^T\boldsymbol{x}_i)$ versus the "response" $\sqrt{Z_i}\log(Z_i)$ should have points that scatter about the identity line. The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞ while the Poisson regression MLE tends to be consistent if the sample size $n \to \infty$. See Agresti [1, pp. 611-612]. Since the two estimators are often close for many data sets, the plotted points in the weighted forward response plot should scatter about the identity line if $\hat{E}(Y|SP) = \exp(ESP)$ is a good approximation to the mean function $E(Y|SP)$.

The Poisson GAM analogs for the two plots will plot $\sqrt{Z_i}$ EAP versus $\sqrt{Z_i}\log(Z_i)$ and $\sqrt{Z_i}$ EAP versus $\sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}$ EAP. Similar plots can be used for the negative binomial GLM and GAM.

2.4 EXAMPLES

Example 1. The ICU data is available from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). Also see Hosmer and Lemeshow [14, pp. 23-25]. The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN= Is cancer part of the present problem? (0 = No, 1 = Yes), CRN= History of chronic renal failure (0 = No, 1 = Yes), INF= Infection probable at ICU admission (0 = No, 1 = Yes), CPR= CPR prior to ICU admission (0 = No, 1 = Yes), SYS= Systolic blood pressure at ICU admission (in mm Hg), HRA= Heart rate at ICU admission (beats/min), PRE= Previous admission to an ICU within 6 months (0 = No, 1 = Yes), TYP= Type of admission (0 = Elective, 1 = Emergency), FRA= Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes), PO2= PO2 from initial blood gases (0 = >60, 1 = 60), PH= PH from initial blood gases (0 = 7.25, 1 <7.25), PCO= PCO2 from initial blood gases (0 = 45, 1 = >45), Bic= Bicarbonate from initial blood gases (0 = 18, 1 = <18), CRE= Creatinine from initial blood gases (0 = 2.0, 1 = >2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma). Factors LOC and RACE had two

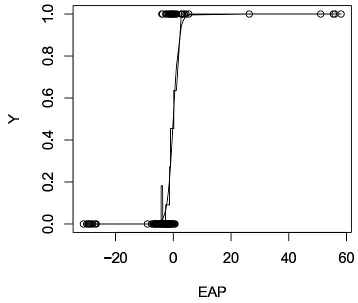

Figure 2.1. Visualizing the ICU GAM

indicator variables.

A binary generalized additive model was fit with unspecified functions for AGE, SYS and HRA and linear functions for the remaining variables. Output suggested that functions for SYS and HRA are linear but the function for AGE may be slightly curved. The response plot in Figure 2.1 shows that the step function of slice proportions tracks the model logistic curve fairly well. To visualize the model with the response plot, use $Y|\boldsymbol{x} \approx$ binomial$[1, \rho(EAP) = e^{EAP}/(1 + e^{EAP})]$. When $\boldsymbol{x}$ is such that $EAP < -5$, $\rho(EAP) \approx 0$. If $EAP > 5$, $\rho(EAP) \approx 1$, and if $EAP = 0$, then $\rho(EAP) = 0.5$. The logistic curve gives $\rho(EAP) \approx P(Y = 1|\boldsymbol{x}) = \rho(AP)$. The different estimated binomial distributions have $\hat{\rho}(AP) = \rho(EAP)$ that increases according to the logistic curve as EAP increases. If the step function tracks the logistic curve closely, the binary GAM gives useful smoothed estimates of $\rho(AP)$ provided that the number of 0's and 1's are both much larger than the model degrees of freedom so that the GAM is not overfitting.

A binary logistic regression was also fit, and Figure 2.2 shows the plot of EAP versus ESP. The plot shows that the near zero and near one probabilities are handled differently by the GAM and GLM, but the estimated success probabilities for the two models are similar: $\hat{\rho}(ESP) \approx \hat{\rho}(EAP)$.

Hence we used the GLM, and the response plot in Figure 2.3 shows that the

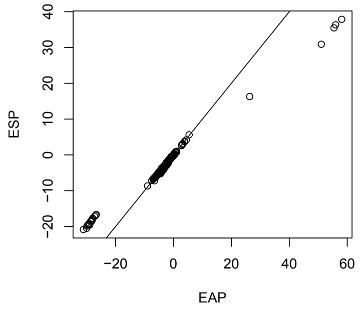

Figure 2.2. GAM and GLM give Similar Success Probabilities

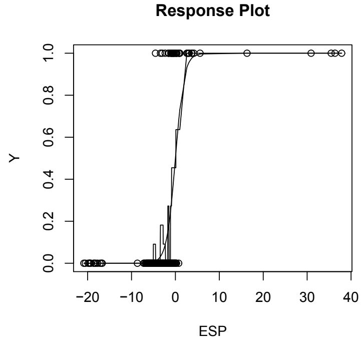

Figure 2.3. Visualizing the ICU GLM

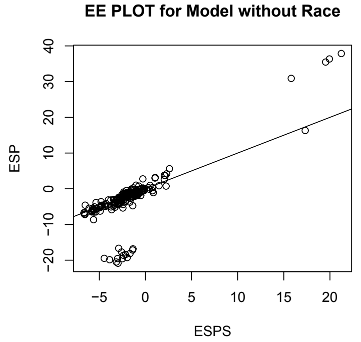

Figure 2.4. EE Plot Suggests Race is an Important Predictor

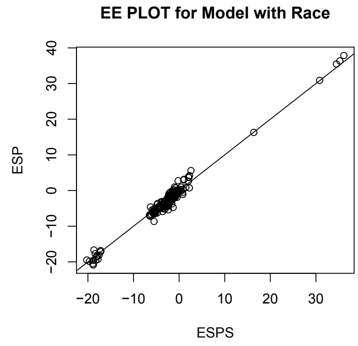

Figure 2.5. EE Plot Suggests Race is an Important Predictor

logistic regression model using the 19 predictors is useful for predicting survival. Note that the step function of slice proportions tracks the model logistic curve fairly well. Variable selection suggested the submodel using AGE, CAN, SYS, TYP and LOC. The EE plot of ESP(sub) versus ESP(full) is shown in Figure 2.4. Olive and Hawkins [24] show that the plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. This clustering did not occur in Figure 2.4. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black. Figure 2.5 shows the EE plot when RACE is added to the submodel. Then all of the points cluster about the identity line. Although variable selection did not suggest that RACE is important, the two EE plots suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black. This example shows the plots can be used to quickly improve and check the models obtained from variable selection.

Example 2. Chambers and Hastie [7, pp. 251, 516] examine an environmental study that measured the four variables Z = ozone concentration, solar radiation, temperature, and wind speed for 111 consecutive days. Generalized additive models are fit using Z and $Z^{1/3}$ as the response. Figure 2.6 shows the four best transformation plots. The residual plots in Figure 2.7 suggest that no transformation, $Y = Z$ may be best since the other transformations fit the case in the lower left corner poorly.

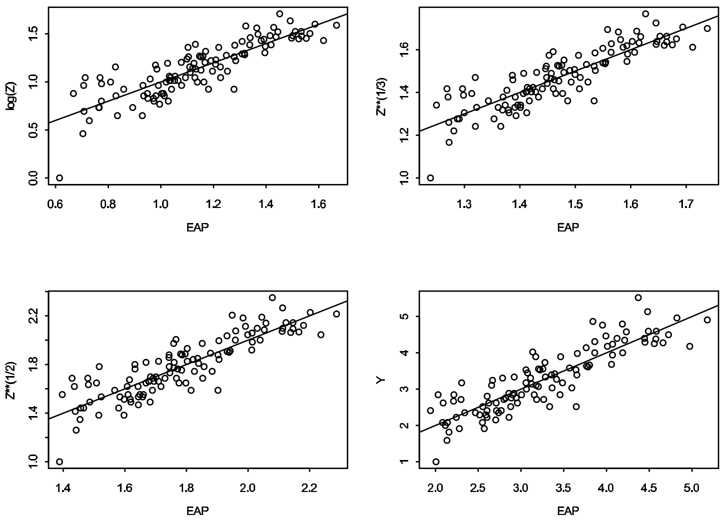

Figure 2.6. Transformation Plots for Ozone Data

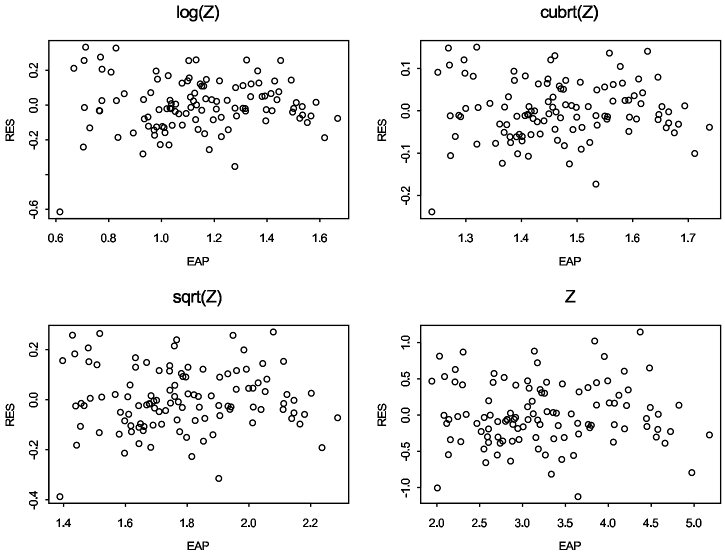

Figure 2.7. Residual Plots for Ozone Data

Example 3. For binary data, Kay and Little [16] suggest examining the two distributions $x|Y = 0$ and $x|Y = 1$. Use predictor x if the two distributions are roughly symmetric with similar spread. Use x and x^2 if the distributions are roughly symmetric with different spread. Use x and $\log(x)$ if one or both of the distributions are skewed. The log rule says add $\log(x)$ to the model if $\min(x) > 0$ and $\max(x)/\min(x) > 10$. The Gladstone [11] data is useful for illustrating these suggestions. The response was *gender* and a GLM with predictors *age, log(age), height* and the head measurements *circumference, length, size* and $\log(size)$ was used. The log rule suggested adding $log(age)$, and $log(size)$ was added because *size* is skewed. The GAM with these terms had plots of $\hat{S}_j(x_j)$ that were fairly linear. When the GAM was fit without $log(age)$ or $log(size)$, the $\hat{S}_j$ for *age*, *height* and *circumference* were nonlinear.

Example 4. Wood [29, p. 82-86] describes heart attack data where the response Y is the *number of heart attacks* for n_i patients suspected of suffering a heart attack. The enzyme *ck* (creatine kinase) was measured for the patients and it was determined whether the patient had a heart attack or not. A binomial GLM with predictors $x_1 = ck$, $x_2 = [ck]^2$ and $x_3 = [ck]^3$ was fit and had AIC = 33.66. Figure 2.8 shows that the EE plot for this model was not too good. The log rule suggests using *ck* and $\log(ck)$, but *ck* was not significant. Hence a GLM with the single predictor $\log(ck)$ was fit. Figure 2.9 shows the EE plot and Figure 2.10 shows the response plot where the $Z_i = Y_i/m_i$ track the logistic curve closely. There was no evidence of overdispersion and the model had AIC = 33.45.

Example 5. The species data is from Cook and Weisberg [10, pp. 285-286] and Johnson and Raven [15]. The response variable is the total number of species recorded on each of 29 islands in the Galápagos Archipelago. Predictors include area of island, *areanear* = the area of the closest island, the distance to the closest island, the elevation, and *endem* = the number of endemic species (those that were not

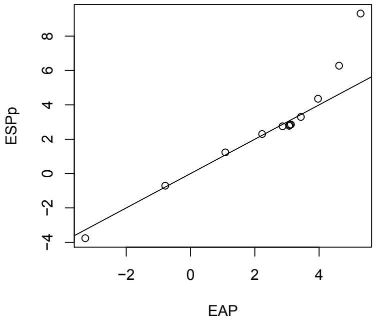

Figure 2.8. EE plot for cubic GLM Data

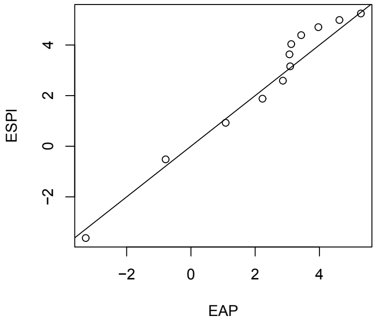

Figure 2.9. EE plot with log(ck) in the GLM

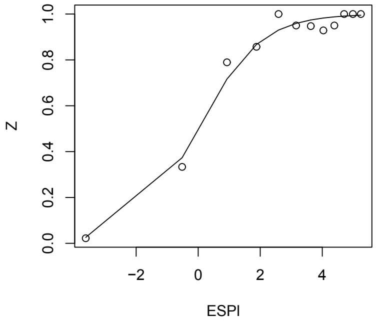

Figure 2.10. Response Plot for Heart Attack Data

introduced from elsewhere). A scatterplot matrix of the predictors suggested that log transformations should be taken. Poisson regression suggested that $log(endem)$ and $log(areanear)$ were the important predictors, but the deviance and Pearson X^2 statistics suggested overdispersion was present since both statistics were near 71.4 with 26 degrees of freedom. The residual plot also suggested increasing variance with increasing fitted value. A negative binomial regression suggested that only $log(endem)$ was needed in the model, and had a deviance of 26.12 on 27 degrees of freedom. The residual plot for this model was roughly ellipsoidal. The negative binomial GAM with $log(endem)$ had an $\hat{S}$ that was linear and the plotted points in the EE plot had correlation near 1.

The response plot with the exponential and lowess curves added as visual aids is shown in Figure 2.11. The interpretation is that $Y|\boldsymbol{x} \approx$ negative binomial with $E(Y|\boldsymbol{x}) \approx \exp(EAP)$. Hence if EAP $= 0$, $E(Y|\boldsymbol{x}) \approx 1$. The negative binomial and Poisson GAM and GLM have the same conditional mean function. If the plot was for a Poisson GAM, the interpretation would be that $Y|\boldsymbol{x} \approx$ Poisson($\exp(EAP)$). Hence if EAP $= 0$, $Y|\boldsymbol{x} \approx$ Poisson(1).

Figure 2.12 shows the OD plot for the negative binomial GAM with the identity line and slope 4 line through the origin added as visual aids. The plotted points fall within the "slope 4 wedge," suggesting that the negative binomial regression model has successfully dealt with overdispersion. Here $\hat{E}(Y|AP) = \exp(EAP)$ and $\hat{V}(Y|AP) = \exp(EAP) + \hat{\tau}\exp(2EAP)$ where $\hat{\tau} = 1/37$.

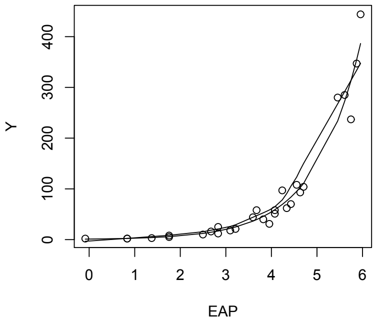

Figure 2.11. Response Plot for Negative Binomial GAM

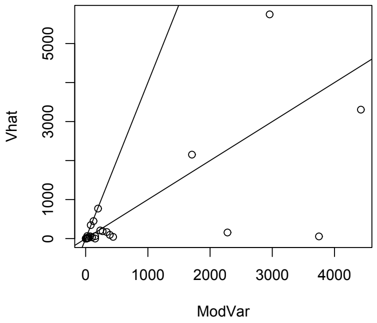

Figure 2.12. OD Plot for Negative Binomial GAM

CHAPTER 3

PREDICTION INTERVALS FOR GENERALIZED ADDITIVE MODELS

This chapter follows Olive [20] closely.

An important regression model is

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{3.1}$$

for $i = 1, ..., n$ where m is a function of $\boldsymbol{x}_i$ and the errors e_i are continuous and iid. Many of the most important regression models have this form, including the multiple linear regression model and many time series, nonlinear, nonparametric and semiparametric models. If $\hat{m}$ is an estimator of m, then the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$.

Olive [22] showed how to form asymptotically optimal prediction intervals for such models when the errors are iid from a continuous unimodal distribution. A problem with these intervals is that for many regression models and estimators, large n is needed for the intervals to perform well. Prediction intervals derived for multiple linear regression using least squares (OLS) did perform well. Olive [20] derives asymptotically optimal prediction intervals that perform well for many models for moderate n.

A large sample $100(1-\alpha)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1 - \alpha$ as the sample size $n \to \infty$. Following Olive [22], let ξ_α be the α percentile of the error e, i.e., $P(e \leq \xi_\alpha) = \alpha$. Let $\hat{\xi}_\alpha$ be the sample α percentile of the residuals. Consider predicting a future observation Y_f given a vector of predictors $\boldsymbol{x}_f$ where $(Y_f, \boldsymbol{x}_f)$ comes from the same population as the past data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let $1 - \alpha_2 - \alpha_1 = 1 - \alpha$ with $0 < \alpha < 1$ and $\alpha_1 < 1 - \alpha_2$ where $0 < \alpha_i < 1$. Then $P[Y_f \in (m(\boldsymbol{x}_f) + \xi_{\alpha_1}, m(\boldsymbol{x}_f) + \xi_{1-\alpha_2})] = 1 - \alpha$.

Assume that $\hat{m}$ is consistent: $\hat{m}(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ as $n \to \infty$. Then $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) \xrightarrow{P} Y_i - m(\boldsymbol{x}_i) = e_i$ and $\hat{\xi}_\alpha \xrightarrow{\mathrm{P}} \xi_\alpha$. If $a_n \xrightarrow{P} 1$ and $b_n \xrightarrow{P} 1$, then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + a_n \hat{\xi}_{\alpha_1}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\alpha_2}) \tag{3.2}$$

is a large sample $100(1-\alpha)\%$ PI for Y_f.

As an example, consider the multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown iid zero mean errors e_i with variance σ^2. Let the "leverage" $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$ and use the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{OLS}$ to find $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}_{OLS}$. Let $\hat{\xi}_\alpha$ be the sample quantile of the residuals. Following Olive (2007), let

$$a_n = b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}. \tag{3.3}$$

Then a large sample semiparametric $100(1-\alpha)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n \hat{\xi}_{\alpha/2}, \hat{Y}_f + a_n \hat{\xi}_{1-\alpha/2}). \tag{3.4}$$

A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. The PI (3.4) is asymptotically optimal on a large class of unimodal continuous symmetric error distributions. For more general distributions, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\alpha) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. See Grübel [12]. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\hat{\xi}_{\alpha_1}, \hat{\xi}_{1-\alpha_2})$ correspond to the interval with the smallest distance. Following Olive [22], a 100 $(1-\alpha)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n \hat{\xi}_{\alpha_1}, \hat{Y}_f + a_n \hat{\xi}_{1-\alpha_2}) \tag{3.5}$$

where a_n is given by (3.3). This prediction interval performs well for moderate n for multiple linear regression and least squares.

A problem with prediction intervals is choosing a_n and b_n so that the intervals have short length and coverage close to or higher than the nominal coverage for a wide variety of regression models when n is moderate. Section 3.1 shows how to modify (3.4) and (3.5) to achieve these goals.

3.1 ASYMPTOTICALLY OPTIMAL PREDICTION INTERVALS

The technique used to produce asymptotically optimal PIs that perform well for moderate samples is simple. Find $\hat{Y}_f$ and the residuals from the regression model. For a wide range of regression models, extrapolation occurs if $h_f > 2p/n$: if $\boldsymbol{x}_f$ is too far from the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, then the model may not hold and prediction can be arbitrarily bad. This result suggests replacing (3.3) by

$$a_n = b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n + 2p}{n - p}}. \tag{3.6}$$

Let $q_n = \min(1 - \alpha + 0.05, 1 - \alpha + p/n)$ for $\alpha > 0.1$ and

$$q_n = \min(1 - \alpha/2, 1 - \alpha + 10\alpha p/n), \quad \text{otherwise.} \tag{3.7}$$

Let $q_n = 1 - \alpha_n$. Then

$$(\hat{L}_n, \hat{U}_n) = (\hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{\alpha_n/2}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\alpha_n/2}) \tag{3.8}$$

is a large sample $100(1 - \alpha)\%$ PI for Y_f that is similar to (3.2) and (3.4).

Let $c = \lceil n q_n \rceil$. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\hat{\xi}_{\alpha_1}, \hat{\xi}_{1-\alpha_2})$ correspond to the interval with the smallest distance. Then the asymptotically optimal 100 $(1 - \alpha)\%$ large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{\alpha_1}, \hat{m}(\boldsymbol{x}_f) + b_n \hat{\xi}_{1-\alpha_2}), \tag{3.9}$$

and is similar to (3.5).

For asymptotic optimality, can not have extrapolation. If $\hat{m}$ is consistent so that $r_i - e_i \xrightarrow{P} 0$, then the coverage will converge to the nominal coverage, but the length need not be asymptotically shortest unless the highest $1-\alpha$ density region of the probability density function of the iid errors is an interval. Thus asymptotic optimality happens for unimodal distributions, but need not occur for multimodal distributions for fixed α. Also see Cai, Tian, Solomon and Wei [4].

Notice that the technique computes an asymptotically optimal PI for coverage $q_n > 1-\alpha$ which converges to the nominal coverage $1-\alpha$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% PI uses $q_n = 0.975$ while the nominal 50% PI uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variablity of the estimator $\hat{m}$. This variability is typically unknown but converges to 0 as $n \to \infty$. Letting the "coverage" q_n decrease to the nominal coverage $1-\alpha$ inflates the length of the PI for small n, compensating for the unknown variability of $\hat{m}$.

The geometry of the "asymptotically optimal prediction region" is simple. The region is the area between two parallel lines with unit slope. Consider a plot of $m(\boldsymbol{x}_i)$ versus Y_i on the vertical axis. The identity line with zero intercept and unit slope is $E(Y_i) = m(\boldsymbol{x}_i)$. Let (L_i, U_i) be the asymptotically optimal 95% prediction interval containing $m(\boldsymbol{x}_i)$. For example, if the errors are iid $N(0, \sigma^2)$, then $Y_i | m(\boldsymbol{x}_i) \sim N(m(\boldsymbol{x}_i), \sigma^2)$, and $(L_i, U_i) = (m(\boldsymbol{x}_i) - 1.96\sigma, m(\boldsymbol{x}_i) + 1.96\sigma)$. Then the upper line has unit slope and passes through $(m(\boldsymbol{x}_i), U_i)$ while the lower line has unit slope and passes through $(m(\boldsymbol{x}_i), L_i)$.

A response plot of $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ versus Y_i has identity line $\hat{E}(Y_i) = \hat{m}(\boldsymbol{x}_i)$. The region corresponding to pointwise prediction intervals is between two lines with unit slope passing through the points $(\hat{m}(\boldsymbol{x}_i), \hat{U}_i)$ and $(\hat{m}(\boldsymbol{x}_i), \hat{L}_i)$, respectively, where $(\hat{L}_i, \hat{U}_i)$ is the asymptotically optimal prediction interval (3.9) for Y_f if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Olive [22] suggested a similar plot for PIs (3.4) and (3.5), but the region was not between two parallel lines since the length of PIs (3.4) and (3.5) depends on h_f.

For the multiple linear regression model, expect the points in the response plot

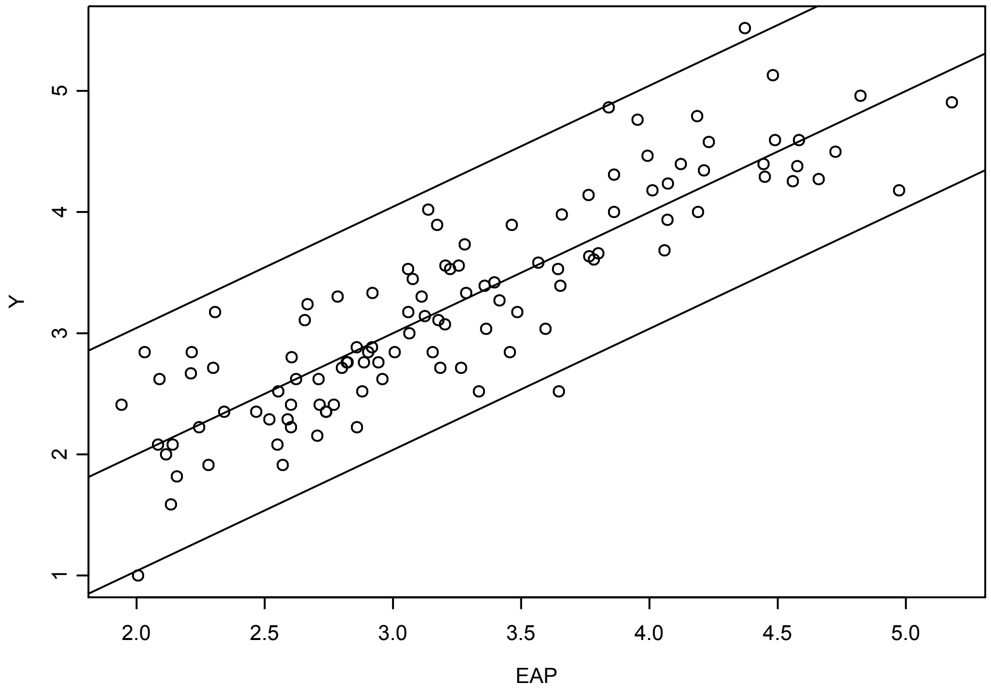

Figure 3.1. Pointwise Prediction Interval Bands for Ozone Data

to scatter in an evenly populated band for $n > 5p$. Other regression models, such as generalized additive models, may need a much larger sample size n.

Example 6. Chambers and Hastie [7, pp. 251, 516] examine an environmental study that measured the four variables Y = ozone concentration, solar radiation, temperature, and wind speed for $n = 111$ consecutive days. Figure 3.1 shows the response plot with the pointwise large sample 95% PI bands for the generalized additive model. Here $\hat{m}(\boldsymbol{x})$ = estimated additive predictor (EAP). Note that the plotted points scatter about the identity line in a roughly evenly populated band, and that 3 of the 111 PIs (3.9) corresponding to the observed data do not contain Y.

Three small simulation studies compares the PI lengths and coverages for sample sizes $n = 50, 100$ and 1000 for PIs (3.8) and (3.9). Values for PI (3.8) were denoted by scov and slen while values for PI (3.9) were denoted by ocov and olen. The five error distributions in the simulation were 1) N(0,1), 2) t_3, 3) exponential(1) -1, 4) uniform$(-1,1)$ and 5) $0.9N(0,1)+0.1N(0,100)$. The value $n=\infty$ gives the asymptotic coverages and lengths and does not depend on the model. So these values are same for multiple linear and nonlinear regression as well as generalized additive models.

The multiple linear regression model with $E(Y_i)=1+x_{i1}+\cdots+x_{i7}$ was used. The vectors $(x_1,...,x_7)^T$ were iid $N_7(\mathbf{0},\boldsymbol{I}_7)$ where $\boldsymbol{I}_p$ is the $p\times p$ identity matrix. For nonlinear regression $Y_i=m(\boldsymbol{x}_i)+e_i$, $E(Y_i)=m(\boldsymbol{x}_i)=\beta_1 x_{i1}+\beta_2 x_{i1}^2+\beta_3 x_{i2}+\beta_4 x_{i2}^2+\beta_5 x_{i3}+\beta_6 x_{i3}^2$. For the first generalized additive model, $m(\boldsymbol{x}_i)=\alpha+\sum_{j=1}^3 S_j(x_{ij})$. Both the nonlinear regression and generalized additive model had the same mean function $m(\boldsymbol{x}_i)=x_{i1}+x_{i1}^2$. Thus $\boldsymbol{\beta}=(1,1,0,0,0,0)^T$, $\alpha=0$, $S_1(x_{i1})=x_{i1}+x_{i1}^2$, $S_2(x_{i2})=0$ and $S_3(x_{i3})=0$. For these two models, the vectors $(x_1,x_2,x_3)^T$ were iid $N_3(\mathbf{0},\boldsymbol{I}_3)$. For the second generalized additive model, $m(\boldsymbol{x}_i)=\sin(x_{i1})+\cos(x_{i2})+\log(|x_{i3}|), \alpha=0, S_1(x_{i1})=\sin(x_{i1}), S_2(x_{i2})=\cos(x_{i2})$, and $S_3(x_{i3})=\log(|x_{i3}|)$. For the third generalized additive model, $m(\boldsymbol{x}_i)=\sqrt{|x_{i1}|}+\sqrt{|x_{i2}|}+\sqrt{|x_{i3}|}, \alpha=0, S_1(x_{i1})=\sqrt{|x_{i1}|}, S_2(x_{i2})=\sqrt{|x_{i2}|}$, and $S_3(x_{i3})=\sqrt{|x_{i3}|}$.

The Olive [22] PIs (3.4) and (3.5) are tailored for multiple linear regression but are liberal (too short) for moderate n for many other techniques. The new PIs (3.8) and (3.9) are meant to have coverage near or higher than the nominal coverage for moderate n and for a wide variety of techniques and are longer than PIs (3.4) and (3.5). For multiple linear regression, the new PIs (3.8) and (3.9) were conservative (too long with roughly 98% coverage for the 95% PI and 70% or 60% coverage for the 50% PI) for $n=50$ and 100 compared to (3.4) and (3.5) for least squares.

The PIs (3.8) and (3.9) for nonlinear regression and generalized additive models appear to have coverage near the nominal values in the simulations. For $n=50$ and 100, the PIs for nonlinear regression were usually roughly 10% longer than those for

Table 3.1. PIs for First Generalized Additive Model

error		95%	PI	95%	PI	50%	PI	50%	PI
type	n	slen	olen	scov	ocov	slen	olen	scov	ocov
1	50	5.123	4.997	0.959	0.958	1.852	1.668	0.586	0.521
1	100	4.702	4.524	0.963	0.956	1.656	1.523	0.548	0.498
1	1000	3.994	3.944	0.954	0.950	1.378	1.349	0.496	0.491
1	∞	3.920	3.920	0.95	0.950	1.349	1.349	0.50	0.50
2	50	9.351	8.567	0.955	0.946	2.373	2.147	0.572	0.528
2	100	8.273	7.625	0.963	0.953	2.041	1.877	0.565	0.518
2	1000	6.523	6.390	0.951	0.949	1.584	1.552	0.519	0.512
2	∞	6.365	6.365	0.950	0.950	1.530	1.530	0.50	0.50
3	50	5.157	4.800	0.956	0.947	1.562	1.273	0.605	0.525
3	100	4.647	4.148	0.965	0.955	1.381	1.062	0.593	0.544
3	1000	3.778	3.227	0.956	0.949	1.122	0.774	0.502	0.514
3	∞	3.664	2.996	0.950	0.950	1.099	0.693	0.50	0.50
4	50	2.626	2.589	0.959	0.954	1.228	1.078	0.590	0.491
4	100	2.318	2.271	0.972	0.964	1.156	1.027	0.555	0.492
4	1000	1.936	1.926	0.963	0.958	1.014	0.969	0.511	0.499
4	∞	1.900	1.900	0.950	0.950	1.00	1.00	0.50	0.50
5	50	19.766	17.835	0.949	0.938	2.962	2.678	0.597	0.533
5	100	18.724	16.169	0.951	0.940	2.342	2.157	0.576	0.530
5	1000	13.810	12.877	0.952	0.949	1.603	1.571	0.504	0.493
5	∞	13.490	13.490	0.950	0.950	1.507	1.507	0.50	0.50

generalized additive models.

The PIs for the generalized additive models were computed using the R function gam. See Hastie and Tibshirani [13] and Wood [29]. The PIs are asymptotically optimal for the five error distributions except for PI (3.8) with error type 3.

The simulations used 5000 runs and gave the proportion $\hat{p}$ of runs where Y_f fell within the nominal $100(1-\alpha)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1 - \delta_n)$ distribution where $1 - \delta_n$ converges to the asymptotic coverage $(1 - \delta)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0031$ and 0.0071 for $p = 0.05$ and 0.5, respectively. Hence an observed coverage $\hat{p} \in (.941, .959)$ for 95% and $\hat{p} \in (.479, .521)$ for 50% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Table 3.1, 3.2, and 3.3 show that for $n = 1000$, the coverages and lengths are near the asymptotic $n = \infty$ values. For tables 3.1 and 3.3, the 95% PI (3.9) coverages were in or near (.94,.96) while the 50% PI (3.9) was sometimes slightly conservative. The coverage for the 50% PI (3.8) was near 60% for $n = 50$. For table 3.2, the (3.9) coverage was sometimes a bit low for $n = 50$. PI (3.9) is recommended since its asymptotic optimality does not depend on the symmetry of the error distribution.

Simulations were done in Splus and R. See MathSoft [17] and R Development Core Team [25]. The programs in the collection of functions rpack.txt are available at (www.math.siu.edu/olive/ol-bookp.htm). For multiple linear regression, pisim simulates PIs (3.4) and (3.5) while the Splus function pisim4 simulates PIs (3.8) and (3.9) using OLS, L1 and M-estimators. The function pisim3 was used to create Tables 3.1, 3.2, and 3.3.

Table 3.2. PIs for Second Generalized Additive Model

error		95%	PI	95%	PI	50%	PI	50%	PI
type	n	slen	olen	scov	ocov	slen	olen	scov	ocov
1	50	5.546	5.318	0.940	0.931	1.853	1.671	0.526	0.466
1	100	5.060	4.822	0.950	0.942	1.701	1.562	0.531	0.480
1	1000	4.348	4.277	0.952	0.945	1.455	1.424	0.507	0.494
1	∞	3.920	3.920	0.95	0.950	1.349	1.349	0.50	0.50
2	50	9.502	8.771	0.946	0.940	2.514	2.270	0.565	0.505
2	100	8.469	7.843	0.952	0.942	2.149	1.980	0.540	0.503
2	1000	6.810	6.667	0.948	0.944	1.684	1.649	0.502	0.487
2	∞	6.365	6.365	0.950	0.950	1.530	1.530	0.50	0.50
3	50	5.923	5.597	0.942	0.929	1.551	1.367	0.528	0.473
3	100	5.377	5.002	0.948	0.940	1.388	1.203	0.535	0.506
3	1000	4.304	4.203	0.949	0.944	1.155	0.978	0.504	0.488
3	∞	3.664	2.996	0.950	0.950	1.099	0.693	0.50	0.50
4	50	3.504	3.320	0.926	0.914	1.180	1.058	0.509	0.445
4	100	3.168	2.867	0.952	0.942	1.142	1.040	0.529	0.470
4	1000	2.576	2.461	0.950	0.946	1.043	1.015	0.508	0.493
4	∞	1.900	1.900	0.950	0.950	1.00	1.00	0.50	0.50
5	50	19.765	17.906	0.949	0.939	3.244	2.930	0.579	0.527
5	100	18.776	16.338	0.954	0.942	2.606	2.396	0.568	0.530
5	1000	13.919	13.048	0.950	0.947	1.725	1.690	0.497	0.485
5	∞	13.490	13.490	0.950	0.950	1.507	1.507	0.50	0.50

Table 3.3. PIs for Third Generalized Additive Model

error		95%	PI	95%	PI	50%	PI	50%	PI
type	n	slen	olen	scov	ocov	slen	olen	scov	ocov
1	50	5.198	5.066	0.956	0.951	1.847	1.666	0.567	0.505
1	100	4.711	4.529	0.959	0.950	1.656	1.519	0.549	0.493
1	1000	3.998	3.947	0.950	0.946	1.380	1.352	0.505	0.487
1	∞	3.920	3.920	0.95	0.950	1.349	1.349	0.50	0.50
2	50	9.407	8.595	0.958	0.951	2.436	2.206	0.594	0.532
2	100	8.290	7.650	0.956	0.945	2.097	1.928	0.560	0.516
2	1000	6.523	6.387	0.950	0.947	1.601	1.569	0.509	0.498
2	∞	6.365	6.365	0.950	0.950	1.530	1.530	0.50	0.50
3	50	5.304	4.984	0.950	0.945	1.581	1.362	0.561	0.501
3	100	4.787	4.341	0.962	0.954	1.361	1.139	0.560	0.516
3	1000	3.849	3.409	0.950	0.948	1.112	0.830	0.505	0.487
3	∞	3.664	2.996	0.950	0.950	1.099	0.693	0.50	0.50
4	50	2.773	2.719	0.946	0.937	1.144	1.022	0.535	0.481
4	100	2.439	2.373	0.952	0.944	1.080	0.979	0.523	0.472
4	1000	1.998	1.985	0.950	0.948	1.002	0.963	0.499	0.478
4	∞	1.900	1.900	0.950	0.950	1.00	1.00	0.50	0.50
5	50	19.850	17.978	0.951	0.939	2.984	2.702	0.598	0.539
5	100	18.835	16.257	0.953	0.947	2.415	2.225	0.572	0.526
5	1000	13.748	12.840	0.954	0.949	1.646	1.613	0.512	0.499
5	∞	13.490	13.490	0.950	0.950	1.507	1.507	0.50	0.50

REFERENCES

[1] Agresti, A., *Categorical Data Analysis*, Second edition, Wiley, Hoboken, NJ, 2002.

[2] Box, G.E.P., and Cox, D.R., *An Analysis of Transformations,* J. Roy. Stat. Soc. B, **26** (1964), 211-246.

[3] Brillinger, D.R., *A Generalized Linear Model with "Gaussian" Regressor Variables,* in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 1983, 97-114.

[4] Cai, T., Tian, L., Solomon, S.D., Wei, L.J., *Predicting Future Responses Based on Possibly Misspecified Working Models,* Biomet. **95** (2008), 75-92.

[5] Cameron, A.C., Trivedi, P.K., *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, UK, 1998.

[6] Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P., *Graphical Methods for Data Analysis,* Duxbury Press, Boston, 1983.

[7] Chambers, J.M., and Hastie, T.J. (eds.), *Statistical Models in S*, Chapman & Hall, New York, NY, 1993.

[8] Collett, D., *Modelling Binary Data*, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[9] Cook, R.D., and Weisberg, S., *Graphics for Assessing the Adequacy of Regression Models,* J. Amer. Stat. Assoc. **92** (1997), 490-499.

[10] Cook, R.D., and Weisberg, S., *Applied Regression Including Computing and Graphics*, Wiley, New York, NY, 1999.

[11] Gladstone, R.J., *A Study of the Relations of the Brain to the Size of the Head,* Biomet. **4** (1905-6), 105-123.

[12] Grübel, R., *The Length of the Shorth,* Ann. Stat. **16** (1988), 619-628.

[13] Hastie, T.J., Tibshirani, R.J., *Generalized Additive Models,* Chapman & Hall, London, UK, 1990.

[14] Hosmer, D.W., and Lemeshow, S., *Applied Logistic Regression*, Second edition, Wiley, New York, NY, 2000.

[15] Johnson, M.P., and Raven, P.H., *Species Number and Endemism, the Galápagos Archipelago Revisited,* Sci. **179** (1973), 893-895.

[16] Kay, R., and Little, S., *Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data,* Biomet. **74** (1987), 495-501.

[17] MathSoft, *S-Plus 2000 Guide to Statistics, Vol. 1*, Data Analysis Products Division, MathSoft, Seattle, WA, 1999.

[18] Mosteller, F., and Tukey, J.W. *Data Analysis and Regression,* Addison-Wesley, Reading, MA, 1977.

[19] Olive, D.J., *Applied Robust Statistics.* Preprint, see (www.math.siu.edu/olive/ol-bookp.htm) (2008).

[20] Olive, D.J., *Asymptotically Optimal Prediction.* Preprint, see (www.math.siu.edu/olive/ppapred.pdf) (2011).

[21] Olive, D.J., *Plots for Generalized Additive Models.* Preprint, see (www.math.siu.edu/olive/ppgam.pdf) (2010).

[22] Olive, D.J., *Prediction Intervals for Regression Models,* Comp. Stat. Dat. Analy. **51** (2007), 3115-3122.

[23] Olive, D.J., *Visualizing 1D Regression,* in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst S., Series: Statistics for Industry and Technology, Birkhäuser, Basel 2004.

[24] Olive, D.J., Hawkins, D.M., *Variable Selection for 1D Regression Models,* Techno. **47** (2005), 43-50.

[25] R Development Core Team, *R: a Language and Environment for Statistical Computing,* R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org) 2008.

[26] Tukey, J.W., *Comparative Anatomy of Transformations,* Ann. Math. Stat. **28** (1957), 602-632.

[27] Venables, W.N., and Ripley, B.D., *Modern Applied Statistics with S*, Fourth edition, Springer-Verlag, New York, NY, 2002.

[28] Winkelmann, R., *Econometric Analysis of Count Data*, Third edition, Springer-Verlag, New York, NY, 2000.

[29] Wood, S.N., *Generalized Additive Models: an Introduction with R*, Chapman & Hall/CRC, Boca Rotan, FL, 2006.

[30] Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M., *Mixed Effects Models and Extensions in Ecology with R*, Springer-Science, New York, NY, 2009.

VITA

Graduate School
Southern Illinois University

Joshua Powers Date of Birth: January 13, 1987

603 South Minnesota Street, Cape Girardeau, MO 63703

jepowers113@gmail.com

Southeast Missouri State University
Bachelor of Science, Mathematics, May 2009

Research Paper Title:
PLOTS AND PREDICTION INTERVALS FOR GENERALIZED ADDITIVE MODELS

Major Professor: Dr. D. Olive