3
Logistic Regression

We have seen from Examples 1.1–1.6 that the status of many health conditions is represented by a binary response. Because of its practical importance, analyzing a binary response has been the subject of countless works; see, e.g., the books of Cox and Snell (1989), Agresti (1990), and the references therein. For comparison purposes, we give a brief introduction to logistic regression.

3.1 Logistic Regression Models

Logistic regression is a standard approach to the analysis of binary data. For every study subject i we assume that the response Y_i has the Bernoulli distribution

$$I\!P\{Y_i = y_i\} = \theta_i^{y_i}(1-\theta_i)^{1-y_i}, \quad y_i = 0, 1, \ i = 1, \ldots, n, \tag{3.1}$$

where the parameters

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$$

must be estimated from the data. Here, a prime denotes the transpose of a vector or matrix.

To model these data, we generally attempt to reduce the n parameters in $\boldsymbol{\theta}$ to fewer degrees of freedom. The unique feature of logistic regression is to accomplish this by introducing the logit link function:

$$\theta_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}, \tag{3.2}$$

H. Zhang and B.H. Singer, *Recursive Partitioning and Applications*,
Springer Series in Statistics, DOI 10.1007/978-1-4419-6824-1_3,

where

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$$

is the new $(p+1)$-vector of parameters to be estimated and $(x_{i1}, \ldots, x_{ip})$ are the values of the p covariates included in the model for the ith subject $(i = 1, \ldots, n)$.

To estimate $\boldsymbol{\beta}$, we make use of the likelihood function

$$\begin{aligned} & L(\boldsymbol{\beta}; \mathbf{y}) \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})} \right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})} \right]^{1-y_i} \\ &= \frac{\prod_{y_i=1} \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{\prod_{i=1}^{n} [1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})]}. \end{aligned}$$

By maximizing $L(\boldsymbol{\beta}; \mathbf{y})$, we obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Although the solution for $\hat{\boldsymbol{\beta}}$ is unique, it does not have a closed form. The Newton–Raphson method, an iterative algorithm, computes $\hat{\boldsymbol{\beta}}$ numerically; see, e.g., Agresti (1990, Section 4.7).

The interpretation of the parameter $\boldsymbol{\beta}$ is the most attractive feature of the logit link function. Based on (3.2), the odds that the ith subject has an abnormal condition is

$$\frac{\theta_i}{1 - \theta_i} = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}).$$

Consider two individuals i and k for whom $x_{i1} = 1$, $x_{k1} = 0$, and $x_{ij} = x_{kj}$ for $j = 2, \ldots, p$. Then, the odds ratio for subjects i and k to be abnormal is

$$\frac{\theta_i/(1 - \theta_i)}{\theta_k/(1 - \theta_k)} = \exp(\beta_1).$$

Taking the logarithm of both sides, we see that β_1 is the log odds ratio of the response resulting from two such subjects when their first covariate differs by one unit and the other covariates are the same. In the health sciences, $\exp(\beta_1)$ is referred to as the adjusted odds ratio attributed to x_1 while controlling for $x_2, \ldots, x_p$. The remaining β's have similar interpretations. This useful interpretation may become invalid, however, in the presence of interactive effects among covariates.

3.2 A Logistic Regression Analysis

In this section we analyze the Yale Pregnancy Outcome data using logistic regression. Most statistical packages include procedures for logistic regression. We used SAS to perform the analysis. First, we start with a model that

includes all predictors in Table 2.1 as main effects and use the backward stepwise procedure to select variables that have significant (at the level of 0.05) main effects. Recall that preterm delivery is our response variable. For the selected variables, we then consider their second-order interactions.

In Table 2.1, three predictors, x_2 (marital status), x_3 (race), and x_{12} (hormones/DES use), are nominal and have five levels. To include them in logistic regression models, we need to create four (dichotomous) dummy variables for each of them. For instance, Table 2.1 indicates that the five levels for x_2 are currently married, divorced, separated, widowed, and never married. Let

$$z_1 = \begin{cases} 1 & \text{if a subject was currently married,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{if a subject was divorced,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{if a subject was separated,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_4 = \begin{cases} 1 & \text{if a subject was widowed,} \\ 0 & \text{otherwise.} \end{cases}$$

Likewise, let

$$z_5 = \begin{cases} 1 & \text{for a Caucasian,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_6 = \begin{cases} 1 & \text{for an African-American,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_7 = \begin{cases} 1 & \text{for a Hispanic,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_8 = \begin{cases} 1 & \text{for an Asian,} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$z_9 = \begin{cases} 1 & \text{if a subject's mother did not use hormones or DES,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_{10} = \begin{cases} 1 & \text{if a subject's mother used hormones only,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_{11} = \begin{cases} 1 & \text{if a subject's mother used DES only,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_{12} = \begin{cases} 1 & \text{if a subject's mother used both hormones and DES,} \\ 0 & \text{otherwise.} \end{cases}$$

Note here that the subject refers to a pregnant woman. Thus, z_9 through z_{12} indicate the history of hormones and DES uses for the mother of a pregnant woman.

TABLE 3.1. MLE for an Initially Selected Model

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.172	0.6912	0.0017
x_1(age)	1	0.046	0.0218	0.0356
z_6(Black)	1	0.771	0.2296	0.0008
x_6(educ.)	1	−0.159	0.0501	0.0015
z_{10}(horm.)	1	1.794	0.5744	0.0018

TABLE 3.2. MLE for a Revised Model

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.334	0.4583	0.0001
x_6(educ.)	1	−0.076	0.0313	0.0151
z_6(Black)	1	0.705	0.1688	0.0001
x_{11}(grav.)	1	0.114	0.0466	0.0142
z_{10}(horm.)	1	1.535	0.4999	0.0021

Due to missing information, 1,797 of the 3,861 observations are not used in the backward deletion step by `SAS PROC LOGISTIC`. Table 3.1 provides the key information for the model that is selected by the backward stepwise procedure. In this table as well as the next two, the first column refers to the selected predictors, and the second column is the degrees of freedom (DF). The third column contains the estimated coefficients corresponding to the selected predictors, followed by the standard errors of the estimated coefficients. The last column gives the p-value for testing whether or not each coefficient is zero. We should note that our model selection used each dummy variable as an individual predictor in the model. As a consequence, the selected model may depend on how the dummy variables are coded. Alternatively, one may want to include or exclude a chunk of dummy variables that are created for the same nominal variable.

The high proportion of the removed observations due to the missing information is an obvious concern. Note that the model selection is based on the observations with complete information in all predictors even though fewer predictors are considered in later steps. We examined the distribution of missing data and removed x_7 (employment) and x_8 (smoking) from further consideration because they were not selected in the first place and they contained most of the missing data. After this strategic adjustment, only 24 observations are removed due to missing data, and the backward deletion process produces another set of variables as displayed in Table 3.2.

We have considered the main effects, and next we examine possible (second-order) interactions between the selected variables. For the two se-

TABLE 3.3. MLE for the Final Model

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	−2.344	0.4584	0.0001
x_6(educ.)	1	−0.076	0.0313	0.0156
z_6(Black)	1	0.699	0.1688	0.0001
x_{11}(grav.)	1	0.115	0.0466	0.0137
z_{10} (horm.)	1	1.539	0.4999	0.0021

lected dummy variables, we include their original variables, race and hormones/DES uses, into the backward stepwise process to open our eyes a little wider. It turns out that none of the interaction terms are significant at the level of 0.05. Thus, the final model includes the same four variables as those in Table 3.2. However, the estimates in Table 3.2 are based on 3,837 (i.e., $3861 - 24$) observations with complete information for 13 predictors. Table 3.3 presents the information for the final model for which only 3 observations are removed due to missing information in the four selected variables. The different numbers of used observations explain the minor numerical discrepancy between Tables 3.2 and 3.3.

From Table 3.3, we see that the odds ratio for a Black woman (z_6) to deliver a premature infant is doubled relative to that for a White woman, because the corresponding odds ratio equals $\exp(0.699) \approx 2.013$. The use of DES by the mother of the pregnant woman (z_{10}) has a significant and enormous effect on the preterm delivery. Years of education (x_6), however, seems to have a small, but significant, protective effect. Finally, the number of previous pregnancies (x_{11}) has a significant, but low-magnitude negative effect on the preterm delivery.

We have witnessed in our analysis that missing data may lead to serious loss of information. As a potential consequence, we may end up with imprecise or even false conclusions. For example, by reviewing Tables 3.1 and 3.3, we realize that x_1 is replaced with x_{11} in Table 3.3 and the estimated coefficients for the remaining three predictors are notably different. The difference could be more dramatic if we had a smaller sample. Therefore, precaution should be taken in the presence of missing data. In Section 4.8, we will see that the tree-based method handles the missing data efficiently by either creating a distinct category for the missing value or using surrogate variables. These strategies prevent the tragic consequence of missing data.

Although it is not frequently practiced, we find it useful and important to evaluate the predictive performance of the final logistic model. To this end, we make use of ROC (receiver operating characteristic) curves (see, e.g., Hanley, 1989). We know that we cannot always make perfect classifications or predictions for the outcome of interest. For this reason, we want to

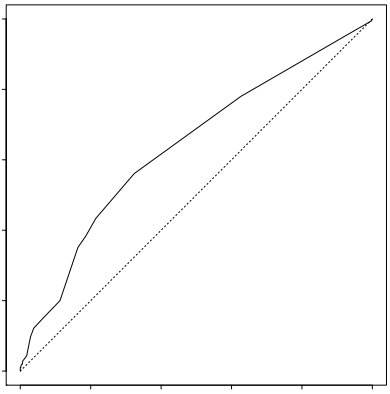

FIGURE 3.1. ROC curve for the final logistic regression model

make as few mistakes as possible. Two kinds of mistakes can occur when we predict an ill-conditioned outcome as normal or a normal condition as abnormal. To distinguish them, statisticians refer to these mistakes as type I and type II errors, respectively. In medical-decision making, they are called false-positive and false-negative diagnoses, respectively. In reasonable settings, these errors oppose each other. That is, reducing the rate of one type of error elevates the rate of the other type of error. ROC curves reflect both rates and quantify the accuracy of the prediction through a graphical presentation.

For subject i, we estimate her risk of having preterm delivery by

$$\hat{\theta}_i = \frac{\exp(-2.344 - 0.076x_{i6} + 0.699z_{i6} + 0.115x_{i,11} + 1.539z_{i,10})}{1 + \exp(-2.344 - 0.076x_{i6} + 0.699z_{i6} + 0.115x_{i,11} + 1.539z_{i,10})}, \tag{3.3}$$

$i = 1, \ldots, 3861$, using the estimates in Table 3.3. For any risk threshold r $(0 \leq r \leq 1)$, we calculate the empirical true and false-positive probabilities respectively as

$$TPP = \frac{\text{the number of preterm deliveries for which } \hat{\theta}_i > r}{\text{the total number of preterm deliveries}}$$

and

$$FPP = \frac{\text{the number of term deliveries for which } \hat{\theta}_i > r}{\text{the total number of term deliveries}}.$$

As r varies continuously, the trace of (TPP, FPP) constitutes the ROC curve as shown in Figure 3.1. In the medical literature, the true positive and negative probabilities are commonly referred to as sensitivity and specificity.

Figure 3.1 indicates that the final logistic regression model improves the predictive precision over a random prediction model. The latter predicts

the risk of 1 and 0 by tossing a fair coin. The ROC curve for this random prediction is featured by the dotted straight line. It is evident from Figure 3.1 that a great deal of variation is not explained and hence that further improvement should be sought.

Note also that the ROC curve is drawn from the resubstitution estimate of the risk, which tends to be optimistic in the sense that the ROC curve may have an upward-biased area. The reason is as follows. The prediction in (3.3) was derived to "maximize" the area under the ROC curve based on the Yale Pregnancy Outcome Study data. If we conduct another similar, independent study, which we call a validation study, it is almost sure that we will end up with an optimal prediction that differs from equation (3.3), although the difference may not be substantial. The other side of the coin is that if we make predictions for the subjects in the validation study from equation (3.3), the quality of the prediction is usually downgraded as compared to the prediction made for the original Yale Pregnancy Outcome Study. In some applications, validation studies are available, e.g., Goldman et al. (1982, 1996). In most cases, investigators have only one set of data. To assess the quality of the prediction, certain sample reuse techniques such as the cross-validation procedure are warranted (e.g., Efron, 1983). The cross-validation procedure will be heavily used in this book, specifically in Chapters 4 and 9–12. The basic idea is that we build our models using part of the available data and reserve the left-out observations to validate the selected models. This is a way to create an artificial validation study at the cost of reducing the sample size for estimating a model. The simplest strategy is to cut the entire sample into two pieces of equal size. While one piece is used to build a model, the other piece tests the model. It is a sample reuse mechanism because we can alternate the roles for the two pieces of sample.