Chapter 13

Generalized Linear Models

13.1 Introduction

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression and loglinear Poisson regression. Assume that there is a response variable Y and a $k \times 1$ 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 13.1. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{13.1}$$

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

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{13.2}$$

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

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{13.3}$$

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

The GLM parameterization (13.3) can be written in several ways. By Equation (13.2),

$$\begin{aligned} f(y_i|\theta(\boldsymbol{x}_i)) &= \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y) \end{aligned}$$

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

$$\mu(\boldsymbol{x}_i) = g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{13.4}$$

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

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

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

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

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

iv) Check for goodness of fit of the GLM with an estimated sufficient summary plot.

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

vi) Check for overdispersion with an OD plot.

vii) Check whether Y is independent of $\boldsymbol{x}$; ie, check whether $\boldsymbol{\beta} = \boldsymbol{0}$.

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

ix) Use variable selection to find a good submodel.

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

13.2 Multiple Linear Regression

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

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

Since

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

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

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

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

for $i = 1, ..., n$. This model can be written as

$$Y_i \equiv Y_i|\boldsymbol{x}_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$$

where $e_i \sim N(0, \sigma^2)$.

When the predictor variables are quantitative, the above model is called a multiple linear regression (MLR) model. When the predictors are categorical, the above model is called an analysis of variance (ANOVA) model, and when the predictors are both quantitative and categorical, the model is called an MLR or analysis of covariance model. The MLR model is discussed in detail in Chapter 5, where the normality assumption and the assumption that σ is known can be relaxed.

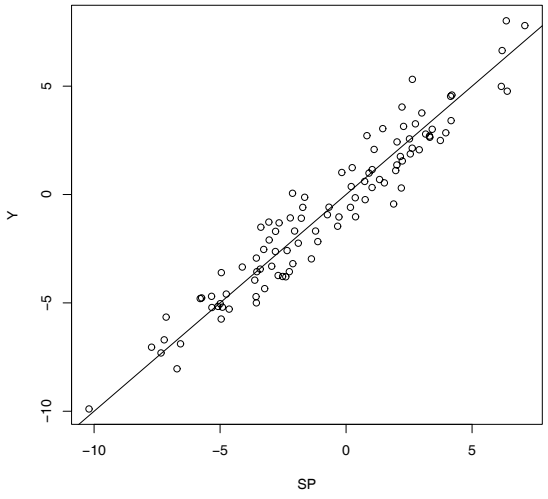

Figure 13.1: SSP for MLR Data

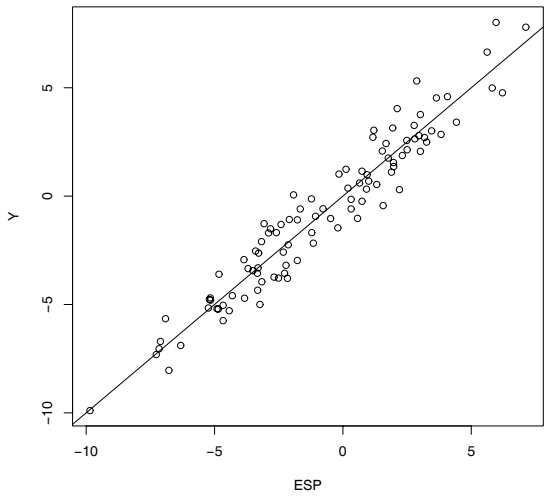

Figure 13.2: ESSP = Response Plot for MLR Data

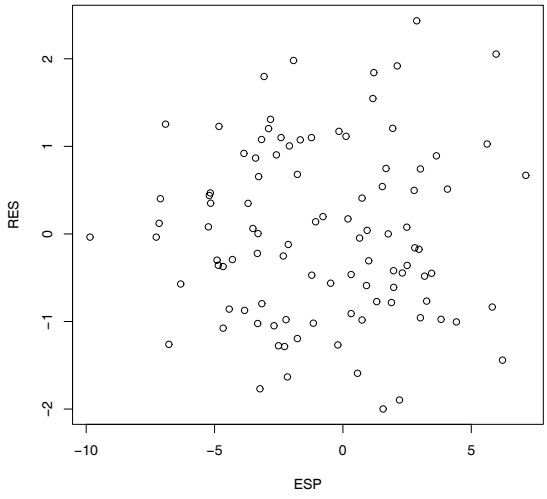

Figure 13.3: Residual Plot for MLR Data

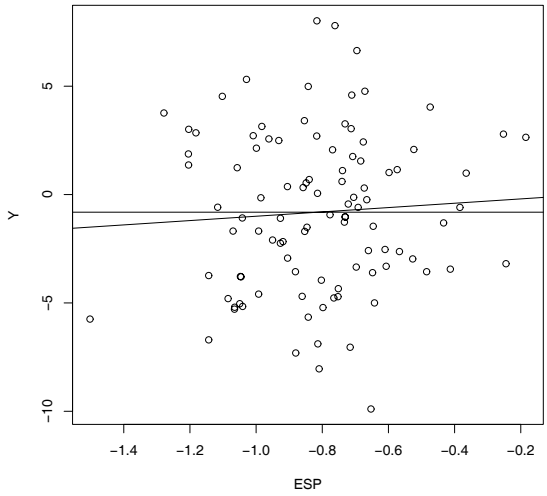

Figure 13.4: Response Plot when Y is Independent of the Predictors

A sufficient summary plot (SSP) of the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the multiple linear regression model. This plot can not be used for real data since α and $\boldsymbol{\beta}$ are unknown. The artificial data used to make Figure 13.1 used $n = 100$ cases with $k = 5$ nontrivial predictors. The data used $\alpha = -1$, $\boldsymbol{\beta} = (1, 2, 3, 0, 0)^T$, $e_i \sim N(0, 1)$ and $\boldsymbol{x} \sim N_5(\mathbf{0}, \boldsymbol{I})$.

In Figure 13.1, notice that the identity line with unit mean and zero intercept corresponds to the mean function since the identity line is the line $Y = SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = g(\mu(\boldsymbol{x}))$. The vertical deviation of Y_i from the line is equal to $e_i = Y_i - (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. For a given value of SP, $Y_i \sim N(SP, \sigma^2)$. For the artificial data, $\sigma^2 = 1$. Hence if $SP = 0$ then $Y_i \sim N(0, 1)$, and if $SP = 5$ the $Y_i \sim N(5, 1)$. Imagine superimposing the $N(SP, \sigma^2)$ curve at various values of SP. If all of the curves were shown, then the plot would resemble a road through a tunnel. For the artificial data, each Y_i is a sample of size 1 from the normal curve with mean $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$.

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

as that of the SSP, but now the mean SP is estimated by the estimated sufficient predictor (ESP). This plot is used as a goodness of fit diagnostic. The residual plot is a plot of the ESP versus r_i and is used as a lack of fit diagnostic. These two plots should be made immediately after fitting the MLR model and before performing inference. Figures 13.2 and 13.3 show the response plot and residual plot for the artificial data.

The response plot is also a useful visual aid for describing the ANOVA F test (see p. 174) which tests whether $\boldsymbol{\beta} = \mathbf{0}$, that is, whether the predictors $\boldsymbol{x}$ are needed in the model. If the predictors are not needed in the model, then Y_i and $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the sample mean $\overline{Y}$. If the predictors are needed, then Y_i and $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the ESP $\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. The fitted value $\hat{Y}_i$ is the maximum likelihood estimator computed using ordinary least squares. If the identity line clearly fits the data better than the horizontal line $Y = \overline{Y}$, then the ANOVA F test should have a small p–value and reject the null hypothesis H_o that the predictors $\boldsymbol{x}$ are not needed in the MLR model. Figure 13.4 shows the response plot for the artificial data when only X_4 and X_5 are used as predictors with the identity line and the line $Y = \overline{Y}$ added as visual aids. In this plot the horizontal line fits the data about as well as the identity line which was expected since Y is independent of X_4 and X_5.

It is easy to find data sets where the response plot looks like Figure 13.4, but the p–value for the ANOVA F test is very small. In this case, the MLR model is statistically significant, but the investigator needs to decide whether the MLR model is practically significant.

13.3 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. Often the labelling is arbitrary, eg, 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 13.3. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i})).$$

The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{13.5}$$

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

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})),$$

or

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{13.6}$$

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

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

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

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

Hence this family is a 1-parameter exponential family with $\theta = \rho = E(Y)$ and canonical link

$$c(\rho) = \log\left(\frac{\rho}{1 - \rho}\right).$$

This link is known as the *logit link*, and if $g(\mu(\boldsymbol{x})) = g(\rho(\boldsymbol{x})) = c(\rho(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ then the inverse link satisfies

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

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

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1 - \rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression,

$$\rho(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}).$$

If this population proportion $\rho = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1 + \exp(x))$ which is the cdf of the logistic $L(0, 1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the Normal $N(0, 1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1 - \exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 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,

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

Definition 13.4. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{13.7}$$

and

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

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 $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Consider the ESS plot of the ESP versus Y. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. In this case the maximum likelihood estimator for the logistic regression parameters $(\alpha, \boldsymbol{\beta})$ does not exist because the logistic curve can not approximate a step function perfectly. See Atkinson and Riani (2000, p. 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, change in SS F test, and OLS t tests are often asymptotically valid when the conditions in Definition 13.4 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 13.1 below. Assume that $\text{Cov}(\boldsymbol{x}) \equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and that $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. Let $\boldsymbol{\mu}_j = E(\boldsymbol{x}|Y = j)$ for $j = 0, 1$. Let N_i be the number of Ys that are equal to i for $i = 0, 1$. Then

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

for $i = 0, 1$ while $\hat{\pi}_i = N_i/n$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Notice that Theorem 13.1 holds as long as $\text{Cov}(\boldsymbol{x})$ is nonsingular and Y is binary with values 0 and 1. The LR and discriminant function models need not be appropriate.

Theorem 13.1. Assume that Y is binary and that $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular. Let $(\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS})$ be the OLS estimator found from regressing Y on a constant and $\boldsymbol{x}$ (using software originally meant for multiple linear regression). Then

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

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

Proof. From Section 12.5,

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

and

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

Thus

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

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

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

and the result follows. QED

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

$$\hat{\boldsymbol{\beta}}_D = \frac{n(n-1)}{N_0 N_1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}.$$

Now when the conditions of Definition 13.4 are met and if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is small enough so that there is not perfect classification, then

$$\boldsymbol{\beta}_{LR} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Empirically, the OLS ESP and LR ESP are highly correlated for many LR data sets where the conditions are not met, eg when some of the predictors are factors. This suggests that $\boldsymbol{\beta}_{LR} \approx d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ for many LR data sets where d is some constant depending on the data.

Using Definition 13.4 makes simulation of logistic regression data straightforward. Set $\pi_0 = \pi_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$. Then $\alpha = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. The artificial data set used in the following discussion used $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$ and hence $\alpha = -1.5$. Let N_i be the number of cases where $Y = i$ for $i = 0, 1$. For the artificial data, $N_0 = N_1 = 100$, and hence the total sample size $n = N_1 + N_0 = 200$.

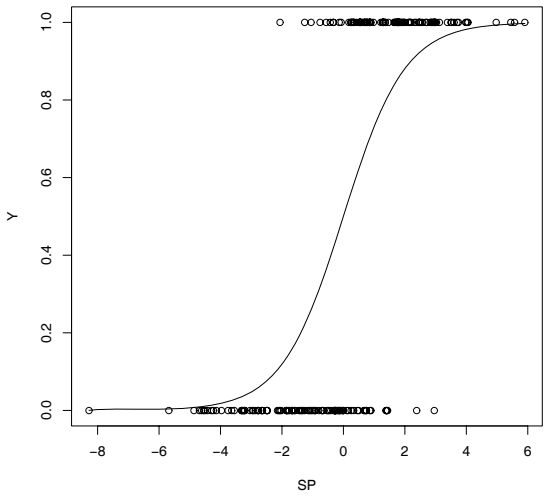

Figure 13.5: SSP for LR Data

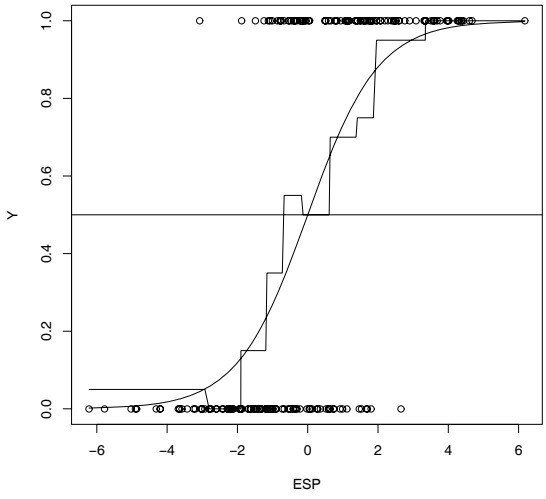

Figure 13.6: ESS Plot for LR Data

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

Unlike the SSP for multiple linear regression where the mean function is always the identity line, the mean function in the SSP for LR can take a variety of shapes depending on the range of the SP. For the LR SSP, the mean function is

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

If the SP = 0 then $Y|SP \sim$ binomial(1,0.5). If the SP = -5, then $Y|SP \sim$ binomial(1,$\rho \approx 0.007$) while if the SP = 5, then $Y|SP \sim$ binomial(1,$\rho \approx 0.993$). Hence if the range of the SP is in the interval $(-\infty, -5)$ then the mean function is flat and $\rho(SP) \approx 0$. If the range of the SP is in the interval $(5, \infty)$ then the mean function is again flat but $\rho(SP) \approx 1$. If $-5 < SP < 0$ then the mean function looks like a slide. If $-1 < SP < 1$ then the mean function looks linear. If $0 < SP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < SP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 13.5.

The estimated sufficient summary plot (ESSP or ESS plot) is a plot of $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function

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

added as a visual aid. The interpretation of the ESS plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

This plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Y's in each slice and add the resulting step function to the ESS plot. This is done in Figure 13.6 with $J = 10$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147–156).

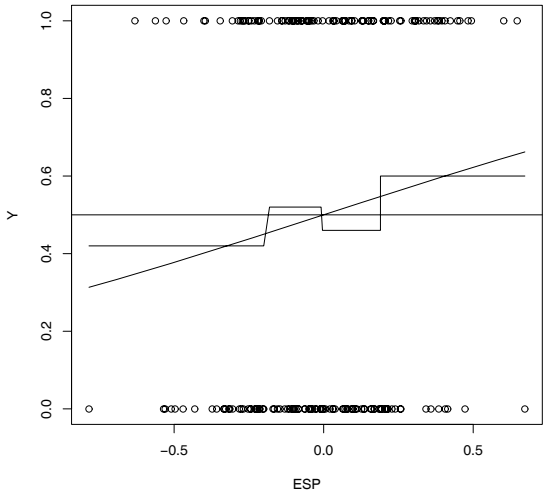

Figure 13.7: ESS Plot When Y Is Independent Of The Predictors

The deviance test described in Section 13.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the 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

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

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_o will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be independent of the predictors. Figure 13.7 shows the ESS plot when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 13.7 where the p–value for the deviance test is very small. Then the LR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

For binary data the Y_i only take two values, 0 and 1, and the residuals do

not behave very well. Hence the ESS plot will be used both as a goodness of fit plot and as a lack of fit plot.

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

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

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, p. 93-94) and Agresti (2002, p. 554-555), let $\delta = \rho/\theta$ and $\nu = (1 - \rho)/\theta$, so $\rho = \delta/(\delta + \nu)$ and $\theta = 1/(\delta + \nu)$. Let

$$B(\delta, \nu) = \frac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta + \nu)}.$$

If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m}, \rho, \theta)$, then the probability mass function of Y is

$$P(Y = y) = \binom{m}{y} \frac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$$

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 $\mathrm{V}(Y) = m\rho(1-\rho)[1+(m-1)\theta/(1+\theta)]$. If $Y|\pi \sim$ binomial(m, π) and $\pi \sim$ beta(δ, ν), then $Y \sim \mathrm{BB(m}, \rho, \theta)$.

Definition 13.5. The BBR model states that $Y_1, ..., Y_n$ are independent random variables where $Y_i|SP_i \sim \mathrm{BB(m_i}, \rho(\mathrm{SP_i}), \theta)$.

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

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

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

For both the LR and BBR models, the conditional distribution of $Y|\boldsymbol{x}$ can still be visualized with an ESS plot of the ESP versus Y_i/m_i with the estimated mean function

$$\hat{\rho}(ESP)$$

and a step function or lowess curve added as visual aids.

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

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

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

The ESS plot is a powerful method for assessing the adequacy of the binary LR regression model. Suppose that both the number of 0s and the number of 1s is large compared to the number of predictors k, that the ESP takes on many values and that the binary LR model is a good approximation to the data. Then $Y|ESP \approx$ Binomial(1, $\hat{\rho}(ESP)$. For example if the ESP

$= 0$ then $Y|ESP \approx$ Binomial(1,0.5). If $-5 < ESP < 5$ then the estimated mean function has the characteristic "ESS" shape of the logistic curve.

Combining the ESS 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 small, 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 small, then the plotted points in the OD plot will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line.

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

Suppose the bulk of the plotted points in the OD plot fall in a wedge. Then the identity line, slope 4 line and OLS line will be added to the plot as visual aids. It is easier to use the OD plot to check the variance function than the ESS 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 $\hat{V}_{mod} = \hat{V}(Y_i|ESP) \approx m_i\rho(ESP)(1 - \rho(ESP))$ 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 $\approx$

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

The first example is for binary data. For binary data, G^2 is not approximately χ^2 and some plots of residuals have a pattern whether the model is

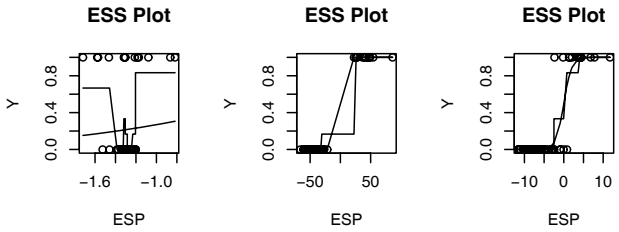

Figure 13.8: Plots for Museum Data

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 ESS plot is very useful if the logistic curve and step function of observed proportions are added as visual aids. The logistic curve gives the estimated LR probability of success. For example, when ESP = 0, the estimated probability is 0.5.

Example 13.1. 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 left plot in Figure 13.8 uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The middle plot uses the predictor *head height* and perfectly classifies the data since the ape skulls can be separated from the human skulls with a vertical line at ESP = 0. Christmann and Rousseeuw (2001) also used the ESS plot to visualize overlap. The right plot 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.

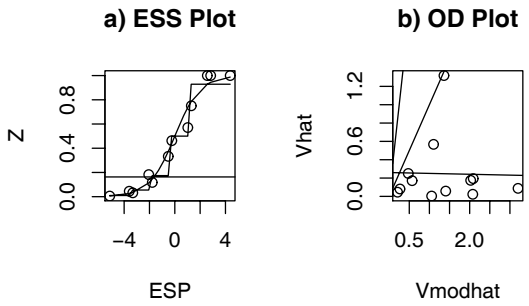

Figure 13.9: Visualizing the Death Penalty Data

Example 13.2. Abraham and Ledolter (2006, p. 360-364) describe death penalty sentencing in Georgia. The predictors are *aggravation level* from 1 to 6 (treated as a continuous variable) and *race of victim* coded as 1 for white and 0 for black. There were 362 jury decisions and 12 level race combinations. The response variable was the number of death sentences in each combination. The ESS plot in Figure 13.9a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve). The step function based on 5 slices also tracks the logistic curve well. The OD plot is shown in Figure 13.9b with the identity, slope 4 and OLS lines added as visual aids. The vertical scale is less than the horizontal scale and there is no evidence of overdispersion.

Example 13.3. Collett (1999, p. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. Figure 13.10a shows the ESS plot. 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

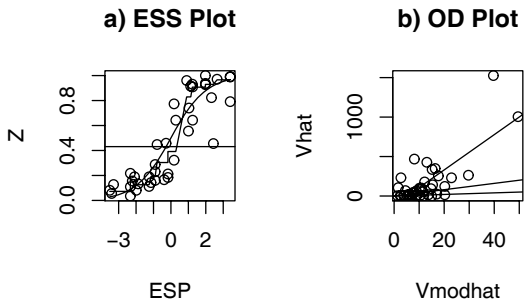

Figure 13.10: Plots for Rotifer Data

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

13.4 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 13.6. The **Poisson regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i})).$$

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

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{13.8}$$

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

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

for $y = 0, 1, \ldots,$ where $\mu > 0$. Hence this family is a 1-parameter exponential family with $\theta = \mu = E(Y)$, and the canonical link is the log link

$$c(\mu) = \log(\mu).$$

Since $g(\mu(\boldsymbol{x})) = c(\mu(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, the inverse link satisfies

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

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

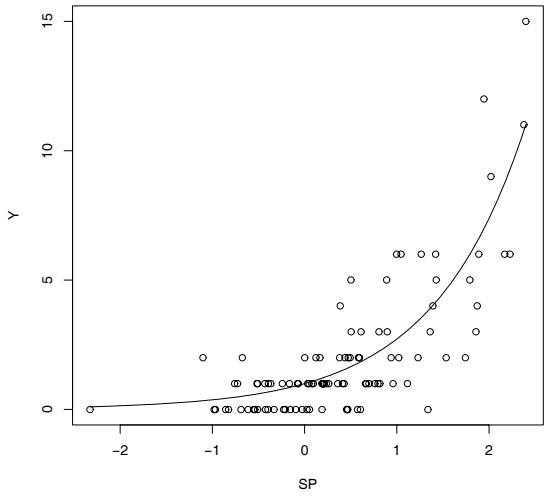

Figure 13.11: SSP for Loglinear Regression

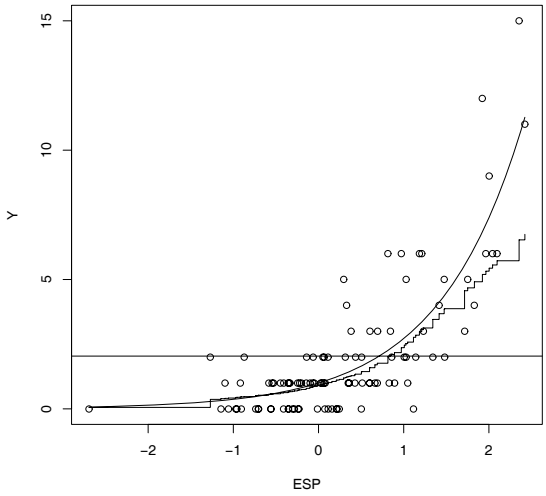

Figure 13.12: Response Plot for Loglinear Regression

A sufficient summary plot of the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the loglinear regression (LLR) model. Artificial data needs to be used because the plot can not be used for real data since α and $\boldsymbol{\beta}$ are unknown. The data used in the discussion below had $n = 100$, $\boldsymbol{x} \sim N_5(\mathbf{1}, \boldsymbol{I}/4)$ and

$$Y_i \sim \text{Poisson}(\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}))$$

where $\alpha = -2.5$ and $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$.

Model (13.8) can be written compactly as $Y|SP \sim$ Poisson(exp(SP)). Notice that $Y|SP = 0 \sim$ Poisson(1). Also note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$. The shape of the mean function $\mu(SP) = \exp(SP)$ for loglinear regression depends strongly on the range of the SP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence the range of the SP is narrow, then the exponential function will be rather flat. If the range of the SP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot. Figure 13.11 shows the SSP for the artificial data.

The estimated sufficient summary plot (ESSP or response plot or EY plot) is a plot of the $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function

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

added as a visual aid. The interpretation of the EY plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function called a "scatterplot smoother." The lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 13.12. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR model may fit the data well. **A useful lack of fit plot** is a plot of the ESP versus the *deviance residuals* that are often available from the software.

The deviance test described in Section 13.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LLR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the LLR estimator

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

If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_o should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. Figure 13.13 shows the ESSP when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 13.13 where the p–value for the deviance test is very small. Then the LLR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

Warning: For many count data sets where the LLR mean function is correct, the LLR model is not appropriate but the LLR MLE is still a consistent estimator of $\boldsymbol{\beta}$. The problem is that for many data sets where

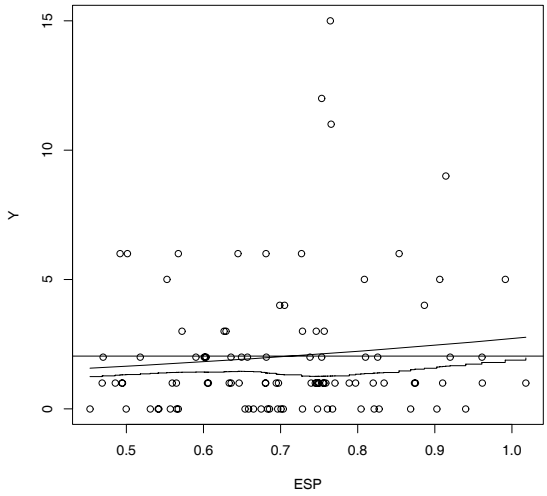

Figure 13.13: Response Plot when Y is Independent of the Predictors

$E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, 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 EY plot can be useful. See Cook and Weisberg (1999a, p. 401-403). Alternatively, if the EY plot looks good and $G^2/(n-k-1) \approx 1$, then the LLR model is likely useful. If $G^2/(n-k-1) > 1 + 3/\sqrt{n-k-1}$, then a more complicated count model may be needed. Here the deviance G^2 is described in Section 13.5.

A useful alternative to the LLR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of Y is

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

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\mathrm{V}(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ, ρ) distribution with $\rho = \kappa/(\mu+\kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

Definition 13.7. The **negative binomial regression (NBR) model** states that $Y_1, ..., Y_n$ are independent random variables where $Y_i \sim \mathrm{NB}(\mu(\boldsymbol{x}_i), \kappa)$

with $\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. Hence $Y|SP \sim \text{NB}(\exp(\text{SP}), \kappa)$, $E(Y|SP) = \exp(SP)$ and

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

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

Since the Poisson regression model is simpler than the NBR model, graphical diagnostics for the goodness of fit of the LLR model would be useful. To check for overdispersion, we suggest using the OD plot of $\exp(SP)$ versus $\hat{V} = [Y - \exp(SP)]^2$ Combining the EY plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model.

To motivate the OD plot, recall that if a count Y is not too small, then a normal approximation is good for both the Poisson and negative binomial distributions. 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, 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.

It is easier to use the OD plot to check the variance function than the EY 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.

Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the LLR model is appropriate. 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%. (A percentage greater than $5\% + 43\%/\sqrt{n}$ would be unusual.)

Judging the mean function from the EY plot may be rather difficult for large counts since the mean function is curved and lowess does not track the exponential function very well for large counts. Simple diagnostic plots for the Poisson regression model can be made using weighted least squares

(WLS). To see this, assume that all n of the counts Y_i are large. Then

$$\log(\mu(\boldsymbol{x}_i)) = \log(\mu(\boldsymbol{x}_i)) + \log(Y_i) - \log(Y_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i,$$

or

$$\log(Y_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$$

where

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

The error e_i does not have zero mean or constant variance, but if $\mu(\boldsymbol{x}_i)$ is large

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

by the central limit theorem. Recall that $\log(1+x) \approx x$ for $|x| < 0.1$. Then, heuristically,

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

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

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

Definition 13.8. The **minimum chi–square estimator** of the parameters $(\alpha, \boldsymbol{\beta})$ in a loglinear regression model are $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$, and are found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$.

The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞ while the loglinear regression maximum likelihood estimator tends to be consistent if the sample size $n \to \infty$. See

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

Definition 13.9. For a loglinear Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i} ESP = \sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus the WMLR residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$.

If the loglinear regression model is appropriate and if the minimum chi–square estimators are reasonable, then the plotted points in the weighted fit response plot should follow the identity line. Cases with large WMLR residuals may not be fit very well by the model. When the counts Y_i are small, the WMLR residuals can not be expected to be approximately normal. Notice that a resistant estimator for $(\alpha, \boldsymbol{\beta})$ can be obtained by replacing OLS (in Definition 13.9) with a resistant MLR estimator.

Example 13.4. For the Ceriodaphnia data of Myers, Montgomery and Vining (2002, p. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$ and seven concentrations of jet fuel (x_1) and an indicator for two strains of organism (x_2) were used as predictors. The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 13.14 shows the 4 plots for this data. In the EY plot of Figure 13.14a, the lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The OD plot in Figure 13.14b suggests that there is little evidence of overdispersion. These two plots as well as Figures 13.14c and 13.14d suggest that the LLR Poisson regression model is a useful approximation to the data.

Example 13.5. For the crab data, the response Y is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the color, spine condition, caparice width and weight of the female crab. Agresti (2002, p. 126-131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 13.15a suggests that there is one case with an unusually large value of the ESP. The lowess curve does not track the exponential curve all that well. Figure 13.15b suggests

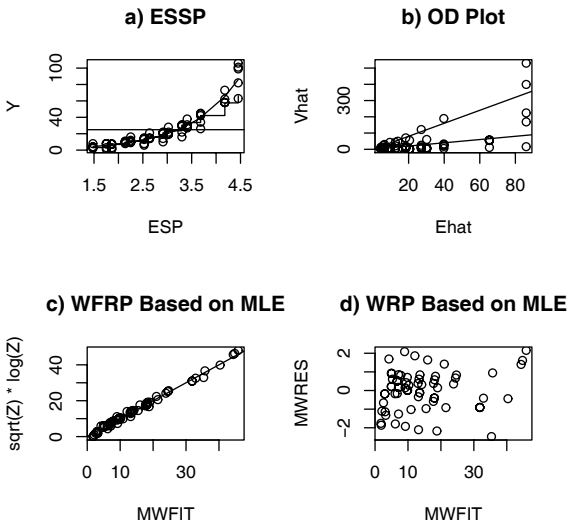

Figure 13.14: Plots for Ceriodaphnia Data

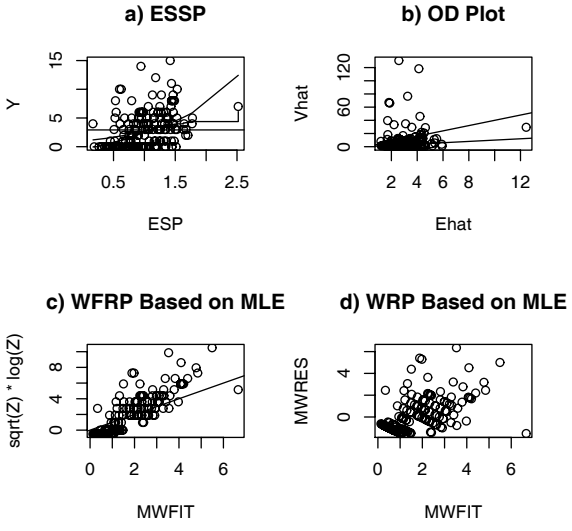

Figure 13.15: Plots for Crab Data

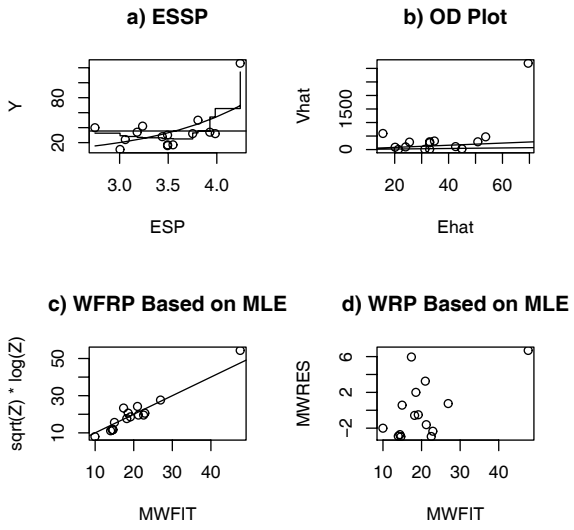

Figure 13.16: Plots for Popcorn Data

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 13.15c also suggests that the Poisson regression mean function is a rather poor fit since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \overline{Y}$, an alternative model to the NBR model may fit the data better. In later chapters, Agresti uses binomial regression models for this data.

Example 13.6. For the popcorn data of Myers, Montgomery and Vining (2002, p. 154), the response variable Y is the number of inedible popcorn kernels. The sample size was $n = 15$ and the predictor variables were temperature (coded as 5, 6 or 7), amount of oil (coded as 2, 3 or 4) and popping time (75, 90 or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier. Ignoring the outlier in Figure 13.16a suggests that the line $Y = \overline{Y}$ will fit the data and lowess curve better than the exponential curve. Hence Y seems to be independent of the predictors. Notice that the outlier sticks out in Figure 13.16b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected, then the Poisson regression model would suggest that tem-

perature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated.

13.5 Inference

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

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

Response = Y
Coefficient Estimates

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1 / se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

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

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

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

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

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{13.9}$$

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

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{13.10}$$

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

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k - 1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.

ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

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

The Wald confidence interval (CI) for β_j can also be obtained from 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. Recall that 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 k 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 $k+1$ parameters $\alpha, \beta_1, ..., \beta_k$ 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 $l_{FULL}(\alpha, \boldsymbol{\beta})$ 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

$$L_{FULL} = \log\ l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k + 1$ is the number of parameters for the full model.

The saturated model for logistic regression states that $Y_1, ..., Y_n$ 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 loglinear regression states that $Y_1, ..., Y_n$ 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 loglinear regression model is ok if $G^2 \leq n - k - 1$ (or if $G^2 \leq n - k - 1 + 3\sqrt{n - k - 1}$). For binary LR, the χ^2_{n-k+1} approximation for G^2 is rarely good even for large sample sizes n. For LR, the ESS plot is often a much better diagnostic for goodness of fit, especially when $ESP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ takes on many values and when $k + 1 << n$. For LLR, both the EY plot and $G^2 \leq n - k - 1 + 3\sqrt{n - k - 1}$ should be checked.

The *Arc* output on the following page, shown in symbols and for a real data set, is used for the deviance test described below. Assume that the estimated sufficient summary plot has been made and that the logistic or loglinear 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 $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then for loglinear regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression

$$\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$$

should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** is

i) $H_o : \beta_1 = \cdots = \beta_k = 0 \quad H_A : \ not \ \ H_o$

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

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

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_1, ..., X_k$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a GLM relationship between Y and the predictors $X_1, ..., X_k$.

Response = Y
Terms = $(X_1, ..., X_k)$
Sequential Analysis of Deviance

Predictor	df	Total Deviance	df	Change Deviance
Ones	$n - 1 = df_o$	G^2_o		
X_1	$n - 2$		1	
X_2	$n - 3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_k	$n - k - 1 = df_{FULL}$	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 output shown on the following page, both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced

Response = Y Terms = $(X_1, ..., X_k)$ (Full Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k/se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

Degrees of freedom: n - k - 1 = 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	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_k/se(\hat{\beta}_r)$	for Ho: $\beta_r = 0$

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

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

Degrees of freedom:          196
Deviance:                  0.009

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

Degrees of freedom:          198
Deviance:                 21.109
```

model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This test is a competitor of the Wald test. This change in deviance test is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

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

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $k - r$ 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 loglinear regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

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

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

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

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

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

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

Interpretation of coefficients: if $x_1, ..., x_{i-1}, x_{i+1}, ..., x_k$ 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

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

In logistic regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated odds of success by a factor of $\exp(\hat{\beta}_i)$.

13.6 Variable Selection

This section gives some rules of thumb for variable selection for logistic and loglinear regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x, sometimes x is not used by itself in the full model. Suppose that Y is binary. Then to decide what functions of x should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. The rules shown in Table 13.1 are used if x is an indicator variable or if x is a continuous variable. See Cook and Weisberg (1999a, p. 501) and Kay and Little (1987).

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

A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots and is used to examine the marginal relationships of the predictors and response. Place

Table 13.1: 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 $\log(1 - x)$

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

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

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{13.11}$$

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

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

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{13.12}$$

Definition 13.10. The model with $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model*. A model with $SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel*.

Suppose that S is a subset of I and that model (13.11) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{13.13}$$

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

Definition 13.11. An **EE plot** is a plot of $ESP(I)$ versus ESP.

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

Backward elimination starts with the full model with k nontrivial variables, and the predictor that optimizes some criterion is deleted. Then there are $k-1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $k-2, k-3, ..., 2$ and 1 predictors.

Forward selection starts with the model with 0 variables, and the predictor that optimizes some criterion is added. Then there is 1 variable in the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., k-2$ and $k-1$ predictors. Both forward selection and backward elimination result in a sequence, often different, of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{k-1}^*\}, \{x_1^*, x_2^*, ..., x_k^*\}$ = 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 |corr(ESP, OLS ESP)| ≥ 0.95. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ where the subset I has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 5.1, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg 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 11 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more predictors than submodel I_l, which has no more predictors than the minimum AIC model. It is possible that $I_l = I_{min} = I_{full}$. Then the submodel I is good if
i) the EY plot for the submodel looks like the EY plot for the full model.
ii) corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For LR want $r_I + 1 \leq \min(N_1, N_0)/10$. For LLR, want $r_I + 1 \leq n/10$.

vi) The plotted points in the VV plot cluster tightly about the identity line.
vii) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
viii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

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

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the smallest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

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

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

13.7 Complements

GLMs were introduced by Nelder and Wedderburn (1972). Books on generalized linear models (in roughly decreasing order of difficulty) include McCullagh and Nelder (1989), Fahrmeir and Tutz (2001), Myers, Montgomery and Vining (2002), Dobson and Barnett (2008) and Olive (2007d). Also see Hardin and Hilbe (2007), Hilbe (2007), Hoffman (2003), Hutcheson and Sofroniou (1999) and Lindsey (2000). Cook and Weisberg (1999, ch. 21-23) also has an excellent discussion. Texts on categorical data analysis that have useful discussions of GLMs include Agresti (2002), Le (1998), Lindsey (2004), Simonoff (2003) and Powers and Xie (2000) who give econometric applications. Collett (1999) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression. See Christensen (1997) for a Bayesian approach and see Cramer (2003) for econometric applications. Cameron and Trivedi (1998) and Winkelmann (2008) cover Poisson regression.

Barndorff-Nielsen (1982) is a very readable discussion of exponential families. Also see Olive (2007e, 2008ab). Many of the distributions in Chapter 3 belong to a 1-parameter exponential family.

The EY and ESS plots are a special case of model checking plots. See Cook and Weisberg (1997, 1999a, p. 397, 514, and 541). Cook and Weisberg (1999, p. 515) add a lowess curve to the ESS plot.

The ESS plot is essential for understanding the logistic regression model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ takes on many values. Some other diagnostics include Cook (1996), Eno and Terrell (1999), Hosmer and Lemeshow (1980), Landwehr, Pregibon and Shoemaker (1984), Menard (2000), Pardoe and Cook (2002), Pregibon (1981), Simonoff (1998), Su and Wei (1991), Tang (2001) and Tsiatis (1980). Hosmer and Lemeshow (2000) has additional references. Also see Cheng and Wu (1994), Kauermann and Tutz (2001) and Pierce and Schafer (1986).

The EY plot is essential for understanding the Poisson regression model and for checking goodness and lack of fit if the estimated sufficient predictor

$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ takes on many values. Goodness of fit is also discussed by Spinelli, Lockart and Stephens (2002).

Olive (2007bc) discusses plots for Binomial and Poisson regression. The ESS plot can also be used to measure overlap in logistic regression. See Christmann and Rousseeuw (2001) and Rousseeuw and Christmann (2003).

For Binomial regression and BBR, and for Poisson regression and NBR, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Breslow (1990), Cameron and Trevedi (1998), Collett (1999, ch. 6), Dean (1992), Ganio and Schafer (1992), Lambert and Roeder (1995) and Winkelmann (2000).

Olive and Hawkins (2005) give a simple all subsets variable selection procedure that can be applied to logistic regression and Poisson regression using readily available OLS software. The procedures of Lawless and Singhai (1978) and Nordberg (1982) are much more complicated.

Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004), Cook and Weisberg (1999a) and Hastie (1987).

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

Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$.

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

Some robust and resistant methods include Cantoni and Ronchetti (2001), Christmann (1994), Morgenthaler (1992), Pregibon (1982),

13.8 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

```
Output for problem 13.1: 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
```

13.1. 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 13.2
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
```

13.2*. Now the data is as in Problem 13.1, 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 13.3
Response      = ape
Terms         = (lower jaw, upper jaw, face length)
Trials        = Ones
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones           59   62.7188      |
lower jaw      58   51.9017      |     1   10.8171
upper jaw      57   17.1855      |     1   34.7163
face length    56   13.5325      |     1   3.65299
```

13.3*. A museum has 60 skulls of apes and humans. Lengths of the lower jaw, upper jaw and face are the explanatory variables. The response variable is *ape* (= 1 if ape, 0 if human). Using the output above, perform the four step deviance test for whether there is a LR relationship between the response variable and the predictors.

```
Output for Problem 13.4.
Full Model
Response      = ape
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant     11.5092        5.46270        2.107     0.0351
lower jaw   -0.360127       0.132925      -2.709     0.0067
upper jaw    0.779162       0.382219       2.039     0.0415
face length -0.374648       0.238406      -1.571     0.1161

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

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

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

13.4*. Suppose the full model is as in Problem 13.3, but the reduced model omits the predictor *face length*. Perform the 4 step change in deviance test to examine whether the reduced model can be used.

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

```
Output for Problem 13.5
Data set = Possums, Response      = possums
Terms         = (Habitat Stags)
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -0.652653        0.195148       -3.344     0.0008
Habitat    0.114756        0.0303273       3.784     0.0002
Stags      0.0327213       0.00935883      3.496     0.0005

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

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

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.

```
Output for Problem 13.6
Response      = possums Terms         = (Habitat Stags)
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones          150   187.490      |
Habitat       149   149.861      |     1   37.6289
Stags         148   138.685      |     1   11.1759
```

13.6*. Perform the 4 step deviance test for the same model as in Problem 13.5 using the output above.

```
Output for Problem 13.7
Terms          = (Acacia Bark Habitat Shrubs Stags Stumps)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -1.04276         0.247944       -4.206     0.0000
Acacia     0.0165563       0.0102718       1.612     0.1070
Bark       0.0361153       0.0140043       2.579     0.0099
Habitat    0.0761735       0.0374931       2.032     0.0422
Shrubs     0.0145090       0.0205302       0.707     0.4797
Stags      0.0325441       0.0102957       3.161     0.0016
Stumps    -0.390753        0.286565       -1.364     0.1727
Number of cases:             151
Degrees of freedom:          144
Deviance:                 127.506
```

13.7*. Let the reduced model be as in Problem 13.5 and use the output for the full model be shown above. Perform a 4 step change in deviance test.

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

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

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

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

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

Arc Problems

The following four problems use data sets from Cook and Weisberg (1999a).

13.9. Activate the *banknote.lsp* dataset with the menu commands "File > Load > Data > Arcg > banknote.lsp." Scroll up the screen to read the data description. Twice you will fit logistic regression models and include the coefficients in *Word*. Print out this output when you are done and include the output with your homework.

From *Graph&Fit* select *Fit binomial response.* Select *Top* as the predictor, *Status* as the response and *ones* as the number of trials.

a) Include the output in *Word*.

b) Predict $\hat{\rho}(x)$ if $x = 10.7$.

c) Find a 95% CI for β.

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

e) From *Graph&Fit* select *Fit binomial response.* Select *Top* and *Diagonal* as predictors, *Status* as the response and *ones* as the number of trials. Include the output in *Word*.

f) Predict $\hat{\rho}(\boldsymbol{x})$ if $x_1 =$ Top $= 10.7$ and $x_2 =$ Diagonal $= 140.5$.

g) Find a 95% CI for β_1.

h) Find a 95% CI for β_2.

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

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

13.10*. Activate *banknote.lsp* in *Arc.* with the menu commands "File > Load > Data > Arcg > banknote.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit binomial response.* Select *Top* and *Diagonal* as predictors, *Status* as the response and *ones* as the number of trials.

a) Include the output in *Word*.

b) From *Graph&Fit* select *Fit linear LS.* Select *Diagonal* and *Top* for predictors, and *Status* for the response. From *Graph&Fit* select *Plot of* and select *L2:Fit-Values* for *H*, *B1:Eta'U* for *V*, and *Status* for *Mark by.* Include the plot in *Word.* Is the plot linear? How are $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$ and $\hat{\alpha}_{logistic} + \hat{\boldsymbol{\beta}}_{logistic}^T \boldsymbol{x}$ related (approximately)?

13.11*. Activate *possums.lsp* in *Arc* with the menu commands "File > Load > Data > Arcg > possums.lsp." Scroll up the screen to read the data description.

a) From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *Acacia*, *bark*, *habitat*, *shrubs*, *stags* and *stumps* as the predictors. Include the output in *Word*. This is your full model.

b) EY plot: From *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the EY plot in *Word*.

c) From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *bark*, *habitat*, *stags* and *stumps* as the predictors. Include the output in *Word*.

d) EY plot: From *Graph&Fit* select *Plot of.* Select *P2:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the EY plot in *Word*.

e) Deviance test. From the *P2* menu, select *Examine submodels* and click on OK. Include the output in *Word* and perform the 4 step deviance test.

f) Perform the 4 step change of deviance test.

g) EE plot. From *Graph&Fit* select *Plot of.* Select *P2:Eta'U* for the H box and *P1:Eta'U* for the V box. Move the OLS slider bar to 1. Click on the *Options* popup menu and type "y=x". Include the plot in *Word*. Is the plot linear?

13.12*. In this problem you will find a good submodel for the *possums* data.

Activate *possums.lsp* in *Arc* with the menu commands
"File > Load > Data > Arcg> possums.lsp." Scroll up the screen to read the data description.

From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *Acacia*, *bark*, *habitat*, *shrubs*, *stags* and *stumps* as the predictors.

In Problem 13.11, you showed that this was a good full model.

a) Using what you have learned in class find a good submodel and include the relevant output in *Word*.

(Hints: Use forward selection and backward elimination and find a model that discards a lot of predictors but still has a deviance close to that of the full model. Also look at the model with the smallest AIC. Either of these models could be your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. You may have several models, say P2, P3, P4 and P5 to look at. Make a scatterplot matrix of the Pi:ETA'U from these models and from the full model P1. Make the EE and EY plots for each model. The correlation in the EE plot should be at least 0.9 and preferably greater than 0.95. As a very rough guide for Poisson regression, the number of predictors in the full model should be less than $n/5$ and the number of predictors in the final submodel should be less than $n/10$.)

b) Make an EY plot for your final submodel, say P2. From *Graph&Fit* select *Plot of.* Select *P2:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the EY plot in *Word*.

c) Suppose that P1 contains your full model and P2 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the V box and *P2:Eta'U*, for the H box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Also move the OLS slider bar to 1. Include the plot in *Word*.

d) Using a), b), c) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

Warning: The following problems use data from the book's webpage. Save the data files on a disk. Get in Arc and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:)*. Then click twice on the data set name.

13.13*. (ESS Plot): Activate *cbrain.lsp* in *Arc* with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit binomial response.* Select *brnweight*, *cephalic*, *breadth, cause*, *size,* and *headht* as predictors, *sex* as the response and *ones* as the number of trials. Perform the logistic regression and from *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) very well?

13.14*. Suppose that you are given a data set, told the response, and asked to build a logistic regression model with no further help. In this problem, we use the *cbrain* data to illustrate the process.

a) Activate *cbrain.lsp* in *Arc* with the menu commands
"File > Load > 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Scatterplot-matrix of.* Place *sex* in the *Mark by* box. Then select *age*, *breadth, cause*, *cephalic*, *circum*, *headht*, *height*, *length*, *size,* and *sex*. Include the scatterplot matrix in *Word*.

b) Use the menu commands "cbrain>Make factors" and select *cause*.

This makes *cause* into a factor with 2 degrees of freedom. Use the menu commands "cbrain>Transform" and select *age* and the log transformation.

Why was the log transformation chosen?

c) From *Graph&Fit* select *Plot of* and select *size* in **H**. Also place *sex* in the **Mark by** box. A plot will come up. From the *GaussKerDen* menu (the triangle to the left) select *Fit by marks*, move the sliderbar to 0.9, and include the plot in *Word*.

d) Use the menu commands "cbrain>Transform" and select *size* and the log transformation. From *Graph&Fit* select *Fit binomial response.* Select *age*, *log(age)*, *breadth*, *{F}cause*, *cephalic*, *circum*, *headht*, *height*, *length*, *size* and *log(size)* as predictors, *sex* as the response and *ones* as the number of trials. This is the full model *B1*. Perform the logistic regression and include the relevant output for testing in *Word*.

e) From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

f) From *B1* select *Examine submodels* and select *Add to base model (Forward Selection).* Include the output with the header "Base terms: ..." and from "Add: length 259" to "Add: {F}cause 258" in *Word*.

g) From *B1* select *Examine submodels* and select *Delete from full model (Backward Elimination).* Include the output with df corresponding to the minimum AIC model in *Word*. What predictors does this model use?

h) As a final submodel *B2*, use the model from f): from *Graph&Fit* select *Fit binomial response.* Select *age*, *log(age)*, *circum*, *height*, *length*, *size* and *log(size)* as predictors, *sex* as the response and *ones* as the number of trials. Perform the logistic regression and include the relevant output for testing in *Word*.

i) Put the EE plot H B2:ETA'U versus V B1:ETA'U in *Word*. Is the plot linear?

j) From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B2:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1.

From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

k) Perform the 4 step change in deviance test using the full model in d) and the reduced submodel in h).

Now act as if the final submodel is the full model.

l) From *B2* select *Examine submodels* click OK and include the output in *Word*. Then use the output to perform a 4 step deviance test on the submodel.

m) From *Graph&Fit* select *Inverse regression.* Select *age*, *log(age)*, *circum*, *height*, *length*, *size,* and *log(size)* as predictors, and *sex* as the response. From *Graph&Fit* select *Plot of.* Place *I3.SIR.p1* on the H axis and *B2.Eta'U* on the V axis. Include the plot in *Word*. Is the plot linear?

13.15*. In this problem you will find a good submodel for the *ICU* data obtained from STATLIB.

Activate *ICU.lsp* in *Arc* with the menu commands
"File > Load > 1/2 Floppy(A:) > ICU.lsp." Scroll up the screen to read the data description.

Use the menu commands "ICU>Make factors" and select *loc* and *race*.

a) From *Graph&Fit* select *Fit binomial response.* Select *STA* as the response and *ones* as the number of trials. The full model will use every predictor except ID, LOC and RACE (the latter 2 are replaced by their factors): select *AGE*, *Bic*, *CAN*, *CPR*, *CRE*, *CRN*, *FRA*, *HRA*, *INF*, $\{F\}LOC$, *PCO*, *PH*, *PO2* , *PRE* , $\{F\}RACE$, *SER*, *SEX*, *SYS* and *TYP* as predictors. Perform the logistic regression and include the relevant output for testing in *Word*.

b) Make the ESS plot for the full model: from *Graph&Fit* select *Plot of.* Place *STA* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Is the full model good?

c) Using what you have learned in class find a good submodel and include

the relevant output in *Word*.

[Hints: Use forward selection and backward elimination and find a model that discards a lot of predictors but still has a deviance close to that of the full model. Also look at the model with the smallest AIC. Either of these models could be your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. WARNING: do not delete part of a factor. Either keep all 2 factor dummy variables or delete all I-1=2 factor dummy variables. You may have several models, say B2, B3, B4 and B5 to look at. Make the EE and ESS plots for each model. WARNING: if a factor is in the full model but not the reduced model, then the EE plot may have I = 3 lines. See part f) below.]

d) Make an ESS plot for your final submodel.

e) Suppose that B1 contains your full model and B5 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of*. Select *B1:Eta'U* for the V box and *B5:Eta'U*, for the H box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Include the plot in *Word*.

If the EE plot is good and there are one or more factors in the full model that are not in the final submodel, then the bulk of the data will cluster tightly about the identity line, but some points may be far away from the identity line (often lying on some other line) due to the deleted factors.

f) Using c), d), e) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

13.16. In this problem you will examine the *museum* skull data.

Activate *museum.lsp* in *Arc* with the menu commands
"File > Load > 3 1/2 Floppy(A:) > museum.lsp." Scroll up the screen to read the data description.

a) From *Graph&Fit* select *Fit binomial response*. Select *ape* as the response and *ones* as the number of trials. Select *x5* as the predictor. Perform the logistic regression and include the relevant output for testing in *Word*.

b) Make the ESS plot and place it in *Word* (the response variable is *ape*

not y). Is the LR model good?

Now you will examine logistic regression when there is perfect classification of the sample response variables. Assume that the model used in c)–g) is in menu *B2*.

c) From *Graph&Fit* select *Fit binomial response.* Select *ape* as the response and *ones* as the number of trials. Select *x3* as the predictor. Perform the logistic regression and include the relevant output for testing in *Word*.

d) Make the ESS plot and place it in *Word* (the response variable is *ape* not y). Is the LR model good?

e) Perform the Wald test for $Ho : \beta = 0$.

f) From *B2* select *Examine submodels* and include the output in *Word*. Then use the output to perform a 4 step deviance test on the submodel used in part c).

g) The tests in e) and f) are both testing $Ho : \beta = 0$ but give different results. Why are the results different and which test is correct?

13.17. In this problem you will find a good submodel for the *credit* data from Fahrmeir and Tutz (2001).

Activate *credit.lsp* in *Arc* with the menu commands
"File > Load > Floppy(A:) > credit.lsp." Scroll up the screen to read the data description. This is a big data set and computations may take several minutes.

Use the menu commands "credit>Make factors" and select x_1, x_3, x_4, x_6, $x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{14}, x_{15}, x_{16}$, and x_{17}. Then click on *OK*.

a) From *Graph&Fit* select *Fit binomial response.* Select y as the response and *ones* as the number of trials. Select {F}x_1, x_2, {F}x_3, {F}x_4, x_5, {F}x_6, {F}x_7, {F}x_8, {F}x_9, {F}x_{10}, {F}x_{11}, {F}x_{12}, x_{13}, {F}x_{14}, {F}x_{15}, {F}x_{16}, {F}x_{17}, x_{18}, x_{19} and x_{20} as predictors. Perform the logistic regression and include the relevant output for testing in *Word*. You should get 1000 cases, df = 945, and a deviance of 892.957

b) Make the ESS plot for the full model: from *Graph&Fit* select *Plot of.* Place y on *V* and *B1:Eta'U* on *H*. From the *OLS* popup menu, select

Logistic and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Is the full model good?

c) Using what you have learned in class find a good submodel and include the relevant output in *Word*.

[Hints: Use forward selection and backward elimination and find a model that discards a lot of predictors but still has a deviance close to that of the full model. Also look at the model with the smallest AIC. Either of these models could be your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. WARNING: do not delete part of a factor. Either keep all 2 factor dummy variables or delete all I-1=2 factor dummy variables. You may have several models, say B2, B3, B4 and B5 to look at. Make the EE and ESS plots for each model. WARNING: if a factor is in the full model but not the reduced model, then the EE plot may have I = 3 lines. See part f) below.]

d) Make an ESS plot for your final submodel.

e) Suppose that B1 contains your full model and B5 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of.* Select *B1:Eta'U* for the V box and *B5:Eta'U*, for the H box. Place y in the *Mark by* box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Also move the OLS slider bar to 1. Include the plot in *Word*.

f) Using c), d), e) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

13.18*. a) This problem uses a data set from Myers, Montgomery and Vining (2002). Activate *popcorn.lsp* in *Arc* with the menu commands "File > Load > Floppy(A:) > popcorn.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit Poisson response.* Use *oil, temp* and *time* as the predictors and y as the response. From *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the

EY plot in *Word*.

b) From the *P1* menu select *Examine submodels*, click on *OK* and include the output in *Word.*

c) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

d) From the *popcorn* menu, select *Transform* and select y. Put 1/2 in the p box and click on *OK*. From the *popcorn* menu, select *Add a variate* and type *yt = sqrt(y)*log(y)* in the resulting window. Repeat three times adding the variates *oilt = sqrt(y)*oil*, *tempt = sqrt(y)*temp* and *timet = sqrt(y)*time*. From *Graph&Fit* select *Fit linear LS* and choose $y^{1/2}$, *oilt, tempt* and *timet* as the predictors, *yt* as the response and click on the *Fit intercept* box to remove the check. Then click on *OK*. From *Graph&Fit* select *Plot of.* Select *L2:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word.*

e) From *Graph&Fit* select *Plot of.* Select *L2:Fit-Values* for the H box and *L2:Residuals* for the V box. Include the weighted residual response plot in *Word.*

f) For the plot in e), highlight the case in the upper right corner of the plot by using the mouse to move the arrow just above and to the left the case. Then hold the rightmost mouse button down and move the mouse to the right and down. From the *Case deletions* menu select *Delete selection from data set*, then from *Graph&Fit* select *Fit Poisson response.* Use *oil, temp* and *time* as the predictors and y as the response. From *Graph&Fit* select *Plot of.* Select *P3:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the EY plot in *Word*.

g) From the *P3* menu select *Examine submodels*, click on *OK* and include the output in *Word.*

h) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

i) From *Graph&Fit* select *Fit linear LS*. Make sure that $y^{1/2}$, *oilt, tempt* and *timet* are the predictors, *yt* is the response, and that the *Fit intercept* box does not have a check. Then click on *OK* From *Graph&Fit* select *Plot*

of. Select *L4:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word.*

j) From *Graph&Fit* select *Plot of.* Select *L4:Fit-Values* for the H box and *L4:Residuals* for the V box. Include the weighted residual response plot in *Word.*

k) Is the deleted point influential? Explain briefly.

l) From *Graph&Fit* select *Plot of.* Select *P3:Eta'U* for the H box and *P3:Dev-Residuals* for the V box. Include the deviance residual response plot in *Word.*

m) Is the weighted residual plot from part j) a better lack of fit plot than the deviance residual plot from part m)? Explain briefly.

R/Splus problems

Download functions with the command *source("A:/rpack.txt").* **See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *lrdata*, will display the code for the function. Use the `args` command, eg *args(lrdata)*, to display the needed arguments for the function.

13.19.

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

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

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

13.20. Obtain the function `llrdata` from `rpack.txt`. Enter the commands

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

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

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

The following problem uses SAS and Arc.

13.21*. SAS–all subsets: On the webpage (http://www.math.siu.edu/olive/students.htm) there are 2 files *cbrain.txt* and *hw10d2.sas* that will be used for this problem. The first file contains the *cbrain* data (that you have analyzed in *Arc* several times) without the header that describes the data.

i) Using *Netscape* or *Internet Explorer*, go to the webpage and click on *cbrain.txt*. After the file opens, copy and paste the data into *Notepad*. (In *Netscape*, the commands "Edit>Select All" and "Edit>copy" worked.) Then open *Notepad* and enter the commands "Edit>paste" to make the data set appear.

ii) SAS needs an "end of file" marker to determine when the data ends. SAS uses a period as the end of file marker. Add a period on the line after the last line of data in *Notepad* and save the file as *cbrain.dat* on your disk using the commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *cbrain.dat*. In the *Save as type* box, click on the right of the box and select *All Files*. **Warning: make sure that the file has been saved as** *cbrain.dat*, **not as** *cbrain.dat.txt*.

iii) As described in i), go to the webpage and click on *hw10d2.sas*. After the file opens, copy and paste the SAS program for 13.21 into *Notepad*. Use the commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *hw13d21.sas*. In the *Save as type* box, click on the right of the box and select *All Files*, and the file will be saved on your disk. **Warning: make sure that the file has been saved as** *hw13d21.sas*, **not as** *hw13d21.sas.txt*.

iv) Get into SAS, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "3 1/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 3 1/2 Floppy(A:).) Double click on **hw13d21.sas**. (Alternatively cut and paste the program into the SAS editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. **Warning: if you do not have the two files on A drive, then you need to**

change the *infile* command in **hw13d21.sas** to the drive that you are using, eg change *infile "a:cbrain.dat";* to *infile "f:cbrain.dat";* if you are using F drive.

a) To copy and paste relevant output into *Word,* click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

Interesting models have $C(p) \leq 2k$ where $k =$ "number in model."

The only SAS output for this problem that should be included in Word are two header lines (Number in model, R-square, C(p), Variables in Model) and the first line with Number in Model = 6 and C(p) = 7.0947. You may want to copy all of the SAS output into *Notepad*, and then cut and paste the relevant two lines of output into *Word*.

b) Activate *cbrain.lsp* in *Arc* with the menu commands "File > Load > Data > mdata > cbrain.lsp." From *Graph&Fit* select *Fit binomial response.* Select *age* = X2, *breadth* = X6, *cephalic* = X10, *circum* = X9, *headht* = X4, *height* = X3, *length* = X5 and *size* = X7 as predictors, *sex* as the response and *ones* as the number of trials. This is the full logistic regression model. Include the relevant output in *Word*. (A better full model was used in Problem 13.14.)

c) ESS plot. From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

d) From *Graph&Fit* select *Fit binomial response.* Select *breadth* = X6, *cephalic* = X10, *circum* = X9, *headht* = X4, *height* = X3, and *size* = X7 as predictors, *sex* as the response and *ones* as the number of trials. This is the "best submodel." Include the relevant output in *Word*.

e) Put the EE plot H B2 ETA'U versus V B1 ETA'U in *Word*. Is the plot linear?

f) From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B2:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?