

Chapter 3

Binary Response and Logistic Regression Analysis

February 7, 2001

Part of the **Iowa State University** NSF/ILI project *Beyond Traditional Statistical Methods*

Copyright 2000 D. Cook, P. Dixon, W. M. Duckworth, M. S. Kaiser, K. Koehler, W. Q. Meeker and W. R. Stephenson.

Developed as part of NSF/ILI grant DUE9751644.

Objectives

This chapter explains

- the motivation for the use of logistic regression for the analysis of binary response data.
- simple linear regression and why it is inappropriate for binary response data.
- a curvilinear response model and the logit transformation.
- the use of maximum likelihood methods to perform logistic regression.
- how to assess the fit of a logistic regression model.
- how to determine the significance of explanatory variables.

Overview

Modeling the relationship between explanatory and response variables is a fundamental activity encountered in statistics. Simple linear regression is often used to investigate the relationship between a single explanatory (predictor) variable and a single response variable. When there are several explanatory variables, multiple regression is used. However, often the response is not a numerical value. Instead, the response is simply a designation of one of two possible outcomes (a binary response) e.g. alive or dead, success or failure. Although responses may be accumulated to provide the number of successes and the number of failures, the binary nature of the response still remains.

Data involving the relationship between explanatory variables and binary responses abound in just about every discipline from engineering to, the natural sciences, to medicine, to education, etc. How does one model the relationship between explanatory variables and a binary response variable? This chapter looks at binary response data and its analysis via logistic regression. Concepts from simple and multiple linear regression carry over to logistic regression. Additionally, ideas of maximum likelihood (ML) estimation, seen in other chapters, are central to the analysis of binary data.

3.1 Introduction

Data involving the relationship between explanatory variables and binary responses abound in just about every discipline from engineering, to the natural sciences, to medicine, to education, etc. What sorts of engineering, science and medical situations lead to binary data?

3.1.1 The *Challenger* Disaster

On January 28, 1986 the space shuttle, *Challenger*, had a catastrophic failure due to burn through of an O-ring seal at a joint in one of the solid-fuel rocket boosters. Millions of television viewers watched in horror as a seemingly perfect launch turned to tragedy as the rocket booster exploded shortly after lift off. A special commission was appointed by President Reagan to investigate the accident and to come up with a probable cause and recommendations for insuring the future safety of shuttle flights.

The launch of *Challenger* on that day was the 25th launch of a shuttle vehicle. After each launch the solid rocket boosters are recovered from the ocean and inspected. Of the previous 24 shuttle launches, 7 had incidents of damage to the joints, 16 had no incidents of damage and 1 was unknown because the boosters were not recovered after launch. Here was a classic case of binary response data, damage or no damage to the solid rocket booster joints. What variables might explain the existence of damage or no damage? January 28, 1986 was a particularly cold day at the launch site in Florida. Could temperature at the time of launch be a contributing factor?

The following data on the temperature at the time of launch and whether or not the rocket boosters on that launch had damage to the field joints is derived from data from the Presidential Commission on the Space Shuttle *Challenger* Accident (1986). A 1 represents damage to field joints, and a 0 represents no damage. Multiple dots at a particular temperature represent multiple launches.

Incidence of Booster Field Joint Damage vs. Temperature

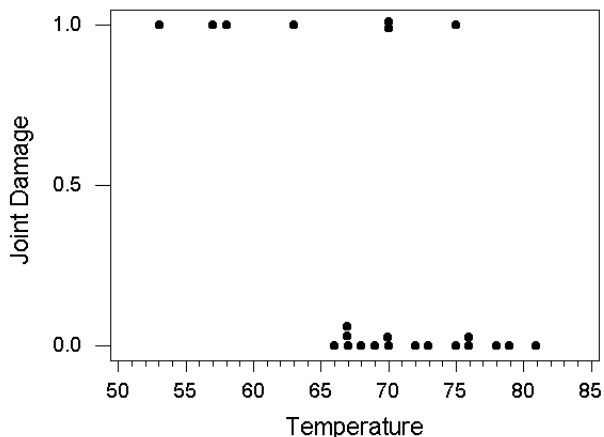


Figure 3.1: The *Challenger* Disaster

Although not conclusive, the graph indicates that launches at temperatures below 65° F appear to have a higher proportion (4 out of 4 or 100%) with damage than those at temperatures above 65° F (only 3 out of 19 or 16%). Is there some way to predict the chance of booster rocket field joint damage given the temperature at the time of launch? Not all booster rocket field joint damage results in a catastrophe like the *Challenger*, however, being able to predict the probability of joint damage would contribute to making future launches safer.

3.1.2 The sex of turtles

Did you ever wonder what determines the sex (male or female) of a turtle? With humans, the sex of a child is a matter of genetics. With turtles, environment during the period eggs are incubated plays a significant role in the sex of the hatchlings. How does temperature during incubation affect the proportion of male/female turtles? The following experiment and data, which come from a statistical consulting project of Professor Kenneth Koehler at Iowa State University, attempts to answer that question for one species of turtle.

In the experiment, turtle eggs were collected from nests in Illinois. Several eggs were placed in each of several boxes. Boxes were incubated at different temperatures, that is the entire box was placed in a incubator that was set at a specified temperature. There were 5 different temperatures all between 27 and 30 degrees centigrade. There were 3 boxes of eggs for each temperature. Different boxes contain different numbers of eggs. When the turtles hatched, their sex was determined.

Temp	male	female	% male	Temp	male	female	% male
27.2	1	9	10%	28.4	7	3	70%
	0	8	0%		5	3	63%
	1	8	11%		7	2	78%
27.7	7	3	70%	29.9	10	1	91
	4	2	67%		8	0	100%
	6	2	75%		9	0	100%
28.3	13	0	100%				
	6	3	67%				
	7	1	88%				

A further summary of the data reveals that the proportion of males hatched tends to increase with temperature. When the temperature is less than 27.5 C only 2 of 25 or 8% of hatchlings are male. This increases to 19 of 51 or 37% male for temperatures less than 28 C. For temperatures less than 28.5 C, 64 of 108 or 59% are male. For all the boxes at all the temperatures the proportion of males is 91 of 136 or 67%.

Proportion of male turtles vs. incubation temperature

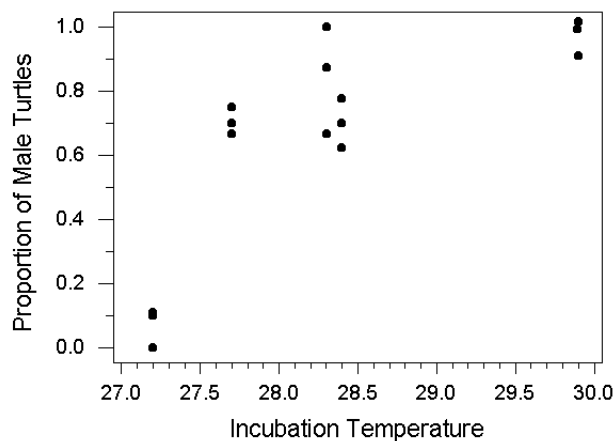


Figure 3.2: The sex of turtles

Is there some way to predict the proportion of male turtles given the incubation temperature? At what temperature will you get a 50:50 split of males and females?

3.1.3 Bronchopulmonary displasia in newborns

The following example comes from *Biostatistics Casebook*, by Rupert Miller, *et. al.*, (1980), John Wiley & Sons, New York. The data we will consider is a subset of a larger set of data presented in the casebook.

Bronchopulmonary displasia (BPD) is a deterioration of the lung tissue. Evidence of BPD is given by scars in the lung as seen on a chest X-ray or from direct examination of lung tissue at the time of death. BPD is seen in newborn babies with respiratory distress syndrome (RDS) and oxygen therapy. BPD also occurs in newborns who do not have RDS but who have gotten high levels of oxygen for some other reason. Having BPD or not having BPD is a binary response. Since incidence of BPD may be tied to the administration of oxygen to newborns, exposure to oxygen, O_2 , could be a significant predictor. Oxygen is administered at different levels; Low (21 to 39% O_2), Medium (40 to 79% O_2), and High (80 to 100% O_2). The number of hours of exposure to different levels of oxygen, O_2 is recorded for each newborn in the study. The natural logarithm of the number of hours of exposure at each level; $\ln L$, $\ln M$ and $\ln H$ will be used as the explanatory variables in our attempts to model the incidence of BPD.

Incidence of BPD vs. $\ln(\text{hours})$ at various levels of O_2

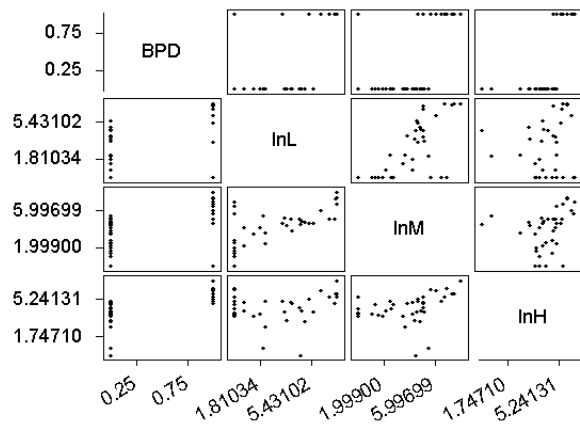


Figure 3.3: Bronchopulmonary displasia

Is there some why to predict the chance of developing BPD given the hours (or the natural logarithm of hours) of exposure to the various levels of oxygen? Do the differing levels of oxygen have differing effects on the chance of developing BPD?

3.1.4 Other examples

There are many other examples of binary response data. For instance,

- College mathematics placement: Using ACT or SAT scores to predict whether individuals would “succeed” (receive a grade of C or better) in an entry level mathematics course and so should be placed in a higher level mathematics course.
- Grades in a statistics course: Do things like interest in the course, feeling of pressure/stress and gender relate to the grade (A, B, C, D, or F) one earns in a statistics course?
- Credit card scoring: Using various demographic and credit history variables to predict if individuals will be good or bad credit risks.
- Market segmentation: Using various demographic and purchasing information to predict if individuals will purchase from a catalog sent to their home.

All of the examples mentioned above have several things in common. They all have a binary (or categorical) response (damage/no damage, male/female, BPD/no BPD). They all involve the idea of prediction of a chance, probability, proportion or percentage. Unlike other prediction situations, what we are trying to predict is bounded below by 0 and above by 1 (or 100%).

These common features turn out to present special problems for prediction techniques that the reader may already be familiar with, simple linear and multiple regression. A different statistical technique, logistic regression, can be used in the analysis of binary response problems. Logistic regression is different from simple linear and multiple regression but there are many similarities. It is important to recognize the similarities and differences between these two techniques.

In order to refresh the reader's memory, the next section starts with a review of simple linear regression.

3.2 Simple linear regression

Simple linear regression is a statistical technique that fits a straight line to a set of (X,Y) data pairs. The slope and intercept of the fitted line are chosen so as to minimize the sum of squared differences between observed response values and fitted response values. That is, a method of ordinary least squares is used to fit a straight line model to the data.

In thinking about simple linear regression it is important to keep in mind the type of data suited for this statistical technique as well as the model for the relationship between the explanatory (X) and response (Y) variables.

In simple linear regression, the data consist of pairs of observations. The explanatory (X) variable is numerical (continuous measurement) that will be used to predict the response. The response (Y) variable is numerical (continuous measurement). The simple linear model consists of two parts; a structure on the means and an error structure. It is assumed that the mean response is a linear function of the explanatory (predictor) variable. The error structure in the model attempts to describe how individual measurements vary around the mean value. It is assumed that individual responses vary around the mean according to a normal distribution with variance σ^2 . This model can be expressed symbolically as

- Structure on the means:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

- Error structure:

$$\epsilon_i \sim N(0, \sigma^2)$$

and shown graphically in Figure 3.4.

In contrast with this structure on the data, continuous measurement response, let's look at the *Challenger* data. The basic response for the *Challenger* data is binary. $Y_i = 1$ if there is damage to rocket booster field joint, and $Y_i = 0$ if there is no damage to rocket booster field joint. This can be re-expressed in terms of a chance or probability of damage.

$$\begin{aligned} \text{Prob}(Y_i = 1) &= \pi_i \\ \text{Prob}(Y_i = 0) &= 1 - \pi_i \end{aligned}$$

With a binary response one has, in general,

$$E(Y_i) = 0 * (1 - \pi_i) + 1 * \pi_i = \pi_i$$

or with an explanatory variable, X_i

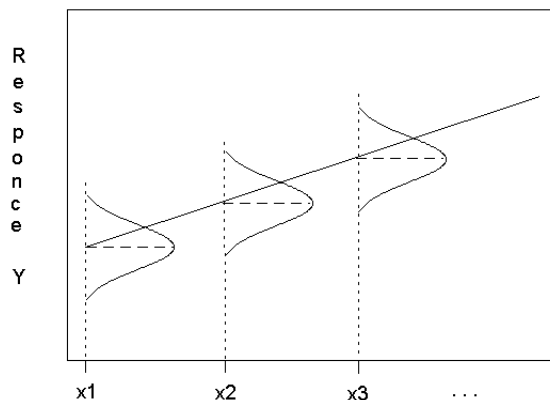


Figure 3.4: Simple linear regression

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i = \pi_i \quad (3.1)$$

With binary response data, if we try to use a simple linear model we are saying that the probability of damage to a rocket booster field joint is a linear function of the temperature. Since probabilities are always between zero and one, this model has the constraint that reasonable values for the response, and therefore the predicted response, are between zero and one. The binary nature of the response also creates difficulties in how we view the variability of individual values around the mean. The variance of a binary response is a function of the probability, π_i . Explicitly

$$Var(Y_i) = \pi_i(1 - \pi_i)$$

From equation 5.1 above π_i is a function of X_i and so the variance of Y_i is also a function of π_i . That is, an assumption of constant variance σ^2 is violated. Additionally, since binary responses can take on only two values, 0 and 1, it is obvious that binary responses cannot vary about the mean according to a normal distribution.

One should remember at this point that the error structure, our assumptions on how individuals vary about the mean, is necessary for the proper application and interpretation of the formal statistical inferences made for simple linear regression. In other words, we need these assumptions to construct confidence intervals and test for significance of the linear relationship. What many people may not realize is that we do not need these assumptions to perform ordinary least squares and come up with estimates of the slope and intercept in the model that relates the mean response to the explanatory variable. We can, and will, use ordinary least squares to fit a linear relationship between the binary response, damage/no damage, to the temperature at launch for the *Challenger* data.

Example 3.1 The following S+ commands plots the *Challenger* data, performs a simple linear regression, and creates a plot of residuals. The data (given below) should be stored in a data frame called **challenger.dat**.

Temp	Damage	Temp	Damage	Temp	Damage
66	0	57	1	70	0
70	1	63	1	81	0
69	0	70	1	76	0
68	0	78	0	79	0
67	0	67	0	75	1
72	0	53	1	76	0
73	0	67	0	58	1
70	0	75	0		

```

challenger <- read.table("challenger.dat",head=T)
challenger
attach(challenger)
plot(Temp,Damage,xlab="Temperature",ylab="Damage",
      main="Incidence of Booster Field Joint Damage vs. Temperature")
challenger.fit1 <- lm(Damage~Temp)
summary(challenger.fit1)
resid1 <- residuals(challenger.fit1)
plot(Temp,resid1,xlab="Temperature",ylab="Residuals:SLR model",
      main="Challenger Disaster: Plot of Residuals (SLR Model)")

```

The ordinary least squares, simple linear regression model, prediction equation is given by

$$\text{Predicted damage} = 2.905 - 0.0374 * \text{Temp}$$

The predicted damage can be interpreted as the predicted probability of damage for the given temperature. Plugging various temperatures into this equation we get somewhat reasonable predicted probabilities. For a temperature of 77, the predicted probability of damage is 0.025. For a temperature of 65, the predicted probability is 0.474. Finally, for a temperature of 51, the predicted probability is 0.998. As temperature decreases the predicted probability of damage increases. Figure 3.5 shows the fit of the simple linear regression model.

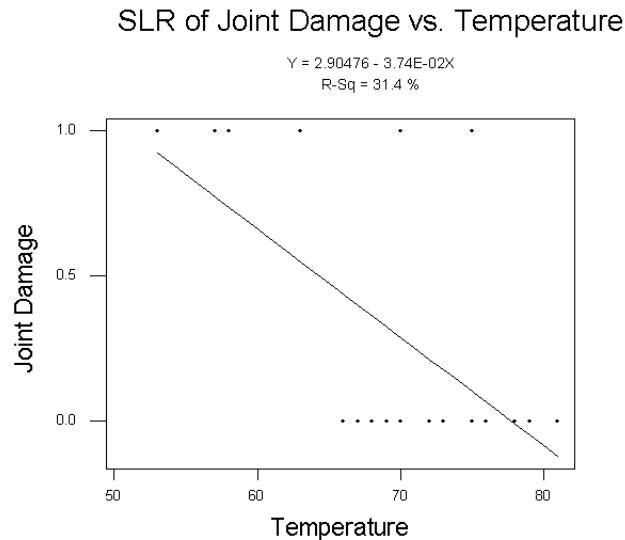


Figure 3.5: Fitted line plot

Although this appears to be a reasonable prediction equation, it can lead to nonsensical predictions. For example, if the temperature is 85, the predicted probability of damage is -0.274. This is nonsense since probabilities can never be less than zero. Similarly, any temperature 50 or below will give a predicted probability of more than one. Again, this is nonsense. The bounded nature of probabilities is not incorporated into the simple linear regression model.

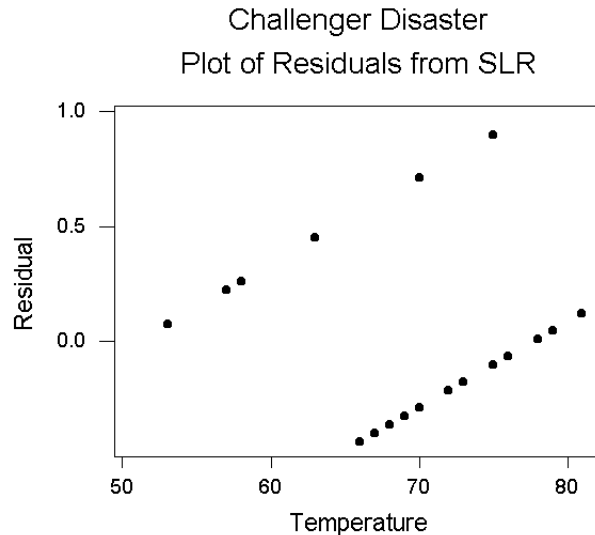


Figure 3.6: Plot of residuals

Another clue to the inadequacy of the simple linear model for the *Challenger* data is shown in the pattern of the residuals. Rather than exhibit a random scatter of points, the plot of residuals versus temperature (see Figure 3.6) shows two slanted lines. ■

When ordinary least squares is used to fit a simple linear regression model, the estimates of the intercept and slope will be unbiased. That is, on average the intercept and slope estimates give the correct values. Since we are violating the equal variance assumption (recall the variance of a binary response will change with changes in the mean), the standard errors for the estimates of intercept and slope will be larger than they should be. Additionally, the unusual pattern in the residuals indicates that the simple linear model does not provide a good fit to the data.

When variances are unequal, weighted least squares (an extension of ordinary least squares) can be used. In essence, observations with large variances are given lower weights and thus have less influence on the fit of the simple linear model. Ordinarily the weights are inversely proportional to the variances. That is

$$w_i = \frac{1}{\pi_i(1 - \pi_i)}$$

Since π_i is not known, estimated weights are used.

$$\hat{w}_i = \frac{1}{\hat{Y}_i(1 - \hat{Y}_i)}$$

The procedure can be summarized as follows

- Fit Ordinary Least Squares (simple linear model)
 - Obtain estimates, \hat{Y}_i

- If an estimate is less than 0 or more than 1, set it to 0.001 or 0.999, respectively.
- Compute weights, \hat{w}_i
- Fit Weighted Least Squares (simple linear model)

Example 3.2 The results of ordinary least squares and weighted least squares fits to the *Challenger* data are given in the following table.

Coefficient	Ordinary Least Squares		Weighted Least Squares	
	Estimate	Std Error	Estimate	Std Error
Intercept	2.905	0.8421	2.344	0.5324
Slope	-0.0374	0.0120	-0.0295	0.0067

Note that estimates and standard errors differ between ordinary least squares and weighted least squares. It is true that the standard errors of the estimates for weighted least squares are less than those for ordinary least squares. This however is only a small improvement. We are still dealing with a simple linear model and so the possibility that predicted values will be less than zero or greater than one still exists. Also, the peculiar pattern in the residuals still remains even with the weighted least squares procedure. The main problem lies in the choice of a linear model instead of some curvilinear model for the relationship between an explanatory variable and a binary response. ■

When the response variable is binary, or a binomial proportion, the expected response is more appropriately modeled by some curved relationship with the predictor variable. One such curved relationship is given by the logistic model.

$$E(Y_i|X_i) = \pi_i = \frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}} \quad (3.2)$$

There are several things that should be noted about this function. First of all it is bounded between zero and one. This will eliminate the possibility of getting nonsensical predictions of proportions or probabilities. Secondly, there is a linear model hidden in the function that can be revealed with a proper transformation of the response. Finally, the sign associated with the coefficient, β_1 indicates the direction of the curve. A positive value for β_1 indicates an increasing function while a negative value indicates a decreasing function like the one given in Figure 3.7.

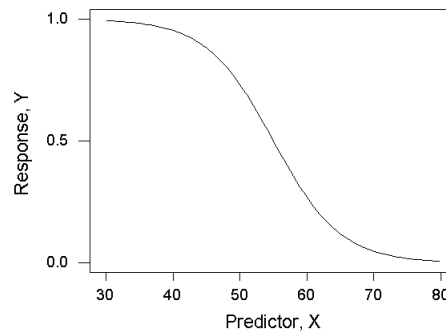


Figure 3.7: Logistic Model

As mentioned above there is a linear model hidden within the logistic model. An appropriate transformation of the response can reveal that linear model. Such a transformation is called the logit transformation. From equation (5.2), the natural logarithm of the ratio of π_i to $(1 - \pi_i)$, gives a linear model in X_i . Specifically,

$$\pi'_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_o + \beta_1 X_i \quad (3.3)$$

Since π_i is not known, use the observed proportion, p_i and the logit transformation, $\ln\left(\frac{p_i}{1-p_i}\right)$. The logit transformation requires that we have observed proportions that are not zero or one. The *Challenger* data is made up of observed proportions that are almost all zeros or ones. In order to see how the logit transformation works, the combined data for the sex of turtles will be used.

Example 3.3 The data for this example comes from combining the numbers of males and females for the three boxes at each incubation temperature.

Temp	male	female	total	pmale, p_i
27.2	2	25	27	0.0741
27.7	17	7	24	0.7083
28.3	26	4	30	0.8667
28.4	19	8	27	0.7037
29.9	27	1	28	0.9643

In order to see the advantages of the logit transformation, a simple linear model relating the observed proportions to the incubation temperature is fit using ordinary least squares. The fitted equation is given as

$$\hat{\pi} = -6.902 + 0.2673Temp \quad (3.4)$$

SLR of proportion male on incubation temperature

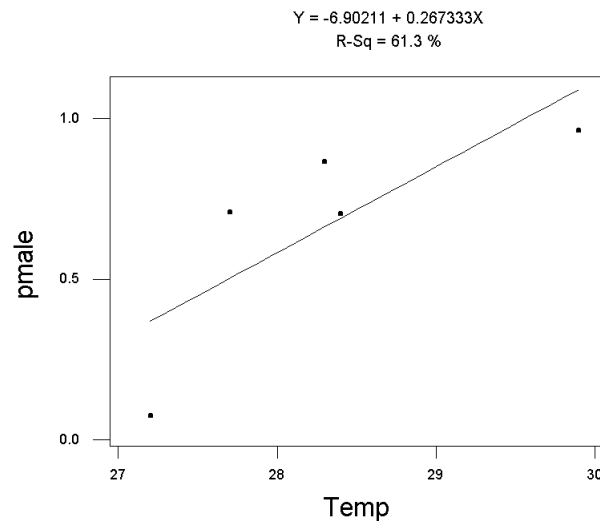


Figure 3.8: Simple linear model on observed proportions

The five incubation temperatures and fitted proportions, using equation $\hat{\pi} = -6.902 + 0.2673Temp$ are summarized below and in Figure 3.8.

Temp	27.2	27.7	28.3	28.4	29.9
Fit prop., $\hat{\pi}$	0.369	0.503	0.663	0.690	1.091

Note that the simple linear model on the observed proportions gives a fitted value that is greater than one.

Next, the logit transformation is applied to the observed proportions.

Temp	pmale, p_i	$\ln\left(\frac{p_i}{1-p_i}\right)$
27.2	0.0741	-2.5257
27.7	0.7083	0.8873
28.3	0.8667	1.8718
28.4	0.7037	0.8650
29.9	0.9643	3.2958

Ordinary least squares is used to fit a linear model to the observed logit transformed proportions. The resulting prediction equation is

$$\hat{\pi}' = -51.1116 + 1.8371Temp \quad (3.5)$$

SLR of logit(pmale) on incubation temperature

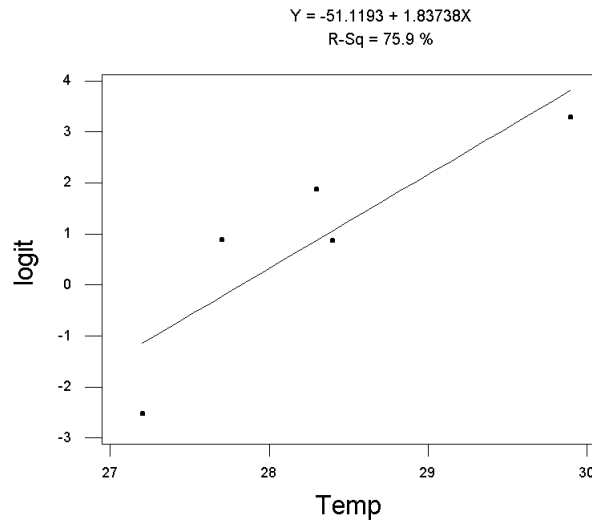


Figure 3.9: Simple linear model on logits

The five incubation temperatures, fitted logits and fitted proportions, using the equation 3.5 are summarized below and in Figures 3.9 and 3.10.

Temp	27.2	27.7	28.3	28.4	29.9
Fit logit, $\hat{\pi}'$	-1.1420	-0.2334	0.8788	1.0626	3.8182
Fit prop., $\hat{\pi}$	0.242	0.444	0.707	0.743	0.979

Note how the curved nature of the logistic model keeps the predicted proportions between zero and one.

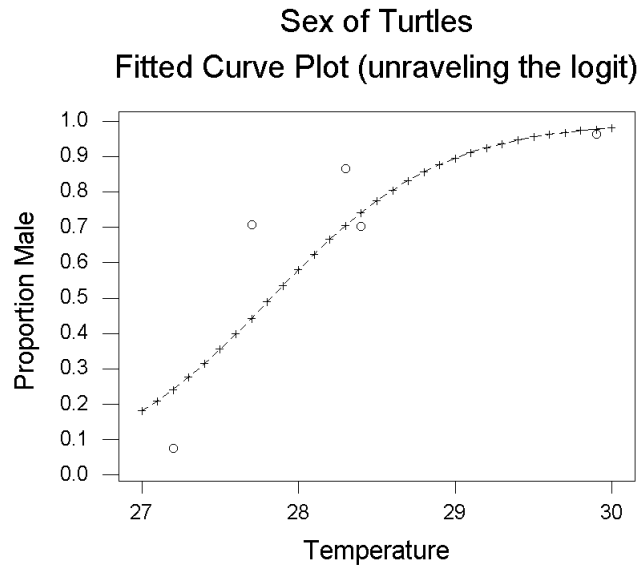


Figure 3.10: Fitted logistic curve

Researchers are often interested in the value of the explanatory variable that produces a 50:50 split. Using the simple linear model on the observed proportions for the combined turtle data, an incubation temperature of 27.69 degrees Celsius would yield a predicted proportion male of 0.50. For the simple linear model on the logit transformed proportions an incubation temperature of 27.82 degrees Celsius will give a predicted proportion of male of 0.50.

■

The logit transformation has adjusted for the curved nature of the response and restricts all predicted proportions to sensible values (between zero and one). Although progress has been made in terms of fitting a model (structure on the means) to the data, there are still problems with the error structure. Because of the binary nature of the response, the equal variance assumption is not valid nor is the assumption that responses vary about the mean according to a normal distribution. Even if we can obtain reasonable fitted and predicted values, we have no way of assessing that fit nor do we have a way to assess the statistical significance of the parameter estimates.

By doing simple linear regression, even on the logit transformed proportions, we are trying to force binary response data into a familiar method of analysis. What we really need is a new way of looking at this problem and a new means of analyzing the data. Instead of using least squares, we can use maximum likelihood methods to estimate the parameters of a logistic model.

3.3 Maximum Likelihood Estimation

The likelihood is a function of the data and the parameters of the model. The likelihood is maximized by finding estimates of the model parameters that are most likely to give us the data. With binary data the form of the likelihood is relatively simple.

Recall that with a binary response

- $Y_i = 1$ with probability π_i
- $Y_i = 0$ with probability $1 - \pi_i$

Provided observed responses are independent, the likelihood is a product of π_i 's and $1 - \pi_i$'s. For each $Y_i = 1$ the probability π_i appears in the product. Similarly, for each $Y_i = 0$ the probability $1 - \pi_i$ appears in the product. The general form of the likelihood for binary data is

$$L((\beta_o, \beta_1); Data) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \quad (3.6)$$

So far the likelihood only deals with the product of the probabilities. There is no mention of how those probabilities will be modeled. We could put in a linear model for the π_i 's but we have seen how a linear model for the probabilities has problems. Instead we use a logistic function to model the relationship between the probability π_i and the explanatory variable, X_i . Equation 3.2 gives the logistic model for π_i . It, along with the corresponding model for $1 - \pi_i$, are given below for convenience.

$$\pi_i = \frac{e^{(\beta_o + \beta_1 X_i)}}{1 + e^{(\beta_o + \beta_1 X_i)}}$$

$$(1 - \pi_i) = \frac{1}{1 + e^{(\beta_o + \beta_1 X_i)}}$$

Replacing these in equation 3.6 we get

$$\begin{aligned} L((\beta_o, \beta_1); Data) &= \prod_{i=1}^n \left(\frac{e^{(\beta_o + \beta_1 X_i)}}{1 + e^{(\beta_o + \beta_1 X_i)}} \right)^{Y_i} \left(\frac{1}{1 + e^{(\beta_o + \beta_1 X_i)}} \right)^{1-Y_i} \\ &= \prod_{i=1}^n \frac{(e^{(\beta_o + \beta_1 X_i)})^{Y_i}}{(1 + e^{(\beta_o + \beta_1 X_i)})} \end{aligned} \quad (3.7)$$

Finding values of β_o and β_1 that maximize equation 3.7 does not appear to be a simple task. Often working with the natural logarithm of the likelihood is easier. The log-likelihood for binary data with a logistic model is given by

$$\log [L((\beta_o, \beta_1); Data)] = \sum_{i=1}^n Y_i (\beta_o + \beta_1 X_i) - \sum_{i=1}^n \log [1 + e^{(\beta_o + \beta_1 X_i)}] \quad (3.8)$$

We want to choose β_o and β_1 so as to maximize the log-likelihood. These choices will also maximize the likelihood. Similar to ordinary least squares, there will be two equations that must be solved for two unknowns (the solutions provide the estimates of β_o and β_1). Unlike ordinary least squares, the two equations will not be linear and so must be solved by iteration (start with initial values for β_o and β_1 , evaluate the log-likelihood, choose a new value for β_o or β_1 that reduces the log-likelihood, repeat until the log-likelihood does not change).

The whole process of maximum likelihood estimation for binary response data with a logistic model is available in the general linear model (glm) procedure in S+. The option **family=binomial** uses the likelihood given in equation 3.6. The option **link=logit** uses the logistic model given in equation 3.2. The following example uses the glm procedure in S+ to fit a logistic regression model using maximum likelihood estimation.

Example 3.4 In this example we use the glm procedure in S+ to fit a logistic regression model to the combined turtle data that appear in Example 3.3. Because of the small size of the data set, we choose to put the data directly into S+.

```
t <- c(27.2, 27.7, 28.3, 28.4, 29.9)
nmale <- c(2, 17, 26, 19, 27)
nfemale <- c(25, 7, 4, 8, 1)
```

```

ntur <- nmale+nfemale
pmale <- nmale/ntur

#-----
#
# fit logistic regression model using the proportion male as the
# response and the number of turtles as the weights in glm. The
# logit transformation is the default for the family binomial.
#
#-----

tur.logit.wls <- glm(pmale~t,family=binomial,weights=ntur)
summary(tur.logit.wls)
coefficients(tur.logit.wls)

#-----
#
# make a plot of the summarized data and the logistic regression fit
#
#-----

tx <- c(270:300/10)

tyhat <- coefficients(tur.logit.wls)[c(1)] +
  coefficients(tur.logit.wls)[c(2)]*tx

tpihat <- exp(tyhat)/(1+exp(tyhat))

plot(t,pmale,xlab="Temperature",ylab="Proportion Male",
     main="Proportion of Male Turtles vs. Incubation Temperature")

lines(tx,tpihat)

#-----
#
# fit logistic regression model using the number male and the
# number female. S+ automatically creates the number of
# turtles and includes these in the weighted glm.
#
#-----

response <- cbind(nmale,nfemale)
tur.logit <- glm(response~t,family=binomial)
summary(tur.logit)

```

There are actually two ways of obtaining a logistic regression analysis of binary response data. The first uses the response as the observed proportion of male turtles. Because the proportions of males do not carry information on how many total turtles went into calculating the proportions, a weights option is needed to tell S+ the total number of turtles used to make each proportion. The second way uses a two-variable response (nmale,nfemale). S+ is smart enough to create the proportion male and the total number of turtles from this information. In both glm() statements the family is specified as binomial. The default link for a binomial family is the logit link.

The fitted logistic regression equation is given by

$$\hat{\pi}' = -61.3183 + 2.2110Temp \quad (3.9)$$

The fitted logits and fitted proportions from this equation are given in the following table and in Figure 3.11.

Temp	27.2	27.7	28.3	28.4	29.9
Fit logit, $\hat{\pi}'$	-1.1791	-0.0736	1.2530	1.4741	4.7906
Fit prop., $\hat{\pi}$	0.235	0.482	0.778	0.814	0.992

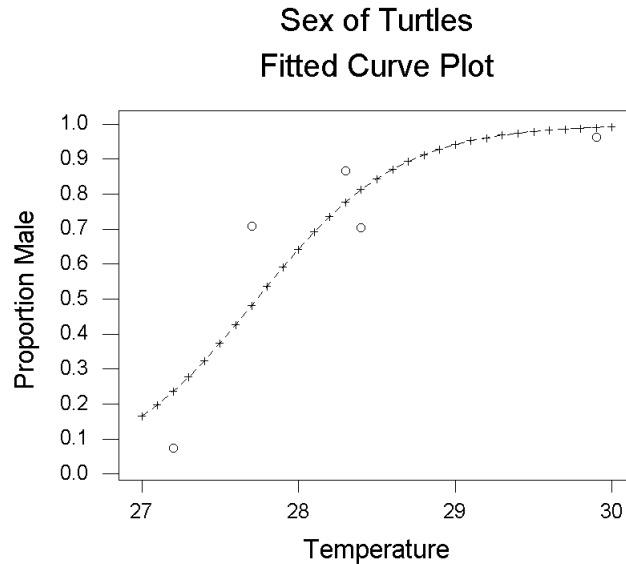


Figure 3.11: S+: `glm(binomial,logit)`

Note that the fit is very similar to that given in Figure 3.10. Again for researchers who are interested in the temperature for a 50:50 split, this equation indicates that a temperature of 27.73 degrees Celsius will give a predicted proportion of males of 0.50. ■

A few comments are in order concerning the way logistic regression models are fit in S+ and other statistical computing packages. There is always the possibility that an observed proportion is either zero or one. This creates problems with the logistic model and the logit transformation. When a p_i equals 0 or 1, it is replaced by $p_i = \frac{1}{2n_i}$ or $p_i = 1 - \frac{1}{2n_i}$, respectively. Some computer packages may adjust all the p_i by some small amount, *e.g.* $p_i + \frac{(0.5-p_i)}{n_i}$.

Different computer packages may use different adjustments, starting values, round off, and algorithms than S+. Even with S+, the two ways to run a binary response logistic regression give slightly different parameter estimates. Below are the logistic regression fits for various computer packages for the combined turtle data.

- S+: $\pi'_i = -61.31828 + 2.21103X_i$
or $\pi'_i = -61.318299 + 2.211031X_i$
- Minitab: $\pi'_i = -61.32 + 2.2110X_i$
- SAS: $\pi'_i = -61.3183 + 2.2110X_i$

The coefficients in a logistic regression are often difficult to interpret because the effect of increasing X by one unit varies depending on where X is. This is the essence of a nonlinear model. Consider first the interpretation of the odds:

$$\frac{\pi_i}{(1-\pi_i)}$$

If $\pi_i = 0.75$, then the odds of getting a male turtle are 3 to 1. That is, a male turtle is 3 times as likely as a female turtle.

In logistic regression we model the log-odds. The predicted log-odds, $\hat{\pi}'_i$ are given by the linear equation 3.9. The predicted odds are

$$e^{\hat{\pi}'_i} = \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$$

If we increase X_i by 1 unit, we multiply the predicted odds by $e^{\hat{\beta}_1}$. The intercept can be thought of as the predicted log-odds when X_i is zero. The anti-log of the intercept may have some meaning as a baseline odds, especially if zero is within the range of the data for the predictor variable, X .

Example 3.5 Continuing our analysis of the proportion of male turtles we see

- At 27° C the predicted odds for a male turtle are 0.20, about 1 in 5. That is, it is 5 times more likely to get a female than a male at this temperature.
- At 28° C the predicted odds for a male are $e^{2.2110} = 9.125$ times the odds at 27° C, or 1.825. At 28° C getting a male is almost twice as likely as getting a female.
- At 29° C the predicted odds for a male are $e^{2.2110} = 9.125$ times the odds at 28° C, or 16.65. At 29° C getting a male is over 16 times more likely than getting a female.

Since all the data come from temperatures, values of X , between 27° C and 30° C. The values $X = 0$ is well outside the range of the data. In the turtle example, the intercept, or the anti-log of the intercept, has no practical interpretation. ■

So far we have only looked at fitting a model by estimating parameters using maximum likelihood techniques. As in all of statistics, estimates of model parameters are subject to variation. We must be able to quantify this variation in order to make inferences; tests of hypotheses and confidence intervals for model parameters.

Some of the inferences we would like to be able to make involve testing the significance of the slope parameters. Such tests attempt to answer the question; Is there a statistically significant relationship between the explanatory variable(s) and the binary response. Additionally, we would like to be able to assess the “goodness” of the fit of the logistic model or to be able to tell if there is a significant lack of fit.

It is easy to use the fitted model equations to predict, say, the proportion of male turtles for a temperature not investigated in the original experiment. However, a point estimate does not give the whole story. Rather a confidence interval for the proportion of male turtles at that temperature would be more informative.

Before we explore statistical inference for logistic regression in the next section, a few words of caution. Inference techniques for logistic regression appear to be similar to, or at least analogous to, inference techniques for ordinary least squares regression with a linear model. We will use these similarities to try to motivate these new techniques. However, there is one considerable difference. Inference techniques in simple linear regression rely on the assumptions of independent observations, normally distributed errors all with the same variance. If these assumptions are reasonable for your data, then the inference techniques are exact. Inference for logistic regression is based on asymptotic theory. That is, the inference techniques are approximate at best with the approximations getting better when you have larger amounts of data (larger sample sizes).

3.4 Inference for logistic regression

Just as with ordinary least squares regression we need some means of determining the significance of the estimates of the model parameters. We also need a means of assessing the fit, or lack of fit, of the logistic model. Inference for logistic regression is often based on the deviance (also known as the residual deviance). The deviance is twice the log-likelihood ratio statistic. The deviance for a logistic model can be likened to the residual sum of squares in ordinary least squares regression for the linear model. The smaller the deviance the better the fit of the logistic model. A large value for the deviance is an indication that there is a significant lack of fit for the logistic model and some other model may be more appropriate.

Asymptotically, the deviance has a χ^2 distribution. Therefore, to perform tests of hypotheses regarding the fit of the model the deviance is compared to the percentiles of a χ^2 distribution. The degrees of freedom is determined by the number of observations less the number of parameters estimated. Keep in mind that this is an asymptotic (large sample size) procedure and the P-values calculated using the χ^2 distribution are approximate.

Example 3.6 The `summary()` command in S+ gives the residual deviance as well as the parameter estimates and other important information for the logistic regression fit. For the combined turtle data of example 3.3 and their analysis in example 3.4. Output from the S+ general linear model analysis with `family=binomial` is given below.

Coefficients:

	Value	Std. Error	t value
(Intercept)	-61.31828	12.0155501	-5.103244
t	2.21103	0.4306079	5.134672

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 64.4285 on 4 degrees of freedom

Residual Deviance: 14.86295 on 3 degrees of freedom <-----

The residual deviance for the logistic model fit to the combined turtle data is 14.863 on 3 degrees of freedom. We have only 5 observations when we combine the data, one proportion for each temperature. We fit two parameters, an intercept and a slope in the logistic regression model. The P-value, the chance that a χ^2 with 3 degrees of freedom exceeds 14.863, is 0.0019. Small P-values represent significant departure from the null hypothesis. In this case the null hypothesis states that the logistic regression model provides an adequate fit to the data. Since the residual deviance is large (the P-value is small), this indicates a significant lack of fit for the logistic model. There is room for improvement and other models should be considered. ■

What went wrong? It could be that the dip in the number of males at 28.4° C compared to the number of males at 28.3° C is causing the logistic model not to fit well. Alternatively, it could be that the curvilinear relationship between temperature and the proportion of male turtles is not symmetric. The logistic model is symmetric in its rise and then leveling off. If the true nature is one of a faster rise and a slower leveling off, or *vice versa*, then the logistic model will not capture this nuance in the behavior of the proportions and not give an adequate fit.

However, all is not lost. Although the logistic model may not be the best, there still may be some worthwhile information about the probability of a male hatchling and the incubation temperature. The question still remains; Does temperature and the logistic model give us statistically significant information about the proportion of male turtles? We need to be able to measure if we are doing better predicting the proportion of male turtles using temperature and the logistic model than if we were to make predictions ignoring the temperature.

The logistic regression analysis calculates a so called null deviance as well as the residual deviance. The null deviance summarizes the fit of a logistic model that just includes an intercept.

$$\pi_i = \frac{e^{\beta_o}}{1+e^{\beta_o}}$$

Such a model would predict a constant value for the response proportion regardless of the value of the explanatory variable. By looking at the change in the deviance when an explanatory variable is added, we can determine whether or not that explanatory variable is adding significantly to the predictive ability of the model.

The difference between the null deviance and the residual deviance represents the effect of adding the single explanatory variable to the logistic model. This is analogous to the change in the sum of squared residuals (sum of squares for error) in ordinary least squares regression. When an explanatory variable is added in ordinary least squares regression, the change in the sum of squares for error represents the amount of variability explained by that variable. The change in deviance in logistic regression can be compared to a χ^2 distribution to determine statistical significance. The degrees of freedom for the χ^2 is equal to the number of predictor variables added to the model, in this case, 1. Keep in mind that this test, like all the others, requires a large sample size and any results are approximate.

Example 3.7 Returning to our sex of turtles combined data. Output from the S+ general linear model analysis with family=binomial is given below.

Coefficients:

	Value	Std. Error	t value
(Intercept)	-61.31828	12.0155501	-5.103244
t	2.21103	0.4306079	5.134672

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 64.4285 on 4 degrees of freedom <-----

Residual Deviance: 14.86295 on 3 degrees of freedom <-----

The null deviance (model with a constant proportion of males) is 64.4285 on 4 degrees of freedom (we are only fitting one parameter (the constant proportion of male turtles) in the null logistic model). The residual deviance (logistic model relating proportion of males to temperature) is 14.8630 on 3 degrees of freedom. The change in deviance (“Importance” of temperature in the logistic model) is 49.5655 on 1 degrees of freedom. Comparing the change in deviance, 49.5655, to a χ^2 with 1 degree of freedom, the P-value is virtually zero. This change in deviance is not attributable to chance alone, rather including temperature in the logistic model is adding significantly to your ability to predict the proportion of male turtles.

In summary, temperature is statistically significant in the logistic regression model for the sex of turtles. Using temperature and the fitted logistic model will give you better predictions for the proportion of males than using a constant proportion as your prediction. Although the logistic model using temperature is better than a constant proportion model, it may not give the best predictions. There is a significant lack of fit for the logistic model. This indicates that other curvilinear models may provide a better fit.

The analysis we have done is on the combined data. Combining the three separate observations at each temperature into one is analogous to averaging observations in ordinary least squares linear regression. We can do logistic regression on the 15, three for each temperature, separate observations (see exercise 3.2). The equation of the fitted logistic regression will be the same, however the deviances and degrees of freedom will change. The conclusions for this analysis, however, should be similar to those for the analysis of the combined data. ■

An alternative to the change in deviance for determining statistical significance of predictor variables in logistic regression is given by an approximate z-test statistic.

$$z = \frac{\text{estimated parameter}}{\text{standard error}}$$

This z-test statistic has an approximate standard normal distribution for large samples. For very large samples (another asymptotic result) the change in deviance and the square of the z-test statistic should give approximately the same value. In small to moderate size samples, the two statistics can give different results. When in doubt, use the change in deviance.

Example 3.8 For the logistic regression on the combined turtle data, output from the S+ general linear model analysis with family=binomial is given below.

Coefficients:

	Value	Std. Error	t value	
(Intercept)	-61.31828	12.0155501	-5.103244	
t	2.21103	0.4306079	5.134672	<-----

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 64.4285 on 4 degrees of freedom

Residual Deviance: 14.86295 on 3 degrees of freedom

The z-test statistic for the significance of temperature is $z = \frac{2.211}{0.4306} = 5.13$. This has an associated P-value of practically zero and indicates that temperature is a significant variable in the logistic regression model. Note, however, that the square of the z-test statistic is only about half that of the change in deviance. Obviously, we don't have large enough samples to get the agreement between these two test statistics.

In summary, both the z-test statistic and the change in deviance indicate that temperature is a highly significant variable in the logistic regression model for the proportion of male turtles. Sample sizes for the combined turtle data (the number of eggs at each temperature) are moderate, between 25 and 30. The P-values for either test will be approximate, at best. ■

3.5 Multiple logistic regression

Just as with multiple regression using ordinary least squares, we often have several explanatory (predictor) variables with a binary response. The ideas of logistic regression can be extended to the case with multiple explanatory (predictor) variables. As with multiple regression using ordinary least squares, we must be aware of problems introduced by multicollinearity (correlated predictor variables).

The general multiple logistic regression model is given in equation 3.10.

$$\pi_i = \frac{e^{(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}}{1 + e^{(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}} \quad (3.10)$$

The object of multiple logistic regression is to discover what combination of explanatory variables provides the best fit for the observed proportions. This task is complicated by correlations that may exist between explanatory variables. There are several selection procedures for meeting this objective.

Forward selection looks at each explanatory variable individually and selects the single explanatory variable that fits the data the best on its own as the first variable included in the model. Given the first variable, other variables are examined to see if they will add significantly to the overall fit of the model. Among the remaining variables the one that adds the most is included. This latter step (examining remaining variables in light of those already in the model and adding those that add significantly to the overall fit) is repeated until none of the remaining variables would add significantly or there are no variables remaining.

Backward selection starts with a model that contains all of the explanatory variables available. Each variable is examined to see if removing it from the model would significantly deteriorate the overall fit. The variable that, if removed, would cause the smallest change in the overall fit of the model is then removed. This continues until all variables in the model are significant.

A stepwise selection procedure is a combination of the forward and backward in that one can start with a forward selection procedure but after every step one checks to see if variables added early on have been made redundant (and so can be dropped) by variables or combinations of variables added later. Similarly, one can start with a backward selection procedure but after every step check to see if a variable that has been dropped should be added back into the model.

None of these procedures is guaranteed to find the overall “best” model. In fact, they can arrive at different final models with the same data. In order to exhaust all possibilities, all possible models can be fit and then candidate “best” models can be examined further.

Example 3.9 Returning to the broncopulmonary displasia data, recall that the binary response is 1 if broncopulmonary dysplasia (BPD) is present and 0 if BPD is absent. The explanatory (predictor) variables are the number of hours of exposure to Low, Medium and High levels of O_2 . Since these numbers are quite spread out, a natural log transformation is used. Since some values are zero, the log transformation is applied to the number of hours plus 1.

Low	Medium	High	BPD	Low	Medium	High	BPD
102	89	0	0	790	161	231	1
7	233	1	0	1142	157	131	1
0	4	41	0	0	2	49	0
8	37	13	0	0	0	50	0
40	79	26	0	5	68	49	0
0	625	156	1	0	0	48	0
0	12	79	0	0	6	40	0
0	3	119	0	1	8	64	0
115	136	65	0	0	998	551	1
428	416	435	1	253	99	60	0
34	174	56	0	1395	799	244	1
0	0	37	0	0	0	50	0
97	162	89	0	1	68	145	0
56	47	132	0	1318	1724	331	1
1214	1515	324	1	0	0	79	0
30	103	161	1	3	31	37	0
8	11	158	0	195	108	206	1
52	155	144	0	0	15	121	0
142	119	24	0	0	278	513	1
1370	2968	1083	1	0	0	253	1

Examination of the matrix plot in Figure 3.3 indicates that increasing the hours of oxygen, especially at the High level corresponds to a greater incidence of BPD. Further, the three explanatory variables (on the log scale) appear to be correlated.

Single predictor models are fit for each of the three explanatory variables. The results are summarized in the table below.

Predictor	$\ln L = \ln(\text{Low} + 1)$	$\ln M = \ln(\text{Medium} + 1)$	$\ln H = \ln(\text{High} + 1)$
Equation, $\hat{\pi}_i =$	$-1.9193 + 0.3822\ln L$	$-4.8411 + 0.9103\ln M$	$-55.6682 + 11.0679\ln H$
Null deviance	50.4465, df = 39	50.4465, df = 39	50.4465, df = 39
Residual deviance	42.4022, df = 38	34.1814, df = 38	10.0584, df = 38
Change in deviance	8.0443, df = 1	16.2651, df = 1	40.3881, df = 1
P-value	0.005	0.000	0.000

The single best predictor is $\ln H = \ln(\text{High} + 1)$. This results in the largest change in deviance, leaving the smallest residual deviance. Is this the “best” model? If one of the other explanatory variables is added to a single predictor models does this improve the overall fit of the model; reduce the residual

deviance significantly? Here is where the various selection procedures, such as forward selection and backward selection come in.

For forward selection, we start with a null (intercept only) model and add variables one at a time looking at the change in deviance to see if adding the variable causes a significant change. We can use S+ and the `anova()` command to fit and summarize the sequential addition of variables. By listing the variables in order, `lnL + lnM + lnH`, the following summary is obtained.

Model	Deviance	Change	P-value
Null	50.4465		
lnL	42.4022	8.0443	0.005
lnL + lnM	34.0134	8.3888	0.004
lnL + lnM + lnH	1.3409	32.6725	0.000

The final model contains all three explanatory variables. Each variable, when added to the model, reduces the residual deviance significantly. The form of the `anova()` command is

```
attach(bpd)
bpd.logistic.full <- glm(BPD ~ lnL+lnM+lnH,family=binomial)
anova(bpd.logistic.full,test='Chisq')
```

Rather than start with `lnL`, one might start with `lnH`, since it is the best single predictor. If `lnH` is entered first then adding either `lnL` or `lnM` will not significantly change the residual deviance. This forward selection would stop at this point. One of the difficulties with the automated selection procedures when you have multicollinear explanatory variables is that they often do not arrive at the same “best” model.

Model	Deviance	Change	P-value
Null	50.4465		
lnH	10.0584	40.3881	0.000
lnH + lnL	8.2717	1.7867	0.181
lnH + lnM	6.4784	3.5800	0.058

One could attempt a backward elimination procedure by putting all three explanatory variables into a model and then look what would happen to the fit if each of the variables was removed.

Model	Deviance	Change	P-value
lnL + lnM + lnH	1.3409		
lnM + lnH	6.4784	5.1375	0.023
lnL + lnH	8.2717	6.9308	0.008
lnL + lnM	34.0134	32.6725	0.000

Dropping any variable from the full model will significantly change the residual deviance giving a worse fit. This indicates that the model with all three variables is the best fit.

A stepwise selection procedure applied to these data will run into the same problems a forward selection procedure will in terms of stopping too soon. If one does a stepwise selection procedure starting with a backward elimination and checking to see if any of the previously eliminated variables might be added at a later time, one ends up with a final “best” model containing all three variables

■

3.6 Summary

In this chapter we have examined the analysis of binary response data. Binary response data abounds in many application areas and presents a unique problem because ordinary least squares simple linear regression is an inappropriate means of analysis. Performing simple linear regression on the logit transformed data corrects for the non-linear nature of the binary response but does not address the violation of equal variance and normality assumptions. The use of maximum likelihood estimation provides a means of working with binary response data. The general linear model procedure with a binomial family stipulation in S+ provides a convenient means of fitting simple and multiple logistic regression models. The fit of such models can be assessed by examining the residual deviance. Additionally, the significance of individual variables in the model can be assessed by looking at the change in deviance or a z-test. Caution is advised in interpreting the results of such tests, especially the P-values, as they are approximate, at best.

References

Miller, R. *et.al.*, (1980), *Biostatistics Casebook*, John Wiley & Sons.

Exercises

3.1 The data on the temperature at the time of launch and whether or not there were booster rocket field joint data for 23 shuttle launches is given below.

Temp	Damage	Temp	Damage	Temp	Damage
66	0	57	1	70	0
70	1	63	1	81	0
69	0	70	1	76	0
68	0	78	0	79	0
67	0	67	0	75	1
72	0	53	1	76	0
73	0	67	0	58	1
70	0	75	0		

- Plot the data and fit a simple linear regression model to the binary response.
- Fit a logistic regression model to these data and obtain a plot of the data and fitted curve. The S+ code below can be used to accomplish this.

```

challenger <- read.table("challenger.dat",head=T)
challenger
attach(challenger)
chal.logit <- glm(Damage~Temp,family=binomial)
summary(chal.logit)
cx <- c(50:80/1)
cyhat <- coefficients(chal.logit)[c(1)] +
         coefficients(chal.logit)[c(2)]*cx
cpihat <- exp(cyhat)/(1+exp(cyhat))
plot(Temp,Damage,xlab="Temperature",ylab="Damage",
      main="Incidence of Booster Field Joint Damage vs. Temperature")
lines(cx,cpihat)

```

- How well does the logisitc regression model fit these data? Consider the goodness of fit as indicated by the residual deviance. Also look at a plot of residuals.

3.2 The data on incubation temperature and number of male and female turtles from eggs collected in Illinois are given below.

Temp	male	female	% male	Temp	male	female	% male
	1	9	10%		7	3	70%
27.2	0	8	0%	28.4	5	3	63%
	1	8	11%		7	2	78%
	7	3	70%		10	1	91
27.7	4	2	67%	29.9	8	0	100%
	6	2	75%		9	0	100%
	13	0	100%				
28.3	6	3	67%				
	7	1	88%				

- (a). In example 3.4 we fit a logistic regression model to the combined turtle data. Use the complete data, 3 observed proportions for each incubation temperature, and fit a logistic regression model. How does this fit compare to that of the combined data? Look at the residual deviance as well as the fitted equation.
- (b). Is temperature significant in the logistic regression model using the complete data? Justify your answer statistically.

3.3 Example 3.9 discusses the data relating the incidence of bronchopulmonary displasia (BPD) to hours of oxygen treatment in 40 newborns. The data is given again below.

Low	Medium	High	BPD	Low	Medium	High	BPD
102	89	0	0	790	161	231	1
7	233	1	0	1142	157	131	1
0	4	41	0	0	2	49	0
8	37	13	0	0	0	50	0
40	79	26	0	5	68	49	0
0	625	156	1	0	0	48	0
0	12	79	0	0	6	40	0
0	3	119	0	1	8	64	0
115	136	65	0	0	998	551	1
428	416	435	1	253	99	60	0
34	174	56	0	1395	799	244	1
0	0	37	0	0	0	50	0
97	162	89	0	1	68	145	0
56	47	132	0	1318	1724	331	1
1214	1515	324	1	0	0	79	0
30	103	161	1	3	31	37	0
8	11	158	0	195	108	206	1
52	155	144	0	0	15	121	0
142	119	24	0	0	278	513	1
1370	2968	1083	1	0	0	253	1

- (a). Fit logistic regression models for each of the three explanatory variables; $\ln L = \ln(\text{Low} + 1)$, $\ln M = \ln(\text{Medium} + 1)$ and $\ln H = \ln(\text{High} + 1)$, separately.
- (b). Fit a logistic regression model that sequentially adds terms in the following order; $\ln M$, $\ln L$ and $\ln H$. According to this analysis what would be the “best” model? Justify your answer statistically.