Chapter 10

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.

10.1 Binary Regression

Definition 10.1. The **binary regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(1, \rho(\boldsymbol{x}_{\text{i}})).$$

The **binary logistic regression (LR) model** is the special case of binary regression where

$$P(\text{success}|\boldsymbol{x}_{\text{i}}) = \rho(\boldsymbol{x}_{\text{i}}) = \frac{\exp(\alpha + \boldsymbol{\beta}^{\text{T}}\boldsymbol{x}_{\text{i}})}{1 + \exp(\alpha + \boldsymbol{\beta}^{\text{T}}\boldsymbol{x}_{\text{i}})}. \quad (10.1)$$

Definition 10.2. The **sufficient predictor** SP $= \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ while the **estimated sufficient predictor** ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.

Thus the binary regression model says that

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

where

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

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

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

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

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

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{10.2}$$

and

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

Using Definitions 10.1 and 10.3 makes simulation of logistic regression data straightforward. To use Definition 10.3, set $\pi_0 = \pi_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\mu}_0 = \mathbf{0}$. Then $\alpha = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. The artificial data set used to make Figure 1.6 had $\boldsymbol{\beta} = (1,1,1,0,0)^T$ 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 $n = N_1 + N_0 = 200$. 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.

To visualize the LR model, the response plot will be useful.

Definition 10.4. The *response plot* or *estimated sufficient summary plot* or *ESS plot* is the plot of the 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.

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

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

The response plot is a powerful method for assessing the adequacy of the binary LR regression model. Suppose that both the number of 0s and the number of 1s is large compared to the number of predictors 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, $\hat{\rho}(ESP)$. Unlike the response plot for multiple linear regression where the mean function is always the identity line, the mean function in the response plot for LR can take a variety of shapes depending on the range of the ESP. For LR, the (estimated) mean function is

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

If the ESP $= 0$ then $Y|SP \approx$ binomial(1,0.5). If the ESP $= -5$, then $Y|SP \approx$

binomial(1,$\rho \approx 0.007$) while if the ESP = 5, then $Y|SP \approx$ binomial(1,$\rho \approx 0.993$). Hence if the range of the ESP is in the interval $(-\infty, -5)$ then the mean function is flat and $\hat{\rho}(ESP) \approx 0$. If the range of the ESP is in the interval $(5, \infty)$ then the mean function is again flat but $\hat{\rho}(ESP) \approx 1$. If $-5 < ESP < 0$ then the mean function looks like a slide. If $-1 < ESP < 1$ then the mean function looks linear. If $0 < ESP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < ESP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 1.6.

This plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Ys in each slice and add the resulting step function to the ESS plot. This is done in Figure 1.6 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).

The deviance test described in Section 10.3 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 1.7 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 1.7 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

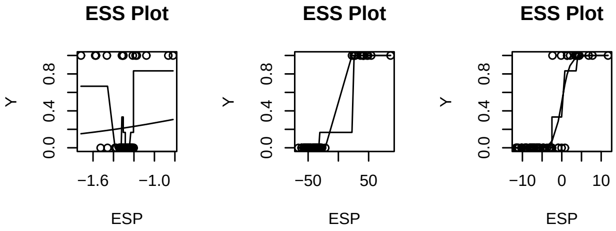

Figure 10.1: Plots for Museum Data

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.

The logistic regression (maximum likelihood) estimator also tends to perform well the discriminant function model above Definition 10.3. 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.”

Example 10.1. Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response variable *ape* is 1 for an ape skull. The left plot in Figure 10.1 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 $= 0$. 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.

Example 10.2. Is There a Gender Gap? In the United States, there does not appear to be a gender gap in math and science ability in that the average score and the percentage passing standardized tests appear to be about the same for both genders for math and science until after 8th grade. For example, in Illinois all students take standardized exams at various times, and the Nov. 16, 2001 *Chicago Tribune* reported that the percentage of Illinois students meeting or exceeding state standards for math was 61% for M and 62% for F 5th graders. For science it was 72% for both M and F 7th graders. After 8th grade, differences in gender scores are likely due to different gender choices (males take more math in high school) rather than to differences in ability. In recent years, the gap for high school juniors has greatly decreased in the United States, and may not have been statistically significant in 2008.

In many other countries, there does seem to be a difference in average gender scores. The TIMSS data is from Beaton, Martin, Mullis, Gonzales, Smith, and Kelly (1996). The variable Y was a 1 if there was a statistically significant gender difference in the nation's TIMSS test, and Y was 0 otherwise. Two predictors were $x_1 =$ percent of 8th graders whose friends think it is important to do well in science and $x_2 =$ percent of 8th graders taught by female teachers. The horizontal axis is the ESP $= 6.9668 - 0.05684 x_1 - 0.03609 x_2$.

Logistic regression was used to estimate the probability that Y $= 1$ given the values of the predictors. The estimated probability is given by the smooth curve in Figure 10.2. For example, in Japan 83% of the students thought that it was important to do well in the sciences and 20% of the 8th grade science teachers were female. Hence Japan had Y $= 1$, $x_1 = 83$ and $x_2 = 20$. This corresponds to ESP $= 1.527$ and an estimated probability of 0.8216. In contrast, the USA had Y $= 0$, $x_1 = 69$ and $x_2 = 54$. Then the ESP $= 1.096$ and an estimated probability of 0.7495. In general, draw a vertical line to

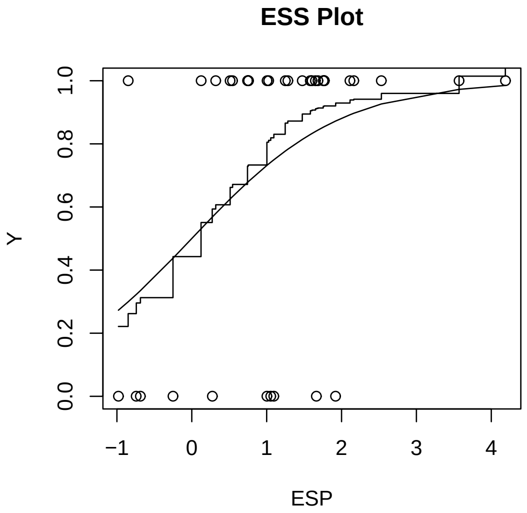

Figure 10.2: Visualizing TIMSS Data

the smooth curve and then a horizontal line to the vertical axis to estimate the probability.

The jagged curve is the scatterplot smoother lowess. Since it is close to the solid line, then the LR model is likely to be useful. Hence nations with low percentages of female science teachers and of motivated students were more likely to have a gender difference in the TIMSS science scores than nations with high percentages.

10.2 Binomial Regression

Definition 10.5. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with

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

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

$$P(\text{success}|\boldsymbol{x}_\mathrm{i}) = \rho(\boldsymbol{x}_\mathrm{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\mathrm{T}\boldsymbol{x}_\mathrm{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\mathrm{T}\boldsymbol{x}_\mathrm{i})}. \qquad (10.3)$$

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

$$Y_i | SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{10.4}$$

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

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

For binomial regression, the ESS plot needs to be modified and a check for overdispersion (described on the following page) is needed.

Definition 10.6. 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* or *response 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, $Y \sim \mathrm{BB(m,}\rho,\theta)$, 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 10.7. 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.

The following plot was suggested by Olive (2007b) to check for overdispersion.

Definition 10.8. To check for overdispersion, use the *OD plot* of the estimated model variance $\hat{V}_{mod} \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LR model, $\hat{V}(Y_i|SP) = m_i \rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i \rho(ESP_i)$.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the logistic regression model much as R^2 is used for multiple linear regression. When the 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.

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. 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

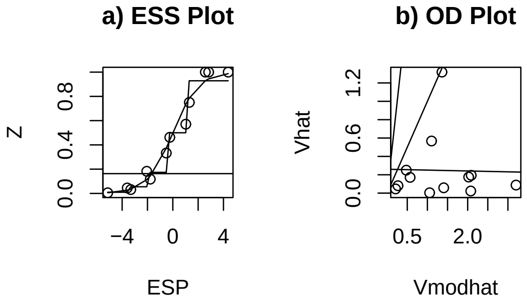

Figure 10.3: Visualizing the Death Penalty Data

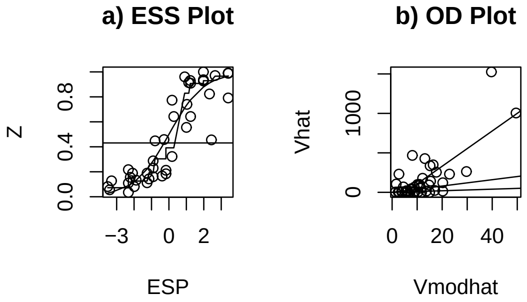

Figure 10.4: Plots for Rotifer Data

that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

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

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

Example 10.3. 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 10.3a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve). The step function based on 5 slices also tracks the logistic curve well. The OD plot is shown in Figure 10.3b with the identity, slope 4 and OLS lines added as visual aids. The vertical scale is less than the horizontal scale and there is no evidence of overdispersion.

Example 10.4. Collett (1999, p. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. Figure 10.4a 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 scale is about 30 times the horizontal scale. The OLS line has slope much larger than 4 and two outliers seem to be present.

10.3 Inference

This section gives a brief discussion of inference for the logistic regression (LR) model. Inference for this model is very similar to inference for the multiple linear regression, survival regression and Poisson regression models. For all 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 $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$:

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

To perform inference for LR, computer output is needed. The following page shows 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

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{10.5}$$

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.

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
```

If Ho is rejected, then conclude that X_j is needed in the LR 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 LR model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful LR 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 LR, 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.

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 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 $k+1 << n$.

The *Arc* output on the following page, shown in symbols and for a real data set, is used for the deviance test described after the output. Assume that the ESS plot has been made and that the logistic regression model fits the data well in that the nonparametric step function follows the estimated model mean function closely. The deviance test is used to test whether $\boldsymbol{\beta} = \boldsymbol{0}$. If this is the case, then the predictors are not needed in the LR model. If $H_o : \boldsymbol{\beta} = \boldsymbol{0}$ is not rejected, then for logistic regression

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

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

The 4 step **deviance test** is

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

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

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with k 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 LR 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 LR relationship between Y and the predictors $X_1, ..., X_k$.

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

		Total		Change
Predictor	df	Deviance	df	Deviance
Ones	$n - 1 = df_o$	G_o^2		
X_1	$n - 2$		1	
X_2	$n - 3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_k	$n - k - 1 = df_{FULL}$	G_{FULL}^2	1	

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

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 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 likelihood ratio test is a competitor of the Wald test. The likelihood ratio 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.

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
```

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$(m_i, \rho(\boldsymbol{x}_{Ri}))$.

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 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. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P}(\text{failure}|\boldsymbol{x})$ 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 $\exp(\hat{\beta}_i)$.

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

Logistic Regression Output for Reduced Model,
Response      = gender, Terms        = (height  size)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -6.26111         1.34466        -4.656     0.0000
height    -0.0536078       0.0239044      -2.243     0.0249
size       0.00282146      0.000507935     5.555     0.0000

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

Example 10.5. Let the response variable $Y = gender = 0$ for F and 1 for M. Let $x_1 = height$ (in inches) and $x_2 = size$ of head (in mm^3). Logistic regression is used, and data is from Gladstone (1905-6).

a) Predict $\hat{\rho}(\boldsymbol{x})$ if height $= x_1 = 65$ and size $= x_2 = 3500$.

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

Solution: a) $ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -6.26111 - 0.0536078(65) + 0.0028215(3500) = 0.1296$. So

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

b) i) Ho the reduced model is good H_A use the full model

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

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

iv) Reject Ho, use the full model.

Example 10.6. Suppose that Y is a 1 or 0 depending on whether the person is or is not credit worthy. Let $x1$ through $x6$ be the predictors and use the following output to perform a 4 step deviance test. The credit data is available from the text's website as file *credit.lsp*, and is from Fahrmeir and Tutz (1996).

```
Response      = y
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones          999   1221.73      |
x1            998   1177.11      |     1   44.6148
x2            997   1176.55      |     1   0.561629
x3            996   1168.33      |     1   8.21723
x4            995   1168.20      |     1   0.137583
x5            994   1163.44      |     1   4.75625
x6            993   1158.22      |     1   5.21846
```

Solution: i) Ho $\beta_1 = \cdots = \beta_6 \quad H_A$ not H0

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

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

iv) Reject Ho, there is a LR relationship between $Y =$ credit worthiness and the predictors $x_1, ..., x_6$.

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

Example 10.7. A museum has 60 skulls, some of which are human and some of which are from apes. Consider trying to estimate whether the *skull type* is human or ape from the *height of the lower jaw*. Use the above logistic regression output to answer the following problems. The museum data is available from the text's website as file *museum.lsp*, and is from Schaaffhausen (1878).

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

b) Find a 95% CI for β.

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

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

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

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

c) i) Ho $\beta = 0 \quad H_A \quad \beta \neq 0$

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

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

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

10.4 Variable Selection

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

The full model will often contain factors and interaction. 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

Table 10.1: Building the Full Logistic Regression Model

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

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

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 LR can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{10.6}$$

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

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{10.7}$$

Definition 10.9. 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.* The full model is always a submodel.

Suppose that S is a subset of I and that model (10.6) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \boldsymbol{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \qquad (10.8)$$

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, $\boldsymbol{\beta}_O = \boldsymbol{0}$ 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 $(\alpha, \boldsymbol{\beta})$ 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 10.10. 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_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. I_I is the initial submodel to examine.

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 $k - 2, k - 3, ..., 3$ and 2 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 $2, 3, ..., k-1$ and k predictors. Both forward selection and backward elimination result in a sequence of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{k-1}^*\}, \{x_1^*, x_2^*, ..., x_k^*\} =$ full model, and the two sequences need not be the same.

All subsets variable selection can be performed with the following procedure. Compute the LR ESP and the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that $|\text{corr(LR ESP, OLS ESP)}| \geq 0.95$. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ where the subset I has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 3.2, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and LR 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 10 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Then the submodel I is good if
i) the ESS plot for the submodel looks like the ESS plot for the full model.
ii) Want 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) Want $r_I + 1 \leq \min(N_1, N_0)/10$.
vi) 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).
vii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.

viii) Want hardly any predictors with p-values > 0.05.
ix) Want few predictors with p-values between 0.01 and 0.05.
x) 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, et cetera. 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 ESS 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.

Example 10.8. The following output is for forward selection and backward elimination. All models use a constant. For forward selection, the min AIC model uses {F}LOC, TYP, AGE, CAN, SYS, PCO, and PH. Model I_I uses {F}LOC, TYP, AGE, CAN, and SYS. Let model I use {F}LOC, TYP, AGE, and CAN. This model may be good, so for forward selection, models I_I and I are the first models to examine.

```
Forward Selection                                      comment

Base terms: ({F}LOC TYP)
          df   Deviance Pearson X2 |    k    AIC  > min AIC + 7
Add: AGE  195   141.873  187.84    |    5  151.873

Base terms: ({F}LOC TYP AGE)
          df   Deviance  Pearson X2 |    k    AIC < min AIC + 7
Add: CAN  194   134.595  170.367    |    6  146.595
                     ({F}LOC TYP AGE CAN) could be a good model

Base terms: ({F}LOC TYP AGE CAN)
         df  Deviance Pearson X2 |    k    AIC   < min AIC + 2
Add: SYS 193  128.441      179.753  | 7  142.441
                 ({F}LOC TYP AGE CAN SYS) could be a good model

Base terms: ({F}LOC TYP AGE CAN SYS)
         df   Deviance  Pearson X2 |    k    AIC  < min AIC + 2
Add: PCO 192   126.572  186.71     |    8  142.572
                      PCO not important since AIC < min AIC + 2

Base terms: ({F}LOC TYP AGE CAN SYS PCO)
         df   Deviance      Pearson X2 |    k    AIC
Add: PH  191   123.285      191.264    |    9  141.285 min AIC
                       PH not important since AIC < min AIC + 2
```

```
Backward Elimination

Current terms: (AGE CAN {F}LOC PCO PH PRE SYS TYP)
             df   Deviance  Pearson X2| k    AIC
Delete: PRE  191   123.285  191.264  | 9  141.285 min AIC model

Current terms: (AGE CAN {F}LOC PCO PH SYS TYP)
           df  Deviance Pearson X2 | k    AIC     < min AIC + 2
Delete: PH 192  126.572  186.71    |8  142.572  PH not important

Current terms: (AGE CAN {F}LOC PCO SYS TYP)
             df  Deviance Pearson X2 |k    AIC < min AIC + 2
Delete: PCO  193  128.441 179.753  | 7 142.441 PCO not important
                    (AGE CAN {F}LOC SYS TYP) could be good model

Current terms: (AGE CAN {F}LOC SYS TYP)
             df  Deviance Pearson X2| k    AIC < min AIC + 7
Delete: SYS  194  134.595 170.367  |6  146.595
                                   SYS may not be important
                    (AGE CAN {F}LOC TYP) could be good model

Current terms: (AGE CAN {F}LOC TYP)
             df   Deviance Pearson X2 | k    AIC > min AIC + 7
Delete: CAN  195   141.873 187.84     | 5  151.873   AIC
```

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

Example 10.9. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. One predictor was a factor, and a factor was considered to have a bad Wald p-value

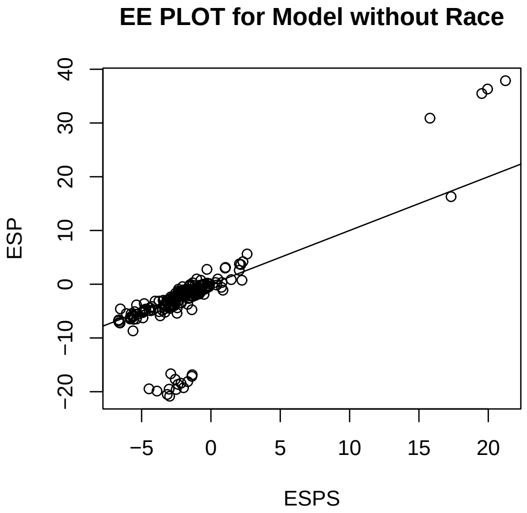

Figure 10.5: EE Plot Suggests Race is an Important Predictor

> 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 267 cases: for the response, 113 were 0's and 154 were 1's.

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

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

Example 10.10. The ICU data studies the survival of 200 patients following admission to an intensive care unit. The response variable was STA (0 = Lived, 1 = Died). The 19 predictors were primarily indicator variables describing the health of the patient at time of admission, but two factors had 3 levels including RACE (1 = White, 2 = Black, 3 = Other). The response plot showed that the full model using the 19 predictors was useful for predicting survival. Variable selection suggested a submodel using five predictors. The EE plot of the submodel ESP vs. full model ESP is shown in Figure 10.5. The plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. This clustering did not occur in Figure 10.5. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black. When RACE is added to the submodel, all of the points cluster about the identity line. Although variable selection did not suggest that RACE is important, the above results suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black.

10.5 Complements

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. Also see Allison (2001), Cox and Snell (1989), Hilbe (2009), Kleinbaum and Klein (2005a) and Pampel (2000).

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. The ESS plot and OD plot are examined in Olive (2009e). 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 ESS plot can also be used to measure overlap in logistic regression. See Rousseeuw and Christmann (2003).

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

Olive and Hawkins (2005) give the simple all subsets variable selection procedure that can be applied to logistic 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 (1999) and Hastie (1987).

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

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, partial F test, and OLS t tests are often asymptotically valid when the conditions in Definition 10.3 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 10.1 below. Assume that Cov($\boldsymbol{x}$) $\equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and that 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 10.1 holds as long as 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 10.1. Assume that Y is binary and that 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. We have that

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

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 estimator

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

Now when the conditions of Definition 10.3 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. Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$. So a rough approximation is LR ESP $\approx$ (OLS ESP)$/MSE$.

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 generalized linear model 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, $g(\rho(\boldsymbol{x})) = \log[-\log(1-\rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

10.6 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

```
Output for problem 10.1: Response = sex
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -18.3500         3.42582        -5.356     0.0000
circum     0.0345827       0.00633521      5.459     0.0000
```

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

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

b) Find a 95% CI for β.

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

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

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

a) Predict $\hat{\rho}(\boldsymbol{x})$ if circumference $= x_1 = 550.0$ 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 10.3
Response      = ape
Terms         = (lower jaw, upper jaw, face length)
Trials        = Ones
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones           59   62.7188      |
lower jaw      58   51.9017      |     1   10.8171
upper jaw      57   17.1855      |     1   34.7163
face length    56   13.5325      |     1   3.65299
```

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

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

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


Reduced Model
Response      = ape
Coefficient Estimates
```

```
Label       Estimate         Std. Error     Est/SE    p-value
Constant   8.71977          4.09466          2.130     0.0332
lower jaw -0.376256         0.115757        -3.250     0.0012
upper jaw  0.295507         0.0950855        3.108     0.0019

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

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

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

10.5*. 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.

```
Response     = pass   Terms         = (hscalc survey)
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant   0.875469        0.532291        1.645     0.1000
hscalc     10.3274         54.7562         0.189     0.8504
survey    -2.26176         1.23828        -1.827     0.0678
```

10.6. The response variable *pass* was a 1 if the Math 150 (intro calc) student got a C or higher on the combined final and a 0 (withdrew, D or F) otherwise. Data was collected at the beginning of the semester on 31 students who took a section of Math 150 in Fall, 2002. Here $x_1 = hscalc$ was coded as a 1 if the student said that their last math class was high school calculus and as a 0 otherwise. Here $x_2 = survey$ was coded as a 1 if the student failed to turn in the survey, 0 otherwise.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if $hscalc = x_1 = 1.0$ and $survey = x_2 = 0.0$.
b) Perform the 4 step Wald test for $Ho: \beta_1 = 0$.
c) Perform the 4 step Wald test for $Ho: \beta_2 = 0$.

Arc Problems

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

10.7. 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 $x_1 = \text{Top} = 10.7$ and $x_2 = \text{Diagonal} = 140.5$.

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 $Ho : \beta_2 = 0$.

10.8*. 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 $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}^T_{OLS}\boldsymbol{x}$ and $\hat{\alpha}_{logistic} + \hat{\boldsymbol{\beta}}^T_{logistic}\boldsymbol{x}$ related (approximately)?

10.9*. (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?

10.10*. 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.* Select *age*, *breadth, cephalic*, *circum*, *headht*, *height*, *length*, *size,* and *sex.* Also place *sex* in the *Mark by* box.

Include the scatterplot matrix in *Word.*

b) 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.* 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, cephalic*, *circum*, *headht*, *height*, *length*, *size, log(size)*, as predictors, *sex* as the response and *ones* as the number of trials. This is the full model. 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 (For-*

ward Selection). Include the output with df = 259 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, 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.

10.11*. In this problem you will find a good submodel for the *ICU* data obtained from STATLIB. Get the file *ICU.lsp* from the text's website.

a) 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.

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

c) 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.*

d) 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?

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

(Hints: Create a full model. The full model has a deviance at least as small as that of any submodel. Consider forward selection and backward elimination. For each method, find the submodel I_{min} with the smallest AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$, and find submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I that have $\Delta(I) \leq 7$. 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 corr$(ESP(I), ESP) \geq 0.95$. Submodel I_I is 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 $J-1$ factor dummy variables or delete all $J-1$ factor dummy variables. You may have several models, B2, B3, B4 and B5 to examine. Let B1 be the full model. Make the EE and ESS plots for each model. WARNING: if an important factor is in the full model but not the reduced model, then the plotted points in the EE plot may follow more than 1 line. See part g) below.)

f) Make an ESS plot for your final submodel.

g) 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 full model is good and the EE plot is good, then the plotted points should cluster tightly about the identity line. If the full model is good and an important factor is deleted, then the bulk of the data will cluster tightly about the identity line, but some points may cluster about different lines. If the deleted factor was important and had J levels, there could be clusters about J lines, but there could be clusters about as few as two lines if only two groups of levels differ. Such clustering in the EE plot suggests that the deleted factor is probably important.

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

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

a) 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.

b) 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.*

c) 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 d)–h) is in menu *B2.*

d) 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.*

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

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

g) 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 d).

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

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

a) 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.

b) 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.*

c) 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

d) 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?

e) Using what you have learned in class find a good submodel and include the relevant output in *Word.*
See the hints give below Problem 10.11e.

f) Make an ESS plot for your final submodel.

g) 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.*

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

R/Splus problems

Download functions with the command *source("A:/regpack.txt").* **See Preface or Section 17.1.** Typing the name of the `regpack` function, eg *binrplot*, will display the code for the function. Use the `args` command, eg *args(lressp)*, to display the needed arguments for the function.

10.14.

a) Obtain the function `lrdata` from `regpack.txt`. Enter the commands

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

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

The following problem uses SAS and Arc.

10.15*. SAS–all subsets: On the webpage (www.math.siu.edu/olive/students.htm) there are 2 files *cbrain.txt* and *hwbrain.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.

a) 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.

b) 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.*

c) As described in a), go to the webpage and click on *hwbrain.sas.* After the file opens, copy and paste the data 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 *hwbrain.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** *hwbrain.sas*, **not as** *hwbrain.sas.txt.*

d) 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 **hwbrain.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 **hwbrain.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.

e) 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".

The model should be good if $C(p) \leq 2k$ 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.*

f) 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 10.10.)

g) (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?

h) 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.*

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?

Binomial Regression in SAS

```
options ls = 70;
data crabs;
* Agresti, p. 272;
input width cases satell;
cards;
22.69  14  5
23.84  14  4
24.77  28  17
25.84  39  21
26.79  22  15
27.74  24  20
28.67  18  15
30.41  14  14
;
proc logistic; model satell/cases = width;
     output out = predict p = pi_hat;
     proc print data = predict
run;
```

10.16. a) Enter the above SAS program (or get the program from the webpage (www.math.siu.edu/olive/reghw.txt)). Then to copy and paste the program into SAS and save it on your disk. Then run the program in SAS. Click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy". In *Word*, use the commands "Edit>Paste". Most of the output is irrelevant. Then cut out all of the output except *the Model Fit Statistics* the output for testing *BETA = 0* and the *coefficient estimates* from Proc Logistic. (All of this output should fit on about half a page.) Print out the output.

The crab data is from Agresti (1996, p. 105–107, 272). Use the estimates from the output (which differ slightly from those in the text).

b) Predict $\hat{\rho}(x)$ if $x = 21.0$.

c) Find a 95% CI for β.

d) Perform the 4 step Wald test for $Ho : \beta = 0$.
(SAS output gives z_o^2 as the Wald chi-square. You need to use $z_o = \hat{\beta}/\mathrm{se}(\hat{\beta}) = \sqrt{\mathrm{z}_\mathrm{o}^2}$. Recall that $z^2 \sim \chi^2_1$ if $z \sim N(0, 1)$).