Chapter 10
GLMs and GAMs

10.1 Introduction

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression and Poisson regression. Assume that there is a response variable Y and a $k \times 1$ vector of nontrivial predictors $\boldsymbol{u}$. Before defining 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. Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ and $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_s^T)^T$.

Definition 10.1. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{10.1}$$

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

In the definition, it is crucial that k and w do not depend on y and that h and t do not depend on θ. The parameterization is not unique since, for example, w could be multiplied by 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

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{10.2}$$

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

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{10.3}$$

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

The GLM parameterization (10.3) can be written in several ways. By Equation (10.2), $f(y_i|\theta(\boldsymbol{x}_i)) = \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) =$

$$\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

$$= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

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

$$\mu(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{10.4}$$

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

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

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

The following three sections illustrate three of the most important generalized linear models. After selecting a GLM, the investigator will often want to check whether the model is useful and to perform inference. Several things to consider are listed below.

i) Show that the GLM provides a simple, useful approximation for the relationship between the response variable Y and the predictors $\boldsymbol{x}$.

ii) Estimate $\boldsymbol{\beta}$ using maximum likelihood estimators.

iii) Estimate $\mu(\boldsymbol{x}_i) = d_i\tau(\boldsymbol{x}_i)$ or estimate $\tau(\boldsymbol{x}_i)$ where the d_i are known constants.

iv) Check for goodness of fit of the GLM with a response plot = estimated sufficient summary plot.

v) Check for lack of fit of the GLM (e.g. with a residual plot).

vi) Check for overdispersion with an OD plot.

vii) Check whether Y is independent of $\boldsymbol{u}$; i.e., check whether $\boldsymbol{\beta}_s = \mathbf{0}$.

viii) Check whether a reduced model can be used instead of the full model.

ix) Use variable selection to find a good submodel.

x) Predict Y_i given $\boldsymbol{x}_i$.

10.2 Multiple Linear Regression

Suppose that the response variable Y is quantitative. Then the multiple linear regression model is often a very useful model and is closely related to the GLM based on the normal distribution. To see this claim, let $f(y|\mu)$ be the $N(\mu, \sigma^2)$ family of pdfs where $-\infty < \mu < \infty$ and $\sigma > 0$ is known. Recall that μ is the mean and σ is the standard deviation of the distribution. Then the pdf of Y is

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

Since

$$f(y|\mu) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-1}{2\sigma^2}\mu^2)}_{k(\mu)\geq 0}\underbrace{\exp(\frac{-1}{2\sigma^2}y^2)}_{h(y)\geq 0}\exp(\underbrace{\frac{\mu}{\sigma^2}}_{c(\mu)/a(\sigma^2)} y),$$

this family is a 1-parameter exponential family. For this family, $\theta = \mu = E(Y)$, and the known dispersion parameter $\phi = \sigma^2$. Thus $a(\sigma^2) = \sigma^2$ and the canonical link is the **identity link** $c(\mu) = \mu$.

Hence the GLM corresponding to the $N(\mu, \sigma^2)$ distribution with canonical link states that $Y_1, ..., Y_n$ are independent random variables where

$$Y_i \sim N(\mu(\boldsymbol{x}_i), \sigma^2) \text{ and } E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i$$

for $i = 1, ..., n$. This model can be written as $Y_i \equiv Y_i|\boldsymbol{x}_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i$ where $e_i \sim N(0, \sigma^2)$.

When the predictor variables are quantitative, the above model is called a multiple linear regression (MLR) model. When the predictors are categorical, the above model is called an analysis of variance (ANOVA) model, and when the predictors are both quantitative and categorical, the model is called an

MLR or analysis of covariance model. The MLR model is discussed in detail in Chapter 5, where the normality assumption and the assumption that σ is known can be relaxed.

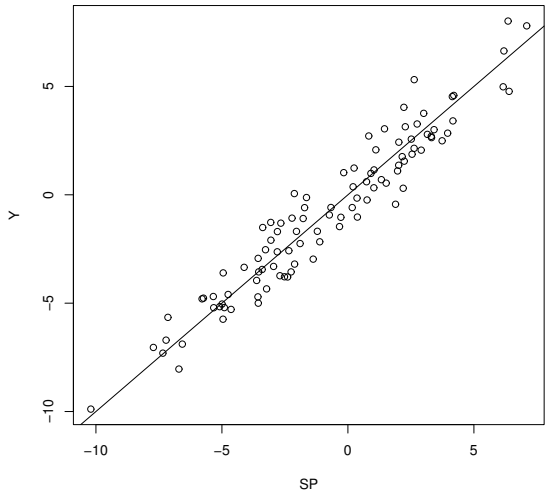

Fig. 10.1 SSP for MLR Data

A sufficient summary plot (SSP) of the sufficient predictor $SP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the multiple linear regression model. This plot can not be used for real data since $\boldsymbol{\beta}$ is unknown. The artificial data used to make Figure 10.1 used $n = 100$ cases with $k = 5$ nontrivial predictors. The data used $\boldsymbol{\beta} = (-1, 1, 2, 3, 0, 0)^T$, $e_i \sim N(0, 1)$ and $\boldsymbol{u} \sim N_5(\boldsymbol{0}, \boldsymbol{I})$.

In Figure 10.1, notice that the identity line with unit mean and zero intercept corresponds to the mean function since the identity line is the line $Y = SP = \boldsymbol{\beta}^T \boldsymbol{x} = g(\mu(\boldsymbol{x}))$. The vertical deviation of Y_i from the line is equal to $e_i = Y_i - (\boldsymbol{\beta}^T \boldsymbol{x}_i)$. For a given value of SP, $Y_i \sim N(SP, \sigma^2)$. For the artificial data, $\sigma^2 = 1$. Hence if $SP = 0$ then $Y_i \sim N(0, 1)$, and if $SP = 5$ the $Y_i \sim N(5, 1)$. Imagine superimposing the $N(SP, \sigma^2)$ curve at various values of SP. If all of the curves were shown, then the plot would resemble a road through a tunnel. For the artificial data, each Y_i is a sample of size 1 from the normal curve with mean $\boldsymbol{\beta}^T \boldsymbol{x}_i$.

The estimated sufficient summary plot (ESSP), also called a **response plot**, is a plot of $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the identity line added as a visual aid. Now the vertical deviation of Y_i from the line is equal to the residual

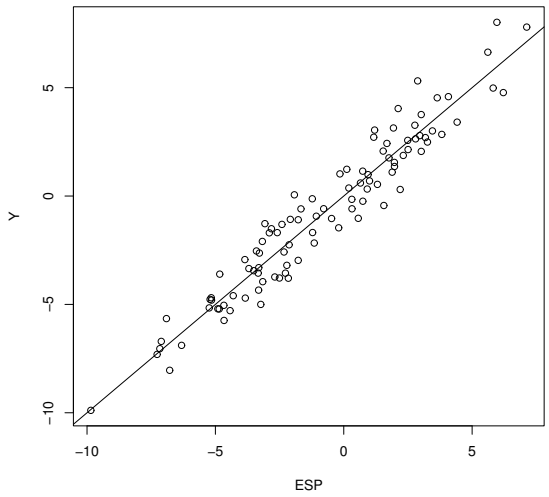

Fig. 10.2 ESSP = Response Plot for MLR Data

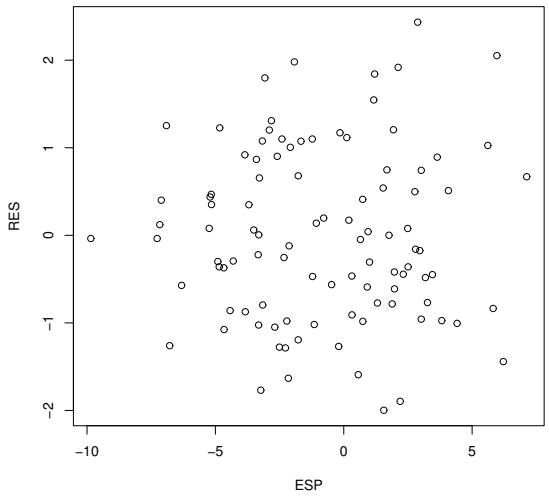

Fig. 10.3 Residual Plot for MLR Data

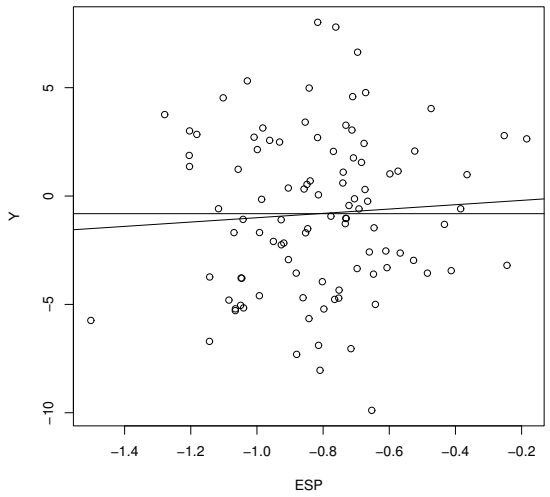

Fig. 10.4 Response Plot when Y is Independent of the Predictors

$r_i = Y_i - (\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. The interpretation of the ESSP is almost the same as that of the SSP, but now the mean SP is estimated by the estimated sufficient predictor (ESP). This plot is used as a goodness of fit diagnostic. The residual plot is a plot of the ESP versus r_i and is used as a lack of fit diagnostic. These two plots should be made immediately after fitting the MLR model and before performing inference. Figures 10.2 and 10.3 show the response plot and residual plot for the artificial data.

The response plot is also a useful visual aid for describing the ANOVA F test (see Section 5.5) which tests whether $\boldsymbol{\beta} = \mathbf{0}$, that is, whether the predictors $\boldsymbol{x}$ are needed in the model. If the predictors are not needed in the model, then Y_i and $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the sample mean $\overline{Y}$. If the predictors are needed, then Y_i and $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the ESP $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. The fitted value $\hat{Y}_i$ is the maximum likelihood estimator computed using ordinary least squares. If the identity line clearly fits the data better than the horizontal line $Y = \overline{Y}$, then the ANOVA F test should have a small p–value and reject the null hypothesis H_o that the predictors $\boldsymbol{x}$ are not needed in the MLR model. Figure 10.4 shows the response plot for the artificial data when only X_4 and X_5 are used as predictors with the identity line and the line $Y = \overline{Y}$ added as visual aids. In this plot the horizontal line fits the data about as well as the identity line which was expected since Y is independent of X_4 and X_5.

It is easy to find data sets where the response plot looks like Figure 10.4, but the p–value for the ANOVA F test is very small. In this case, the MLR

model is statistically significant, but the investigator needs to decide whether the MLR model is practically significant.

10.3 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for 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 labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. Often the labelling is arbitrary, e.g., if the response variable is *gender* taking on the two categories female and male. If males are counted then $Y = 1$ if the subject is male and $Y = 0$ if the subject is female. If females are counted then this labelling is reversed. For a binary response variable, a binary regression model is often appropriate.

Definition 10.3. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i}))$. The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{10.5}$$

If the sufficient predictor $SP = \boldsymbol{\beta}^T\boldsymbol{x}$, then the most used binomial regression models are such that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i}))$, or

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{10.6}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i\rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$. Note that the LR model has

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

To see that the binary logistic regression model is a GLM, assume that Y is a binomial$(1, \rho)$ random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \binom{1}{y}\rho^y(1 - \rho)^{1-y} = \underbrace{\binom{1}{y}}_{h(y)\geq 0}\underbrace{(1 - \rho)}_{k(\rho)\geq 0}\exp[\underbrace{\log(\frac{\rho}{1 - \rho})}_{c(\rho)}y].$$

Hence this family is a 1-parameter exponential family with $\theta = \rho = E(Y)$ and canonical link $c(\rho) = \log\left(\frac{\rho}{1-\rho}\right)$. This link is known as the *logit link*, and if $g(\mu(\boldsymbol{x})) = g(\rho(\boldsymbol{x})) = c(\rho(\boldsymbol{x})) = \boldsymbol{\beta}^T\boldsymbol{x}$ then the inverse link satisfies

$$g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}^T\boldsymbol{x})} = \rho(\boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the binomial$(1, \rho)$ distribution with canonical link is the binary logistic regression model.

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1-\rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression, $\rho(\boldsymbol{x}) = P(Y=1|\boldsymbol{x}) = 1-P(Y=0|\boldsymbol{x})$. If this population proportion $\rho = \rho(\boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\boldsymbol{\beta}^T\boldsymbol{x})) = \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\boldsymbol{\beta}^T\boldsymbol{x})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1+\exp(x))$ which is the cdf of the logistic $L(0,1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the Normal $N(0,1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1-\exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1-\rho(\boldsymbol{x}))] = \boldsymbol{\beta}^T\boldsymbol{x}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 43–44).let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_s^T)^T$ and $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$. Assume that $\pi_j = P(Y=j)$ and that $\boldsymbol{u}|Y=j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{u}$ given $Y=j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{u}|Y) \neq \mathrm{Cov}(\boldsymbol{u})$. Then as for the binary logistic regression model,

$$P(Y=1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}^T\boldsymbol{x})}.$$

Definition 10.4. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta}_s = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{10.7}$$

$$\text{and} \quad \beta_1 = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

The logistic regression (maximum likelihood) estimator also tends to perform well for this type of data. An exception is when the $Y=0$ cases and $Y=1$ cases can be perfectly or nearly perfectly classified by the ESP. Let

the logistic regression ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Consider the response plot of the ESP versus Y. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. In this case the maximum likelihood estimator for the logistic regression parameters $(\alpha, \boldsymbol{\beta})$ does not exist because the logistic curve can not approximate a step function perfectly. If only a few cases need to be deleted in order for the data set to have perfect classification, then the amount of "overlap" is small and there is nearly "perfect classification."

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, partial F test, and OLS t tests are often asymptotically valid when the conditions in Definition 10.4 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 10.1 below. Assume that $\text{Cov}(\boldsymbol{u}) \equiv \boldsymbol{\Sigma_u}$ and that $\text{Cov}(\boldsymbol{u}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. Let $\boldsymbol{\mu}_j = E(\boldsymbol{u}|Y = j)$ for $j = 0, 1$. Let N_i be the number of Ys that are equal to i for $i = 0, 1$. Then

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i} \sum_{j:Y_j=i} \boldsymbol{u}_j$$

for $i = 0, 1$ while $\hat{\pi}_i = N_i/n$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Notice that Theorem 10.1 holds as long as $\text{Cov}(\boldsymbol{u})$ is nonsingular and Y is binary with values 0 and 1. The LR and discriminant function models need not be appropriate.

Theorem 10.1. Assume that Y is binary and that $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$ is nonsingular. Let $(\hat{\beta}_{OLS,1}, \hat{\boldsymbol{\beta}}_{OLS,s})$ be the OLS estimator found from regressing Y on a constant and $\boldsymbol{u}$ (using software originally meant for multiple linear regression). Then

$$\hat{\boldsymbol{\beta}}_{OLS,s} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{n}{n-1} \hat{\pi}_0 \hat{\pi}_1 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

$$\xrightarrow{D} \boldsymbol{\beta}_{OLS,s} = \pi_0 \pi_1 \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \text{ as } \text{n} \to \infty.$$

Proof. From Section 5.5,

$$\hat{\boldsymbol{\beta}}_{OLS,s} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \text{n} \to \infty$$

$$\text{and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\, \overline{Y}.$$

$$\text{Thus } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}\text{Y}} = \frac{1}{\text{n}} \left[\sum_{\text{j:Y}_\text{j}=1} \boldsymbol{u}_\text{j}(1) + \sum_{\text{j:Y}_\text{j}=0} \boldsymbol{u}_\text{j}(0) \right] - \overline{\boldsymbol{u}}\, \hat{\pi}_1 =$$

$$\frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1 + N_0 \hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\boldsymbol{\mu}}_0 =$$

$$\hat{\pi}_1(1-\hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

and the result follows. QED

The discriminant function estimators $\hat{\beta}_{D,1}$ and $\hat{\boldsymbol{\beta}}_{D,s}$ are found by replacing the population quantities π_1, π_0, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ by sample quantities. Also

$$\hat{\boldsymbol{\beta}}_{D,s} = \frac{n(n-1)}{N_0 N_1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}\hat{\boldsymbol{\beta}}_{OLS,s}.$$

Now when the conditions of Definition 10.4 are met and if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is small enough so that there is not perfect classification, then $\boldsymbol{\beta}_{LR} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Empirically, the OLS ESP and LR ESP are highly correlated for many LR data sets where the conditions are not met, e.g. when some of the predictors are factors. This suggests that $\boldsymbol{\beta}_{LR} \approx d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ for many LR data sets where d is some constant depending on the data.

Using Definition 10.4 makes simulation of logistic regression data straightforward. Set $\pi_0 = \pi_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$. Then $\beta_1 = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta}_s = \boldsymbol{\mu}_1$. The artificial data set used in the following discussion used $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$ and hence $\beta_1 = -1.5$. Let N_i be the number of cases where $Y = i$ for $i = 0, 1$. For the artificial data, $N_0 = N_1 = 100$, and hence the total sample size $n = N_1 + N_0 = 200$.

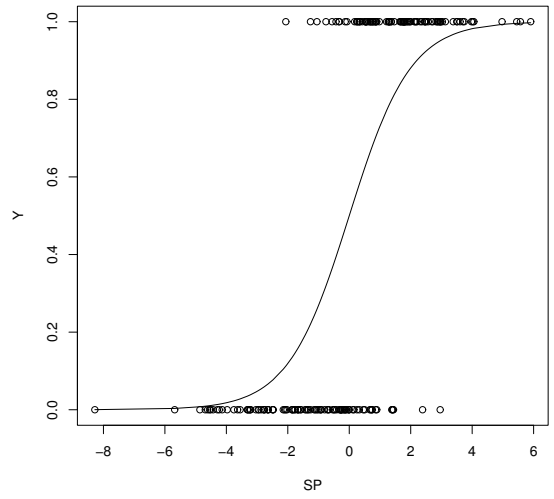

Fig. 10.5 SSP for LR Data

Again a sufficient summary plot of the sufficient predictor $SP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the binary logistic regression (LR) model. The artificial data described above was used because the plot can not be used for real data since $\boldsymbol{\beta}$ are unknown.

Unlike the SSP for multiple linear regression where the mean function is always the identity line, the mean function in the SSP for LR can take a variety of shapes depending on the range of the SP. For the LR SSP, the mean function is $\rho(SP) = \dfrac{\exp(SP)}{1 + \exp(SP)}$. If the SP $= 0$ then $Y|SP \sim$ binomial(1,0.5). If the SP $= -5$, then $Y|SP \sim$ binomial(1,$\rho \approx 0.007$) while if the SP $= 5$, then $Y|SP \sim$ binomial(1,$\rho \approx 0.993$). Hence if the range of the SP is in the interval $(-\infty, -5)$ then the mean function is flat and $\rho(SP) \approx 0$. If the range of the SP is in the interval $(5, \infty)$ then the mean function is again flat but $\rho(SP) \approx 1$. If $-5 < SP < 0$ then the mean function looks like a slide. If $-1 < SP < 1$ then the mean function looks linear. If $0 < SP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < SP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 10.5.

The estimated sufficient summary plot (ESSP or response plot) is a plot of $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function $\hat{\rho}(SP) = \rho(ESP) = \dfrac{\exp(ESP)}{1 + \exp(ESP)}$ added as a visual aid. The interpretation of the

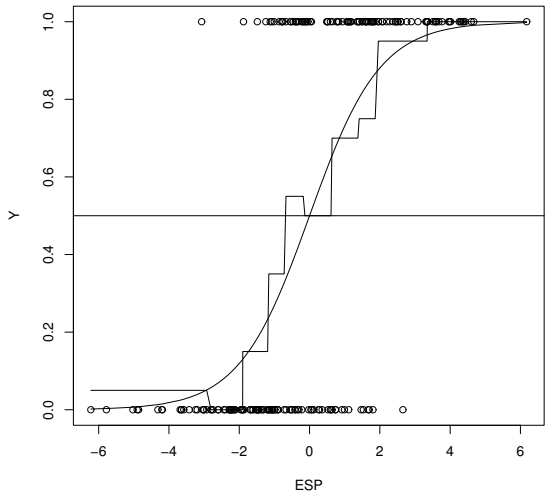

Fig. 10.6 Response Plot for LR Data

response plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

This plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Y's in each slice and add the resulting step function to the response plot. This is done in Figure 10.6 with $J = 10$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147–156).

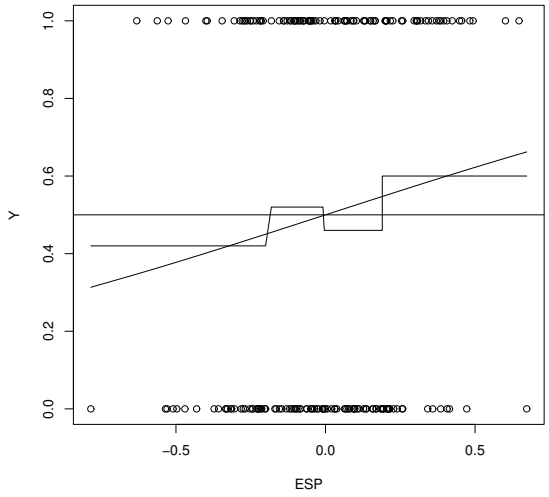

Fig. 10.7 Response Plot When Y Is Independent Of The Predictors

The deviance test described in Section 10.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the binary LR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator $\hat{\rho}(\boldsymbol{x}_i) = \dfrac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}$. If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_o will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be

independent of the predictors. Figure 10.7 shows the response plot when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 10.7 where the p–value for the deviance test is very small. Then the LR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

For binary data the Y_i only take two values, 0 and 1, and the residuals do not behave very well. Hence the response plot will be used both as a goodness of fit plot and as a lack of fit plot.

For binomial regression, the response plot needs to be modified and a check for overdispersion is needed. Let $Z_i = Y_i/m_i$. Then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the LR binomial regression model can be visualized with a response plot of the ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Z_i with the estimated mean function $\hat{\rho}(SP) = \rho(ESP) = \dfrac{\exp(ESP)}{1 + \exp(ESP)}$ added as a visual aid. 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. Either the step function or the lowess curve could be added to the response plot. Both the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated mean) closely, then the LR mean function is a reasonable approximation to the data.

Checking the LR model in the nonbinary case is more difficult because the binomial distribution is not the only distribution appropriate for data that takes on values $0, 1, ..., m$ if $m \geq 2$. Hence both the mean and variance functions need to be checked. Often the LR mean function is a good approximation to the data, the LR MLE is a consistent estimator of $\boldsymbol{\beta}$, but the LR model is not appropriate. The problem is that for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. This phenomenon is called *overdispersion*.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, p. 93-94) and Agresti (2002, p. 554-555), let $\delta = \rho/\theta$ and $\nu = (1 - \rho)/\theta$, so $\rho = \delta/(\delta + \nu)$ and $\theta = 1/(\delta+\nu)$. Let $B(\delta, \nu) = \dfrac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta + \nu)}$. If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m}, \rho, \theta)$, then the probability mass function of Y is $P(Y = y) = \binom{m}{y} \dfrac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$ for $y = 0, 1, 2, ..., m$ where $0 < \rho < 1$ and $\theta > 0$. Hence $\delta > 0$ and $\nu > 0$. Then $E(Y) = m\delta/(\delta + \nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1 + (m-1)\theta/(1+\theta)]$. If $Y|\pi \sim \text{binomial}(m, \pi)$ and $\pi \sim \text{beta}(\delta, \nu)$, then $Y \sim \text{BB}(\text{m}, \rho, \theta)$.

Definition 10.5. The BBR model states that $Y_1, ..., Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{SP}_\text{i}), \theta)$.

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. Note that $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))[1 + (m_i - 1)\theta/(1 + \theta)].$$

As $\theta \to 0$, it can be shown that $V(\pi) \to 0$ and the BBR model converges to the binomial regression model.

For both the LR and BBR models, the conditional distribution of $Y|\boldsymbol{x}$ can still be visualized with a response plot of the ESP versus $Z_i = Y_i/m_i$ with the estimated mean function $\hat{E}(Z_i|\boldsymbol{x}_i) = \hat{\rho}(SP) = \rho(ESP)$ and a step function or lowess curve added as visual aids.

Since binomial regression is the study of $Z_i|\boldsymbol{x}_i$ (or equivalently of $Y_i|\boldsymbol{x}_i$), the response plot is crucial for analyzing LR models. The response plot is a special case of the model checking plot and emphasizes goodness of fit.

Since the binomial regression model is simpler than the BBR model, graphical diagnostics for the goodness of fit of the LR model would be useful. To check for overdispersion, we suggest using the OD plot of $\hat{V}(Y|SP)$ versus $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. This plot was suggested by Winkelmann (2000, p. 110) to check overdispersion for Poisson regression.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the logistic regression model much as R^2 is used for multiple linear regression. When the m_i are small, G^2 may not be reliable but the response plot is still useful. If the Y_i are not too close to 0 or m_i, if the response and OD plots look good, and the deviance G^2 satisfies $G^2/(n - k - 1) \approx 1$, then the LR model is likely useful. If $G^2 > (n - k - 1) + 3\sqrt{n - k + 1}$, then a more complicated count model may be needed.

The response plot is a powerful method for assessing the adequacy of the binary LR regression model. Suppose that both the number of 0s and the number of 1s is large compared to the number of predictors k, that the ESP takes on many values and that the binary LR model is a good approximation to the data. Then $Y|\boldsymbol{x} \approx$ Binomial(1, $\rho(ESP)$. For example if the ESP = 0 then $Y|\boldsymbol{x} \approx$ Binomial(1,0.5). If $-5 < ESP < 5$ then the estimated mean function has the characteristic "ESS" shape of the logistic curve.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the LR model. To motivate the OD plot, recall that if a count Y is not too close to 0 or m, then a normal approximation is good for the binomial distribution. Notice that if $Y_i = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y_i - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, and if the counts are not too close to 0 or m_i, then the plotted points in the OD plot 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.

If the data are binary, the response plot is enough to check the binomial regression assumption. When the counts are small, the OD plot is not wedge shaped, but if the LR model is correct, the least squares (OLS) line should be close to the identity line through the origin with unit slope.

Suppose the bulk of the plotted points in the OD plot fall in a wedge. Then the identity line, slope 4 line and OLS line will be added to the plot as visual aids. 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 the variability about the logistic curve. Also outliers are often easier to spot with the OD plot. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$. The evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. 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%.

If the binomial LR OD plot is used but the data follows a beta–binomial regression model, then $\hat{V}_{mod} = \hat{V}(Y_i|SP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1 - \rho(ESP))[1 + (m_i - 1)\theta/(1 + \theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx 1 + (m - 1)\dfrac{\theta}{1 + \theta} = \dfrac{1 + m\theta}{1 + \theta}$.

The first example is for binary data. For binary data, G^2 is not approximately χ^2 and some plots of residuals have a pattern whether the model is correct or not. For binary data the OD plot is not needed, and the plotted points follow a curve rather than falling in a wedge. The response plot is very useful if the logistic curve and step function of observed proportions are added as visual aids. The logistic curve gives the estimated LR probability of success. For example, when ESP = 0, the estimated probability is 0.5.

Example 10.1. Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response variable *ape* is 1 for an ape skull. The response plot in Figure 10.8a) uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The response plot in Figure 10.8b) uses the predictor *head height* and perfectly classifies the data since the ape skulls can be separated from the human skulls with a vertical line at ESP = 0. Christmann and Rousseeuw (2001) also used the response plot to visualize overlap. The response plot in Figure 10.8c uses predictors *lower jaw length, face length,* and *upper jaw length*. None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely. The OD plot in Figure 10.8d) is curved and is not needed for a binary response.

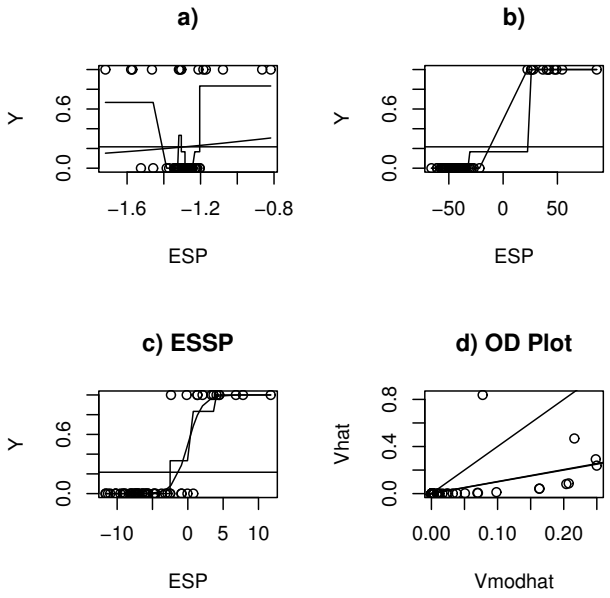

Fig. 10.8 Response Plots for Museum Data

Example 10.2. Abraham and Ledolter (2006, p. 360-364) describe death penalty sentencing in Georgia. The predictors are *aggravation level* from 1 to 6 (treated as a continuous variable) and *race of victim* coded as 1 for white and 0 for black. There were 362 jury decisions and 12 level race combinations. The response variable was the number of death sentences in each combination. The response plot (ESSP) in Figure 10.9a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve). The step function based on 5 slices also tracks the logistic curve well. The OD plot is shown in Figure 10.9b with the identity, slope 4 and OLS lines added as visual aids. The vertical scale is less than the horizontal scale and there is no evidence of overdispersion.

Example 10.3. Collett (1999, p. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. Figure 10.10a shows the response plot (ESSP). Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is overdispersion

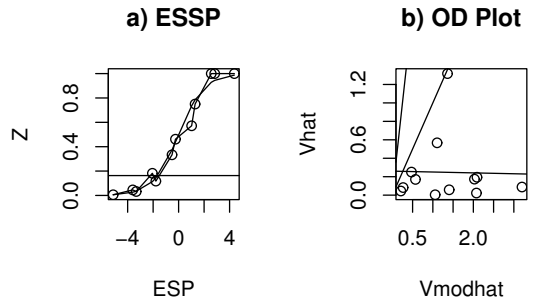

Fig. 10.9 Visualizing the Death Penalty Data

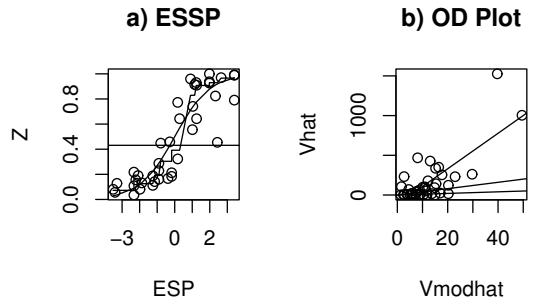

Fig. 10.10 Plots for Rotifer Data

since the vertical scale is about 30 times the horizontal scale. The OLS line has slope much larger than 4 and two outliers seem to be present.

10.4 Poisson Regression

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 10.6. The **Poisson regression (PR) model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_i))$. The **Poisson regression model** is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{10.8}$$

To see that the PR model is a GLM, assume that Y is a Poisson(μ) random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} = \underbrace{e^{-\mu}}_{k(\mu)\geq 0} \underbrace{\frac{1}{y!}}_{h(y)\geq 0} \exp[\underbrace{\log(\mu)}_{c(\mu)} y]$$

for $y = 0, 1, \ldots$, where $\mu > 0$. Hence this family is a 1-parameter exponential family with $\theta = \mu = E(Y)$, and the canonical link is the log link $c(\mu) = \log(\mu)$. Since $g(\mu(\boldsymbol{x})) = c(\mu(\boldsymbol{x})) = \boldsymbol{\beta}^T \boldsymbol{x}$, the inverse link satisfies

$$g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the Poisson(μ) distribution with canonical link is the Poisson regression model.

A sufficient summary plot of the sufficient predictor $SP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the Poisson regression (PR) model. Artificial data needs to be used because the plot can not be used for real data since $\boldsymbol{\beta}$ is unknown. The data used in the discussion below had $n = 100$, $\boldsymbol{u} \sim N_5(\boldsymbol{1}, \boldsymbol{I}/4)$ and $Y_i \sim \text{Poisson}(\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}))$ where $\boldsymbol{\beta} = (-2.5, 1, 1, 1, 0, 0)^T$.

Model (10.8) can be written compactly as $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$. Notice that $Y|SP = 0 \sim \text{Poisson}(1)$. Also note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$. The shape of the mean function $\mu(SP) = \exp(SP)$ for Poisson regression depends

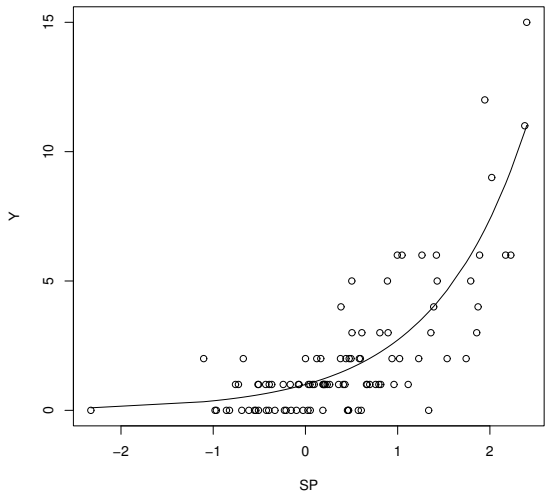

Fig. 10.11 SSP for Poisson Regression

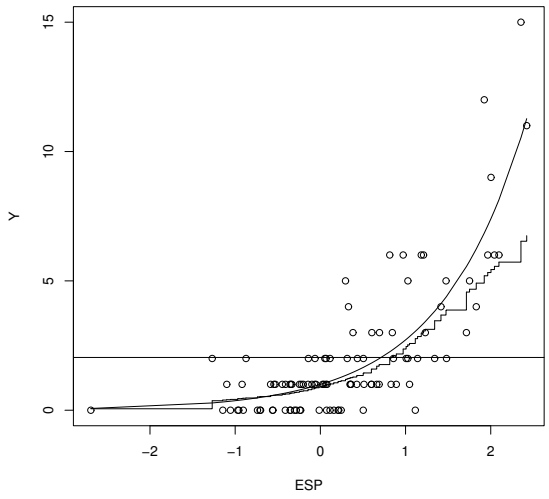

Fig. 10.12 Response Plot for Poisson Regression

strongly on the range of the SP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence the range of the SP is narrow, then the exponential function will be rather flat. If the range of the SP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot. Figure 10.11 shows the SSP for the artificial data.

The estimated sufficient summary plot (ESSP or response plot) is a plot of the $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ added as a visual aid. The interpretation of the response plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function called a "scatterplot smoother." The lowess curve is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve) in Figure 10.12. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the PR model may fit the data well. **A useful lack of fit plot** is a plot of the ESP versus the *deviance residuals* that are often available from the software.

The deviance test described in Section 10.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the PR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the PR estimator $\hat{\mu}(\boldsymbol{x}_i) = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_o should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. Figure 10.13 shows the ESSP when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 10.13 where the p–value for the deviance test is very small. Then the PR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

Warning: For many count data sets where the PR mean function is correct, the PR model is not appropriate but the PR MLE is still a consistent estimator of $\boldsymbol{\beta}$. The problem is that for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, it turns out that $V(Y|\boldsymbol{x}) > \exp(SP)$. This phenomenon is called **overdispersion**. Adding parametric and nonparametric estimators of the standard deviation function to the response plot can be useful. See Cook and Weisberg (1999a, p. 401-403). Alternatively, if the response plot looks good and $G^2/(n-k-1) \approx 1$, then the PR model is likely

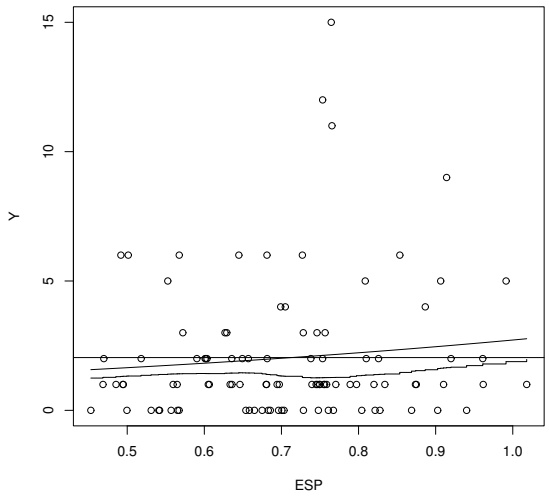

Fig. 10.13 Response Plot when Y is Independent of the Predictors

useful. If $G^2/(n-k-1) > 1+3/\sqrt{n-k-1}$, then a more complicated count model may be needed. Here the deviance G^2 is described in Section 10.5.

A useful alternative to the PR model is a negative binomial regression (NBR) 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 $\mathrm{V}(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ, ρ) distribution with $\rho = \kappa/(\mu+\kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

Definition 10.7. The **negative binomial regression (NBR) model** states that $Y_1, ..., Y_n$ are independent random variables where $Y_i \sim \mathrm{NB}(\mu(\boldsymbol{x}_i), \kappa)$ with $\mu(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$. Hence $Y|SP \sim \mathrm{NB}(\exp(\mathrm{SP}), \kappa)$, $E(Y|SP) = \exp(SP)$ and

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

The NBR model has the same mean function as the PR model but allows for overdispersion. As $\kappa \to \infty$, the NBR model converges to the PR model. See Section 10.8.

Judging the mean function from the response plot may be rather difficult for large counts since the mean function is curved and lowess does not track the exponential function very well for large counts. Simple diagnostic plots for the Poisson regression model can be made using weighted least squares (WLS). To see this, assume that all n of the counts Y_i are large. Then $\log(\mu(\boldsymbol{x}_i)) = \log(\mu(\boldsymbol{x}_i)) + \log(Y_i) - \log(Y_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$, or $\log(Y_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$ where $e_i = \log\left(\frac{Y_i}{\mu(\boldsymbol{x}_i)}\right)$. The error e_i does not have zero mean or constant variance, but if $\mu(\boldsymbol{x}_i)$ is large $\frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N(0,1)$ by the central limit theorem. Recall that $\log(1+x) \approx x$ for $|x| < 0.1$. Then, heuristically,

$$e_i = \log\left(\frac{\mu(\boldsymbol{x}_i) + Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)}\right) \approx \frac{Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)} =$$

$$\frac{1}{\sqrt{\mu(\boldsymbol{x}_i)}} \frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N\left(0, \frac{1}{\mu(\boldsymbol{x}_i)}\right).$$

This suggests that for large $\mu(\boldsymbol{x}_i)$, the errors e_i are approximately 0 mean with variance $1/\mu(\boldsymbol{x}_i)$. If the $\mu(\boldsymbol{x}_i)$ were known, and all of the Y_i were large, then a weighted least squares of $\log(Y_i)$ on $\boldsymbol{x}_i$ with weights $w_i = \mu(\boldsymbol{x}_i)$ should produce good estimates of $\boldsymbol{\beta}$. Since the $\mu(\boldsymbol{x}_i)$ are unknown, the estimated weights $w_i = Y_i$ could be used. Since $P(Y_i = 0) > 0$, the estimators given in the following definition are used. Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$.

Definition 10.8. The **minimum chi–square estimator** of $\boldsymbol{\beta}$ in a Poisson regression model is $\hat{\boldsymbol{\beta}}_M$, and is found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}\boldsymbol{x}_i$.

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 maximum likelihood estimator tends to be consistent if the sample size $n \to \infty$. See Agresti (2002, p. 611-612). However, the two estimators are often close for many data sets. This result and the equivalence of the minimum chi–square estimator to an OLS estimator suggest the following diagnostic plots. Let $\tilde{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$.

Definition 10.9. For a Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP = \sqrt{Z_i}\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus the WLS residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$.

If the Poisson regression model is appropriate and if the minimum chi–square estimators are reasonable, then the plotted points in the weighted fit response plot should follow the identity line. Cases with large WLS residuals may not be fit very well by the model. When the counts Y_i are small, the WLS residuals can not be expected to be approximately normal. Notice that a resistant estimator for $\boldsymbol{\beta}$ can be obtained by replacing OLS (in Definition 10.9) with a resistant MLR estimator.

Example 10.4. For the Ceriodaphnia data of Myers, Montgomery and Vining (2002, p. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$ and seven concentrations of jet fuel (x_1) and an indicator for two strains of organism (x_2) were used as predictors. The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 10.14 shows the 4 plots for this data. In the response plot of Figure 10.14a, the lowess curve is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The OD plot in Figure 10.14b suggests that there is little evidence of overdispersion. These two plots as well as Figures 10.14c and 10.14d suggest that the Poisson regression model is a useful approximation to the data.

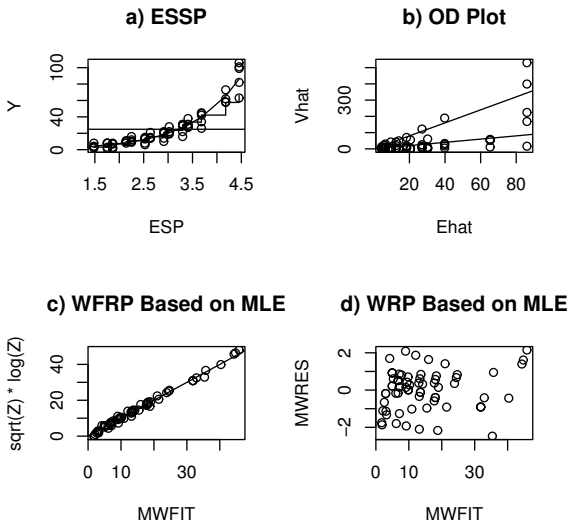

Fig. 10.14 Plots for Ceriodaphnia Data

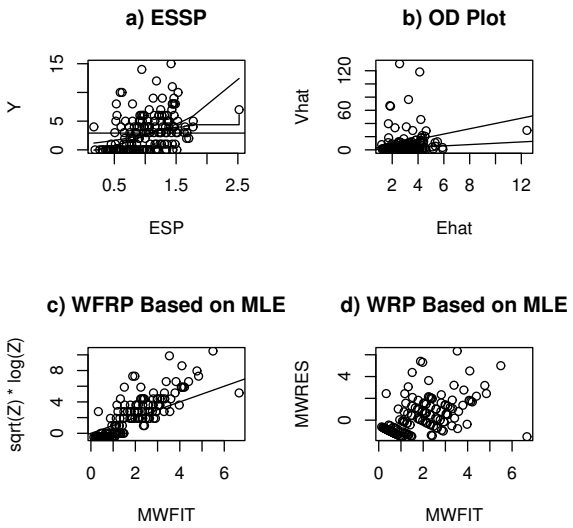

Fig. 10.15 Plots for Crab Data

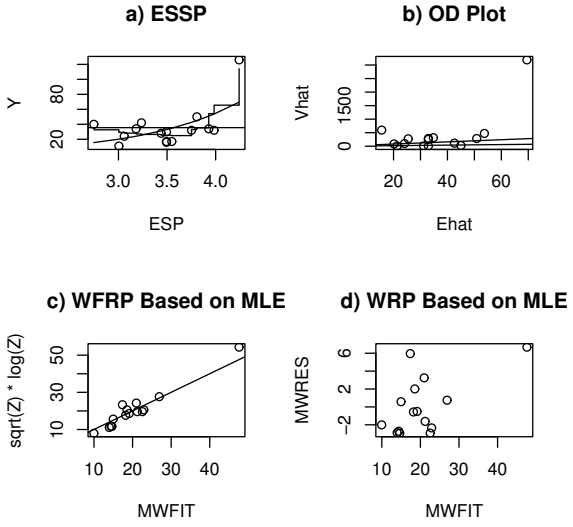

Fig. 10.16 Plots for Popcorn Data

Example 10.5. For the crab data, the response Y is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the color, spine condition, caparice width and weight of the female crab. Agresti (2002, p. 126-131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 10.15a suggests that there is one case with an unusually large value of the ESP. The lowess curve does not track the exponential curve all that well. Figure 10.15b suggests that overdispersion is present since the vertical scale is about 10 times that of the horizontal scale and too many of the plotted points are large and greater than the slope 4 line. Figure 10.15c also suggests that the Poisson regression mean function is a rather poor fit since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \overline{Y}$, an alternative model to the NBR model may fit the data better. In later chapters, Agresti uses binomial regression models for this data.

Example 10.6. For the popcorn data of Myers, Montgomery and Vining (2002, p. 154), the response variable Y is the number of inedible popcorn kernels. The sample size was $n = 15$ and the predictor variables were temperature (coded as 5, 6 or 7), amount of oil (coded as 2, 3 or 4) and popping time (75, 90 or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier. Ignoring the outlier in Figure 10.16a suggests that the line $Y = \overline{Y}$ will fit the data and lowess curve better than the exponential curve. Hence Y seems to be independent of the predictors. Notice that the outlier sticks out in Figure 10.16b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected, then the Poisson regression model would suggest that temperature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated.

10.5 Inference

This section gives a very brief discussion of inference for the logistic regression (LR) and Poisson regression (PR) models. Inference for these two models is very similar to inference for the multiple linear regression (MLR) model. For all three of these models, Y is independent of the $k \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_k)^T$ given the sufficient predictor $\boldsymbol{\beta}^T\boldsymbol{x}$: $Y \perp\!\!\!\perp \boldsymbol{x} | (\boldsymbol{\beta}^T\boldsymbol{x})$.
Response = Y
Coefficient Estimates

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
Number of cases:                 n
Degrees of freedom:              n - k - 1
Pearson X2:
Deviance:                        D = G^2
------------------------------------
Binomial Regression
Kernel mean function = Logistic
Response      = Status
Terms         = (Bottom Left)
Trials        = Ones
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -389.806         104.224        -3.740     0.0002
Bottom     2.26423         0.333233        6.795     0.0000
Left       2.83356         0.795601        3.562     0.0004

Scale factor:                  1.
Number of cases:             200
Degrees of freedom:          197
Pearson X2:               179.809
Deviance:                  99.169
```

To perform inference for LR and PR, computer output is needed. Above is shown output using symbols and *Arc* output from a real data set with $k = 2$ nontrivial predictors. This data set is the *banknote* data set described in Cook and Weisberg (1999a, p. 524). There were 200 Swiss bank notes of which 100 were genuine ($Y = 0$) and 100 counterfeit ($Y = 1$). The goal of the analysis was to determine whether a selected bill was genuine or counterfeit from physical measurements of the bill.

Point estimators for the mean function are important. Given values of $\boldsymbol{x} = (x_1, ..., x_k)^T$, a major goal of binary logistic regression is to estimate the success probability $P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{10.9}$$

Similarly, a major goal of Poisson regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{10.10}$$

For tests, the p–value is an important quantity. Recall that H_o is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_o should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_o should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the GLM model for Y given that the other $k-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the GLM model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful GLM predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained using the output: the large sample 100 $(1-\delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1-\delta$ refers to the coverage of the CI. A 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a GLM, often 3 models are of interest: the **full model** that uses all p of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the p parameters $\beta_1, ..., \beta_p$ are estimated while the reduced model has r parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\boldsymbol{\beta})$ be the likelihood function for the full model. Let $L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$ be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let $L_{FULL} = \log\ l_{FULL}(\hat{\boldsymbol{\beta}})$ be the log likelihood function for the full model evaluated at the MLE $\hat{\boldsymbol{\beta}}$. Then

the **deviance** $D = G^2 = -2(L_{FULL} - L_{SAT})$. The degrees of freedom for the deviance $= df_{FULL} = n - p$ where n is the number of parameters for the saturated model and p is the number of parameters for the full model.

The saturated model for logistic regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model is usually not very good for binary data (all $m_i = 1$) or if the m_i are small. The saturated model can be good if all of the m_i are large or if ρ_i is very close to 0 or 1 whenever m_i is not large.

The saturated model for Poisson regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if n is fixed and all of the counts Y_i are large.

If $X \sim \chi^2_d$ then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $X > d + 3\sqrt{d}$ is unusually large and an observed value of $X < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the logistic or Poisson regression model is ok if $G^2 \leq n - p$ (or if $G^2 \leq n - p + 3\sqrt{n-p}$). For binary LR, the χ^2_{n-p+3} approximation for G^2 is rarely good even for large sample sizes n. For LR, the response plot is often a much better diagnostic for goodness of fit, especially when $ESP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ takes on many values and when $p << n$. For PR, both the response plot and $G^2 \leq n - p + 3\sqrt{n-p}$ should be checked.

The *Arc* output on the following two pages, shown in symbols and for a real data set, is used for the deviance test described below. Assume that the estimated sufficient summary plot has been made and that the logistic or Poisson regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta}_s = \mathbf{0}$. If this is the case, then the nontrivial predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta}_s = \mathbf{0}$ is not rejected, then for Poisson regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression $\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$ should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** is

i) $H_o : \boldsymbol{\beta}_s = \mathbf{0}$ $\quad H_A : \boldsymbol{\beta}_s \neq \mathbf{0}$

ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$.

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with $k = p - 1$ degrees of freedom. Note that $k = k + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_2, ..., X_p$. If p–value $\geq \delta$, then fail to

reject H_o and conclude that there is not a GLM relationship between Y and the predictors $X_2, ..., X_p$.

This test can be performed in R by obtaining output from the full and null model.

```
outf <- glm(Y~x2 + x3 + ... + xp, family = binomial)
outn <- glm(Y~1,family = binomial); anova(outn,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     k   G^2(0|F)     pvalue
```

Response = Y
Terms = $(X_2, ..., X_p)$
Sequential Analysis of Deviance

Predictor	df	Total Deviance	df	Change Deviance
Ones	$n-1=df_o$	G_o^2		
X_2	$n-2$		1	
X_3	$n-3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_p	$n-p=df_{FULL}$	G_{FULL}^2	1	

```
----------------------------------------
Data set = cbrain, Name of Fit = B1
Response      = sex
Terms         = (cephalic size log[size])
Sequential Analysis of Deviance
                     Total             Change
Predictor      df    Deviance     |    df   Deviance
Ones          266   363.820      |
cephalic      265   363.605      |     1   0.214643
size          264   315.793      |     1   47.8121
log[size]     263   305.045      |     1   10.7484
```

Response = Y Terms = $(X_2, ..., X_p)$ (Full Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

Degrees of freedom: n - p = df_{FULL}
Deviance: $D = G^2_{FULL}$

Response = Y Terms = $(X_2, ..., X_r)$ (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	for Ho: $\beta_r = 0$

Degrees of freedom: n - r - 1 = df_{RED}
Deviance: $D = G^2_{RED}$

```
(Full Model) Response = Status, Terms = (Diagonal Bottom Top)
Label      Estimate        Std. Error     Est/SE    p-value
Constant   2360.49         5064.42         0.466     0.6411
Diagonal  -19.8874         37.2830        -0.533     0.5937
Bottom     23.6950         45.5271         0.520     0.6027
Top        19.6464         60.6512         0.324     0.7460

Degrees of freedom:          196
Deviance:                  0.009

(Reduced Model) Response = Status, Terms = (Diagonal)
Label      Estimate        Std. Error     Est/SE    p-value
Constant   989.545         219.032         4.518     0.0000
Diagonal  -7.04376         1.55940        -4.517     0.0000

Degrees of freedom:          198
Deviance:                 21.109
```

The above output, shown both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This test is a competitor of the Wald test. This change in deviance test is usually better than the Wald test if the sample size n is not large, but the Wald test is often easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP(red) = \beta_{R1} + \beta_{R2} x_{R2} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p - r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$ while for Poisson regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good. Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** is

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$.

iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{p-r}$ has a chi–square distribution with $p - r$ degrees of freedom. Note that p is the number of predictors in the full model while r is the number of predictors in the reduced model. Also notice that $p - r = df_{RED} - df_{FULL} = n - r - (n - p)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

This test can be performed in R by obtaining output from the full and reduced model.

```
outf <- glm(Y~x2 + x3 + ... + xp, family = binomial)
outr <- glm(Y~ x3 + x5 + x7,family = binomial)
```

```
anova(outr,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     p-r  G^2(R|F)     pvalue
```

Interpretation of coefficients: if $x_2, ..., x_{i-1}, x_{i+1}, ..., x_p$ can be held fixed, then increasing x_i by 1 unit increases the sufficient predictor SP by β_i units. As a special case, consider logistic regression. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P(failure}|\boldsymbol{x})$ where a "success" is what is counted and a "failure" is what is not counted (so if the Y_i are binary, $\rho(\boldsymbol{x}) = P(Y_i = 1|\boldsymbol{x})$). Then the **estimated odds of success** is $\hat{\Omega}(\boldsymbol{x}) = \dfrac{\hat{\rho}(\boldsymbol{x})}{1 - \hat{\rho}(\boldsymbol{x})} = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})$. In logistic regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated odds of success by a factor of $\exp(\hat{\beta}_i)$.

10.6 Variable Selection

This section gives some rules of thumb for variable selection for logistic and Poisson regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x, sometimes x is not used by itself in the full model. Suppose that Y is binary. Then to decide what functions of x should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. The rules shown in Table 10.1 are used if x is an indicator variable or if x is a continuous variable. Replace normality by "symmetric with similar spreads" and "symmetric with different spreads" in the second and third lines of the table. See Cook and Weisberg (1999a, p. 501) and Kay and Little (1987).

The full model will often contain factors and interactions. If w is a nominal variable with J levels, make w into a factor by using use $J - 1$ (indicator or) dummy variables $x_{1,w}, ..., x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

As in Chapter 5, a **scatterplot matrix** is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary

Table 10.1 Building the Full Logistic Regression Model

distribution of $x\|y = i$	variables to include in the model
$x\|y = i$ is an indicator	x
$x\|y = i \sim N(\mu_i, \sigma^2)$	x
$x\|y = i \sim N(\mu_i, \sigma_i^2)$	x and x^2
$x\|y = i$ has a skewed distribution	x and $\log(x)$
$x\|y = i$ has support on (0,1)	$\log(x)$ and $\log(1 - x)$

logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make a response plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i$ = the number of 1's and $N_0 = n - N_1$ = the number of 0's. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$. For Poisson regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for a GLM can be described by

$$SP = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{10.11}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of r terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{10.12}$$

Definition 10.10. The model with $SP = \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model*. A model with $SP = \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel*. The full model is a submodel.

Suppose that S is a subset of I and that model (10.11) holds. Then

$$SP = \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{10.13}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ if the set of predictors S is a subset of I. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ be the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_I$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

Definition 10.11. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Find the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel I_I is the initial submodel to examine. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$.

Backward elimination starts with the full model and always contains the constant $x_1 = x_1^*$, and the predictor that optimizes some criterion is deleted. Then there are $p-1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $p-2, p-3, ..., 2$ and 1 predictors. The last model just has the constant $x_1 = x_1^*$.

Forward selection starts with the model with a constant $x_1 = x_1^*$ variables, and the predictor that optimizes some criterion is added. Then there is 2 variables in the model, and the predictor that optimizes some criterion is added. This process continues for models with $3, ..., p-1$ and p predictors. Both forward selection and backward elimination result in a sequence, often different, of p models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\} =$ full model.

All subsets variable selection can be performed with the following procedure. Compute the ESP of the GLM and compute the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that $|\text{corr(ESP, OLS ESP)}| \geq 0.95$. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection

with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ where the subset I has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 5.1, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (e.g. forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 12 rules of thumb to hold simultaneously. Let submodel I have r_I+1 predictors, including a constant. Do not use more predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Assume the response plot for the full model is good. Then the submodel I is good if
i) the response plot for the submodel looks like the response plot for the full model.
ii) corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For binary LR want $r_I+1 \leq \min(N_1, N_0)/10$. For PR, want $r_I+1 \leq n/10$.
vi) The plotted points in the VV plot cluster tightly about the identity line.
vii) Want the deviance $G^2(I) \geq G^2(full)$ but close. ($G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance.)
viii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.
xii) The OD plot should look good.

Heuristically, backward elimination tries to delete the variable that will increase the deviance the least. An increase in deviance greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with j predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the biggest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the model with $j+1$ terms from the previous step (using the j predictors in I and the variable x^*_{j+1}) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has

1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the smallest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, etc. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

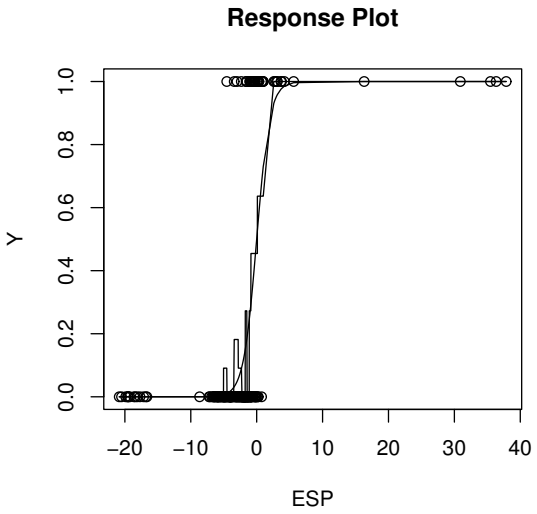

Fig. 10.17 Visualizing the ICU Data

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), a good response plot and an EE plot that clusters tightly about the identity line. If a factor has $I - 1$ dummy variables, either keep all $I - 1$ dummy variables or delete all $I - 1$ dummy variables, do not delete some of the dummy variables.

Some logistic regression output can be unreliable if $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly. Then $ESP = \infty$ or $ESP = -\infty$ respectively. Some binary logistic regression output can also be unreliable if there is perfect classification of 0's and 1's so that the 0's are to the left and the 1's to the right of $ESP = 0$ in the response plot. Then the logistic regression MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist, and variable selection rules of thumb may fail. Note that when there is perfect classification, the logistic regression model is very useful, but the logistic curve can not approximate a step function rising from 0 to 1 at $ESP = 0$, arbitrarily closely.

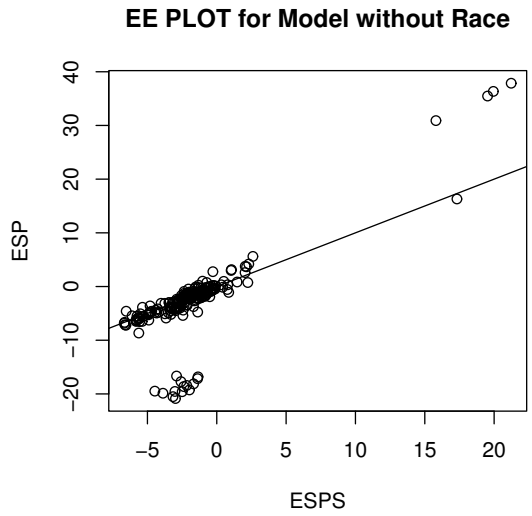

Fig. 10.18 EE Plot Suggests Race is an Important Predictor

Example 10.7. The ICU data is available from the text's website and from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). Also see Hosmer and Lemeshow (2000, p. 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).

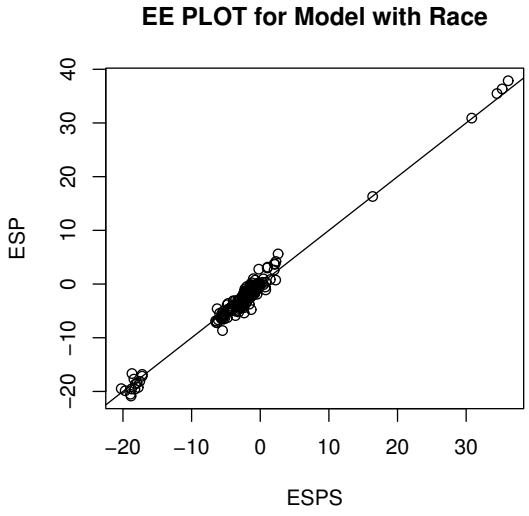

Fig. 10.19 EE Plot Suggests Race is an Important Predictor

Factors LOC and RACE had two indicator variables to model the three levels. The response plot in Figure 10.17 shows that the logistic regression model using the 19 predictors is useful for predicting survival, although the output has $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases. Note that the step function of slice proportions tracks the model logistic curve fairly well. Variable selection, using forward selection and backward elimination with the AIC criterion, suggested the submodel using AGE, CAN, SYS, TYP and LOC. The EE plot of ESP(sub) versus ESP(full) is shown in Figure 10.18. The plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. Since this clustering did not occur, the submodel seems to be poor. 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 10.19 shows the EE plot when RACE is added to the submodel. Then all of the points cluster about the identity line. Although numerical variable selection did not suggest that RACE is important, perhaps since output had $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases, the two EE plots suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black. This example illustrates how the plots can be used to quickly improve and check the models obtained by following logistic regression with variable selection even if the MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist.

10.7 Generalized Additive Models

There are many alternatives to the binomial and Poisson regression GLMs. Alternatives to the binomial GLM of Definition 10.3 include the discriminant function model of Definition 10.4, the quasi-binomial model, the binomial generalized additive model (GAM) and the beta-binomial model of Definition 10.5.

Alternatives to the Poisson GLM of Definition 10.6 include the the quasi-Poisson model, the Poisson GAM and the negative binomial regression model of Definition 10.7. Other alternatives include the zero truncated Poisson model, the zero truncated negative binomial model, the hurdle or zero inflated Poisson model, the hurdle or zero inflated negative binomial model, the hurdle or zero inflated additive Poisson model, and the hurdle or zero inflated additive negative binomial model. See Zuur, Ieno, Walker, Saveliev and Smith (2009), Simonoff (2003) and Hilbe (2011).

Many of these models can be visualized with response plots. An interesting research project would be to make response plots for these models, adding the conditional mean function and lowess to the plot. Also make OD plots to check whether the model handled overdispersion. This section will examine several of the above models, especially GAMs.

Definition 10.12. In a *1D regression*, Y is independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$ where $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ for a GLM. In a *generalized additive model*, Y is independent of $\boldsymbol{x} = (x_2, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=2}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated sufficient predictor* $\text{ESP} = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. The *estimated additive predictor* $\text{EAP} = \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(\boldsymbol{x}_j)$. An *ESP–response plot* is a plot of ESP versus Y while an *EAP–response plot* is a plot of EAP versus Y.

Note that a GLM is a special case of the GAM using $\beta_1 = \alpha$ and $S_j(x_j) = \beta_j x_j$ for $j = 2, ..., p$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3$ is a special case of a GAM with $x_4 \equiv x_2 x_3$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3$ is a special case of a GAM with $S_2(x_2) = \beta_2 x_2 + \beta_3 x_2^2$ and $S_3(x_3) = \beta_4 x_3$.

A GLM with p terms may be equivalent to a GAM with k terms $w_1, ..., w_k$ where $k < p$.

The plotted points in the EE plot defined below should scatter tightly about the identity line if the GLM is appropriate and if the sample size is large enough so that the ESP is a good estimator of the SP and the EAP is a good estimator of the AP. If the clustering is not tight but the GAM gives a reasonable approximation to the data, as judged by the EAP–response plot, then examine the $\hat{S}_j$ of the GAM to see if some simple terms such as x_i^2 can be added to the GLM so that the modified GLM has a good ESP–response plot. (This technique is easiest if the GLM and GAM have the same p terms $x_1 \equiv 1, x_2, ..., x_p$. The technique is more difficult, for example, if the GLM has terms x_2, x_2^2 and x_3 while the GAM has terms x_2 and x_3.)

Definition 10.13. An *EE plot* is a plot of EAP versus ESP.

Definition 10.14. Recall the binomial GLM

$$Y_i|SP_i \sim \textit{binomial}\left(m_i, \frac{\exp(SP_i)}{1+\exp(SP_i)}\right).$$

Let $\rho(w) = \exp(w)/[1+\exp(w)]$.

i) The *binomial GAM* is $Y_i|AP_i \sim \textit{binomial}\left(m_i, \frac{\exp(AP_i)}{1+\exp(AP_i)}\right)$. The EAP–response plot adds the estimated mean function $\rho(EAP)$ and a step function to the plot as done for the ESP–response plot of Section 10.3.

ii) The *quasi-binomial model* is a 1D regression model with $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$ and $V(Y_i|\boldsymbol{x}_i) = \phi\ m_i\ \rho(SP_i)(1-\rho(SP_i))$ where the dispersion parameter $\phi > 0$. Note that this model and the binomial GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

Definition 10.15. Recall the Poisson GLM $Y|SP \sim \textit{Poisson}(\exp(SP))$.

i) The *Poisson GAM* is $Y|AP \sim \textit{Poisson}(\exp(AP))$. The EAP–response plot adds the estimated mean function $\exp(EAP)$ and lowess to the plot as done for the ESP–response plot of Section 10.4.

ii) The *quasi-Poisson model* is a 1D regression model with $E(Y|\boldsymbol{x}) = \exp(SP)$ and $V(Y|\boldsymbol{x}) = \phi\exp(SP)$ where the dispersion parameter $\phi > 0$. Note that this model and the Poisson GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

For the quasi-binomial model, the conditional mean and variance functions are similar to those of the binomial distribution, but it is not assumed that $Y|SP$ has a binomial distribution. Similarly, it is not assumed that $Y|SP$ has a Poisson distribution for the quasi-Poisson model.

Next, some notation is needed to derive the zero truncated Poisson regression model. Y has a zero truncated Poisson distribution, $Y \sim ZTP(\mu)$,

if the probability mass function (pmf) of Y is $f(y) = \dfrac{e^{-\mu}\ \mu^y}{(1-e^{\mu})\ y!}$ for $y = 1, 2, 3, ...$ where $\mu > 0$. The ZTP pmf is obtained from a Poisson distribution where $y = 0$ values are truncated, so not allowed. If $W \sim Poisson(\mu)$ with pmf $f_W(y)$, then $P(W = 0) = e^{-\mu}$, so $\sum_{y=1}^{\infty} f_W(y) = 1 - e^{-\mu} = \sum_{y=0}^{\infty} f_W(y) - \sum_{y=1}^{\infty} f_W(y)$. So the ZTP pmf $f(y) = f_W(y)/(1 - e^{\mu})$ for $y \neq 0$.

Now $E(Y) = \sum_{y=1}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf_W(y)/(1 - e^{-\mu}) = E(W)/(1 - e^{-\mu}) = \mu/(1 - e^{-\mu})$.

Similarly, $E(Y^2) = \sum_{y=1}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f_W(y)/(1 - e^{-\mu}) = E(W^2)/(1 - e^{-\mu}) = [\mu^2 + \mu]/(1 - e^{-\mu})$. So

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{\mu^2 + \mu}{1 - e^{-\mu}} - \left(\frac{\mu}{1 - e^{-\mu}}\right)^2.$$

Definition 10.16. The *zero truncated Poisson regression* model has $Y|SP \sim ZTP(\exp(SP))$. Hence the parameter $\mu(SP) = \exp(SP)$,

$$E(Y|\boldsymbol{x}) = \frac{\exp(SP)}{1 - \exp(-\exp(SP))} \quad \text{and}$$

$$V(Y|SP) = \frac{[\exp(SP)]^2 + \exp(SP)}{1 - \exp(-\exp(SP))} - \left(\frac{\exp(SP)}{1 - \exp(-\exp(SP))}\right)^2.$$

The quasi-binomial, quasi-Poisson and zero truncated Poisson regression models have GAM analogs that replace SP by AP. The following examples are important, and the GLM or 1D regression analog of the GAM can be obtained by replacing AP by SP. Often the notation "GAM" can be replaced by "regression model" to obtain the GLM analog of the GAM. Hence the binary logistic regression model is the GLM analog of the binary logistic GAM.

1) The *additive model*

$$Y|AP = AP + e \tag{10.14}$$

has conditional mean function $E(Y|AP) = AP$ and conditional variance function $V(Y|AP) = \sigma^2 = V(e)$. Response transformations and prediction intervals for this GAM were discussed in Section 5.6. *Linear models*, including the *multiple linear regression model*, are the 1D regression analogs of the additive model.

2) The *response transformation model* is

$$Z = t^{-1}(AP + e) \quad \text{where} \quad \mathrm{Y = t(Z) = AP + e.} \tag{10.15}$$

Here, as is often the case when the error is additive, the conditioning $Y|AP$ is suppressed. See Section 5.6.

3) The *binary logistic GAM* states that $Y_1, ..., Y_n$ are independent with

$$Y|AP \sim \text{binomial}(1, \rho(\text{AP})) \text{ where } \rho(\text{AP}) = \frac{\exp(\text{AP})}{1+\exp(\text{AP})}, \quad (10.16)$$

and $\rho(AP) = P(\text{success}|\text{AP})$. This model has $E(Y|AP) = \rho(AP)$ and $V(Y|AP) = \rho(AP)(1-\rho(AP))$.

4) The *binomial logistic GAM* states that $Y_1, ..., Y_n$ are independent with

$$Y_i|AP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{AP}_\text{i})). \quad (10.17)$$

This model has $E(Y_i|AP_i) = m_i\rho(AP_i)$ and $V(Y_i|AP_i) = m_i\rho(AP_i)(1-\rho(AP_i))$. The binary model is a special case with $m_i \equiv 1$.

5) Following the notation for the beta-binomial distribution above Definition 10.5, the *beta-binomial GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i|AP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{AP}_\text{i}), \theta). \quad (10.18)$$

This model has $E(Y_i|AP_i) = m_i\rho(AP_i)$ and

$$V(Y_i|AP_i) = m_i\rho(AP_i)(1-\rho(AP_i))[1+(m_i-1)\theta/(1+\theta)].$$

Following Agresti (2002, p. 554-555), as $\theta \to 0$, it can be shown that the beta-binomial GAM converges to the binomial GAM.

6) The *Poisson GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|AP \sim \text{Poisson}(\exp(\text{AP})). \quad (10.19)$$

This model has $E(Y|AP) = V(Y|AP) = \exp(AP)$.

7) Following the notation for the negative binomial distribution above Definition 10.7, the *negative binomial GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|AP \sim \text{NB}(\exp(\text{AP}), \kappa). \quad (10.20)$$

This model has $E(Y|AP) = \exp(AP)$ and

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

Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the negative binomial GAM converges to the Poisson GAM.

8) Suppose Y has a gamma $\text{G}(\nu, \lambda)$ distribution so that $E(Y) = \nu\lambda$ and $V(Y) = \nu\lambda^2$. The *gamma GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|AP \sim G(\nu, \lambda = \mu(AP)/\nu). \quad (10.21)$$

Hence $E(Y|AP) = \mu(AP)$ and $V(Y|AP) = [\mu(AP)]^2/\nu$. The choices $\mu(AP) = AP$, $\mu(AP) = \exp(AP)$ and $\mu(AP) = 1/AP$ are common. Since $\mu(AP) > 0$, gamma GAMs that use the identity or reciprocal link run into problems if $\mu(EAP)$ is negative for some of the cases.

10.7.1 Response Plots

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. The response plots of Definition 10.12 are used to visualize the 1D regression model or GAM in the background of the data. 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 for the additive model, 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 (10.14) and (5.1) 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. See Example 5.12 for an additive model example. To avoid overfitting, assume $n > 5d$ where d is the model degrees of freedom. Hence $d = p$ for multiple linear regression.

If $Z_i = Y_i/m_i$, then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the binomial GAM can be visualized with a response plot of the EAP versus Z_i with the estimated mean function of the Z_i, $\hat{E}(Z|AP) = \dfrac{\exp(EAP)}{1 + \exp(EAP)}$, and a scatterplot smoother added to the plot as a visual aids. Instead of adding a lowess curve to the plot, consider the following alternative. Divide the EAP 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 response plot for the beta-binomial GAM is similar.

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 conditional mean function) closely, then the logistic conditional 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 (2000, p. 147-151).

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 response plots, 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.

Following the discussion above Definition 10.9, the *weighted forward response plot* is a plot of $\sqrt{Z_i}EAP$ versus $\sqrt{Z_i}\log(Z_i)$. The *weighted residual plot* is a plot of $\sqrt{Z_i}EAP$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}EAP$. These plots can also be used for the negative binomial GAM. If the counts Y_i are large and $\hat{E}(Y|AP) = \exp(EAP)$ is a good approximation to the conditional mean function $E(Y|AP) = \exp(AP)$, then the plotted points in the weighted forward response plot and weighted residual plot should scatter about the identity line and $r = 0$ lines in roughly evenly populated bands. See Examples 10.4, 10.5 and 10.6.

10.7.2 The EE 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 (2005) make an EE plot of $ESP(I)$ versus ESP where $ESP(I)$ is for a submodel I and ESP is for the full model. This plot can also be used to complement the hypothesis test that the reduced model I (which is selected before gathering data) can be used instead of the full model. The obvious extension to GAMs is to make the EE plot of $EAP(I)$ versus EAP. If the fitted full model and submodel I are good, then the plotted points should follow the identity line with high correlation (use correlation ≥ 0.95 as a benchmark).

To justify this claim, 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

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) + \sum_{k \in E} S_k(x_k) = \alpha + \sum_{j \in S} S_j(x_j). \quad (10.22)$$

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

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) = \alpha + \sum_{k \in I} S_k(x_k) = AP(I),$$

(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 if 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.

10.7.3 An EE Plot for Checking the GLM

One useful application of a GAM is for checking whether the corresponding GLM has the correct form of the predictors x_j in the model. Suppose a GLM and the corresponding GAM are both fit with the same link function where at least one general $S_j(x_j)$ was used. Since the GLM is a special case of the GAM, the plotted points in the EE plot of EAP versus ESP should follow the identity line with very high correlation if the fitted GLM and GAM are roughly equivalent. If the correlation is not very high and the GAM has some nonlinear $\hat{S}_j(x_j)$, update the GLM, and remake the EE plot. For example, update the GLM by adding terms such as x_j^2 and possibly x_j^3, or add $\log(x_j)$ if x_j is highly skewed. Then remake the EAP versus ESP plot.

10.7.4 Examples

For the binary logistic GAM, the EAP will not be a consistent estimator of the AP if the estimated probability $\hat{\rho}(AP) = \rho(EAP)$ is exactly zero or one. The following example will show that GAM output and plots can still be used for exploratory data analysis. The example also illustrates that EE plots are useful for detecting cases with high leverage and clusters of cases. Numerical diagnostics, such as analogs of Cook's distances (Cook 1977), tend to fail if there is a cluster of two or more influential cases.

Example 10.8. For the ICU data of Example 10.7, a binary generalized additive model was fit with unspecified functions for AGE, SYS and HRA and linear functions for the remaining 16 variables. Output suggested that functions for SYS and HRA are linear but the function for AGE may be slightly curved. Several cases had $\hat{\rho}(AP)$ equal to zero or one, but the response plot in Figure 10.20 suggests that the full model is useful for predicting survival. Note that the ten slice step function closely tracks the logistic curve. To visualize the model with the response plot, use $Y|\boldsymbol{x} \approx$ binomial[1,

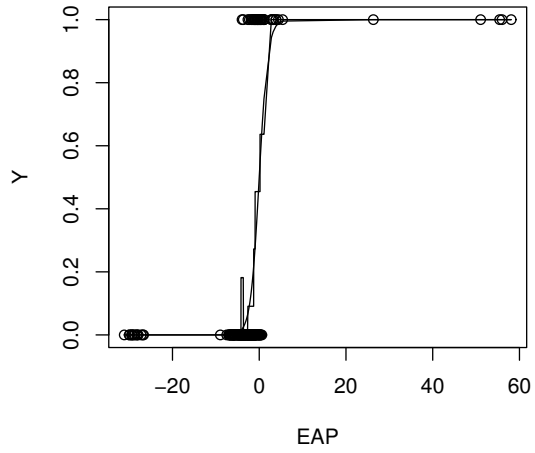

Fig. 10.20 Visualizing the ICU GAM

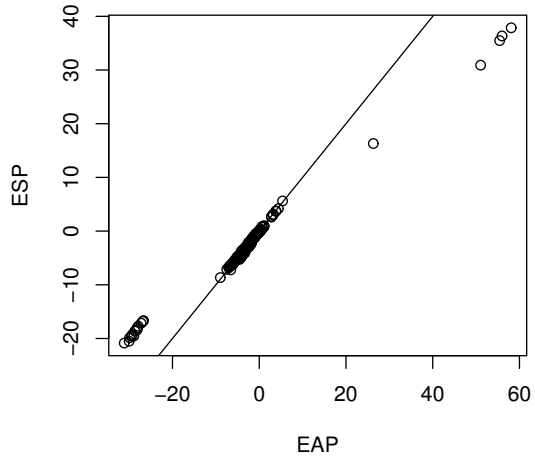

Fig. 10.21 GAM and GLM give Similar Success Probabilities

$\rho(EAP) = e^{EAP}/(1+e^{EAP})]$. 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 10.21 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 perform variable selection as in Example 10.7.

Example 10.9. For binary data, Kay and Little (1987) 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 (1905) data is useful for illustrating these suggestions. The response was *gender* with $Y = 1$ for male and $Y = 0$ for female. The predictors were *age, height* and the head measurements *circumference, length* and *size*. When the GAM was fit without $\log(age)$ or $\log(size)$, the $\hat{S}_j$ for *age*, *height* and $circumference$ were nonlinear. The log rule suggested adding $\log(age)$, and $\log(size)$ was added because *size* is skewed. The GAM for this model had plots of $\hat{S}_j(x_j)$ that were fairly linear. The response plot is not shown but was similar to Figure 10.6, and the step function tracked the logistic curve closely. When $EAP = 0$, the estimated probability of $Y = 1$ (male) is 0.5. When $EAP > 5$ the estimated probability is near 1, but near 0 for $EAP < -5$. The response plot for the binomial GLM, not shown, is similar. See Problem 10.14 for another analysis of this data set.

Example 10.10. Wood (2006, p. 82-86) describes heart attack data where the response Y is the *number of heart attacks* for m_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 $x_2 = ck$, $x_3 = [ck]^2$ and $x_4 = [ck]^3$ was fit and had AIC = 33.66. The binomial GAM with predictor x_2 was fit in *R*, and Figure 10.22 shows that the EE plot for the GLM 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 10.23 shows the EE plot, and Figure 10.24 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. The GAM using $log(ck)$ had a linear $\hat{S}$, and

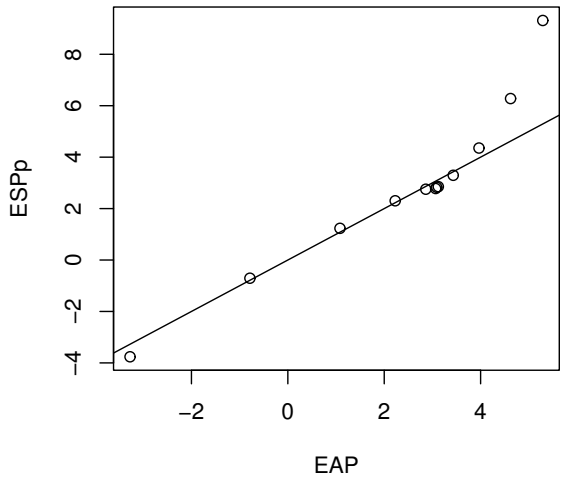

Fig. 10.22 EE plot for cubic GLM for Heart Attack Data

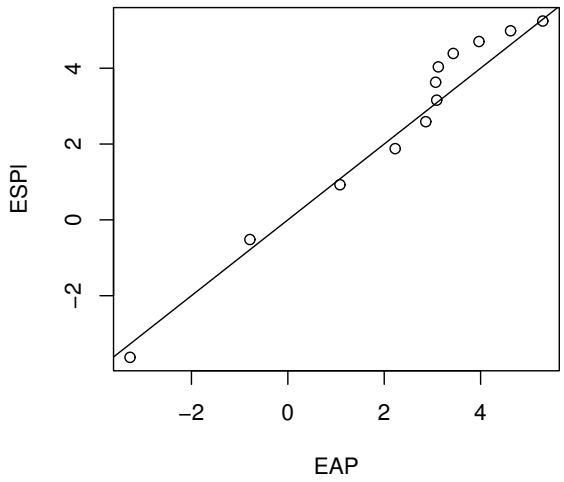

Fig. 10.23 EE plot with log(ck) in the GLM

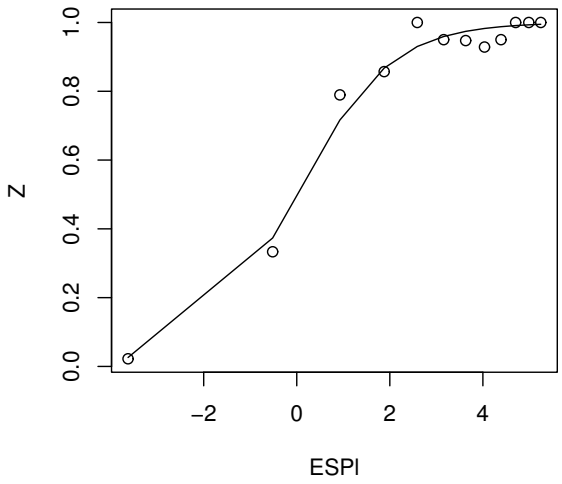

Fig. 10.24 Response Plot for Heart Attack Data

the correlation of the plotted points in the EE plot, not shown, was one. See Problem 10.22.

10.8 Overdispersion

Definition 10.17. Overdispersion occurs when the actual conditional variance function $V(Y|\boldsymbol{x})$ is larger than the model conditional variance function $V_M(Y|\boldsymbol{x})$.

Overdispersion can occur if the model is missing factors, if the response variables are correlated, if the population follows a mixture distribution, or if outliers are present. Typically it is assumed that the model is correct so $V(Y|\boldsymbol{x}) = V_M(Y|\boldsymbol{x})$. Hence the subscript M is usually suppressed. A GAM has conditional mean and variance functions $E_M(Y|AP)$ and $V_M(Y|AP)$ where the subscript M indicates that the function depends on the model. Then overdispersion occurs if $V(Y|\boldsymbol{x}) > V_M(Y|AP)$ where $E(Y|\boldsymbol{x})$ and $V(Y|\boldsymbol{x})$ denote the actual conditional mean and variance functions. Then the assumptions that $E(Y|\boldsymbol{x}) = E_M(Y|\boldsymbol{x}) \equiv m(AP)$ and $V(Y|\boldsymbol{x}) = V_M(Y|AP) \equiv v(AP)$ need to be checked.

First check that the assumption $E(Y|\boldsymbol{x}) = m(SP)$ is a reasonable approximation to the data using the response plot with lowess and the estimated

conditional mean function $\hat{E}_M(Y|\boldsymbol{x}) = \hat{m}(SP)$ added as visual aids. Overdispersion can occur even if the model conditional mean function $E(Y|SP)$ is a good approximation to the data. For example, for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, the binomial regression model is inappropriate since $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. Similarly, for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, the Poisson regression model is inappropriate since $V(Y|\boldsymbol{x}) > \exp(SP)$. If the conditional mean function is adequate, then we suggest checking for overdispersion using the *OD plot*.

Definition 10.18. For 1D regression, the *OD plot* is a plot of the estimated model variance $\hat{V}_M(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}_M(Y|SP)]^2$. Replace SP by AP for a GAM.

The OD plot has been used by Winkelmann (2000, p. 110) for the Poisson regression model where $\hat{V}_M(Y|SP) = \hat{E}_M(Y|SP) = \exp(ESP)$. 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 (2013), Collett (1999, ch. 6), and Winkelmann (2000).

For Poisson regression, Winkelmann (2000, 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 the estimated conditional mean and variance functions are both 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 to binomial regression if the counts are neither too big nor too small. OD plots can also be made for quasi-binomial and quasi-Poisson regression models. Replace SP by AP for the corresponding GAMs.

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.

Section 10.7 gives $E_M(Y|AP) = m(AP)$ and $V_M(Y|AP) = v(AP)$ for several models. Often $\hat{m}(AP) = m(EAP)$ and $\hat{v}(AP) = v(EAP)$, but additional parameters sometimes need to be estimated. Hence $\hat{v}(AP) = m_i\rho(EAP_i)(1-\rho(EAP_i))[1 + (m_i - 1)\hat{\theta}/(1 + \hat{\theta})]$, $\hat{v}(AP) = \exp(EAP) + \hat{\tau}\exp(2\ EAP)$, and $\hat{v}(AP) = [m(EAP)]^2/\hat{\nu}$ for the beta-binomial, negative binomial and gamma GAMs, respectively. The beta-binomial regression model is often used if the binomial regression is inadequate because of overdispersion, and the negative binomial GAM is often used if the Poisson GAM is inadequate.

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.

Since the Poisson regression (PR) model is simpler than the negative binomial regression (NBR) model, and the binomial logistic regression (LR) model is simpler beta-binomial regression (BBR) model, the graphical diagnostics for the goodness of fit of the PR and LR models are very useful. Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson and logistic regression models. NBR and BBR models should also be checked with response and OD plots. OD plots are also discussed in Sections 10.3 and 10.4. See Examples 10.2–10.6.

Example 10.11. The species data is from Cook and Weisberg (1999a, p. 285-286) and Johnson and Raven (1973). 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 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.

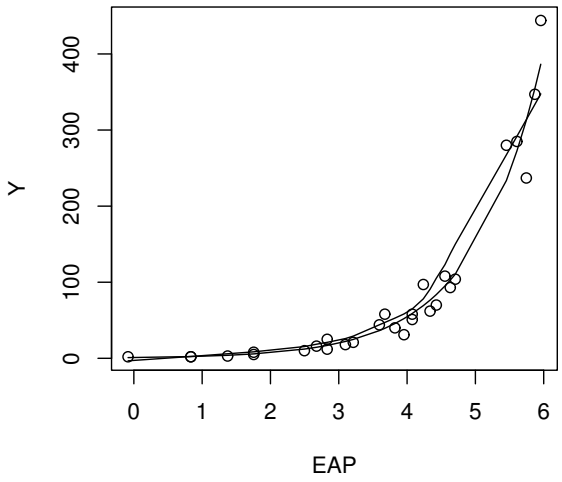

Fig. 10.25 Response Plot for Negative Binomial GAM

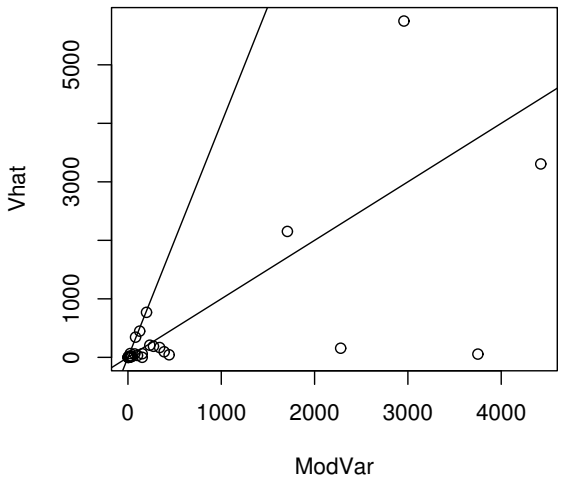

Fig. 10.26 OD Plot for Negative Binomial GAM

The response plot with the exponential and lowess curves added as visual aids is shown in Figure 10.25. 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 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 10.26 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 $\hat{\tau} = 1/37$.

10.9 Complements

GLMs were introduced by Nelder and Wedderburn (1972). Also see McCullagh and Nelder (1989), Myers, Montgomery and Vining (2002), Olive (2010), Andersen and Skovgaard (2010), Agresti (2012), and Cook and Weisberg (1999a, ch. 21-23). Collett (1999) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression while Cameron and Trivedi (2013) and Winkelmann (2008) cover Poisson regression. Alternatives to Poisson regression mentioned in Section 10.7 are covered by Zuur, Ieno, Walker, Saveliev and Smith (2009), Simonoff (2003) and Hilbe (2007).

Following Cook and Weisberg (1999a, p. 396), a residual plot is a plot of a function of the predictors versus the residuals, while a model checking plot is a plot of a function of the predictors versus the response. Hence response plots are a special case of model checking plots. See Cook and Weisberg (1997, 1999a, p. 397, 514, and 541). Cook and Weisberg (1999a, p. 515) add a lowess curve to the response plot. The scatterplot smoother lowess is due to Cleveland (1979).

In a *1D regression model*, $Y \perp\!\!\!\perp \boldsymbol{x}|h(\boldsymbol{x})$ where the real valued function $h : \mathcal{R}^p \to \mathcal{R}$. Then a plot of $\hat{h}(\boldsymbol{x})$ versus Y is a *response plot*. For this model, $Y|\boldsymbol{x}$ can be replace by $Y|h(\boldsymbol{x})$, and the response plot is also called an estimated sufficient summary plot. Note that $h(\boldsymbol{x}) = SP$ or AP and $\hat{h}(\boldsymbol{x}) = ESP$ or EAP for the GLM and the generalized additive model, respectively. The response plot is essential for understanding the model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ takes on many values. See Olive (2013b).

For Binomial regression and BBR, and for Poisson regression and NBR, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi (2013), Collett (1999, ch. 6), Hilbe (2011), Winkelmann (2000) and Zuur, Ieno, Walker, Saveliev and Smith (2009).

Olive and Hawkins (2005) give a simple all subsets variable selection procedure that can be applied to logistic regression and Poisson regression using readily available OLS software.

Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004) and Cook and Weisberg (1999a). Agresti (2012) incorporates some of the ideas from Section 10.6.

The existence of the logistic regression MLE is discussed in Albert and Andersen (1984) and Santer and Duffy (1986).

Results from Cameron and Trivedi (1998, p. 89) suggest that if a Poisson regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{PR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$. So a rough approximation is PR ESP $\approx$ (OLS ESP)$/\overline{Y}$. Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$.

A possible method for resistant binary regression is to use trimmed views but make the response plot for binary regression. This method would work best if $\boldsymbol{x}$ came from an elliptically contoured distribution. Another possibility is to substitute robust estimators for the classical estimators in the discrimination estimator.

Useful references for generalized additive models include Hastie and Tibshirani (1990) and Zuur, Ieno, Walker, Saveliev and Smith (2009). Large sample theory for the GAM is given by Wang, Liu, Liang and Carroll (2011). Olive (2013b) suggests plots for GAMS given in Sections 10.7 and 10.8. Section 5.2 of this book suggested a graphical method for response transformations.

Plots were made in *R* and *Splus*, see R Development Core Team (2011). The Wood (2006) library `mgcv` was used for fitting a GAM, and the Venables and Ripley (2010) library `MASS` was used for the negative binomial family. The Lesnoff and Lancelot (2010) *R* package `aod` has function `betabin` for beta binomial regression and is also useful for fitting negative binomial regression. *SAS* has `proc genmod`, `proc gam` and `proc countreg` which are useful for fitting GLMs such as Poisson regression, GAMs such as the Poisson GAM, and overdispersed count regression models. The *rpack R/Splus* functions include `lrplot` which makes response and OD plots for binomial regression; `lrplot2` which makes the response plot for binary regression; `prplot` which makes the response, weighted forward response, weighted residual and OD plots for Poisson regression; and `prsim` which makes the last 4 plots for simulated Poisson or negative binomial regression models.

10.10 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

`Output for problem 10.1: Response = sex`

```
Coefficient Estimates
Label       Estimate        Std. Error     Est/SE    p-value
Constant  -18.3500         3.42582        -5.356     0.0000
circum      0.0345827      0.00633521      5.459     0.0000
```

10.1. Consider trying to estimate the proportion of males from a population of males and females by measuring the circumference of the head. Use the above logistic regression output to answer the following problems.

a) Predict $\hat{\rho}(x)$ if $x = 550.0$.

b) Find a 95% CI for β.

c) Perform the 4 step Wald test for $Ho : \beta = 0$.

```
Output for Problem 10.2
Response      = sex
Coefficient Estimates
Label       Estimate        Std. Error     Est/SE    p-value
Constant  -19.7762         3.73243        -5.298     0.0000
circum      0.0244688      0.0111243       2.200     0.0278
length      0.0371472      0.0340610       1.091     0.2754
```

10.2*. Now the data is as in Problem 10.1, but try to estimate the proportion of males by measuring the circumference and the length of the head. Use the above logistic regression output to answer the following problems.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if circumference $= x_2 = 550.0$ and length $= x_3 = 200.0$.

b) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

c) Perform the 4 step Wald test for $Ho : \beta_3 = 0$.

```
Output for problem 10.3
Response      = ape
Terms         = (lower jaw, upper jaw, face length)
Trials        = Ones
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones           59   62.7188      |
lower jaw      58   51.9017      |     1   10.8171
upper jaw      57   17.1855      |     1   34.7163
face length    56   13.5325      |     1   3.65299
```

10.3*. A museum has 60 skulls of apes and humans. Lengths of the lower jaw, upper jaw and face are the explanatory variables. The response variable

is *ape* (= 1 if ape, 0 if human). Using the output above, perform the four step deviance test for whether there is a LR relationship between the response variable and the predictors.

```
Output for Problem 10.4.
Full Model
Response      = ape
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant     11.5092        5.46270        2.107     0.0351
lower jaw   -0.360127       0.132925      -2.709     0.0067
upper jaw    0.779162       0.382219       2.039     0.0415
face length -0.374648       0.238406      -1.571     0.1161

Number of cases:              60
Degrees of freedom:           56
Pearson X2:                16.782
Deviance:                  13.532

Reduced Model
Response      = ape
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant   8.71977         4.09466        2.130     0.0332
lower jaw -0.376256        0.115757      -3.250     0.0012
upper jaw  0.295507        0.0950855      3.108     0.0019

Number of cases:              60
Degrees of freedom:           57
Pearson X2:                28.049
Deviance:                  17.185
```

10.4*. Suppose the full model is as in Problem 10.3, but the reduced model omits the predictor *face length*. Perform the 4 step change in deviance test to examine whether the reduced model can be used.

The following three problems use the possums data from Cook and Weisberg (1999a).

```
Output for Problem 10.5
Data set = Possums, Response      = possums
Terms         = (Habitat Stags)
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -0.652653        0.195148      -3.344     0.0008
Habitat    0.114756        0.0303273      3.784     0.0002
Stags      0.0327213       0.00935883     3.496     0.0005
```

```
Number of cases:              151 Degrees of freedom:          148
Pearson X2:               110.187
Deviance:                 138.685
```

10.5*. Use the above output to perform inference on the number of possums in a given tract of land. The output is from a Poisson regression.

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *habitat* $= x_2 = 5.8$ and *stags* $= x_3 = 8.2$.

b) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

c) Find a 95% confidence interval for β_3.

```
Output for Problem 10.6
Response      = possums Terms         = (Habitat Stags)
                     Total            Change
Predictor      df   Deviance     |    df   Deviance
Ones          150   187.490      |
Habitat       149   149.861      |     1   37.6289
Stags         148   138.685      |     1   11.1759
```

10.6*. Perform the 4 step deviance test for the same model as in Problem 10.5 using the output above.

```
Output for Problem 10.7
Terms          = (Acacia Bark Habitat Shrubs Stags Stumps)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -1.04276         0.247944       -4.206     0.0000
Acacia     0.0165563       0.0102718       1.612     0.1070
Bark       0.0361153       0.0140043       2.579     0.0099
Habitat    0.0761735       0.0374931       2.032     0.0422
Shrubs     0.0145090       0.0205302       0.707     0.4797
Stags      0.0325441       0.0102957       3.161     0.0016
Stumps    -0.390753        0.286565       -1.364     0.1727
Number of cases:             151
Degrees of freedom:          144
Deviance:                 127.506
```

10.7*. Let the reduced model be as in Problem 10.5 and use the output for the full model be shown above. Perform a 4 step change in deviance test.

	B1	B2	B3	B4
df	945	956	968	974
# of predictors	54	43	31	25
# with $0.01 \leq$ Wald p-value ≤ 0.05	5	3	2	1
# with Wald p-value > 0.05	8	4	1	0
G^2	892.96	902.14	929.81	956.92
AIC	1002.96	990.14	993.81	1008.912
corr(B1:ETA'U,Bi:ETA'U)	1.0	0.99	0.95	0.90
p-value for change in deviance test	1.0	0.605	0.034	0.0002

10.8*. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. (Several of the predictors were factors, and a factor was considered to have a bad Wald p-value > 0.05 if all of the dummy variables corresponding to the factor had p-values > 0.05. Similarly the factor was considered to have a borderline p-value with $0.01 \leq$ p-value ≤ 0.05 if none of the dummy variables corresponding to the factor had a p-value < 0.01 but at least one dummy variable had a p-value between 0.01 and 0.05.) The response was binary and logistic regression was used. The response plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 1000 cases: for the response, 300 were 0's and 700 were 1's.

a) For the change in deviance test, if the p-value ≥ 0.07, there is little evidence that Ho should be rejected. If $0.01 \leq$ p-value < 0.07 then there is moderate evidence that Ho should be rejected. If p-value < 0.01 then there is strong evidence that Ho should be rejected. For which models, if any, is there strong evidence that "Ho: reduced model is good" should be rejected.

b) For which plot is "corr(B1:ETA'U,Bi:ETA'U)" (using notation from *Arc*) relevant?

c) Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

R Problems Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

Warning: Use a command like *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *regbootsim3*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim3)*, to display the needed arguments for the function.

10.9. Obtain the function `lrdata` from `rpack.txt`. Enter the commands

```
out <- lrdata()
x <- out$x
y <- out$y
```

Obtain the function `lressp` from `rpack.txt`. Enter the commands *lressp(x,y)* and include the resulting plot in *Word*.

10.10. Obtain the function `prdata` from `rpack.txt`. Enter the commands

```
out <- prdata()
x <- out$x
y <- out$y
```

a) Obtain the function `pressp` from `rpack.txt`. Enter the commands *pressp(x,y)* and include the resulting plot in *Word*.

b) Obtain the function `prplot` from `rpack.txt`. Enter the commands *prplot(x,y)* and include the resulting plot in *Word*.

10.11. In a generalized additive model (GAM), $Y \perp\!\!\!\perp \boldsymbol{x}|AP$ where $AP = \alpha + \sum_{i=2}^{p} S_i(x_i)$. In a generalized linear model (GLM), $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Note that a GLM is a special case of a GAM where $S_i(x_i) = \beta_i x_i$. A GAM is useful for showing that the predictors $x_1, ..., x_k$ in a GLM have the correct form, or if predictor transformations or additional terms such as x_i^2 are needed. If the plot of $\hat{S}_i(x_i)$ is linear, do not change x_i in the GLM, but if the plot is nonlinear, use the shape of $\hat{S}_i$ to suggest functions of x_i to add to the GLM, such as $\log(x_i), x_i^2$ and x_i^3. Refit the GAM to check the linearity of the terms in the updated GLM. Wood (2006, p. 82-86) describes heart attack data where the response Y is the *number of heart attacks* for m_i patients suspected of suffering a heart attack. The enzyme *ck* (creatine kinase) was measured for the patients. A binomial logistic regression (GLM) was fit with predictors $x_2 = ck$, $x_3 = [ck]^2$ and $x_4 = [ck]^3$. Call this the Wood model I_2. The predictor *ck* is skewed suggesting $\log(ck)$ should be added to

the model. Then output suggested that ck is not needed in the model. Let the binomial logistic regression model that uses $x = \log(ck)$ as the only predictor be model I_1. a) The R code for this problem from the URL above Problem 10.19 makes 4 plots. Plot a) shows $\hat{S}$ for the binomial GAM using ck as a predictor is nonlinear. Plot b) shows that $\hat{S}$ for the binomial GAM using $\log(ck)$ as a predictor is linear. Plot c) shows the EE plot for the binomial GAM using ck as the predictor and model I_1. Plot d) shows the response plot of ESP versus $Z_i = Y_i/m_i$, the proportion of patients suffering a heart attack for each value of $x_i = ck$. The logistic curve $= \hat{E}(Z_i|x_i)$ is added as a visual aid. Include these plots in *Word*.

Do the plotted proportions fall about the logistic curve closely?

b) The command for b) give AIC(outw) for model I_2 and AIC(out) for model I_1. Include the two AIC values below the plots in a).

A model I_1 with j fewer predictors than model I_2 is "better" than model I_2 if $AIC(I_1) \leq AIC(I_2) + 2j$. Is model I_1 "better" than model I_2?