Chapter 4
1D Regression Models Such as GLMs

... estimates of the linear regression coefficients are relevant to the linear parameters of a broader class of models than might have been suspected.
Brillinger (1977, p. 509)

After computing $\hat{\beta}$, one may go on to prepare a scatter plot of the points $(\hat{\beta}x_j, y_j)$, $j = 1, ..., n$ and look for a functional form for $g(\cdot)$.
Brillinger (1983, p. 98)

This chapter considers 1D regression models including additive error regression (AER), generalized linear models (GLMs), and generalized additive models (GAMs). Multiple linear regression is a special case of these four models.

See Definition 1.2 for the 1D regression model, sufficient predictor ($SP = h(\boldsymbol{x})$), estimated sufficient predictor ($ESP = \hat{h}(\boldsymbol{x})$), generalized linear model (GLM), and the generalized additive model (GAM). When using a GAM to check a GLM, the notation ESP may be used for the GLM, and EAP (estimated additive predictor) may be used for the ESP of the GAM. Definition 1.3 defines the response plot of ESP versus Y.

Suppose the sufficient predictor $SP = h(\boldsymbol{x})$. Often $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. If $\boldsymbol{u}$ only contains the nontrivial predictors, then $SP = \beta_1 + \boldsymbol{u}^T\boldsymbol{\beta}_2 = \alpha + \boldsymbol{u}^T\boldsymbol{\eta}$ is often used where $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$ and $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$.

4.1 Introduction

First we describe some regression models in the following three definitions. The most general model uses $SP = h(\boldsymbol{x})$ as defined in Definition 1.2. The GAM with SP = AP will be useful for checking the model (often a GLM) with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. Thus the additive error regression model with $SP = AP$ is useful for checking the multiple linear regression model. The model with $SP = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$ tends to have the most theory for inference and variable

selection. For the models below, the model estimated mean function and often a nonparametric estimator of the mean function, such as lowess, will be added to the response plot as a visual aid. For all of the models in the following three definitions, $Y_1, ..., Y_n$ are independent, but often the subscripts are suppressed. For example, $Y = SP + e$ is used instead of $Y_i = Y_i|\boldsymbol{x}_i = Y_i|SP_i = SP_i + e_i = h(\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$.

Definition 4.1. i) The **additive error regression (AER) model** $Y = SP + e$ has conditional mean function $E(Y|SP) = SP$ and conditional variance function $V(Y|SP) = \sigma^2 = V(e)$. See Section 4.2. The response plot of ESP versus Y and the residual plot of ESP versus $r = Y - \hat{Y}$ are used just as for multiple linear regression. The estimated model (conditional) mean function is the identity line $Y = ESP$. The *response transformation model* is $Y = t(Z) = SP + e$ where the response transformation $t(Z)$ can be found using a graphical method similar to Section 1.2.

ii) The **binary regression model** is $Y \sim \text{binomial}\left(1, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1 + \mathrm{e}^{\mathrm{SP}}}\right)$. This model has $E(Y|SP) = \rho = \rho(SP)$ and $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. Then $\hat{\rho} = \dfrac{e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function. See Section 4.3.

iii) The **binomial regression model** is $Y_i \sim \text{binomial}\left(\mathrm{m_i}, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1 + \mathrm{e}^{\mathrm{SP}}}\right)$. Then $E(Y_i|SP_i) = m_i\rho(SP_i)$ and $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$, and $\hat{E}(Y_i|\boldsymbol{x}_i) = m_i\hat{\rho} = \dfrac{m_i e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function. See Section 4.3.

iv) The **Poisson regression (PR) model** $Y \sim \text{Poisson}\left(\mathrm{e}^{\mathrm{SP}}\right)$ has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The estimated mean and variance functions are $\hat{E}(Y|\boldsymbol{x}) = e^{ESP}$. See Section 4.4.

v) Suppose Y has a gamma $\mathrm{G}(\nu, \lambda)$ distribution so that $E(Y) = \nu\lambda$ and $V(Y) = \nu\lambda^2$. The **Gamma regression model** $Y \sim G(\nu, \lambda = \mu(SP)/\nu)$ has $E(Y|SP) = \mu(SP)$ and $V(Y|SP) = [\mu(SP)]^2/\nu$. The estimated mean function is $\hat{E}(Y|\boldsymbol{x}) = \mu(ESP)$. The choices $\mu(SP) = SP$, $\mu(SP) = \exp(SP)$ and $\mu(SP) = 1/SP$ are common. Since $\mu(SP) > 0$, Gamma regression models that use the identity or reciprocal link run into problems if $\mu(ESP)$ is negative for some of the cases.

Alternatives to the binomial and Poisson regression models are needed because often the mean function for the model is good, but the variance function is not: there is overdispersion. See Section 4.8.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, pp. 93-94) and Agresti (2002, pp. 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 $Y \sim \text{BB}(\text{m},\rho,\theta)$. As $\theta \to 0$, it can be shown that $V(\pi) \to 0$, and the beta–binomial distribution converges to the binomial distribution.

Definition 4.2. The BBR model states that $Y_1,...,Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{SP}_\text{i}), \theta)$. Hence $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

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

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. As $\theta \to 0$, it can be shown that the BBR model converges to the binomial regression model.

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 $\text{V}(Y) = \mu+\mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ,ρ) distribution where $\rho = \kappa/(\mu+\kappa)$ and $\kappa>0$ is an unknown real parameter rather than a known integer.)

Definition 4.3. The **negative binomial regression (NBR) model** is $Y|SP \sim \text{NB}(\exp(\text{SP}),\kappa)$. Thus $E(Y|SP) = \exp(SP)$ and

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

The NBR model has the same mean function as the PR model but allows for overdispersion. Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the NBR model converges to the PR model.

Several important survival regression models are 1D regression models with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, including the Cox (1972) proportional hazards regression model. The following survival regression models are parametric. The *accelerated failure time model* has $\log(Y) = \alpha + SP_A + \sigma e$ where $SP_A = \boldsymbol{u}^T\boldsymbol{\beta}_A$, $V(e)=1$, and the e_i are iid from a location scale family. If the Y_i are log-

normal, the e_i are normal. If the Y_i are loglogistic, the e_i are logistic. If the Y_i are Weibull, the e_i are from a smallest extreme value distribution. The Weibull regression model is a proportional hazards model using Y_i and an accelerated failure time model using $\log(Y_i)$ with $\boldsymbol{\beta}_P = \boldsymbol{\beta}_A/\sigma$. Let Y hav a Weibull $W(\gamma, \lambda)$ distribution if the pdf of Y is

$$f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^{\gamma}]$$

for $y > 0$. Prediction intervals for parametric survival regression models are for survival times Y, not censored survival times. See Sections 4.10 and 4.11.

Definition 4.4. The *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP)),$$

where $\lambda_0 = \exp(-\alpha/\sigma)$.

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 $q \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. 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.

Definition 4.5. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{4.1}$$

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

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{4.2}$$

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 4.6. 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

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{4.3}$$

Here ϕ is a known constant (often a dispersion parameter), $a(\cdot)$ is a known function, and $\theta(\boldsymbol{x}_i) = \eta(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Let $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$. The GLM also states that $g(\mu(\boldsymbol{x}_i)) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ 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 (4.3) can be written in several ways. By Equation (4.2), $f(y_i|\theta(\boldsymbol{x}_i)) = \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) =$

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

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

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

$$\mu(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}). \tag{4.4}$$

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 $\theta(\boldsymbol{x}_i) = \eta(\boldsymbol{x}_i^T\boldsymbol{\beta})$ depends on the value of $\boldsymbol{x}_i$. Since the model depends on $\boldsymbol{x}$ only through the linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}$, 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. Inference and variable selection for these GLMs are discussed in Sections 4.5 and 4.6. Their generalized additive model analogs are discussed in Section 4.7.

4.2 Additive Error Regression

The linear regression model $Y = SP + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$ includes multiple linear regression (MLR) and many experimental design models as special cases. See Chapter 3 for MLR.

If Y is quantitative, a useful extension is the *additive error regression (AER) model* $Y = SP + e$ where $SP = h(\boldsymbol{x})$. See Definition 4.1 i). If $e \sim N(0, \sigma^2)$, then $Y \sim N(SP, \sigma^2)$. If $e \sim N(0, \sigma^2)$ and $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, then the resulting multiple linear regression model is also a GLM and an additive error regression model. The normality assumption is too restrictive since the error distribution is rarely normal. If m is a smooth function, the *additive error single index model*, where $SP = h(\boldsymbol{x}) = m(\boldsymbol{x}^T\boldsymbol{\beta})$, is an important special case.

Response plots, residual plots, and response transformations for the additive error regression model are very similar to those for the multiple linear regression model. See Olive (2004). To avoid overfitting, assume $n \geq 10d$ where d is the model degrees of freedom, possibly estimated. Hence $d = p$ for multiple linear regression with OLS. Prediction intervals are given in Section 2.3.

The GAM additive error regression model is useful for checking the multiple linear regression (MLR) model. Let $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ be the ESP for MLR where $\boldsymbol{x} = (1, x_2, ..., x_p)^T$. Let $ESP = EAP = \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(x_j)$ be the ESP for the GAM additive error regression model.

After making the usual checks on the MLR model, there are two useful plots that use the GAM. If the plotted points of the EE plot of EAP versus ESP cluster tightly about the identity line, then the MLR and the GAM produce similar fitted values. A plot of x_j versus $\hat{S}_j(x_j)$ can be useful for visualizing whether a predictor transformation $t_j(x_j)$ is needed for the jth predictor x_j. If the plot is linear then no transformation may be needed. If the plot is nonlinear, the shape of the plot, along with the graphical methods of Section 1.2, may be useful for suggesting the transformation t_j. The additive error regression GAM can be fit with all p of the S_j unspecified, or fit p GAMs where S_i is linear except for unspecified S_j where $j = 2, ..., p$. Some of these applications for checking GLMs with GAMs will be discussed in Section 4.7.

Suppose n/p is large and $SP = m(\boldsymbol{x}^T\boldsymbol{\beta})$. Olive (2008: ch. 12, 2010: ch. 15), Olive and Hawkins (2005), and Chang and Olive (2010) show that variable selection methods using C_p and the partial F test, originally meant for multiple linear regression, can be used (under regularity conditions) for the additive error single index model.

4.3 Binary, Binomial, and 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 4.7. 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

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\text{h}(\boldsymbol{x}_\text{i}))}{1 + \exp(\text{h}(\boldsymbol{x}_\text{i}))}. \tag{4.5}$$

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

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{4.6}$$

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

Thus the binary logistic regression model says that

$$Y|SP \sim \text{binomial}(1, \rho(\text{SP}))$$

where

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

for the LR model. Note that the conditional mean function $E(Y|SP) = \rho(SP)$ and the conditional variance function $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. For the LR model, the Y are independent and

$$Y|\boldsymbol{x} \approx \text{binomial}\left(1, \frac{\exp(\text{ESP})}{1 + \exp(\text{ESP})}\right),$$

or $Y|SP \approx Y|ESP \approx \text{binomial}(1, \rho(\text{ESP}))$.

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 $\rho = \rho(h(\boldsymbol{x}))$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\boldsymbol{x}^T\boldsymbol{\beta})) = \boldsymbol{x}^T\boldsymbol{\beta}$ and $g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}) = \rho(\boldsymbol{x}^T\boldsymbol{\beta})$. 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, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \boldsymbol{x}^T\boldsymbol{\beta}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, pp. 43–44). Assume that $\pi_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ 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 $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. Then as for the binary logistic regression model with $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ and $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$,

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

Definition 4.8. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{4.7}$$

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

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 $= \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. 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. See Figure 4.1 b). In this case the maximum likelihood estimator for the logistic regression parameters $\boldsymbol{\beta}$ does not exist because the logistic curve can not approximate a step function perfectly. See Atkinson and Riani (2000, pp. 251-254). 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 4.8 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983). 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.

Definition 4.9. For binary logistic regression, the *response plot* or *estimated sufficient summary plot* is the plot of the ESP $= \hat{h}(\boldsymbol{x}_i)$ versus Y_i with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid.

A scatterplot smoother such as lowess is also added as a visual aid. Alternatively, divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i / \sum_s m_i$ where $m_i \equiv 1$ and the sum is over the cases in slice s. Then plot the resulting step function.

Suppose that $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ is a $p \times 1$ vector of predictors where $q = p - 1$, $N_1 = \sum Y_i$ = the number of 1s and $N_0 = n - N_1$ = the number of 0s. Also assume that $q \leq \min(N_0, N_1)/5$. Then if the parametric estimated mean function $\hat{\rho}(ESP)$ looks like a smoothed version of the step function, then the LR model is likely to be useful. In other words, the observed slice proportions should scatter fairly closely about the logistic curve $\hat{\rho}(ESP) = \exp(ESP)/[1 + \exp(ESP)]$.

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 q, that the ESP takes on many values and that the binary LR model is a good approximation to the data. Then $Y|ESP \approx$ binomial(1, $\hat{\rho}(ESP)$). Unlike the response plot for multiple linear regression where the mean function is always the identity line, the mean function in the response plot for LR can take a variety of shapes depending on the range of the ESP. For LR, the (estimated) mean function is

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

If the ESP = 0 then $Y|SP \approx$ binomial(1,0.5). If the ESP $= -5$, then $Y|SP \approx$ binomial(1,$\rho \approx 0.007$) while if the ESP = 5, then $Y|SP \approx$ binomial(1,$\rho \approx 0.993$). Hence if the range of the ESP is in the interval $(-\infty, -5)$ then the mean function is flat and $\hat{\rho}(ESP) \approx 0$. If the range of the ESP is in the interval $(5, \infty)$ then the mean function is again flat but $\hat{\rho}(ESP) \approx 1$. If $-5 < ESP < 0$ then the mean function looks like a slide. If $-1 < ESP < 1$

then the mean function looks linear. If $0 < ESP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < ESP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 4.1 c).

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 Ys in each slice and add the resulting step function to the response plot. This is done in Figure 4.1 c) with $J = 4$ 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, pp. 147–156).

The deviance test described in Section 4.5 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

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

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_0 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. See Figure 4.1 a).

For binomial logistic regression, the response plot needs to be modified and a check for overdispersion is needed.

Definition 4.10. 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

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

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 or the lowess curve. For binary data the step function is simply the sample proportion in each slice.

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*. The BBR model of Definition 4.2 is a useful alternative to LR.

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 the binomial regression model is simpler than the BBR model, graphical diagnostics for the goodness of fit of the LR model would be useful. The following plot was suggested by Olive (2013b) to check for overdispersion.

Definition 4.11. To check for overdispersion, use the *OD plot* of the estimated model variance $\hat{V}_M \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$.

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 $G^2/(n-p) \approx 1$, then the LR model is likely useful. If $G^2 > (n-p)+3\sqrt{n-p}$, then a more complicated count model may be needed.

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.

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. If the data are binary, the response plot is enough to check the binomial regression assumption.

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 $\approx 1 + (m-1)\dfrac{\theta}{1+\theta} = \dfrac{1+m\theta}{1+\theta}$.

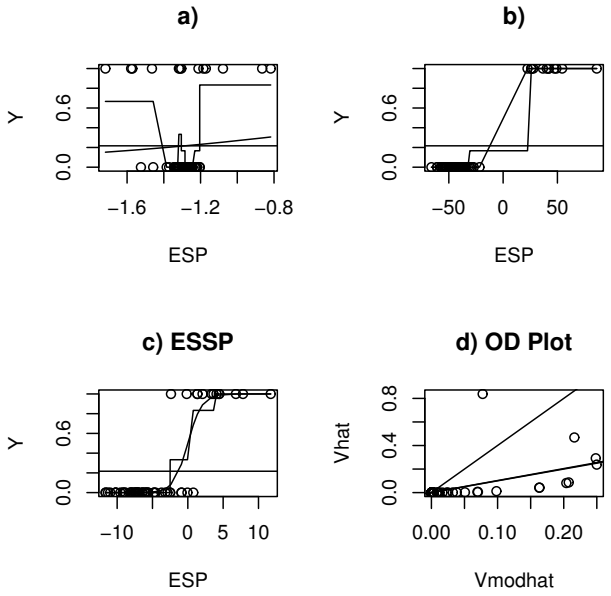

Fig. 4.1 Response Plots for Museum Data

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. The following three examples used $SP = \boldsymbol{x}^T\boldsymbol{\beta}$.

Example 4.1. 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 4.1a) uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The response plot in Figure 4.1b) 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. The response plot in Figure 4.1c 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 4.1d) is curved and is not needed for a binary response.

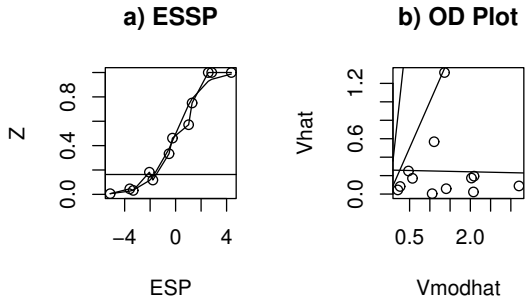

Fig. 4.2 Visualizing the Death Penalty Data

Example 4.2. Abraham and Ledolter (2006, pp. 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 4.2a 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 4.2b 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.

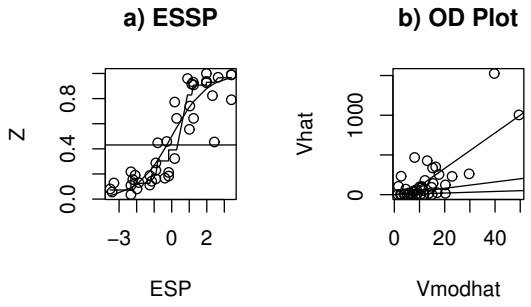

Fig. 4.3 Plots for Rotifer Data

Example 4.3. Collett (1999, pp. 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 4.3a 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 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.

4.4 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 4.12. The **Poisson regression (PR) model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i}))$ where $\mu(\boldsymbol{x}_i) = \exp(h(\boldsymbol{x}_i))$. Thus $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$. Notice that $Y|SP = 0 \sim \text{Poisson}(1)$. Note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$.

In the response plot for Poisson regression, the shape of the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ depends strongly on the range of the ESP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence if the range of the ESP is narrow, then the exponential function will be rather flat. If the range of the ESP 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.

Definition 4.13. The estimated sufficient summary plot (ESSP) or *response plot*, is a plot of the $ESP = \hat{h}(\boldsymbol{x}_i)$ versus Y_i with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. A scatterplot smoother such as lowess is also added as a visual aid.

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function and is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve). See Figure 4.4 a). If the number of notrivial predictors $q < n/10$, if there is no overdispersion, and if the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the PR mean function may be a useful approximation for $E(Y|\boldsymbol{x})$. **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 4.5 is used to test whether $\boldsymbol{\beta} = \boldsymbol{0}$, 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 $\boldsymbol{\beta} = \boldsymbol{0}$, 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

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

If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_0 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. See Figure 4.6 a).

Warning: For many count data sets where the PR mean function is good, 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 (1999, pp. 401-403). The NBR model of Definition 4.3 is a useful alternative to PR.

Since the Poisson regression model is simpler than the NBR model, graphical diagnostics for the goodness of fit of the PR model would be useful. The following plot was suggested by Winkelmann (2000, p. 110).

Definition 4.14. To check for overdispersion, use the **OD plot** of the estimated model variance $\hat{V}_M \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the PR model, $\hat{V}(Y|SP) = \exp(ESP) = \hat{E}(Y|SP)$ and $\hat{V} = [Y - \exp(ESP)]^2$.

Numerical summaries are also available. The deviance G^2, described in Section 4.5, is a statistic 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, G^2 is approximately chi-square with $n - p$ 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 the response and OD plots look good, and $G^2/(n-p) \approx 1$, then the PR model is likely useful. If $G^2 > (n-p) + 3\sqrt{n-p}$, then a more complicated count model than PR may be needed. A good discussion of such count models is in Simonoff (2003).

For PR, Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about the identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the PR model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use for Poisson regression.

First, recall that a normal approximation is good for both the Poisson and negative binomial distributions if the count Y is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the

origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. If the normal approximation is good, only about 5% of the plotted points should be above this line.

Second, 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. (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.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot.

For Poisson regression, 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. Definition 4.16 will give some useful plots. 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 $Y_i = 0$. Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$.

Definition 4.15. The **minimum chi–square estimator** of the parameters $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$ in a Poisson regression model are $(\hat{\alpha}_M, \hat{\boldsymbol{\eta}}_M)$, and are found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{u}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{u}_i^T)^T$.

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 $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$ tends to be consistent if the sample size $n \to \infty$. See Agresti (2002, pp. 611-612). However, the two estimators are often close for many data sets.

The basic idea of the following two plots for Poisson regression is to transform the data towards a linear model, then make the response plot of $\hat{W}$ versus W and residual plot of the residuals $W - \hat{W}$ for the transformed response variable W. The mean function is the identity line and the vertical deviations from the identity line are the WLS residuals. If $ESP = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, The plots are based on weighted least squares (WLS) regression. Use the equivalent OLS regression (without intercept) of $W = \sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{u}_i^T)^T$. Then the plot of the "fitted values" $\hat{W} = \sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\eta}}_M^T\boldsymbol{u}_i)$ versus the "response" $\sqrt{Z_i}\log(Z_i)$ should have points that scatter about the identity line.

These results and the equivalence of the minimum chi–square estimator to an OLS estimator suggest the following diagnostic plots.

Definition 4.16. For a Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}ESP$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}ESP$.

If the Poisson regression model is appropriate and the PR estimator is good, then the plotted points in the weighted fit response plot should follow the identity line. When the counts Y_i are small, the "WLS" residuals can not be expected to be approximately normal. Often the larger counts are fit better than the smaller counts and hence the residual plots have a "left opening megaphone" shape. This fact makes residual plots for Poisson regression rather hard to use, but cases with large "WLS" residuals may not be fit very well by the model. Both the weighted fit response and residual plots perform better for simulated PR data with many large counts than for data where all of the counts are less than 10. The following three examples use $SP = \boldsymbol{x}^T\boldsymbol{\beta}$.

Example 4.4. For the Ceriodaphnia data of Myers et al. (2002, pp. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$, and the predictors were a constant (x_1), seven concentrations of jet fuel (x_2), and an indicator for two strains of organism (x_3). The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 4.4 shows the 4 plots for this data. In the response plot of Figure 4.4a, 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 4.4b suggests that there is little evidence of overdispersion. These two plots as well as Figures 4.4c and 4.4d suggest that the Poisson regression model is a useful approximation to the data.

Example 4.5. 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, pp. 126-131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 4.5a 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 4.5b 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 4.5c 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.

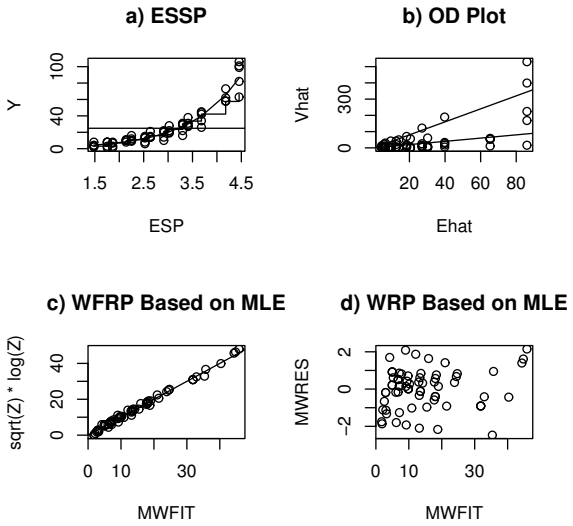

Fig. 4.4 Plots for Ceriodaphnia Data

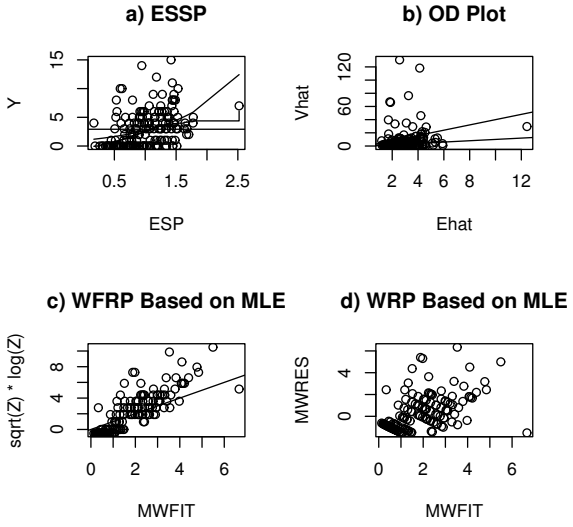

Fig. 4.5 Plots for Crab Data

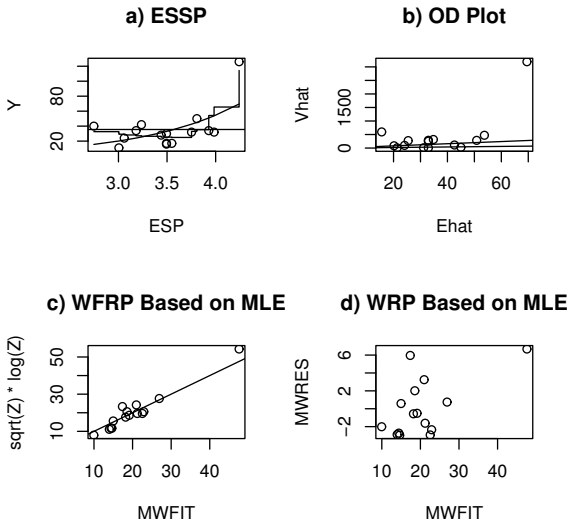

Fig. 4.6 Plots for Popcorn Data

Example 4.6. For the popcorn data of Myers et al. (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 4.6a 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 4.6b 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. However, we probably need to delete the high temperature, low oil, and long popping time combination, to conclude that the response is independent of the predictors.

4.5 GLM Inference, n/p Large

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 $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, x_2, ..., x_p)^T$ given the sufficient predictor $\boldsymbol{x}^T\boldsymbol{\beta}$ where the constant $x_1 \equiv 1$.

To perform inference for LR and PR, computer output is needed. Shown below is output using symbols and output from a real data set with $p = 3$ nontrivial predictors. This data set is the *banknote* data set described in Cook and Weisberg (1999, 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.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for $H_0 : \beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for $H_0 : \beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for $H_0 : \beta_p = 0$

```
Number of cases:                n
Degrees of freedom:             n - p
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
```

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

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})}. \tag{4.8}$$

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

$$\hat{\mu}(\boldsymbol{x}) = \exp(\boldsymbol{x}^T \hat{\boldsymbol{\beta}}). \tag{4.9}$$

For tests, pval, the estimated p–value, is an important quantity. Again what output labels as p–value is typically pval. Recall that H_0 is rejected if the pval $\leq \delta$. A pval between 0.07 and 1.0 provides little evidence that H_0 should be rejected, a pval between 0.01 and 0.07 provides moderate evidence and a pval less than 0.01 provides strong statistical evidence that H_0 should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the pval along with a statement of the strength of the evidence is more informative than stating that the pval 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 $p-1$ predictors are in the model with the following **4 step Wald test of hypotheses**.
i) State the hypotheses $H_0 : \beta_j = 0 \quad H_A : \beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The pval $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the pval from output or use the standard normal table.
iv) State whether you reject H_0 or fail to reject H_0 and give a nontechnical sentence restating your conclusion in terms of the story problem.

If H_0 is rejected, then conclude that x_j is needed in the GLM model for Y given that the other $p-1$ predictors are in the model. If you fail to reject H_0, then conclude that x_j is not needed in the GLM model for Y given that the other $p-1$ predictors are in the model. (Or there is not enough evidence to conclude that x_j is needed 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 $r+1$ 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} 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 $ESP = \boldsymbol{x}_i^T\boldsymbol{\beta}$ 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.

Response = Y
Terms = $(x_1, ..., x_p)$
Sequential Analysis of Deviance

Predictor	df	Total Deviance	Change df	Deviance
Ones	$n - 1 = df_o$	G^2_o		
x_2	$n - 2$		1	
x_3	$n - 3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_p	$n - p = df_{FULL}$	G^2_{FULL}	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
```

The above output, shown in symbols and for a real data set, is used for the deviance test described below. Assume that the response 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 $\boldsymbol{\beta}_2 = \mathbf{0}$ where $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$. If this is the case, then the nontrivial predictors are not needed in the GLM model. If $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ 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 since $m_i \equiv 1$ for $i = 1, ..., n$. This test is similar to the ANOVA F test for multiple liner regression.

The 4 step **deviance test** is

i) $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ $\quad H_A : \boldsymbol{\beta}_2 \neq \mathbf{0}$,

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

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

iv) Reject H_0 if the pval $\leq \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_2, ..., X_p$. If pval $> \delta$, then fail to reject H_0 and conclude that there is not a GLM relationship between Y and the predictors $X_2, ..., X_p$. (Or there is not enough evidence to conclude that there is a GLM relationship between Y and the predictors.)

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
```

The output below, 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 $H_0 : \beta_i = 0$ 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 $ESP(R) = \boldsymbol{x}_{Ri}^T \hat{\boldsymbol{\beta}}_R$ versus $ESP = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ should be highly correlated with the identity line with unit slope and zero intercept.

Response = Y Terms = $(x_1, ..., x_p)$ (Full Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for $H_0 : \beta_1 = 0$
x_2	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for $H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_q$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for $H) : \beta_p = 0$

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

Response = Y Terms = $(x_1, ..., x_r)$ (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for $H_0 : \beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for $H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_k/se(\hat{\beta}_r)$	for $H_0 : \beta_r = 0$

Degrees of freedom: $n - r = df_{RED}$
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
```

After obtaining an acceptable full model where

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

try to obtain a **reduced model**

$$SP(red) = \beta_1 + \beta_{R2} x_{R2} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{x}_R^T\boldsymbol{\beta}_R$$

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_0: 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 next test is similar to the partial F test for multiple linear regression.

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

i) H_0: the reduced model is good H_A: use the full model,

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

iii) The pval $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{p-r}$ has a chi–square distribution with $p - r$ degrees of freedom. Note that $p - 1$ is the number of nontrivial predictors in the full model while $r - 1$ is the number of nontrivial predictors in the reduced model. Also notice that $p - r = df_{RED} - df_{FULL} = n - r - (n - p) = (p - 1) - (r - 1)$.

iv) Reject H_0 if the pval $\leq \delta$ and conclude that the full model should be used. If pval $> \delta$, then fail to reject H_0 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~ x4 + x6 + x8,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 $\hat{\Omega}(\boldsymbol{x}) = \dfrac{\hat{\rho}(\boldsymbol{x})}{1 - \hat{\rho}(\boldsymbol{x})} = \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$. 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 $\exp(\hat{\beta}_i)$.

```
Output for Full Model, Response = gender, Terms =
(age log[age] breadth circum headht
height length size log[size])
Number of cases: 267, Degrees of freedom: 257,
Deviance: 234.792

Logistic Regression Output for Reduced Model,
Response      = gender, Terms      = (height  size)
Label      Estimate  Std. Error   Est/SE    p-value
Constant  -6.26111    1.34466     -4.656    0.0000
height    -0.0536078  0.0239044   -2.243    0.0249
size       0.0028215  0.000507935  5.555    0.0000

Number of cases: 267, Degrees of freedom:  264
Deviance:                313.457
```

Example 4.7. Let the response variable $Y = \textit{gender} = 0$ for F and 1 for M. Let $x_2 = \textit{height}$ (in inches) and $x_3 = \textit{size}$ of head (in mm^3). Logistic regression is used, and data is from Gladstone (1905). There is output above.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if height $= x_2 = 65$ and size $= x_3 = 3500$.

b) The full model uses the predictors listed above to the right of Terms. Perform a 4 step change in deviance test to see if the reduced model can be used. Both models contain a constant.

Solution: a) $ESP = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -6.26111 - 0.0536078(65) + 0.0028215(3500) = 0.1296$. So

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{1.1384}{1 + 1.1384} = 0.5324.$$

b) i) H_0: the reduced model is good $\quad H_A$: use the full model

ii) $G^2(R|F) = 313.457 - 234.792 = 78.665$

iii) Now $df = 264 - 257 = 7$, and comparing 78.665 with $\chi^2_{7,0.999} = 24.32$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject H_0, use the full model.

Example 4.8. Suppose that Y is a 1 or 0 depending on whether the person is or is not credit worthy. Let $x2$ through $x7$ be the predictors and

use the following output to perform a 4 step deviance test. The credit data is available from the text's website as file *credit.lsp*, and is from Fahrmeir and Tutz (2001).

```
Response      = y
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones          999   1221.73      |
x2            998   1177.11      |     1   44.6148
x3            997   1176.55      |     1   0.561629
x4            996   1168.33      |     1   8.21723
x5            995   1168.20      |     1   0.137583
x6            994   1163.44      |     1   4.75625
x7            993   1158.22      |     1   5.21846
```

Solution: i) $H_0 : \beta_2 = \cdots = \beta_7 \quad H_A$: not H_0

ii) $G^2(0|F) = 1221.73 - 1158.22 = 63.51$

iii) Now $df = 999 - 993 = 6$, and comparing 63.51 with $\chi^2_{6,0.999} = 22.46$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject H_0, there is a LR relationship between $Y =$ credit worthiness and the predictors $x_2, ..., x_7$.

```
Coefficient Estimates
Label       Estimate    Std. Error   Est/SE  p-value
Constant  -5.84211     1.74259     -3.353   0.0008
jaw ht     0.103606    0.0383650      ?      ??
```

Example 4.9. A museum has 60 skulls, some of which are human and some of which are from apes. Consider trying to estimate whether the *skull type* is human or ape from the *height of the lower jaw*. Use the above logistic regression output to answer the following problems. The museum data is available from the text's website as file *museum.lsp*, and is from Schaaffhausen (1878). Here $x = x_2$.

a) Predict $\hat{\rho}(x)$ if $x = 40.0$.

b) Find a 95% CI for β_2.

c) Perform the 4 step Wald test for $H_0 : \beta_2 = 0$.

Solution: a) $\exp[ESP] = \exp[\hat{\beta}_1 + \hat{\beta}_2(40)] = \exp[-5.84211 + 0.103606(40)] = \exp[-1.69787] = 0.1830731$. So

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{0.1830731}{1 + 0.1830731} = 0.1547.$$

b) $\hat{\beta}_2 \pm 1.96 SE(\hat{\beta}_2) = 0.103606 \pm 1.96(0.03865) = 0.103606 \pm 0.0751954 = [0.02841, 0.1788]$.

c) i) $H_0 : \beta_2 = 0 \quad H_A : \quad \beta_2 \neq 0$

ii) $Z_0 = \dfrac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \dfrac{0.103606}{0.038365} = 2.7005.$

iii) Using a standard normal table, $pval = 2P(Z < -2.70) = 2(0.0035) = 0.0070$.

iv) Reject H_0, jaw height is a useful LR predictor for whether the skull is human or ape (so is needed in the LR model).

Label	Estimate	Std. Error	Est/SE	p-value
Constant	-0.406023	0.877382	-0.463	0.6435
bombload	0.165425	0.0675296	2.450	0.0143
exper	-0.0135223	0.00827920	-1.633	0.1024
type	0.568773	0.504297	1.128	0.2594

Example 4.10. Use the above output to perform inference on the number of locations where aircraft was damaged. The output is from a Poisson regression. The variable exper = total months of aircrew experience while type of aircraft was coded as 0 or 1. There were $n = 30$ cases. Data is from Montgomery et al. (2001).

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *bombload* $= x_2 = 7.0$, *exper* $= x_3 = 80.2$, and *type* $= x_4 = 1.0$.

b) Perform the 4 step Wald test for $H_0 : \beta_3 = 0$.

c) Find a 95% confidence interval for β_4.

Solution: a) $ESP = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 = -0.406023 + 0.165426(7) - 0.0135223(80.2) + 0.568773(1) = 0.2362$. So $\hat{\mu}(\boldsymbol{x}) = \exp(ESP) = \exp(0.2360) = 1.2665$.

b) i) $H_0 : \beta_3 = 0 \quad H_A : \ \beta_3 \neq 0$

ii) $t_{03} = -1.633$.

iii) pval = 0.1024

iv) Fail to reject H_0, *exper* in not needed in the PR model for number of locations given that *bombload* and *type* are in the model.

c) $\hat{\beta}_4 \pm 1.96 SE(\hat{\beta}_4) = 0.568773 \pm 1.96(0.504297) = 0.568773 \pm 0.9884 = [-0.4196, 1.5572]$.

4.6 Variable and Model Selection

4.6.1 When n/p is Large

This subsection gives some rules of thumb for variable selection for logistic and Poisson regression when $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. 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 4.1 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 (1999, p. 501) and Kay and Little (1987).

The full model will often contain factors and interactions. If w is a nominal variable with K levels, make w into a factor by using $K - 1$ (indicator or) dummy variables $x_{1,w}, ..., x_{K-1,w}$ 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.

Table 4.1 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 $\log(1 - x)$

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 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 1s and $N_0 = n - N_1$ = the number of 0s. 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 is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for many models, including GLMs, is given is Section 2.1. Let ESP correspond to the full model and let $ESP(I)$ correspond to the submodel I.

Definition 4.17. 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 with $q = p - 1$ nontrivial variables, and the predictor that optimizes some criterion is deleted. A constant $x_1^* = x_1 \equiv 1$ is always in the model. Then there are $q - 1$ nontrivial variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $q - 2, q - 3, ..., 2$, and 1 predictors.

Forward selection starts with the model with a constant $x_1^* = x_1 \equiv 1$, and the predictor that optimizes some criterion is added. Then there are 2 variables in the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., p-1$, 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 $C_p(I) \leq 2r$ where the subset I has r variables including a constant, then corr(OLS ESP, OLS ESP(I))

will be high by Olive and Hawkins (2005), 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 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 pval ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For binary LR want $r_I \leq \min(N_1, N_0)/10$. For PR, want $r_I \leq n/10$.
vi) Fit OLS to the full and reduced models. The plotted points in the plot of the OLS residuals from the submodel versus the OLS residuals from the full model should 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 pvals > 0.05.
x) Want few predictors with pvals between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I + 3\sqrt{n - r_I}$.
xii) The OD plot should look good.

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 pval (preferably from a change in deviance test but possibly from a Wald test) in the test $H_0 : \beta_i = 0$ 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.

The final submodel should have few predictors, few variables with large Wald pvals (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 $K-1$ dummy variables, either keep all $K-1$ dummy variables or delete all $K-1$ 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 0s and 1s so that the 0s are to the left and the 1s 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.

Example 4.11. The following output is for forward selection. All models use a constant. For forward selection, the min AIC model uses {F}LOC, TYP, AGE, CAN, SYS, PCO, and PH. Model I_I uses {F}LOC, TYP, AGE, CAN, and SYS. Let model I use {F}LOC, TYP, AGE, and CAN. This model may be good, so for forward selection, models I_I and I are the first models to examine. {F}LOC is notation used for a factor with $K-1=3$ dummy variables, while k is the number of variables in I, including a constant. Output is from the Cook and Weisberg (1999) *Arc* software.

```
Forward Selection                                  comment

Base terms: ({F}LOC TYP)
       Deviance Pearson X2 |  k  AIC  > min AIC + 7
Add:AGE 141.873  187.84    |  5  151.873

Base terms: ({F}LOC TYP AGE)
        Deviance Pearson X2|  k  AIC < min AIC + 7
Add:CAN 134.595  170.367   |  6  146.595
          ({F}LOC TYP AGE CAN) could be a good model

Base terms: ({F}LOC TYP AGE CAN)
        Deviance Pearson X2 | k  AIC < min AIC + 2
Add:SYS 128.441    179.753  | 7  142.441
```

```
       ({F}LOC TYP AGE CAN SYS) could be a good model

Base terms: ({F}LOC TYP AGE CAN SYS)
        Deviance Pearson X2 | k  AIC < min AIC + 2
Add:PCO 126.572  186.71     | 8  142.572
           PCO not important since AIC < min AIC + 2

Base terms: ({F}LOC TYP AGE CAN SYS PCO)
        Deviance Pearson X2 | k    AIC
Add:PH  123.285  191.264    | 9  141.285 min AIC
            PH not important since AIC < min AIC + 2
```

	B1	B2	B3	B4
df	255	258	259	263
# of predictors	11	8	7	3
# with $0.01 \leq$ Wald p-value ≤ 0.05	2	1	0	0
# with Wald p-value > 0.05	4	0	0	0
G^2	233.765	237.212	243.482	278.787
AIC	257.765	255.212	259.482	286.787
corr(ESP,ESP(I))	1.0	0.99	0.97	0.80
p-value for change in deviance test	1.0	0.328	0.045	0.000

Example 4.12. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. One predictor was a factor, 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 267 cases: for the response, 113 were 0's and 154 were 1's.

Which two models are the best candidates for the final submodel? Explain briefly why each of the other 2 submodels should not be used.

Solution: B2 and B3 are best. B1 has too many predictors with rather large p-values. For B4, the AIC is too high and the corr and p-value are too low.

Example 4.13. 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, pp. 23-25). The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN=

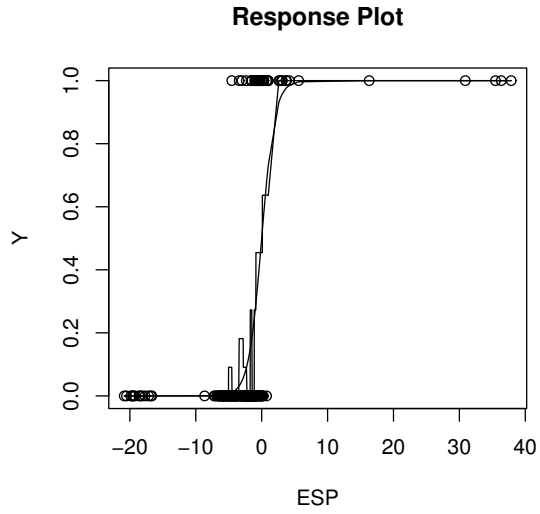

Fig. 4.7 Visualizing the ICU Data

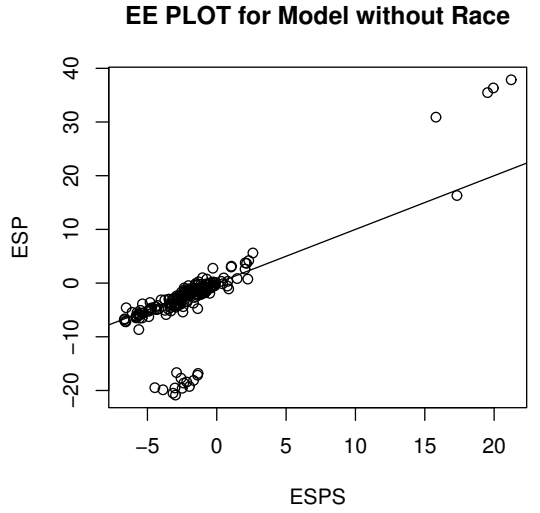

Fig. 4.8 EE Plot Suggests Race is an Important Predictor

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 if >60, 1 if $\leq$ 60), PH= PH from initial blood gases (0 if $\geq$ 7.25, 1 if <7.25), PCO= PCO2 from initial blood gases (0 if $\leq$ 45, 1 if >45), Bic= Bicarbonate from initial blood gases (0 if $\geq$ 18, 1 if <18), CRE= Creatinine from initial blood gases (0 if $\leq$ 2.0, 1 if >2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma).

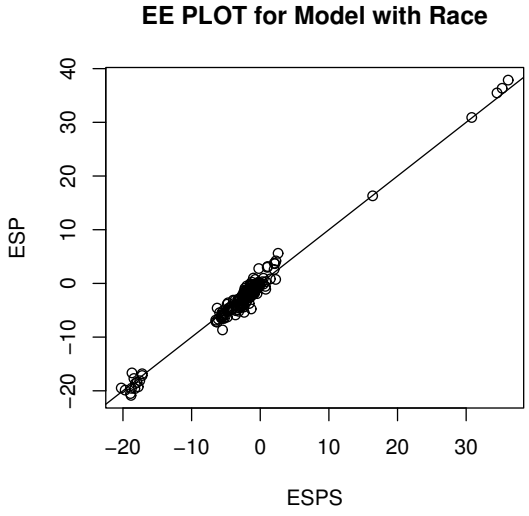

Fig. 4.9 EE Plot Suggests Race is an Important Predictor

Factors LOC and RACE had two indicator variables to model the three levels. The response plot in Figure 4.7 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 4.8. 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 4.9 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.

	P1	P2	P3	P4
df	144	147	148	149
# of predictors	6	3	2	1
# with $0.01 \leq$ Wald p-value ≤ 0.05	1	0	0	0
# with Wald p-value > 0.05	3	0	1	0
G^2	127.506	131.644	147.151	149.861
AIC	141.506	139.604	153.151	153.861
corr(ESP,ESP(I))	1.0	0.954	0.810	0.792
p-value for change in deviance test	1.0	0.247	0.0006	0.0

Example 4.14. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. Poisson regression was used. The response plot for the full model P1 was good. Model P2 was the minimum AIC model found.

Which model is the best candidate for the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Solution: P2 is best. P1 has too many predictors with large pvalues and more predictors than the minimum AIC model. P3 and P4 have corr and pvalue too low and AIC too high.

Warning. Variable selection for GLMs is very similar to that for multiple linear regression. Finding a model I_I from variable selection, and using GLM output for model I_I does not give valid tests and confidence intervals. If there is a good full model that was found before examining the response, and if I_I is the minimum AIC model, then Section 4.9 describes how to do inference after variable selection. If the model needs to be built using the response, use data splitting. A pilot study can also be useful.

4.6.2 When n/p is Not Necessarily Large

Forward selection with EBIC, lasso, and/or elastic net can be used for the Cox proportional hazards regression model and for some GLMs, including binomial and Poisson regression. The relaxed lasso = VS-lasso and relaxed elastic net = VS-elastic net estimators apply the GLM or Cox regression model to the predictors with nonzero lasso or elastic net coefficients. As with multiple linear regression, the population number of active nontrivial predictors = k_S, but for a GLM, model I with $SP = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$ has k active nontrivial predictors. See Section 2.1.

Remark 4.1. Most of the plots in this chapter that use $ESP = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$, and can also be made using $ESP(I) = \boldsymbol{x}_I^T \hat{\boldsymbol{\beta}}_I$. Obtaining a good ESP becomes more difficult as n/p becomes smaller.

Remark 4.2. Suppose the 1D regression model, such as a GLM, has $SP = \boldsymbol{x}^T \boldsymbol{\beta}$. If $n > 10p$, then fit the model using Chapter 3 MLR type methods, such as relaxed lasso and forward selection (using C_p), to find a subset of predictors I. If $n < 10p$, fit the model with MLR lasso. (Limited experience suggests that MLR with EBIC leads to severe underfitting if $n < 10p$ if the 1D regression model is not MLR.) Then fit the 1D regression with Y and $\boldsymbol{x}_I$. Check the model with the response plot and the EE plot of the MLR ESP versus the 1D regression ESP. High correlation in the EE plot suggests MLR model selection may be useful for the 1D regression model selection. For some GLMs, make the OD plot. If $\boldsymbol{x}_I$ is an $a \times 1$ vector, we want $n \geq Ja$ where $J \geq 5$ and preferably $J \geq 10$. For binary logistic regression, we want $a \geq J \min(N_0, N_1)$. Note that if $n < 5p$, the EE plot of the submodel ESP versus the full model ESP should not be used since the full model is overfitting. This method should be best when the predictors are linearly related: there should be no strong nonlinear relationships. See Olive and Hawkins (2005) for this method when $n > 10p$.

Some *R* commands for GLM lasso and Remark 4.2 are shown below. Note that the family command indicates whether a binomial regression (including binary regression) or a Poisson regression is being fit. The default for GLM lasso uses 10-fold CV with a deviance criterion.

```
set.seed(1976)  #Binary regression
library(glmnet)
n<-100
m<-1 #binary regression
q <- 100 #100 nontrivial predictors, 95 inactive
k <- 5 #k_S = 5 population active predictors
y <- 1:n
mv <- m + 0 * y
vars <- 1:q
```

```
beta <- 0 * 1:q
beta[1:k] <- beta[1:k] + 1
beta
alpha <- 0
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
SP <- alpha + x[,1:k] %*% beta[1:k]
pv <- exp(SP)/(1 + exp(SP))
y <- rbinom(n,size=m,prob=pv)
y
out<-cv.glmnet(x,y,family="binomial")
lam <- out$lambda.min
bhat <- as.vector(predict(out,type="coefficients",s=lam))
ahat <- bhat[1] #alphahat
bhat<-bhat[-1]
vin <- vars[bhat!=0] #want 1-5, overfit
 [1]  1  2  3  4  5  6 16 59 61 74 75 76 96
ind <- as.data.frame(cbind(y,x[,vin])) #relaxed lasso GLM
tem <- glm(y~.,family="binomial",data=ind)
tem$coef
(Inter) V2     V3      V4     V5     V6
0.2103  1.0037 1.4304  0.6208 1.8805 0.3831
V7      V8     V9      V10    V11     V12
0.8971  0.4716 0.5196  0.8900 0.6673 -0.7611
V13     V14
-0.5918 0.6926
lrplot3(tem=tem,x=x[,vin]) #binary response plot
#now use MLR lasso
outm<-cv.glmnet(x,y)
lamm <- outm$lambda.min
bm <- as.vector(predict(outm,type="coefficients",s=lamm))
am <- bm[1] #alphahat
bm<-bm[-1]
vm <- vars[bm!=0] #1 more variable than GLM lasso
vm
 [1]  1  2  3  4  5  6 16 35 59 61 74 75 76 96
vin
 [1]  1  2  3  4  5  6 16    59 61 74 75 76 96
inm <- as.data.frame(cbind(y,x[,vm])) #relaxed lasso GLM
tm <- glm(y~.,family="binomial",data=inm)
lrplot3(tem=tm,x=x[,vm]) #binary response plot
#Now use MLR forward selection with EBIC since n < 10p.
library(leaps)
out<-fsel(x,y)
vin<-out$vin
vin  #severe underfit
```

```
[1] 4
inm <- as.data.frame(cbind(y,x[,vin]))
tm <- glm(y~.,family="binomial",data=inm)
lrplot3(tem=tm,x=x[,vin]) #binary response plot

#Poisson regression, using same x and beta as above
y <- rpois(n,lambda=exp(SP))
out<-cv.glmnet(x,y,family="poisson")
lam <- out$lambda.min
bhat <- as.vector(predict(out,type="coefficients",s=lam))
ahat <- bhat[1] #alphahat
bhat<-bhat[-1]
vin <- vars[bhat!=0] #want 1-5, overfit
vin
 [1]  1  2  3  4  5  7  9 10 13 16 17 18 21 23 25
26 27 30 37 39 40 42 44 46 51 53 57 59 62 71 74 84 85 93 95 97 99
ind <- as.data.frame(cbind(y,x[,vin])) #relaxed lasso GLM
out <- glm(y~.,family="poisson",data=ind)
ESP <- predict(out)
prplot2(ESP,x=x[,vin],y) #response and OD plots
#now use MLR lasso
outm<-cv.glmnet(x,y)
lamm <- outm$lambda.min
bm <- as.vector(predict(outm,type="coefficients",s=lamm))
am <- bm[1] #alphahat
bm<-bm[-1]
vm <- vars[bm!=0]
vm #much less overfit than GLM lasso
 [1]  1  2  3  4  5  9 17 21 22 27 29 60 75 95
inm <- as.data.frame(cbind(y,x[,vm])) #relaxed lasso GLM
out <- glm(y~.,family="poisson",data=inm)
ESP <- predict(out)
prplot2(ESP,x=x[,vm],y) #response and OD plots
#Now use MLR forward selection with EBIC since n < 10p.
library(leaps)
out<-fsel(x,y)
vin<-out$vin
vin  #severe underfit causes poor fit and overdispersion
[1] 5
inm <- as.data.frame(cbind(y,x[,vin]))
out <- glm(y~.,family="poisson",data=inm)
ESP <- predict(out)
prplot2(ESP,x=x[,vin],y) #response and OD plots
```

4.7 Generalized Additive Models

There are many alternatives to the binomial and Poisson regression GLMs. Alternatives to the binomial GLM of Definition 4.7 include the discriminant function model of Definition 4.8, the quasi-binomial model, the binomial generalized additive model (GAM), and the beta-binomial model of Definition 4.2.

Alternatives to the Poisson GLM of Definition 4.12 include the quasi-Poisson model, the Poisson GAM, and the negative binomial regression model of Definition 4.3. 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 et al. (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. A GAM is a 1D regression model with SP=AP and ESP=EAP. We may use ESP for a GLM and EAP for a GAM.

Definition 4.18. In a *1D regression*, Y is independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$ where $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ for a GLM. In a *generalized additive model*, Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ 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* $\text{ESP} = \hat{h}(\boldsymbol{x})$ and $\text{ESP} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ for a GLM. 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 $S_j(x_j) = \beta_j x_j$ for $j = 2, ..., p$ with $\alpha = \beta_1$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$ is a special case of a GAM with $x_4 \equiv x_1 x_2$. 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 $x_1, ..., x_p$. The technique is more difficult, for example, if the GLM has terms x_1, x_2, x_2^2, and x_3 while the GAM has terms x_1, x_2 and x_3.)

Definition 4.19. An *EE plot* is a plot of EAP versus ESP.

Definition 4.20. 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 4.3.

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 4.21. Recall the Poisson GLM $Y|SP \sim Poisson(\exp(SP))$.

i) The *Poisson GAM* is $Y|AP \sim Poisson(\exp(AP))$. 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 4.4.

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 $f(y) = \frac{e^{-\mu}\ \mu^y}{(1-e^{\mu})\ y!}$ 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 4.22. 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. Definitions 4.1, 4.2, and 4.3 give important GAM models where SP = AP. Several of these models are GAM analogs of models discussed in Sections 4.2, 4.3, and 4.4.

4.7.1 Response Plots

For a 1D regression model, there are several useful plots using the ESP. A GAM is a 1D regression model with $ESP = EAP$. 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. Similarly, the response plot of ESP or EAP versus the response Y is useful. Assume that the ESP or EAP takes on many values. For a GAM, substitute EAP for ESP for the plots in Definitions 4.9, 4.10, 4.11, 4.13, 4.14, and 4.16.

The response plot for the beta-binomial GAM is similar to that for the binomial GAM. The plots for the negative binomial GAM are similar to those of the Poisson regression GAM, including the plots in Definition 4.16. See Examples 4.4, 4.5, and 4.6.

4.7.2 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

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) + \sum_{k \in E} S_k(x_k) = \alpha + \sum_{j \in S} S_j(x_j). \quad (4.10)$$

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 $\mathrm{corr}(\mathrm{AP}, \mathrm{AP(I)}) = 1$. 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.

4.7.3 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.

4.7.4 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.

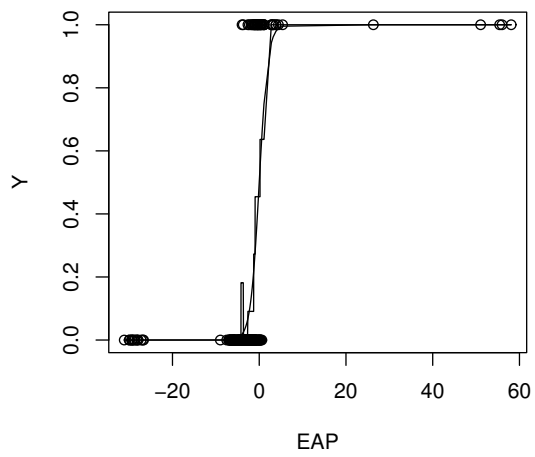

Fig. 4.10 Visualizing the ICU GAM

Example 4.15. For the ICU data of Example 4.13, 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 4.10 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, $\rho(EAP) = e^{EAP}/(1+e^{EAP})$]. 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

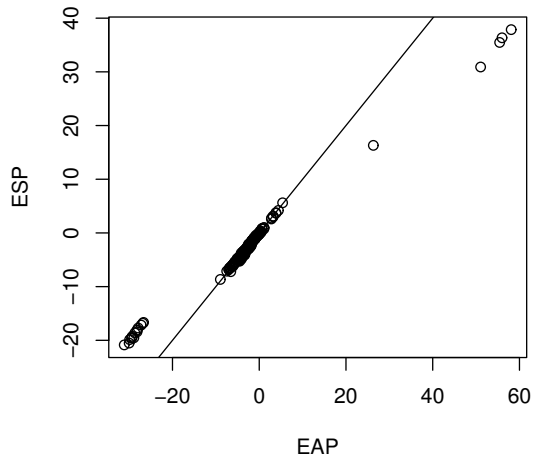

Fig. 4.11 GAM and GLM give Similar Success Probabilities

that the number of 0s and 1s 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 4.11 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 4.13. Some *R* code is below.

```
##ICU data from Statlib or URL
#http://parker.ad.siu.edu/Olive/ICU.lsp
#delete header of ICU.lsp and delete last parentheses
#at the end of the file. Save the file on F drive as
#icu.txt.

icu <- read.table("F:\\icu.txt")

names(icu) <- c("ID", "STA", "AGE", "SEX", "RACE",
"SER", "CAN", "CRN", "INF", "CPR", "SYS", "HRA",
"PRE", "TYP", "FRA", "PO2", "PH", "PCO", "Bic",
"CRE", "LOC")

icu[,5] <- as.factor(icu[,5])
```

```
icu[,21] <- as.factor(icu[,21])
icu2<-icu[,-1]
outf <- glm(formula=STA~.,family=binomial,data=icu2)
ESP <- predict(outf)

library(mgcv)
outgam <- gam(STA ~ s(AGE)+SEX+RACE+SER+CAN+CRN+INF+
CPR+s(SYS)+s(HRA)+PRE+TYP+FRA+PO2+PH+PCO+Bic+CRE+LOC,
family=binomial,data=icu2)
EAP <- predict.gam(outgam)
plot(EAP,ESP)
abline(0,1)
#Figure 4.11

Y <- icu2[,1]
lrplot3(ESP=EAP,Y,slices=18)
#Figure 4.10

lrplot3(ESP,Y,slices=18)
#Figure 4.7
```

Example 4.16. 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 4.10, 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.

Example 4.17. Wood (2017, pp. 125-130) 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 $x_1 = ck$, $x_2 = [ck]^2$, and $x_3 = [ck]^3$ was fit and had AIC = 33.66. The binomial GAM with predictor x_1 was fit in *R*, and Figure 4.12 shows that the EE plot for the GLM was not too good.

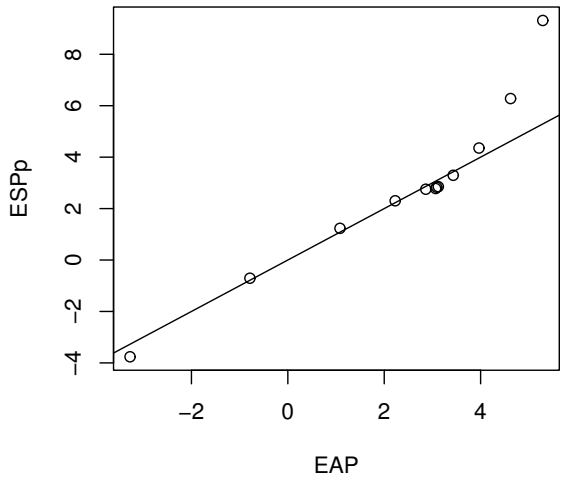

Fig. 4.12 EE plot for cubic GLM for Heart Attack Data

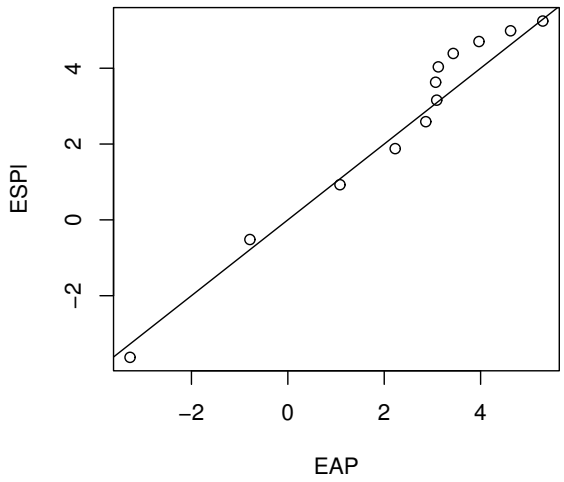

Fig. 4.13 EE plot with log(ck) in the GLM

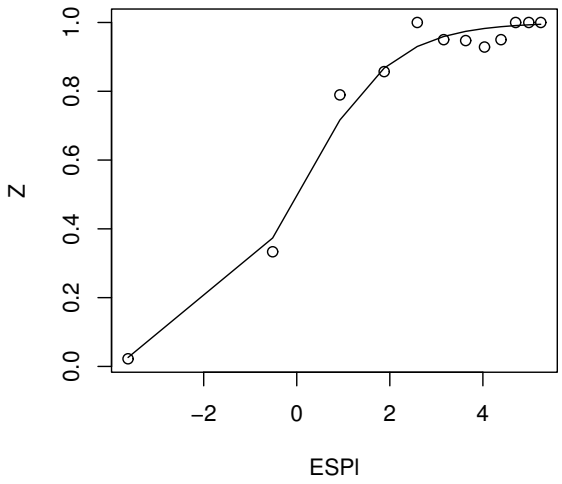

Fig. 4.14 Response Plot for Heart Attack Data

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 4.13 shows the EE plot, and Figure 4.14 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 the correlation of the plotted points in the EE plot, not shown, was one. See Problem 4.8.

4.8 Overdispersion

Definition 4.23. 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 underfits, 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 4.24. 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). See discussion below Definitions 4.11 and 4.14 for how to interpret the OD plot with the identity line, OLS line, and slope 4 line added as visual aids, and for discussion of the numerical summaries G^2 and X^2 for GLMs.

Definition 4.1, with SP = AP, 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.

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. See Examples 4.2–4.6 and the *R* code at the end of Section 4.6 (where $q = p - 1$).

Example 4.18. The species data is from Cook and Weisberg (1999, pp. 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.

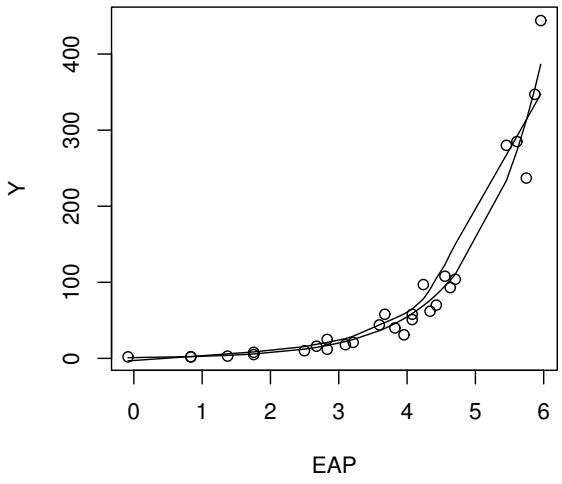

Fig. 4.15 Response Plot for Negative Binomial GAM

The response plot with the exponential and lowess curves added as visual aids is shown in Figure 4.15. 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).

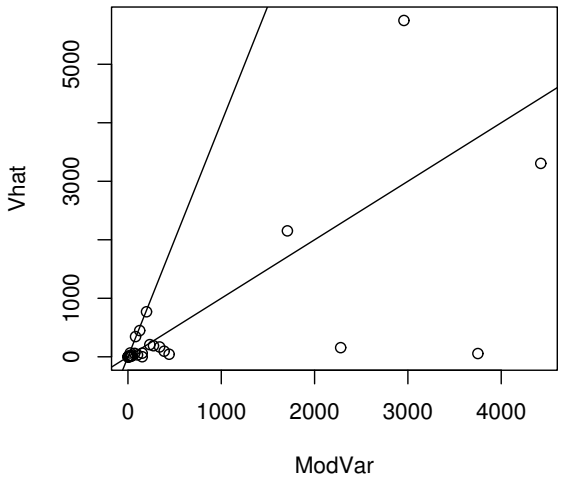

Fig. 4.16 OD Plot for Negative Binomial GAM

Figure 4.16 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 $\hat{\tau} = 1/37$.

4.9 Inference After Variable Selection for GLMs

Inference after variable selection for GLMs is very similar to inference after variable selection for multiple linear regression. AIC, BIC, EBIC, lasso, and elastic net can be used for variable selection. Read Section 4.2 for the large sample theory for $\hat{\boldsymbol{\beta}}_{I_{min},0}$. We assume that $n >> p$. Theorem 4.4, the Variable Selection CLT, still applies, as does Remark 4.4. Hence if lasso or elastic net is consistent, then relaxed lasso or relaxed elastic net is $\sqrt{n}$ consistent. The geometric argument of Theorem 4.5 also applies. We follow Rathnayake and Olive (2019) closely. Read Sections 4.2, 4.5, and 4.6 before reading this section. We will describe the parametric bootstrap, and then consider bootstrapping variable selection.

4.9.1 The Parametric and Nonparametric Bootstrap

Consider a parametric 1D regression model $Y|\boldsymbol{x} \sim D(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$ only through $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters.

Suppose $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and that $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. If $\boldsymbol{I}_n(\boldsymbol{\beta})$ is the Fisher information matrix based on a sample of size n, then $\boldsymbol{I}_n(\boldsymbol{\beta})/n \xrightarrow{P} \boldsymbol{I}(\boldsymbol{\beta})$. For GLMs, see, for example, Sen and Singer (1993, p. 309). For the parametric regression model, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ where the $n \times 1$ vector $\boldsymbol{Y} = (Y_i)$ and the ith row of the $n \times p$ design matrix $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$.

The parametric bootstrap uses $\boldsymbol{Y}_j^* = (Y_i^*)$ where $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for $i = 1, ..., n$. Regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The large sample theory for $\hat{\boldsymbol{\beta}}^*$ is simple. Note that if $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$ where $\boldsymbol{b}$ does not depend on n, then $(\boldsymbol{Y}^*, \boldsymbol{X})$ follows the parametric regression model with parameters $(\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{b}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{b}))$. Now fix large integer n_0, and let $\boldsymbol{b} = \hat{\boldsymbol{\beta}}_{n_o}$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{n_o}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_{n_o}))$. Since $N_p(\boldsymbol{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}})) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta})) \tag{4.11}$$

as $n \to \infty$.

Now suppose $S \subseteq I$. Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}(I)^T, \hat{\boldsymbol{\beta}}(O)^T)^T$. Then $(\boldsymbol{Y}, \boldsymbol{X}_I)$ follows the parametric regression model with parameters $(\boldsymbol{\beta}_I, \boldsymbol{\gamma})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}_I))$. Now $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$ only follows the parametric regression model asymptotically, since $\hat{\boldsymbol{\beta}}(O) \neq \boldsymbol{0}$. However, under regularity conditions, $E(\hat{\boldsymbol{\beta}}_I^*) \approx \hat{\boldsymbol{\beta}}_I$ and $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) - \text{Cov}(\hat{\boldsymbol{\beta}}_I) \to \boldsymbol{0}$ as $n, B \to \infty$.

To see the above claim for GLMs, consider a GLM with $\eta_i = SP_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = g(\mu_i)$ where $\mu_i = E(Y_i|\boldsymbol{x}_i) = g^{-1}(\eta_i)$. Let $V_i = V(Y_i|\boldsymbol{x}_i)$. Let

$$z_i = g(\mu_i) + g'(\mu_i)(Y_i - \mu_i) = \eta_i + \frac{\partial \eta_i}{\partial \mu_i}(Y_i - \mu_i), \quad \boldsymbol{Z} = (z_i),$$

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{V_i}, \quad \boldsymbol{W} = diag(w_i), \quad \hat{\boldsymbol{W}} = \boldsymbol{W}|_{\hat{\boldsymbol{\beta}}}, \quad \text{and} \quad \hat{\boldsymbol{Z}} = \boldsymbol{Z}|_{\hat{\boldsymbol{\beta}}}.$$

Then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\text{I}} = (\boldsymbol{X}_{\text{I}}^{\text{T}}\hat{\boldsymbol{W}}_{\text{I}}\boldsymbol{X}_{\text{I}})^{-1}\boldsymbol{X}_{\text{I}}^{\text{T}}\hat{\boldsymbol{W}}_{\text{I}}\hat{\boldsymbol{Z}}_{\text{I}}$$

while

$$\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T \hat{\boldsymbol{W}}_I^* \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \hat{\boldsymbol{W}}_I^* \hat{\boldsymbol{Z}}_I^* \tag{4.12}$$

where $\hat{\boldsymbol{\beta}}_I^*$ is fit as if $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$ follows the GLM with parameters $(\hat{\boldsymbol{\beta}}(I), \hat{\boldsymbol{\gamma}})$. If $S \subseteq I$, then this approximation is correct asymptotically since $\sqrt{n}\hat{\boldsymbol{\beta}}(O) = O_P(1)$. Hence $\eta_{iI}^* = \boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}(I) = g(\mu_{iI}^*)$, and $V_{iI}^* = V_M(Y_i^* | \boldsymbol{x}_{iI})$ where V_M is the model variance from the GLM with parameters $(\hat{\boldsymbol{\beta}}(I), \hat{\boldsymbol{\gamma}})$. Also, the estimated asymptotic covariance matrices are

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{-1} \text{ and } \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = (\boldsymbol{X}_{\mathrm{I}}^{\mathrm{T}} \hat{\boldsymbol{W}}_{\mathrm{I}} \boldsymbol{X}_{\mathrm{I}})^{-1}.$$

See, for example, Agresti (2002, pp. 138, 147), Hillis and Davis (1994), and McCullagh and Nelder (1989). From Sen and Singer (1994, p. 307), $n(\boldsymbol{X}_I^T \hat{\boldsymbol{W}}_I \boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{I}^{-1}(\boldsymbol{\beta}_I)$ as $n \to \infty$ if $S \subseteq I$.

Let $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Z}$. Then $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ since $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{\beta}$, and $\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{X}) = diag(V_i)$. Since

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)} \text{ and } \frac{\partial \eta_{\mathrm{i}}}{\partial \mu_{\mathrm{i}}} = \mathrm{g}'(\mu_{\mathrm{i}}),$$

$\mathrm{Cov}(\boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{Z}|\boldsymbol{X}) = \boldsymbol{W}^{-1}$. Thus $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1}$. Although $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_P(n^{-1/2})$, we have $n(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{-1} - n(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{I}^{-1}(\boldsymbol{\beta}) - \boldsymbol{I}^{-1}(\boldsymbol{\beta}) = \boldsymbol{0}$ as $n \to \infty$.

Let $\tilde{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T \boldsymbol{W}_I^* \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{W}_I^* \boldsymbol{Z}_I^*$ where $\boldsymbol{W}_i^*$ and $\boldsymbol{Z}_I^*$ are evaluated using $\hat{\boldsymbol{\beta}}(I)$. Then $\mathrm{Cov}(\boldsymbol{Y}^*) = diag(V_i^*) \to diag(V_{iI}^*)$. Hence $\mathrm{Cov}(\boldsymbol{Z}_I^*) \to \boldsymbol{W}_I^{*-1}$ and $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_I^*) \to (\boldsymbol{X}_I^T \boldsymbol{W}_I^* \boldsymbol{X}_I)^{-1}$ as $n, B \to \infty$. Hence $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_I) \to \boldsymbol{0}$ as $n, B \to \infty$ if $S \subseteq I$.

As an example, consider the Poisson regression model from Section 4.4. Then $\mu_{iI}^* = \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}(I)) = \exp(\eta_{iI}^*) = V_{iI}^*$. Hence

$$\frac{\partial \mu_{iI}^*}{\partial \eta_{iI}^*} = \exp(\eta_{iI}^*) = \mu_{iI}^* = V_{iI}^*,$$

$w_{iI}^* = \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}(I))$, and $\hat{w}_{iI}^* = \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^*)$. Similarly, $\eta_{iI}^* = \log(\mu_{iI}^*)$,

$$z_{iI}^* = \eta_{iI}^* + \frac{\partial \eta_{iI}^*}{\partial \mu_{iI}^*}(Y_i^* - \mu_{iI}^*) = \eta_{iI}^* + \frac{1}{\mu_{iI}^*}(Y_i^* - \mu_{iI}^*), \text{ and}$$

$$\hat{z}_{iI}^* = \boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^* + \frac{1}{\exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^*)}(Y_i^* - \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^*)).$$

Note that for $(\boldsymbol{Y}, \boldsymbol{X}_I)$, the formulas are the same with the asterisks removed and $\mu_{iI} = \exp(\boldsymbol{x}_{iI}^T \boldsymbol{\beta}_I)$.

The nonparametric bootstrap samples cases $(Y_i, \boldsymbol{x}_i)$ with replacement to form $(\boldsymbol{Y}_j^*, \boldsymbol{X}_j^*)$, and regresses $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}_j^*$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The

nonparametric bootstrap can be useful even if heteroscedasticity or overdispersion is present, if the cases are an iid sample from some population, a very strong assumption.

4.9.2 Bootstrapping Variable Selection

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. Let the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$. Recall T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. We have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (2.6) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. Hence geometric argument Theorem 2.5 holds: if we had iid data $T_1, ..., T_B$, then the prediction region applied to the iid data and centered at a randomly chosen T_n would be a large sample confidence region for $\boldsymbol{\theta}$.

Next use the argument for multiple linear regression in Section 2.6.4. For the bootstrap, suppose that T_i^* is equal to T_{ij}^* with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

$$T_{1,1}^*, ..., T_{B_{1n},1}^*, ..., T_{1,J}^*, ..., T_{B_{Jn},J}^*$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Denote $T_{1j}^*, ..., T_{B_{jn},j}^*$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $\boldsymbol{S}_{T,j}^*$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Suppose the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T_1^*, ..., T_B^*$ have the same variability asymptotically. Since $E(T_{jn}) \approx \boldsymbol{\theta}$, each component of the iid sample is approximately centered at $\boldsymbol{\theta}$. The bootstrap components are centered at $E(T_{jn}^*)$, and often $E(T_{jn}^*) = T_{jn}$. Geometrically, separating the component clouds so that they are no longer centered at one value makes the overall data cloud larger. Thus the variability of T_n^* is larger than that of T_n for a mixture distribution, asymptotically. Hence the prediction region applied to the bootstrap sample is slightly larger than the prediction region applied to the iid sample, asymptotically (we want $n \geq 20p$). Hence cutoff $\hat{D}_{1,1-\delta}^2 = D_{(U_B)}^2$ gives coverage close to or higher than the nominal coverage for confidence regions (2.30)

and (2.32), using the geometric argument. The deviation $T_i^* - T_n$ tends to be larger in magnitude than the deviation and $T_i^* - \overline{T}^*$. Hence the cutoff $\hat{D}^2_{2,1-\delta} = D^2_{(U_B,T)}$ tends to be larger than $D^2_{(U_B)}$, and region (2.31) tends to have higher coverage than region (2.32) for a mixture distribution.

The full model should be checked with the response plot before doing variable selection inference. Assume p is fixed and $n \geq 20p$. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and that $S \subseteq I_j$. For multiple linear regression with the residual bootstrap that uses residuals from the full OLS model, Chapter 2 showed that the components of the iid sample and bootstrap sample have the same variability asymptotically. The components of the iid sample are centered at $\boldsymbol{A\beta}$ while the components of the bootstrap sample are centered at $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$. Now consider regression models with $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. Assume $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$ where $\boldsymbol{\Sigma}_j = \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. For the nonparametric bootstrap, assume $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_j,0} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$. Then the components of the iid sample and bootstrap sample have the same variability asymptotically. The components of iid sample are centered at $\boldsymbol{A\beta}$ while the components of the bootstrap sample are centered at $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$. For the nonparametric bootstrap, the above results tend to hold if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ and if $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$. Assumptions for the nonparametric bootstrap tend to be rather strong: often one assumption is that the n cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some population. See Shao and Tu (1995, pp. 335-349) for the nonparametric bootstrap for GLMs, nonlinear regression, and Cox's proportional hazards regression. Also see Burr (1994), Efron and Tibshirani (1993), Freedman (1981), and Tibshirani (1997).

For the parametric bootstrap, Section 4.9.1 showed that under regularity conditions, $\text{Cov}(\hat{\boldsymbol{\beta}}^*_I) - \text{Cov}(\hat{\boldsymbol{\beta}}_I) \to \boldsymbol{0}$ as $n, B \to \infty$ if $S \subseteq I$. Hence $\text{Cov}(T_{jn}) - \text{Cov}(T^*_{jn}) \to \boldsymbol{0}$ as $n, B \to \infty$ if $S \subseteq I$. Here $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$, $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, $T^*_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0}$, and $T^*_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_j,0}$. Then $E(T_{jn}) \approx \boldsymbol{A\beta} = \boldsymbol{\theta}$ while the $E(T^*_{jn})$ are more variable than the $E(T_{jn})$ with $E(T^*_{jn}) \approx \boldsymbol{A}\hat{\boldsymbol{\beta}}(I_j, 0)$, roughly, where $\hat{\boldsymbol{\beta}}(I_j, 0)$ is formed from $\hat{\boldsymbol{\beta}}(I_j)$ by adding zeros corresponding to variables not in I_j. Hence the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T^*_1, ..., T^*_B$ have the same variability asymptotically.

In simulations for $n \geq 20p$ for $H_0 : \boldsymbol{A\beta}_S = \boldsymbol{\theta}_0$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}^*_T$ is a good estimator of $\text{Cov}(T^*)$. In the simulations where S is not the full model, inference with backward elimination with I_{min} using AIC was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$. It is possible that $\boldsymbol{S}^*_T$ is singular if a column of the bootstrap sample is equal to $\boldsymbol{0}$. If the regression model has a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$, we may need to replace p by $p + q$.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be

higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \mathbf{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in (2.1) so that H_0 is true. Suppose a nominal 95% confidence region is used and U_B is the 96th percentile. Hence the confidence region (2.30) or (2.31) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}^*_{O,1}, ..., \hat{\boldsymbol{\beta}}^*_{O,B}$, then $\mathbf{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}^*_T$ is singular, but the singleton set $\{\mathbf{0}\}$ is the large sample $100(1-\delta)\%$ confidence region (2.30), (2.31), or (2.32) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is one. (This result holds since $\{\mathbf{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}^*_{O,j}$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from variable selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\mathbf{0}\}$, $n \geq 20p$, and $B \geq 50p$. (Since the pvalue is one, this technique may be useful for data snooping: applying MLE theory to submodel I may have negligible selection bias.)

Remark 4.3. As in Chapter 2, another way to look at the bootstrap confidence region for variable selection estimators is to consider the estimator $T_{2,n}$ that chooses I_j with probability equal to the observed bootstrap proportion $\hat{\rho}_{jn}$. The bootstrap sample $T^*_1, ..., T^*_B$ tends to be slightly more variable than an iid sample $T_{2,1}, ..., T_{2,B}$, and the geometric argument suggests that the large sample coverage of the nominal $100(1-\delta)\%$ confidence region will be at least as large as the nominal coverage $100(1-\delta)\%$.

4.9.3 Examples and Simulations

Pelawa Watagoda and Olive (2019a) have an example and simulations for multiple linear regression using the residual bootstrap. See Chapter 2. We will use Poisson and binomial regression.

Example 4.19. Lindenmayer et al. (1991) and Cook and Weisberg (1999, p. 533) give a data set with 151 cases where Y is the number of possum species found in a tract of land in Australia. The predictors are *acacia*=basal area of acacia + 1, *bark*=bark index, *habitat*=habitat score, *shrubs*=number of shrubs + 1, *stags*= number of hollow trees + 1, *stumps*=indicator for presence of stumps, and a constant. Inference for the full Poisson regression model is shown along with the shorth(c) nominal 95% confidence intervals for β_i computed using the parametric bootstrap with $B = 1000$. As expected, the bootstrap intervals are close to the large sample GLM confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$.

The minimum AIC model from backward elimination used a constant, *bark*, *habitat*, and *stags*. The shorth(c) nominal 95% confidence intervals for β_i using the parametric bootstrap are shown. Note that most of the confidence intervals contain 0 when closed intervals are used instead of open intervals. The Poisson regression output is also shown, but should only be used for inference if the model was selected before looking at the data.

```
       large sample full model inference
           Est.    SE     z    Pr(>|z|)  95% shorth CI
int    -1.0428  0.2480 -4.205  0.0000  [-1.562,-0.538]
acacia  0.0166  0.0103  1.612  0.1070  [-0.004, 0.035]
bark    0.0361  0.0140  2.579  0.0099  [ 0.007, 0.065]
habitat 0.0762  0.0375  2.032  0.0422  [-0.003, 0.144]
shrubs  0.0145  0.0205  0.707  0.4798  [-0.028, 0.056]
stags   0.0325  0.0103  3.161  0.0016  [ 0.013, 0.054]
stumps -0.3907  0.2866 -1.364  0.1727  [-1.010, 0.171]
output and shorth intervals for the min AIC submodel
           Est.    SE     z    Pr(>|z|)  95% shorth CI
int    -0.8994  0.2135 -4.212  0.0000  [-1.438,-0.428]
acacia  0                              [ 0.000, 0.037]
bark    0.0336  0.0121  2.773  0.0056  [ 0.000, 0.060]
habitat 0.1069  0.0297  3.603  0.0003  [ 0.000, 0.156]
shrubs  0                              [ 0.000, 0.060]
stags   0.0302  0.0094  3.210  0.0013  [ 0.000, 0.054]
stumps  0                              [-0.970, 0.000]
```

We tested $H_0 : \beta_2 = \beta_5 = \beta_7 = 0$ with the I_{min} model selected by backward elimination. (Of course this test would be easy to do with the full model using GLM theory.) Then $H_0 : \boldsymbol{A}\boldsymbol{\beta} = (\beta_2, \beta_5, \beta_7)^T = \boldsymbol{0}$. Using the prediction region method with the full model had $[0, D_{(U_B)}] = [0, 2.836]$ with $D_{\boldsymbol{0}} = 2.135$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} backward elimination model had $[0, D_{(U_B)}] = [0, 2.804]$ while $D_{\boldsymbol{0}} = 1.269$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 0.322. (Use (3.35) with $\boldsymbol{S}^*_T$ and D from backward elimination for the numerator, and from the full model for the denominator.) Hence the backward elimination bootstrap test was more precise than the full model bootstrap test.

Example 4.20. For binary logistic regression, the MLE tends to converge if $\max(|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}|) \leq 7$ and if the Y values of 0 and 1 are not nearly perfectly classified by the rule $\hat{Y} = 1$ if $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} > 0.5$ and $\hat{Y} = 0$, otherwise. If there is perfect classification, the MLE does not exist. Let $\hat{\rho}(\boldsymbol{x}) = \hat{P}(Y = 1|\boldsymbol{x})$ under the binary logistic regression. If $|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}|) \geq 10$, some of the $\hat{\rho}(\boldsymbol{x}_i)$ tend to be estimated to be exactly equal to 0 or 1, which causes problems for the MLE. The Flury and Riedwyl (1988, pp. 5-6) banknote data consists of 100 counterfeit and 100 genuine Swiss banknote. The response variable is

an indicator for whether the banknote is counterfeit. The six predictors are measurements on the banknote: *bottom, diagonal, left, length, right,* and *top.* When the logistic regression model is fit with these predictors and a constant, there is almost perfect classification and backward elimination had problems. We deleted *diagonal*, which is likely an important predictor, so backward elimination would run. For this full model, classification is very good, but the $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ run from -20 to 20. In a plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y on the vertical axis (not shown), the logistic regression mean function is tracked closely by the lowess scatterplot smoother. The full model and backward elimination output is below. Inference using the logistic regression normal approximation appears to greatly underestimate the variability of $\hat{\boldsymbol{\beta}}$ compared to the parametric full model bootstrap variability. We tested $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by backward elimination. Using the prediction region method with the full model had $[0, D_{(U_B)}] = [0, 1.763]$ with $D_{\mathbf{0}} = 0.2046$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} backward elimination model had $[0, D_{(U_B)}] = [0, 1.511]$ while $D_{\mathbf{0}} = 0.2297$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 16.2747. Hence the full model bootstrap inference was much more precise. Backward elimination produced many zeros, but also produced many estimates that were very large in magnitude.

```
       large sample full model inference
            Est.    SE      z Pr(>|z|)  95% shorth CI
int  -475.581 404.913 -1.175 0.240 [-83274.99,1939.72]
length  0.375   1.418  0.265 0.791 [  -98.902,137.589]
left   -1.531   4.080 -0.375 0.708 [ -364.814,611.688]
right   3.628   3.285  1.104 0.270 [ -261.034,465.675]
bottom  5.239   1.872  2.798 0.005 [    3.159,567.427]
top     6.996   2.181  3.207 0.001 [    4.137,666.010]
output and shorth intervals for the min AIC submodel
          Est.     SE      z Pr(>|z|)  95% shorth CI
int  -472.999 269.271 -1.757 0.079 [-168131.6,35623.9]
length  0                          [ -110.850,286.265]
left    0                          [ -752.695,724.702]
right   2.725   2.050  1.329 0.184 [-656.1549,906.136]
bottom  5.005   1.657  3.020 0.003 [   2.985,1428.346]
top     6.821   2.071  3.294 0.001 [   4.333,1957.107]
```

Binary regression data sets like the one in Example 4.20 are common: the response plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y suggests that the logistic regression mean function is good, but the range of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is such that the GLM normal approximation to the MLE $\hat{\boldsymbol{\beta}}$ is likely invalid. Since the parametric bootstrap produces datasets very similar to the actual dataset, the bootstrap distribution of the logistic regression MLE may be superior to the GLM normal

approximation. For Example 4.20, the GLM and bootstrap inference for the full model both suggest that *bottom* and *top* are important predictors.

The results of the following simulation are similar to those of Chapter 2 for multiple linear regression using the residual bootstrap with residuals from the OLS full model. This simulation was for Poisson regression and binomial regression, using $B = \max(200, n/10, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$, $n = 50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where k and ψ are defined in the following paragraph. A larger simulation study is in Rathnayake (2019). In the simulations, we used $\theta = \boldsymbol{A}\boldsymbol{\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$ and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_E = \boldsymbol{0}$.

Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $q = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $q \times q$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\mathrm{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(q-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(q-2)\psi^2]$. Hence the correlations are $cor(z_i, z_j) = \rho = (2\psi + (q-2)\psi^2)/(1 + (q-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii}+k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, ..., 1, 0, ..., 0)^T$ with β_1, k ones, and $p - k - 1$ zeros. Binomial regression used $\beta_1 = 0, a = 5/3$, and $m_i = m$ with $m = 1$ or 20. Poisson regression used $\beta_1 = 1 = a$ and $\beta_1 = 5$ with $a = 2$.

The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$ where $\beta_2 = \cdots = \beta_{k+1} = 1$, and $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$ (whether the last $p - k - 1$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 would suggest coverage is close to the nominal value. The parametric bootstrap was used with AIC.

In the tables, there are two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for backward elimination. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (2.30), hybrid region (2.32), and Bickel and Ren region (2.31). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1..., 1)^T$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests for $\boldsymbol{\beta}_S$ if $k = 1$.

Volume ratios of the three confidence regions can be compared using (2.35), but there is not enough information in the tables to compare the volume of the confidence region for the full model regression versus that for the variable selection regression since the two methods have different determinants $|\boldsymbol{S}_T^*|$.

The inference for backward elimination was often as precise or more precise than the inference for the full model. The coverages tended to be near 0.95 for the parametric bootstrap on the full model. Variable selection coverage tended to be near 0.95 unless the $\hat{\beta}_i$ could equal 0. An exception was binary logistic regression with $m = 1$ where variable selection and the full model often had higher coverage than the nominal 0.95 for the hypothesis tests, especially for $n = 25p$. Compare Tables 4.2 and 4.3. For binary regression, the bootstrap confidence regions using smaller a and larger n resulted in coverages closer to 0.95 for the full model, and convergence problems caused the programs to fail for $a > 4$. The Bickel and Ren (2.31) average cutoffs were at least as high as those of the hybrid region (2.32).

If β_i was a component of β_E, then the backward elimination confidence intervals had higher coverage but were shorter than those of the full model due to zero padding. The zeros in $\hat{\boldsymbol{\beta}}_E$ tend to result in higher than nominal coverage for the variable selection estimator, but can greatly decrease the volume of the confidence region compared to that of the full model.

For the simulated data, when $\psi = 0$, the asymptotic covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ is diagonal. Hence $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for I_{min} and the full model by Remark 2.4. For Tables 4.2-4.5, $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$, and β_{p-1} and β_p are components of $\boldsymbol{\beta}_E$. For Table 4.6, $\boldsymbol{\beta}_S = (\beta_1, ..., \beta_9)^T$. Hence β_1, β_2, and β_{p-1} are components of $\boldsymbol{\beta}_S$, while $\boldsymbol{\beta}_E = \beta_{10}$. For the n in the tables and $\psi = 0$, the coverages and "lengths" did tend to be close for the β_i that are components of $\boldsymbol{\beta}_S$, and for pr1, hyb1, and br1.

Table 4.2 Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 200, n = 100, p = 4, k = 1$, and $m = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9516	0.9328	0.9524	0.9504	0.9724	0.9872	0.9920	0.9802	0.9838	0.9888
len	1.1605	1.0953	0.7171	0.7151	2.5225	2.5225	2.5476	2.5173	2.5173	2.6893
vs,0	0.9564	0.9322	0.9976	0.9976	0.9960	0.9964	0.9988	0.9774	0.9794	0.9948
len	1.1483	1.0798	0.6143	0.6204	2.7329	2.7329	3.0386	2.5160	2.5160	2.6899
reg,0.5	0.9538	0.9428	0.9440	0.9544	0.9680	0.9854	0.9896	0.9724	0.9828	0.9858
len	1.1622	1.6737	1.4547	1.4588	2.5221	2.5221	2.5475	2.5165	2.5165	2.6037
vs,0.5	0.9528	0.9662	0.9978	0.9982	0.9948	0.9918	0.9978	0.9760	0.9756	0.9872
len	1.1462	1.6714	1.2879	1.2883	2.7230	2.7230	3.0170	2.5379	2.5379	2.6860
reg,0.9	0.9662	0.9578	0.9520	0.9500	0.9690	0.9846	0.9884	0.9724	0.9848	0.9876
len	1.1606	9.4523	9.4241	9.4379	2.5220	2.5220	2.5454	2.5142	2.5142	2.5389
vs,0.9	0.9566	0.9422	0.9960	0.9974	0.9958	0.9972	0.9982	0.9866	0.9932	0.9956
len	1.1502	8.4654	8.4806	8.4951	2.7700	2.7700	3.0182	2.6176	2.6176	2.7644

Table 4.3 Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 200$, $n = 200$, $p = 4$, $k = 1$, and $m = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9504	0.9440	0.9552	0.9544	0.9584	0.9662	0.9674	0.9580	0.9662	0.9728
len	0.7539	0.6771	0.4583	0.4587	2.4884	2.4884	2.4992	2.4846	2.4846	2.5745
vs,0	0.9552	0.9490	0.9986	0.9978	0.9954	0.9908	0.9968	0.9600	0.9698	0.9762
len	0.7510	0.6736	0.3909	0.3926	2.7226	2.7226	3.0310	2.4814	2.4814	2.5740
reg,0.5	0.9538	0.9508	0.9550	0.9578	0.9590	0.9686	0.9690	0.9578	0.9658	0.9714
len	0.7548	1.0543	0.9337	0.9309	2.4858	2.4858	2.4958	2.4828	2.4828	2.5266
vs,0.5	0.9538	0.9602	0.9984	0.9974	0.9930	0.9922	0.9958	0.9708	0.9786	0.9828
len	0.7501	1.0607	0.8064	0.8047	2.7022	2.7023	2.9948	2.5004	2.5004	2.6164
reg,0.9	0.9462	0.9536	0.9522	0.9496	0.9548	0.9642	0.9658	0.9496	0.9610	0.9626
len	0.7546	6.0844	6.0691	6.0800	2.4888	2.4888	2.4990	2.4860	2.4860	2.4967
vs,0.9	0.9562	0.9520	0.9958	0.9954	0.9936	0.9922	0.9968	0.9822	0.9870	0.9896
len	0.7502	5.3338	5.3737	5.3847	2.7934	2.7934	3.0392	2.5873	2.5873	2.7225

Table 4.4 Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 1$, and $m = 20$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9576	0.9502	0.9520	0.9548	0.9500	0.9528	0.9530	0.9480	0.9496	0.9502
len	0.1428	0.1232	0.0860	0.0860	3.9837	3.9837	3.9876	2.4538	2.4538	2.4653
vs,0	0.9510	0.9510	0.9992	0.9978	0.9980	0.9982	0.9998	0.9412	0.9458	0.9478
len	0.1424	0.1229	0.0706	0.0707	4.3081	4.3081	4.7454	2.4531	2.4531	2.4747
reg,0.32	0.9536	0.9534	0.9514	0.9548	0.9496	0.9524	0.9530	0.9474	0.9490	0.9506
len	0.1426	0.1833	0.1609	0.1610	3.9840	3.9840	3.9884	2.4528	2.4528	2.4589
vs,0.32	0.9534	0.9620	0.9966	0.9976	0.9968	0.9976	0.9988	0.9534	0.9544	0.9582
len	0.1424	0.1837	0.1347	0.1352	4.2607	4.2607	4.6891	2.4527	2.4527	2.5042
reg,0.9	0.9514	0.9432	0.9552	0.9498	0.9434	0.9448	0.9446	0.9430	0.9440	0.9450
len	0.1427	2.2178	2.2170	2.2175	3.9846	3.9846	3.9887	2.4530	2.4530	2.4553
vs,0.9	0.9590	0.9656	0.9982	0.9986	0.9982	0.9978	0.9996	0.9532	0.9478	0.9654
len	0.1425	2.0342	1.8778	1.8862	4.2368	4.2368	4.6742	2.4449	2.4449	2.5661

Table 4.5 Bootstrapping Poisson Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 1$, $a = 1$, $\beta_1 = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9480	0.9526	0.9526	0.9520	0.9502	0.9512	0.9524	0.9432	0.9454	0.9472
len	0.1752	0.1325	0.1275	0.1276	3.9859	3.9859	3.9901	2.4528	2.4528	2.4740
vs,0	0.9552	0.9574	0.9982	0.9982	0.9984	0.9982	0.9998	0.9524	0.9574	0.9628
len	0.1752	0.1323	0.1051	0.1047	4.3004	4.3004	4.7408	2.4543	2.4543	2.5009
reg,0.32	0.9552	0.9518	0.9520	0.9536	0.9538	0.9536	0.9538	0.9510	0.9532	0.9552
len	0.1752	0.2419	0.2390	0.2386	3.9852	3.9852	3.9894	2.4518	2.4518	2.4689
vs,0.32	0.9562	0.9632	0.9986	0.9992	0.9980	0.9982	0.9992	0.9630	0.9644	0.9712
len	0.1750	0.2419	0.2005	0.2004	4.2618	4.2618	4.6811	2.4520	2.4520	2.5384
reg,0.9	0.9478	0.9530	0.9570	0.9554	0.9458	0.9478	0.9484	0.9448	0.9448	0.9476
len	0.1754	3.2873	3.2859	3.2912	3.9831	3.9831	3.9872	2.4536	2.4536	2.4691
vs,0.9	0.9500	0.9574	0.9984	0.9994	0.9970	0.9966	0.9984	0.9638	0.9626	0.9742
len	0.1752	2.8710	2.7922	2.7879	4.2597	4.2597	4.6886	2.4809	2.4809	2.6402

Table 4.6 Bootstrapping Poisson Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 8$, $a = 2$, $\beta_1 = 5$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9522	0.9468	0.9540	0.9518	0.9496	0.9492	0.9488	0.9474	0.9464	0.9478
len	0.0210	0.0146	0.0146	0.0142	1.9593	1.9593	1.9609	4.1633	4.1633	4.1675
vs,0	0.9544	0.9546	0.9518	0.9980	0.9966	0.9374	0.9966	0.9534	0.9524	0.9552
len	0.0210	0.0146	0.0146	0.0117	2.1470	2.1470	2.3955	4.1655	4.1655	4.1880
reg,0.32	0.9522	0.9510	0.9486	0.9540	0.9494	0.9504	0.9516	0.9460	0.9468	0.9472
len	0.0210	0.0664	0.0664	0.0663	1.9595	1.9595	1.9614	4.1636	4.1636	4.1684
vs,0.32	0.9508	0.9596	0.9496	0.9992	0.9986	0.9434	0.9986	0.9634	0.9646	0.9696
len	0.0210	0.0663	0.0662	0.0541	2.1434	2.1434	2.3960	4.1970	4.1970	4.2703
reg,0.9	0.9536	0.9580	0.9550	0.9584	0.9538	0.9538	0.9548	0.9496	0.9512	0.9524
len	0.0210	1.0357	1.0361	1.0336	1.9585	1.9585	1.9605	4.1603	4.1603	4.1643
vs,0.9	0.9486	0.9484	0.9492	0.9988	0.9982	0.9492	0.9982	0.9688	0.9546	0.9676
len	0.0212	1.0742	1.0745	0.8793	2.1387	2.1387	2.3860	4.2883	4.2883	4.3818

4.10 Prediction Intervals

We use two prediction intervals from Olive et al. (2019). The first prediction interval for Y_f applies the shorth prediction interval of Section 2.3 to the parametric bootstrap sample $Y_1^*, ..., Y_B^*$ where the Y_i^* are iid from the distribution $D(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$. If the regression method produces a consistent estimator $(\hat{h}(\boldsymbol{x}), \hat{\boldsymbol{\gamma}})$ of $(h(\boldsymbol{x}), \boldsymbol{\gamma})$, then this new prediction interval is a large sample $100(1-\delta)\%$ PI that is a consistent estimator of the shortest population interval $[L, U]$ that contains at least $1-\delta$ of the mass as $B, n \to \infty$. The new large sample $100(1-\delta)\%$ PI using $Y_1^*, ..., Y_B^*$ uses the shorth(c) PI with

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.13}$$

For models with a linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}$, we will want prediction intervals after variable selection or model selection. Refer to Equation (2.1) and Section 4.6.1. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for GLM variable selection. The Chen and Chen (2008) EBIC criterion can be useful, especially if n/p is not large. GLM model selection with lasso and the elastic net is also common. See Hastie et al. (2015, ch. 3), Tibshirani (1996), Friedman et al. (2007), and Friedman et al. (2010). Relaxed lasso applies the regression method, such as a GLM, to the active predictors with nonzero coefficients selected by lasso. For $n \geq 10p$, Olive and Hawkins (2005) suggested using multiple linear regression variable selection software with the Mallows (1973) C_p criterion to get a subset I, then fit the GLM using Y and $\boldsymbol{x}_I$. If the regression model contains a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$, then we may need $n \geq 10(p+q)$.

The prediction interval (4.13) can have undercoverage if n is small compared to the number of estimated parameters. The modified shorth PI (4.14) inflates PI (4.13) to compensate for parameter estimation and model selection. Let d be the number of variables $x_1^*, ..., x_d^*$ used by the full model, forward selection, lasso, or relaxed lasso. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used. For a GAM full model, suppose the "degrees of freedom" d_i for $S(x_i)$ is bounded by k. We could let $d = 1 + \sum_{i=2}^{p} d_i$ with $p \leq d \leq pk$.) We want $n \geq 10d$, and the prediction interval length will be increased (penalized) if n/d is not large. Let $q_n = \min(1-\delta+0.05, 1-\delta+d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Then compute the shorth PI with

$$c_{mod} = \min(B, \lceil B[q_n + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.14}$$

Olive (2007, 2018) and Pelawa Watagoda and Olive (2019b) used similar correction factors since the maximum simulated undercoverage was about 0.05 when $n = 20d$. If a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$ is also estimated, we may need to replace d by $d_q = d + q$.

If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$ is the estimator that minimized the variable selection criterion, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$.

Hong et al. (2018) explain why classical PIs after AIC variable selection may not work. Fix p and let I_{min} correspond to the predictors used after variable selection, including AIC, BIC, and relaxed lasso. Suppose $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. See Charkhi and Claeskens (2018), Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232), Hastie et al. (2015, pp. 295-302) and Haughton (1988, 1989) for more information and references about this assumption. For relaxed lasso, the assumption holds if lasso is a consistent estimator. Suppose model (2.1) holds, and that if $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{4.15}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j. Then $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (2.1) if the variable selection criterion is used with forward selection, backward elimination, or all subsets. Hence (4.13) and (4.14) are large sample PIs.

Rathnayake and Olive (2019) gave the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta})$, generalizing the Pelawa Watagoda and Olive (2019a) result for multiple linear regression. See Theorem 2.4. Regularity conditions for (4.13) and (4.14) to be large sample PIs when $p > n$ are much stronger.

Prediction intervals (4.13) and (4.14) often have higher than the nominal coverage if n is large and Y_f can only take on a few values. Consider binary regression where $Y_f \in \{0, 1\}$ and the PIs (4.13) and (4.14) are [0,1] with 100% coverage, [0,0], or [1,1]. If [0,0] or [1,1] is the PI, coverage tends to be higher than nominal coverage unless $P(Y_f = 1|\boldsymbol{x}_f)$ is near δ or $1 - \delta$, e.g., if $P(Y_f = 1|\boldsymbol{x}_f) = 0.01$, then [0,0] has coverage near 99% even if $1 - \delta < 0.99$.

Example 4.21. For the Ceriodaphnia data of Example 4.4, Figure 4.17 shows the response plot of ESP versus Y for this data. In this plot, the lowess curve is represented as a jagged curve to distinguish it from the estimated Poisson regression mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The circles correspond to the Y_i and the $\times$'s to the PIs (4.13) with $d = p = 3$. The n large sample 95% PIs contained 97% of the Y_i. There was no evidence of overdispersion: see Example 4.4. There were 5 replications for each of the 14 strain–species combinations, which helps show the bootstrap PI variability when $B = 1000$. This example illustrates a useful goodness of fit diagnostic: if the model D is a useful approximation for the data and n is large enough, we expect the coverage on the training data to be close to or higher than the nominal coverage $1 - \delta$. For example, there may be undercoverage if a Poisson regression model is used when a negative binomial regression model is needed.

Example 4.22. For the banknote data of Example 4.20, after variable selection, we decided to use a constant, right, and bottom as predictors. The response plot for this submodel is shown in the left plot of Figure 4.18 with $Z = Z_i = Y_i/m_i = Y_i$ and the large sample 95% PIs for $Z_i = Y_i$. The circles correspond to the Y_i and the $\times$'s to the PIs (4.13) with $d = 3$, and 199 of the 200 PIs contain Y_i. The PI [0,0] that did not contain Y_i corresponds to the circle in the upper left corner. The PIs were [0,0], [0,1], or [1,1] since the data is binary. The mean function is the smooth curve and the step function gives the sample proportion of ones in the interval. The step function approximates the smooth curve closely, hence the binary logistic regression model seems reasonable. The right plot of Figure 4.18 shows the GAM using right and bottom with $d = 3$. The coverage was 100% and the GAM had many [1,1] intervals.

Example 4.23. For the species data of Examples 4.18, we used a constant and log(endem), log(area), log(distance), and log(areanear). The response plot looks good, but the OD plot (not shown) suggests overdispersion. When the response plot for the Poisson regression model was made, the n large sample 95% PIs (4.13) contained 89.7% of the Y_i.

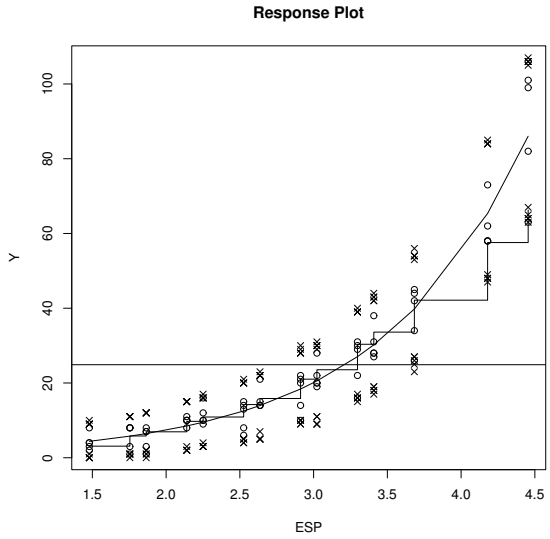

Fig. 4.17 Ceriodaphnia Data Response Plot.

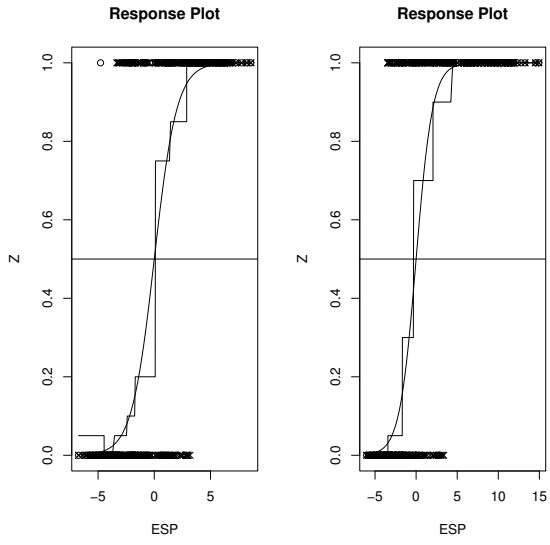

Fig. 4.18 Banknote Data GLM and GAM Response Plots.

For the simulations, generating $\boldsymbol{x}^T\boldsymbol{\beta}$ is important. For example, for binomial logistic regression, typically $-5 \leq \boldsymbol{x}^T\boldsymbol{\beta} \leq 5$ or there can be problems with the MLE. We used the same simulated data as that used for variable selection in Section 4.9.3. Thus $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, .., 1, 0, ..., 0)^T$ with β_1, k ones and $p - k - 1$ zeros. The default settings for Poisson regression use $\beta_1 = 1 = a$. The default settings for binomial regression use $\beta_1 = 0$ and $a = 5/3$.

Table 4.7 Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 4$, $\beta_1 = 1 = a$

n	ψ	k		GLM	GAM	lasso	RL	OHFS	BE
100	0	1	cov	0.9712	0.9714	0.9810	0.9800	0.9792	0.9734
			len	6.6448	6.6118	7.2770	7.2004	7.0680	6.6632
400	0	1	cov	0.9692	0.9694	0.9728	0.9714	0.9722	0.9665
			len	6.6392	6.6474	6.7996	6.7722	6.7588	6.6778
100	0.5	1	cov	0.9642	0.9644	0.9796	0.9786	0.9760	0.9689
			len	6.6922	6.6806	7.3136	7.2824	7.1160	6.7767
400	0.5	1	cov	0.9668	0.9670	0.9722	0.9716	0.9702	0.9754
			len	6.6720	6.6896	6.8342	6.8140	6.7992	6.7802
100	0.9	1	cov	0.9672	0.9674	0.9766	0.9768	0.9738	0.9665
			len	6.6038	6.6186	7.1480	7.1214	7.0002	6.5789
400	0.9	1	cov	0.9660	0.9662	0.9734	0.9700	0.9692	0.9798
			len	6.5838	6.5746	6.7526	6.7196	6.7004	6.7443
100	0	3	cov	0.9696	0.9698	0.9848	0.9834	0.9818	0.9654
			len	6.7080	6.7084	7.5632	7.5442	7.5348	6.7408
400	0	3	cov	0.9728	0.9730	0.9750	0.9746	0.9748	0.9657
			len	6.5718	6.5684	6.7690	6.7356	6.7406	6.7063
100	0.5	3	cov	0.9672	0.9674	0.9842	0.9838	0.9736	0.9592
			len	6.6992	6.7044	7.5804	7.5494	7.3810	6.7128
400	0.5	3	cov	0.9682	0.9684	0.9730	0.9722	0.9702	0.9772
			len	6.6794	6.6890	6.8726	6.8520	6.8466	6.7504
100	0.9	3	cov	0.9664	0.9666	0.9804	0.9810	0.9750	0.9678
			len	6.6704	6.6646	7.2880	7.2672	7.0722	6.7635
400	0.9	3	cov	0.9690	0.9692	0.9744	0.9742	0.9736	0.9667
			len	6.7960	6.8092	6.9696	6.9682	6.9120	6.6987

The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The simulation used $B = 1000$; $p = 4, 50, n$, or $2n$; $\psi = 0, 1/\sqrt{p}$, or 0.9; and $k = 1, 19$, or $p - 1$. The simulated data sets are rather small since the R estimators are rather slow. For binomial and Poisson regression, we only computed the GAM for $p = 4$ with $SP = AP = \alpha + S_2(x_2) + S_2(x_3) + S_4(x_4)$ and $d = p = 4$. We only computed the full model GLM if $n \geq 5p$. Lasso and relaxed lasso were computed for all cases. The regression model was computed from the training data, and a prediction interval was made for the test case Y_f given $\boldsymbol{x}_f$. The "length" and "coverage" were the average length and the

Table 4.8 Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 4$, $\beta_1 = 5$, $a = 2$

n	ψ	k		GLM	GAM	lasso	RL	OHFS	BE
100	0	1	cov	0.9500	0.9440	0.7730	0.9664	0.9654	0.9520
			len	77.6072	77.6306	84.1066	81.8374	82.4752	84.1432
400	0	1	cov	0.9580	0.9564	0.7566	0.9622	0.9628	0.9534
			len	82.0126	82.0212	85.5704	83.2692	83.4374	80.9897
100	0.5	1	cov	0.9456	0.9424	0.7646	0.9634	0.9408	0.9512
			len	83.0236	82.9034	90.5822	88.3060	88.6700	79.6887
400	0.5	1	cov	0.9530	0.9500	0.7584	0.9604	0.9566	0.9678
			len	83.8588	83.8292	87.4336	85.1042	85.1434	79.9855
100	0.9	1	cov	0.9492	0.9452	0.7688	0.9646	0.7712	0.9654
			len	78.3554	78.3798	87.0086	84.6072	83.4980	81.5432
400	0.9	1	cov	0.9550	0.9574	0.7606	0.9606	0.7928	0.9513
			len	76.7028	76.7594	80.5070	78.2308	78.2538	80.1298
100	0	3	cov	0.9544	0.9466	0.7798	0.9708	0.9404	0.9487
			len	80.1476	80.1362	92.1372	89.8532	90.3456	79.4565
400	0	3	cov	0.9560	0.9548	0.7514	0.9582	0.9566	0.9567
			len	80.7868	80.8976	85.0642	82.7982	82.7912	79.4522
100	0.5	3	cov	0.9516	0.9478	0.7848	0.9694	0.3324	0.9515
			len	77.1120	77.1130	88.9346	86.4680	85.8634	81.5643
400	0.5	3	cov	0.9568	0.9558	0.7534	0.9636	0.5214	0.9528
			len	80.4226	80.4932	84.7646	82.5590	83.7526	79.9786
100	0.9	3	cov	0.9492	0.9456	0.7882	0.9620	0.7510	0.9554
			len	79.5374	79.6172	91.2052	89.0692	84.5648	81.8544
400	0.9	3	cov	0.9544	0.9546	0.7638	0.9554	0.7384	0.9586
			len	79.7384	79.6906	83.8318	81.6862	81.0882	80.7521

Table 4.9 Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 50$, $\beta_1 = 5$, $a = 2$

n	ψ	k		GLM	lasso	RL	OHFS	BE
500	0	1	cov	0.9352	0.7564	0.9598	0.9640	0.9476
			len	81.2668	84.3188	81.8934	85.2922	81.1010
500	0.14	1	cov	0.9370	0.7508	0.9580	0.9628	0.9458
			len	81.1820	84.4530	82.1894	85.2304	81.1146
500	0.9	1	cov	0.9368	0.7630	0.9620	0.8994	0.9456
			len	80.4568	86.3506	84.4942	84.1448	80.4202
500	0	19	cov	0.9388	0.7592	0.9756	0.3778	0.9472
			len	81.6922	96.8546	94.6350	99.7436	81.7218
500	0.14	19	cov	0.9368	0.7556	0.9730	0.2770	0.9438
			len	80.0654	95.2964	93.2748	87.3814	80.1276
500	0.9	19	cov	0.9350	0.7544	0.9536	0.9480	0.9352
			len	79.7324	86.3448	84.0674	83.2958	79.6172
500	0	49	cov	0.9386	0.7104	0.9666	0.1004	0.9364
			len	81.1422	96.4304	94.8818	108.0518	81.2516
500	0.14	49	cov	0.9396	0.7194	0.9558	0.2858	0.9402
			len	79.7874	94.8908	93.2538	86.4234	79.8692
500	0.9	49	cov	0.9380	0.7640	0.9480	0.9512	0.9430
			len	78.8146	85.5786	83.2812	82.4104	78.8316

proportion of the 5000 prediction intervals that contained Y_f. Two rows per table were used to display these quantities.

Tables 4.7 to 4.9 show some simulation results for Poisson regression. Lasso minimized 10-fold cross validation and relaxed lasso was applied to the selected lasso model. The full GLM, full GAM and backward elimination (BE in the tables) used PI (4.13) while lasso, relaxed lasso (RL in the tables), and forward selection using the Olive and Hawkins (2005) method (OHFS in the tables) used PI (4.14). For $n \geq 10p$, coverages tended to be near or higher than the nominal value of 0.95, except for lasso and the Olive and Hawkins (2005) method in Tables 4.8 and 4.9. In Table 4.7, coverages were high because the Poisson counts were small and the Poisson distribution is discrete. In Table 4.8, the Poisson counts were not small, so the discreteness of the distribution did not affect the coverage much. For Table 4.9, $p = 50$, and PI (4.13) has slight undercoverage for the full GLM since $n = 10p$. Table 4.9 helps illustrate the importance of the correction factor: PI (4.14) would have higher coverage and longer average length. Lasso was good at choosing subsets that contain S since relaxed lasso had good coverage. The Olive and Hawkins (2005) method is partly graphical, and graphs were not used in the simulation.

Tables 4.10 and 4.11 are for binomial regression where only PI (4.13) was used. For large n, coverage is likely to be higher than the nominal if the binomial probability of success can get close to 0 or 1. For binomial regression, neither lasso nor the Olive and Hawkins (2005) method had undercoverage in any of the simulations with $n \geq 10p$.

For $n \leq p$, good performance needed stronger regularity conditions, and Table 4.12 shows some results with $n = 100$ and $p = 200$. For $k = 1$, relaxed lasso performed well as did lasso except in the second to last column of Table 4.12. With $k = 19$ and $\psi = 0$, there was undercoverage since $n < 10(k + 1)$. For the dense models with $k = 199$ and $\psi = 0$, there was often severe undercoverage, lasso sometimes picked 100 predictors including the constant, and then relaxed lasso caused the program to fail with 5000 runs. Coverage was usually good for $\psi > 0$ except for the second to last column and sometimes the last column of Table 4.12. With $\psi = 0.9$, each predictor was highly correlated with the one dominant principal component.

4.11 Survival Analysis

Regression methods for survival analysis focus on the survival function rather than the mean function, and the data can be right censored.

Definition 10.25. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the **survival time** or time until event. The survival time is **censored** if the event of interest has not been observed. Let Y_i be the ith survival time. Let Z_i be the time the ith observation (possibly an individual or machine)

Table 4.10 Simulated Large Sample 95% PI Coverages and Lengths for Binomial Regression, $p = 4$, $m = 40$

n	ψ	k		GLM	GAM	lasso	RL	OHFS	BE
100	0	1	cov	0.9786	0.9788	0.9774	0.9744	0.9720	0.9726
			len	10.7696	10.7656	10.5332	10.4430	10.1990	10.2016
400	0	1	cov	0.9708	0.9700	0.9696	0.9708	0.9702	0.9688
			len	9.8374	9.8426	9.8292	9.7866	9.7518	9.7548
100	0.5	1	cov	0.9792	0.9720	0.9742	0.9750	0.9724	0.9708
			len	10.6668	10.6426	10.3790	10.3282	10.1060	10.1012
400	0.5	1	cov	0.9678	0.9676	0.9692	0.9670	0.9668	0.9656
			len	9.8352	9.8452	9.8196	9.7890	9.7612	9.7590
100	0.9	1	cov	0.9780	0.9766	0.9762	0.9742	0.9704	0.9714
			len	10.7324	10.7222	10.3774	10.3186	10.1438	10.1602
400	0.9	1	cov	0.9688	0.9672	0.9680	0.9674	0.9684	0.9672
			len	9.7554	9.7646	9.7392	9.7012	9.6778	9.6790
100	0	3	cov	0.9790	0.9750	0.9782	0.9772	0.9780	0.9776
			len	10.6974	10.6960	10.7388	10.7030	10.6956	10.7020
400	0	3	cov	0.9652	0.9652	0.9654	0.9656	0.9650	0.9626
			len	9.7838	9.7878	9.8244	9.7864	9.7800	9.7722
100	0.5	3	cov	0.9780	0.9734	0.9776	0.9766	0.9770	0.9784
			len	10.7224	10.7034	10.7482	10.7042	10.7162	10.7134
400	0.5	3	cov	0.9686	0.9688	0.9726	0.9702	0.9704	0.9706
			len	9.7250	9.7170	9.7460	9.7172	9.7152	9.7290
100	0.9	3	cov	0.9800	0.9798	0.9802	0.9786	0.9698	0.9720
			len	10.6978	10.6994	10.5820	10.5414	10.0660	10.1802
400	0.9	3	cov	0.9682	0.9684	0.9696	0.9674	0.9678	0.9676
			len	9.8146	9.8074	9.8364	9.8190	9.7594	9.7764

Table 4.11 Simulated Large Sample 95% PI Coverages and Lengths for Binomial Regression, $p = 50$, $m = 7$

n	ψ	k		GLM	lasso	RL	OHFS	BE
1000	0	1	cov	0.9896	0.9838	0.9802	0.9798	0.9798
			len	4.0008	3.6666	3.5744	3.5838	3.5842
1000	0.14	1	cov	0.9868	0.9818	0.9782	0.9774	0.9770
			len	4.0422	3.6836	3.6158	3.6226	3.6312
1000	0.9	1	cov	0.9894	0.9794	0.9796	0.9800	0.9798
			len	4.0214	3.5994	3.5794	3.6122	3.6114
1000	0	19	cov	0.9888	0.9870	0.9848	0.9814	0.9812
			len	4.0294	3.9730	3.8438	3.7110	3.7030
1000	0.14	19	cov	0.9872	0.9846	0.9852	0.9804	0.9806
			len	4.0376	3.8350	3.7834	3.7170	3.7066
1000	0.9	19	cov	0.9884	0.9804	0.9808	0.9802	0.9772
			len	4.0348	3.6170	3.5948	3.6226	3.6216
1000	0	49	cov	0.990	0.9904	0.9904	0.9900	0.9904
			len	4.0428	4.0726	4.0528	4.0490	4.0460
1000	0.14	49	cov	0.9866	0.9866	0.9856	0.9806	0.9796
			len	4.0396	3.9044	3.8640	3.7046	3.6988
1000	0.9	49	cov	0.9874	0.9808	0.9792	0.9790	0.9772
			len	4.0660	3.6444	3.6230	3.6556	3.6490

Table 4.12 Simulated Large Sample 95% PI Coverages and Lengths, $n = 100$, $p = 200$

		BR	m=7	BR	m=40	PR,a=1	$\beta_1 = 1$	PR,a=2	$\beta_1 = 5$
ψ,k		lasso	RL	lasso	RL	lasso	RL	lasso	RL
0	cov	0.9912	0.9654	0.9836	0.9602	0.9816	0.9612	0.7620	0.9662
1	len	4.2774	3.8356	11.3482	11.001	7.8350	7.5660	93.7318	91.4898
0.07	cov	0.9904	0.9698	0.9796	0.9644	0.9790	0.9696	0.7652	0.9706
1	len	4.2570	3.9256	11.4018	11.1318	7.8488	7.6680	92.0774	89.7966
0.9	cov	0.9844	0.9832	0.9820	0.9820	0.9880	0.9858	0.7850	0.9628
1	len	3.8242	3.7844	10.9600	10.8716	7.6380	7.5954	98.2158	95.9954
0	cov	0.9146	0.8216	0.8532	0.7874	0.8678	0.8038	0.1610	0.6754
19	len	4.7868	3.8632	12.0152	11.3966	7.8126	7.5188	88.0896	90.6916
0.07	cov	0.9814	0.9568	0.9424	0.9208	0.9620	0.9444	0.3790	0.5832
19	len	4.1992	3.8266	11.3818	11.0382	7.9010	7.7828	92.3918	92.1424
0.9	cov	0.9858	0.9840	0.9812	0.9802	0.9838	0.9848	0.7884	0.9594
19	len	3.8156	3.7810	10.9194	10.8166	7.6900	7.6454	97.744	95.2898
0.07	cov	0.9820	0.9640	0.9604	0.9390	0.9720	0.9548	0.3076	0.4394
199	len	4.1260	3.7730	11.2488	10.9248	8.0784	7.9956	90.4494	88.0354
0.9	cov	0.9886	0.9870	0.9822	0.9804	0.9834	0.9814	0.7888	0.9586
199	len	3.8558	3.8172	10.9714	10.8778	7.6728	7.6602	97.0954	94.7604

leaves the study for any reason other than the event of interest. Then Z_i is the time until the ith observation is censored. Then the **right censored survival time** T_i of the ith observation is $T_i = \min(Y_i, Z_i)$. Let $\delta_i = 0$ if T_i is (right) censored ($T_i = Z_i$) and let $\delta_i = 1$ if T_i is not censored ($T_i = Y_i$).

We will assume that the censoring mechanism is independent of the time to event: Y_i and Z_i are independent. Often censoring occurs because of cost and time constraints. In the definition below, $F(t)$ is the cdf and $f(t)$ is the pdf of a univariate survival time random variable Y that satisfies $P(Y \geq 0) = 1$.

Definition 10.26. i) The **survival function** of Y is $S(t) = P(Y > t) = 1 - F(t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

ii) The **hazard function** of Y is $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

Next, we will consider an important class of survival regression models.

Definition 10.27. The **Cox proportional hazards regression (PH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t)$ is the **unknown baseline function** and $\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ is the **hazard ratio**. The sufficient predictor $\mathbf{SP} = \boldsymbol{\beta}^T\boldsymbol{x}_i = \sum_{j=1}^{p} \beta_j x_{ij}$.

The Cox PH model (= Cox PH regression model = Cox regression model = Cox proportional hazards regression model) is a 1D regression model since

the conditional distribution $Y|\boldsymbol{x}$ is completely determined by the hazard function, and the hazard function only depends on $\boldsymbol{x}$ through $\boldsymbol{\beta}^T\boldsymbol{x}$. Inference for the PH model uses computer output that is used almost exactly as the output for generalized linear models such as the logistic and Poisson regression models. The Cox PH model is semiparametric: the conditional distribution $Y|\boldsymbol{x}$ depends on the sufficient predictor $\boldsymbol{\beta}^T\boldsymbol{x}$, but the parametric form of the hazard function $h_{Y|\boldsymbol{x}}(t)$ is not specified. The Cox PH model is the most widely used survival regression model in survival analysis. For the Cox PH model, often we will use $\boldsymbol{\beta} = \boldsymbol{\beta}_C$.

Survival data is usually right censored so Y is not observed. Instead, the survival time $T_i = \min(Y_i, Z_i)$ where $Y_i \perp\!\!\!\perp Z_i$ and Z_i is the censoring time. Also $\delta_i = 0$ if $T_i = Z_i$ is censored and $\delta_i = 1$ if $T_i = Y_i$ is uncensored. Hence the data is $(T_i, \delta_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$.

The Weibull PH regression model of Definition 4.4 is an important parametric PH regression model. Theorem 4.4 still holds for the Cox PH regression model with AIC. The relaxed lasso estimator is the lasso variable selection model that fits the Cox PH regression model to the predictors with nonzero lasso coefficients. The relaxed lasso estimator is $\sqrt{n}$ consistent by Theorem 4.4 if the lasso estimator is consistent.

4.11.1 Simulations

For variable selection with the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$ where often $\boldsymbol{\theta}_0 = \boldsymbol{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The shorth estimator can be applied to a bootstrap sample $\hat{\beta}^*_{i1}, ..., \hat{\beta}^*_{iB}$ to get a confidence interval for β_i. Here $T_n = \hat{\beta}_i$ and $\theta = \beta_i$.

Next, we describe a small simulation study that was done using $B = \max(1000, n/25, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where k and ψ are defined in the following paragraph. In the simulations, we use $\theta = \boldsymbol{A}\boldsymbol{\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_S = \boldsymbol{1}$ and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_E = \boldsymbol{0}$.

In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I})$ where the p elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $p \times p$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (p-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (p-2)\psi^2]$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{x} = a\boldsymbol{z}/v$. Hence the correlations are $Cor(x_i, x_j) = \rho = (2\psi + (p-2)\psi^2)/(1 + (p-1)\psi^2)$ for $i \neq j$. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let

$SP = \boldsymbol{x}_i^T\boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations use $a = 1$ where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p - k$ zeros.

Then the Y_i were generated such that $Y|\boldsymbol{x}$ follows a Weibull proportional hazards regression model with the above $\boldsymbol{\beta}$. We used R code similar to that of Zhou (2001) to obtain $(Z_i, \delta_i, \boldsymbol{x}_i)$ where some of the Z_i are right censored. Some functions from *survpack* were useful. The function `phdata2` generates a data set as described above. We used the nonparametric bootstrap and the Cox PH model. The function `PHboot` bootstraps the full Cox PH model. The function `PHbootsim` is used to simulate the bootstrap for the full Cox PH model. The functions `LPHboot` and `RLPHboot` bootstraps a Cox PH model with lasso and relaxed lasso. The function `RLPHbootsim` is used to simulate the bootstrap for relaxed lasso. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $g = 1$. Some R code is shown below.

```
library(survival)
library(MASS)
library(glmnet)

out<- phdata2(n=100,p=4,k=1,psi=0,a=1,gam=1,clam = 0.1)
out$beta
$betaP
[1] 1 0 0 0
#out$x gives the matrix of predictors
out$time
$time
  [1] 10.5015  2.5748  2.1266  0.4238  0.4454
  [6]  0.1165  0.0233  0.3108  0.0856  0.3908
 .
 .
 .
 [96]  5.4669  0.1603  0.1510  0.1206  0.6356
out$status
$status  #0 means right censored
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0
      0 1 0 0 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 0 1 0 1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      0 0 1 1

RLPHbootsim(nruns=100,B=200,k=2) #slow 3 runs per minute
$mndd
[1] 3.01  #relaxed lasso used 3 predictors on average
$cicov
 [1] 0.94 0.96 0.97 0.99 0.95 0.97 0.97 0.93 0.95 0.95
$avelen
```

```
 [1] 0.8642748 0.8473142 0.7334978 0.7219106 2.5561583
     2.5561583 2.6622667 2.5124382 2.5124382 2.6253967
$beta
[1] 1 1 0 0
$k
[1] 2
PHbootsim(nruns=100,B=200,k=2) #fairly fast
$cicov
 [1] 0.96 0.95 0.92 0.92 0.91 0.94 0.94 0.95 0.99 0.99
$avelen
 [1] 0.8571470 0.8582906 0.7541797 0.7416362 2.5247451
     2.5247451 2.5558537 2.5021201 2.5021201 2.6243971
$beta
[1] 1 1 0 0
$k
[1] 2
```

The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ (whether first k $\beta_i = 1$) and $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ (whether the last $p - k$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value. The number of runs = 100 is tiny since the relaxed lasso simulation is slow. Using 5000 runs would be much better.

The regression models used the nonparamtric bootstrap on the relaxed lasso estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 4.13 gives results with $n = 100$, $p = 4$, and $k = 1$. Table 4.13 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for variable selection with relaxed lasso. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (2.30), hybrid region (2.32), and Bickel and Ren region (2.31). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests.

Volume ratios of the three confidence regions can be compared using (2.35), but there is not enough information in Table 4.13 to compare the volume of the confidence region for the full model regression versus that for the relaxed lasso since the two methods have different determinants $|\boldsymbol{S}^*_T|$. Table 4.13 corresponds to the above R output with $k = 2$.

The inference for forward selection was often as precise or more precise than the inference for the full model. The coverages were near 0.95 for the

Table 4.13 Bootstrapping Cox PH Regression With Relaxed Lasso

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.96	0.95	0.92	0.92	0.91	0.94	0.94	0.95	0.99	0.99
len	0.857	0.858	0.754	0.742	2.525	2.525	2.556	2.502	2.502	2.624
vs,0	0.94	0.96	0.97	0.99	0.95	0.97	0.97	0.93	0.95	0.95
len	0.864	0.847	0.733	0.722	2.556	2.556	2.662	2.512	2.512	2.625

regression bootstrap on the full model, although there was slight undercoverage for the tests since $(n-p)/n = 0.96$ when $n = 25p$. Suppose $\psi = 0$. Then it may be true that $\hat{\boldsymbol{\beta}}_S$ has the same limiting distribution for I_{min} and the full model. Note that the average lengths and coverages were similar for the full model and forward selection I_{min} for β_1, β_2, and $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$. Forward selection inference was more precise for $\boldsymbol{\beta}_E = (\beta_3, \beta_4)^T$. The Bickel and Ren (2.31) cutoffs and coverages were at least as high as those of the hybrid region (2.32).

See Olive (2020) for results on survival analysis that are similar to the results given in these online notes for MLR and GLMS. In particular, graphs for checking and visualizing the model, prediction intervals, inference, and inference after variable selection, including lasso variable selection, are given. See Tibshirani (1997) and Simon et al. (2011) for lasso and elastic net with the Cox PH regression model.

4.12 Regression Trees

A regression tree is a flexible method for $Y = m(\boldsymbol{x}) + e$ or for $Y_i = m(\boldsymbol{x}_i) + \sigma_i e_i$ where the zero mean errors e_i are iid. The method produces a graph called a tree. Each branch has a label like $x_i > 7.56$ if x_i is quantitative, or $x_j \in \{a, c\}$ (written $x_j = ac$) where x_j is a factor taking on values a, b, c, d, e, f, say. **Unless told otherwise**, go to the left branch if the condition is true, go to the right branch if the condition is false. (Some software switches this. Check the story problem.) The bottom of the tree has leaves that give $\hat{Y} = \hat{Y}|\boldsymbol{x}$. The root is the top node, a leaf is a terminal node, and a split is a rule for creating new branches. Each node has a left and right branch.

Example 4.19. Given a tree and $\boldsymbol{x}$ values, find $\hat{Y}$. The Venables and Ripley (1997, p. 420) and Ein-Dor and Feldmesser (1987) cpu data has $Y = perf$ = central processing unit (CPU) performance with predictor variables $x_1 = cach$ = cache size in kilobytes, $x_2 = mmax$ = maximum main memory in kilobytes, $x_3 = syct$ = cycle time in nanoseconds, and $x_4 = chmin$ = minimum number of channels. The regression tree is shown on the following page.

a) Predict Y if $cach = 30$ and $mmax = 25000$.

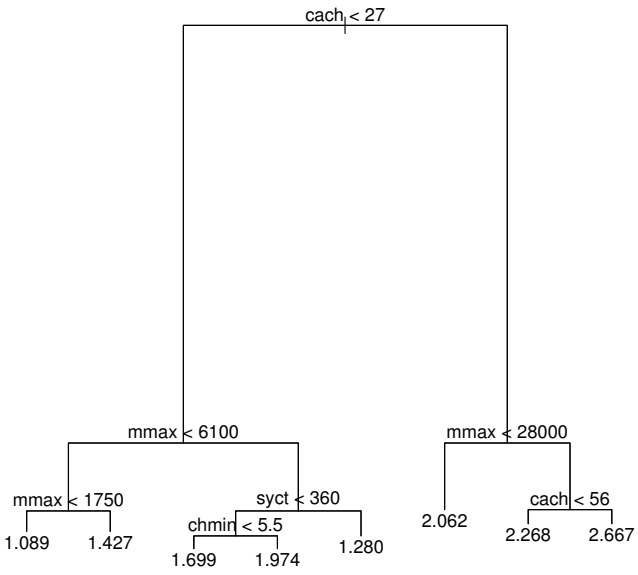

Fig. 4.19 Regression Tree for Example 4.19.

Solution: Since $cach = 30$, the $cach < 27$ condition is false. Go to the right branch. Since $mmax = 25000$, the condition for the next node is true. Go to the left branch where $\hat{Y} = 2.062$.

b) Predict Y if $cach = 25$, $mmax = 7000$, $syct = 200$, and $chmin = 5$.

Solution: Go to the left, then right, then left, then left where $\hat{Y} = 1.699$.

Regression trees have some advantages. Trees can be easier to interpret than competing methods when some predictors are numerical and some are categorical. Trees are invariant to monotone (increasing or decreasing) transformations of the predictor variable x_i. Regression trees can handle missing values better than MLR and can beat MLR if there is nonadditive behavior. Trees can handle complex unknown interactions. Regression trees i) give prediction rules that can be rapidly and repeatedly evaluated, ii) are useful for screening predictors (interactions, variable selection), iii) can be used to assess the adequacy of linear models, and iv) can summarize large multivariate data sets.

Trees that use recursive partitioning for classification and regression trees use the CART algorithm. (Classification trees are very similar to regression

trees. See Section 5.9.) In growing a tree, the binary partitioning algorithm recursively splits the data in each node until either the node is homogeneous ($Y \approx$ constant for a regression tree) or the node contains too few observations (default ≤ 5). The *deviance* is a measure of node homogeneity, and deviance $= 0$ for a perfectly homogeneous node. For a regression tree, often $\hat{Y}$ is the mean of the node observations.

Trees divide the predictor space (set of possible values of the training data $\boldsymbol{x}_i$) into J distinct and nonoverlapping regions $R_1, ..., R_J$ that are high dimensional boxes. Then for every observation that falls in R_j, make the same prediction. Hence $\hat{Y}_{R_j} =$ sample mean of training data Y_i in R_j. Choose R_j so $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{Y}_{R_j})^2$ is small. Let $\{\boldsymbol{x} | x_j < s\}$ be the region in the predictor space such that $x_j < s$ where $\boldsymbol{x} = (x_1, ..., x_p)^T$. Define 2 regions $R_1(j, s) = \{\boldsymbol{x} | x_j < s\}$ and $R_2(j, s) = \{\boldsymbol{x} | x_j \geq s\}$. Then seek cutpoint s and variable x_j to minimize

$$\sum_{i:\boldsymbol{x}_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\boldsymbol{x}_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2.$$

This can be done "quickly" if p is small (could use order statistics). Then repeat the process looking for the best predictor and the best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions. Only split one of the regions, R_1, R_2, and R_3. Continue this process until a stopping criterion is reached such as no region contains more than 5 observations (and stop if the region is homogeneous). If J is too large, the tree overfits.

Since a regression tree uses J regions, the response plot of $ESP = \hat{Y} = \hat{m}(\boldsymbol{x})$ versus Y consists of J dot plots that scatter about the identity line. A dot plot of $z_1, ..., z_m$ consists of an axis and m points corresponding to the vaules of z_i. The regression tree response plot has a dotplot of n_m cases with $\hat{Y} = \hat{Y}_{R_m}$ for each of the J regions. The residual plot consists of J dot plots that scatter about the $r = 0$ line. If $Y = m(\boldsymbol{x}) + e$, we can make prediction intervals for Y_f with the regression tree using $\hat{Y} = ESP = \hat{m}(\boldsymbol{x})$ and $r = Y - \hat{Y}$ as before.

If $Y = \alpha + \sum j = 1^p S_j(x_j) + e$ or $Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e$, then slicing the ESP $\hat{\alpha} + \sum j = 1^p \hat{S}_j(x_j)$ or $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ is more effective than partitioning the predictor space with hyperboxes R_k. Consider the response plot of ESP versus Y with the identity line or lowess added as a visual aid.

4.12.1 Boosting

This subsection follow James et al. (2013) closely. Techniques that can be used to improve both regression and classification trees are discussed in Section

5.9. A technique for improving regression trees is boosting. Like bagging, boosting can be applied to many statistical models, including regression and classification trees.

The boosting algorithm for regression trees follows. i) Set $\hat{f}(\boldsymbol{x}) = 0$ and $r_i - Y_i$ for $i = 1, ..., n$. Hence the step i) residuals are the training data. ii) For $b = 1, ..., B$ repeat: a) fit tree $\hat{f}_b$ with d splits ($d + 1$ teminal nodes) to the training data $(\boldsymbol{X}, \boldsymbol{r})$ where the predictors are collected in matrix $\boldsymbol{X}$. b) Update $\hat{f}(\boldsymbol{x})$ by adding a shrunken version of the new tree: $\hat{f}(\boldsymbol{x}) \leftarrow \hat{f}(\boldsymbol{x}) + \lambda \hat{f}_b(\boldsymbol{x})$, and update the residuals $r_i \leftarrow r_i - \lambda \hat{f}(\boldsymbol{x})$. iii) The boosted model

$$\hat{f}(\boldsymbol{x}) = \sum_{b=1}^{B} \lambda \hat{f}_b(\boldsymbol{x}).$$

The tree is fit to updated residuals rather than Y. This technique slowly improves $\hat{f}$ in areas where it does not perform well, and λ slows the learning process further. As a rule of thumb, iterative techniques that learn slowly tend to perform well. Often $d = 1$ is used where a $d = 1$ tree is called a "stump. The value d is called the interaction depth. The value λ tends to be 0.01 or 0.001. Very small λ tends to need very large B for good performance. Using the $d = 1$ stumps leads to an additive model

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{p} \hat{f}_j(x_j)$$

which is a competitor for the additive error regression GAM.

4.13 Data Splitting

Data splitting is used for inference after model selection. Use a training set to select a full model, and a validation set for inference with the selected full model. Here $p >> n$ is possible. See Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019). Typically when training and validation sets are used, the training set is bigger than the validation set or half sets are used, often causing large efficiency loss.

Let J be a positive integer and let $\lfloor x \rfloor$ be the integer part of x, e.g., $\lfloor 7.7 \rfloor = 7$. Initially divide the data into two sets H_1 with $n_1 = \lfloor n/(2J) \rfloor$ cases and V_1 with $n - n_1$ cases. If the fitted model from H_1 is not good enough, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Continue in this manner, possibly forming sets $(H_1, V_1), (H_2, V_2), ..., (H_J, V_J)$ where H_i has $n_i = in_1$ cases. Stop when H_d gives a reasonable model I_d with a_d predictors if $d < J$. Use $d = J$,

otherwise. Use the model I_d as the full model for inference with the data in V_d.

This procedure is simple for a fixed data set, but it would be good to automate the procedure. For example, if $n = 500000$ and $p = 90$, using $n_1 = 900$ would result in a much smaller loss of efficiency than $n_1 = 250000$.

4.14 Complements

This chapter used material from Chang and Olive (2010), Olive (2013b, 2017a: ch. 13), Olive et al. (2020), and Rathnayake and Olive (2019). GLMs were introduced by Nelder and Wedderburn (1972). Useful references for generalized additive models include Hastie and Tibshirani (1986, 1990), and Wood (2017). Zhou (2001) is useful for simulating the Weibull regression model. Also see McCullagh and Nelder (1989), Agresti (2013, 2015), and Cook and Weisberg (1999, ch. 21-23). Collett (2003) 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 4.7 are covered by Zuur et al. (2009), Simonoff (2003), and Hilbe (2011). Cook and Zhang (2015) show that envelope methods have the potential to significantly improve GLMs. Some GLM large sample theory is given by Claeskens and Hjort (2008, p. 27), Cook and Zhang (2015), and Sen and Singer (1993, p. 309).

An introductions to 1D regression and regression graphics is Cook and Weisberg (1999a, ch. 18, 19, and 20), while Olive (2010) considers 1D regression. A more advanced treatment is Cook (1998). Important papers include Brillinger (1977, 1983) and Li and Duan (1989). Li (1997) shows that OLS F tests can be asymptotically valid for model (4.18) if $\boldsymbol{u}$ is multivariate normal and $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}Y} \neq \mathbf{0}$.

In Section 4.9, the functions `binregbootsim` and `pregbootsim` are useful for the full binomial regression and full Poisson regression models. The functions `vsbrbootsim` and `vsprbootsim` were used to bootstrap backward elimination for binomial and Poisson regression. The functions `LRboot` and `vsLRboot` bootstrap the logistic regression full model and backward elimination. The functions `PRboot` and `vsPRboot` bootstrap the Poisson regression full model and backward elimination.

In Section 4.10, table entries for Poisson regression were made with `prpisim2` while entries for binomial regression were made with `brpisim`. The functions `prpiplot2` and `lrpiplot` were used to make Figures 4.17 and 4.18. The function `prplot` can be used to check the full Poisson regression model for overdispersion. The function `prplot2` can be used to check other Poisson regression models such as a GAM or lasso.

i) *Resistant regression:* Suppose the regression model has an $m \times 1$ response vector $\boldsymbol{y}$, and a $p \times 1$ vector of predictors $\boldsymbol{x}$. Assume that predictor transformations have been performed to make $\boldsymbol{x}$, and that $\boldsymbol{w}$ consists of $k \leq p$ continuous predictor variables that are linearly related. Find the RMVN set based on the $\boldsymbol{w}$ to obtain n_u cases $(\boldsymbol{y}_{ci}, \boldsymbol{x}_{ci})$, and then run the regression method on the cleaned data. Often the theory of the method applies to the cleaned data set since $\boldsymbol{y}$ was not used to pick the subset of the data. Efficiency can be much lower since n_u cases are used where $n/2 \leq n_u \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{w}$.

The method will have the most outlier resistance if $k = p$ (or $k = p - 1$ if there is a trivial predictor $X_1 \equiv 1$). If $m = 1$, make the response plot of $\hat{Y}_c$ versus Y_c with the identity line added as a visual aid, and make the residual plot of $\hat{Y}_c$ versus $r_c = Y_c - \hat{Y}_c$.

In R, assume Y is the vector of response variables, x is the data matrix of the predictors (often not including the trivial predictor), and w is the data matrix of the $\boldsymbol{w}_i$. Then the following R commands can be used to get the cleaned data set. We could use the `covmb2` set B instead of the RMVN set U computed from the $\boldsymbol{w}$ by replacing the command *getu(w)* by `getB(w)`.

```
indx <- getu(w)$indx  #often w = x
Yc <- Y[indx]
Xc <- x[indx,]
#example
indx <- getu(buxx)$indx
Yc <- buxy[indx]
Xc <- buxx[indx,]
outr <- lsfit(Xc,Yc)
MLRplot(Xc,Yc) #right click Stop twice
```

a) *Resistant additive error regression*: An additive error regression model has the form $Y = h(\boldsymbol{x}) + e$ where there is $m = 1$ response variable Y, and the $p \times 1$ vector of predictors $\boldsymbol{x}$ is assumed to be known and independent of the additive error e. An enormous variety of regression models have this form, including multiple linear regression, nonlinear regression, nonparametric regression, partial least squares, lasso, ridge regression, etc. Find the RMVN set (or `covmb2` set) based on the $\boldsymbol{w}$ to obtain n_U cases $(Y_{ci}, \boldsymbol{x}_{ci})$, and then run the additive error regression method on the cleaned data.

b) *Resistant Additive Error Multivariate Regression*

Assume $\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{\epsilon} = E(\boldsymbol{y}|\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $g : \mathbb{R}^p \to \mathbb{R}^m$, $\boldsymbol{y} = (Y_1, ..., Y_m)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_m)^T$. Many models have this form, including multivariate linear regression, seemingly unrelated regressions, partial envelopes, partial least squares, and the models in a) with $m = 1$ response variable. Clean the data as in a) but let the cleaned data be stored in $(\boldsymbol{Z}_c, \boldsymbol{X}_c)$. Again, the theory of the method tends to apply to the method applied to the cleaned data since the response variables were not used to select the cases, but the efficiency is often much lower. In the R code below, assume the $\boldsymbol{y}$ are stored in z.

```
indx <- getu(w)$indx  #often w = x
Zc <- z[indx]
Xc <- x[indx,]
#example
ht <- buxy
t <- cbind(buxx,ht);
z <- t[,c(2,5)];
x <- t[,c(1,3,4)]
indx <- getu(x)$indx
Zc <- z[indx,]
Xc <- x[indx,]
mltreg(Xc,Zc) #right click Stop four times
```

4.15 Problems

```
Output for problem 4.1: 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
```

4.1. 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 4.2        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
```

4.2*. Now the data is as in Problem 4.1, 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 $= x_1 = 550.0$ and length $= x_2 = 200.0$.

b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

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

```
Output for Problem 4.3
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
```

4.3*. 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, and the possums data is from Cook and Weisberg (1999).

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *habitat* $= x_1 = 5.8$ and *stags* $= x_2 = 8.2$.

b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

c) Find a 95% confidence interval for β_2.

	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

4.4*. 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 0s and 700 were 1s.

a) For the change in deviance test, if the p-value ≥ 0.07, there is little evidence that Ho should be rejected. If $0.01 <$ 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*: $\boldsymbol{\eta}^T\boldsymbol{u}$ instead of $\boldsymbol{\beta}^T\boldsymbol{x}$) relevant?

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

4.5. The smoothing spline simulation in Problem 4.7 compares the PI lengths and coverages of 3 large sample 95% PIs for $Y = m(x) + e$ and a single measurement x. Values for the first PI were denoted by scov and slen, values for 2nd PI were denoted by ocov and olen, and values for third PI by dcov and dlen. The average degrees of freedom of the smoothing spline was recorded as *adf*. The number of runs was 5000. The *len* was the average length of the PI and the *cov* was the observed coverage. One student got the following results shown in Table 4.2.

Table 4.14 Results for 3 PIs

error type	n	95% slen	PI olen	95% dlen	PI scov	95% ocov	PI dcov	adf
5	100	18.028	17.300	18.741	0.9438	0.9382	0.9508	9.017

For the PIs with coverage ≥ 0.94, which PI was the most precise (best)?

4.6. James et al. (2013. p.p. 327-328) consider the 1978 Boston housing data where Y_i = *median house price* (in \$1000's so 74 = 74000) in the ith suburb. The predictors are $x_1 = lstat =$ percentage of individuals with lower socioeconomic status, and $x_2 = RM =$ average number of rooms per dwelling. The pruned regression tree shown in Figure 4.6 used a training set of half of the cases.

a) Predict the median price (multiply by 1000) if $x_1 = 7$ and $x_2 = RM = 8$.

b) Predict the median price (multiply by 1000) if $x_1 > 22$.

R Problems

Use the command *source("G:/slpack.txt")* **to download the functions** and the command *source("G:/sldata.txt")* **to download the data. See Preface or Section 8.1.** Typing the name of the `slpack` function, e.g. *lrplot2*, will display the code for the function. Use the `args` command, e.g. *args(lrplot2)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/slrhw.txt) into *R*.

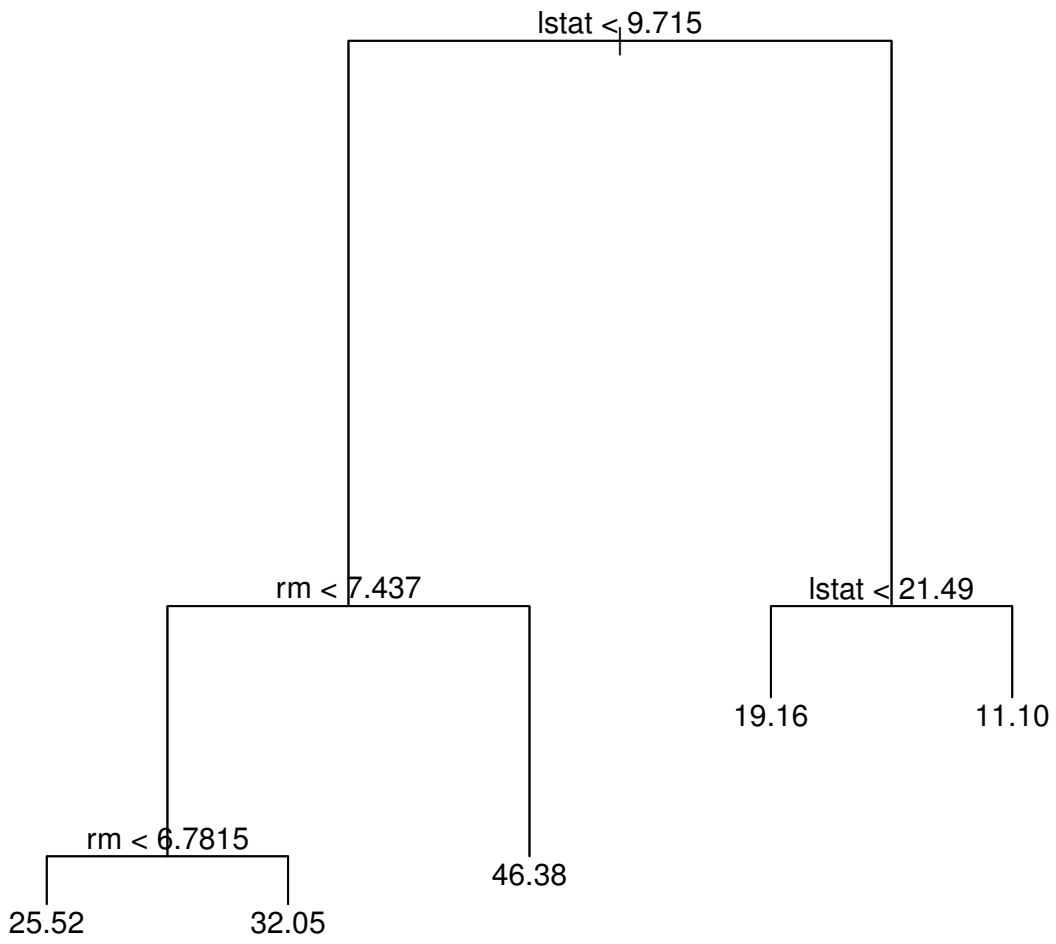

Fig. 4.20 Regression Tree for Problem 4.6.

4.7. The Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response Y = *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable x = year (1950-1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963 and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted.

a) The simple linear regression model is $Y = \alpha + \beta x + e = SP + e$. Copy and paste the *R commands* for this part to make a response plot of $ESP = \hat{Y} = \hat{\alpha} + \hat{\beta} x$ versus Y for this model. Include the plot in *Word*.

b) The additive model is $Y = \alpha + S(x) + e = AP + e$ where S is some unknown function of x. The *R commands* make a response plot of $EAP = \hat{\alpha} + \hat{S}(x)$ versus Y for this model. Include the plot in *Word*.

c) The simple linear regression model is a special case of the additive model with $S(x) = \beta x$. The additive model is a special case of the additive error regression model $Y = m(x) + e$ where $m(x) = \alpha + S(x)$. The response plots for these three models are used in the same way as the response plot for the multiple linear regression model: if the model is good, then the plotted points should cluster about the identity line with no other pattern. Which response plot is better for showing that something is wrong with the model? Explain briefly.

4.8. In a generalized additive model (GAM), $Y \perp\!\!\!\perp \boldsymbol{x} | AP$ where $AP = \alpha + \sum_{i=1}^{k} S_i(x_i)$. 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 (2017, pp. 125-130) 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 $x_1 = ck$, $x_2 = [ck]^2$, and $x_3 = [ck]^3$. 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 4.7 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) gives 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?

4.9. The smoothing spline simulation compares the PI lengths and coverages of 3 PIs for $Y = m(x) + e$ and a single measurement x. Values for the first PI were denoted by scov and slen, values for 2nd PI were denoted by ocov and olen, and values for third PI (2.7) by dcov and dlen. The second PI replaces d by 1 in PI (2.7). Three model types were used 1) $m(x) = x + x^2$, 2) $m(x) = \sin(x) + \cos(x) + \log(|x|)$, and 3) $m(x) = 3\sqrt{|x|}$. The smoothing spline is flexible so the df $> p$. The estimated df is given by adf. Copy and paste the *R* commands for this problem and make a table like the one below. The pimenlen gives slen, olen, and dlen.

Table 4.15 Table for Problem 4.8: PIs for modt = 1,

error type	n	95% slen	PI olen	95% dlen	PI scov	95% ocov	PI dcov	adf
1	100	4.7095	4.6949	5.0585	0.9660	0.9604	0.9736	6.27

a) For Table 4.3, which PI worked best?

b) For the table you make from the *R* output, which PI worked best?

4.10. This problem does lasso for binary regression for artificial data with $n = 100$, $p = 101$ and 5 active population nontrivial predictors. If $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$, then the 100 nontrivial predictors are in $\boldsymbol{x}$ and $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, ..., 0)^T$.

a) Copy and paste the source and library commands into *R*. Then copy and paste the commands for this part into *R*. Relaxed lasso gets the binary logistic regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. Include the plot in *Word*.

Does the step function track the logistic curve?

b) Copy and paste the commands for this part into *R*. These commands to MLR lasso, then the relaxed lasso gets the binary logistic regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. For this data set, one more predictor was used than that in a). Include the plot in *Word*.

Does the step function track the logistic curve?

c) Copy and paste the commands for this part into *R*. The commands for this part use MLR forward selection with EBIC, and only nontrivial predictor x_4 was selected. Then the binary logistic regression if fit using this variable and the response plot is made. Include the plot in *Word*.

Is the plot in c) worse than the plots in a) and b)?

4.11. This problem does lasso for Poisson regression for artificial data with $n = 100$, $p = 101$ and 5 active population nontrivial predictors. If $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$, then the 100 nontrivial predictors are in $\boldsymbol{x}$ and $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, ..., 0)^T$.

a) Copy and paste the source and library commands into *R*. Then copy and paste the commands for this part into *R*. Relaxed lasso gets the Poisson regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. Include the plot in *Word*. The horizontal line is $\overline{Y}$ and the jagged curve is lowess which tracked the exponential curve well until ESP > 3. Lasso overfit using 26 variables instead of 5.

b) Copy and paste the commands for this part into *R*. These commands to MLR lasso, then the relaxed lasso gets the Poisson regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. For this data set, 20 variables were used. Include the plot in *Word*.

c) Copy and paste the commands for this part into *R*. The commands for this part use MLR forward selection with EBIC, and only nontrivial predictor x_5 was selected. Then the Poisson regression if fit using this variable and the response plot is made. Include the plot in *Word*.

If the Poisson regression model is good, we would like the vertical scale to be not more than 10 times the horizontal scale in the OD plot. (This happened in a) and b).) Is the vertical scale more than 10 times the horizontal scale in the OD plot for this model?

4.12. This problem on regression trees is taken from the vignettes for the *R* package `rpart`. See Therneau and Atkinson (2017).

The dataset contains 34 variables on $n = 111$ cars from April, 1990 *Consumer Reports*. The variables "tire size" and "model name" were omitted and "rim size" was also deleted because it was too good a predictor of price. The response Y = price/1000. The four variables used it the tree construction were *Country, Disp, HP.revs* and *Type*.

a) Use the *R* code for this part to print the regression tree. Then predict the car price (in dollars so multiply $\hat{Y}$ by 1000) if $Disp = 200$ and $Hp.res = 5000$.

b) Predict the car price $1000\hat{Y}$ if $Disp = 100$, $Country = a$, and $Type = a$. Note that you go to the left of the tree branch if the label condition is true, and to the right of the tree branch if the label condition is not true.