Chapter 2. Fitting the Model

2.1 The Model

Functional Form

$$logit(p_\mathrm{i}) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

The data consists of $i = 1,2,\ldots,n$ cases. Each case belongs to one of two classes. A binary indicator variable represents the class label for each case

$$\mathrm{y}_i = \begin{cases} 1 & target\ event\ for\ case\ i \\ 0 & no\ target\ event\ for\ case\ i \end{cases}$$

Associated with each case is k-vector of input variables

$$\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$$

The posterior probability of the target event given the inputs is

$$\mathrm{p}_i = E(y_i|\mathbf{x}_i) = \Pr(\mathrm{y}_i = 1|\mathbf{x}_i)$$

The standard logistic regression model assumes that the logit of the posterior probability is a linear combination of the input variables. The parameters, $\beta_0, \ldots, \beta_k$, are unknown constants that must be estimated from the data.

The Logit Link Function

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta \Leftrightarrow p_i = \frac{1}{1+e^{-\eta}}$$

A linear combination can take any value. Probability must be between zero and one. The logit transformation (the log of the odds) is a device for constraining the posterior probability to be between zero and one. The logit function transforms the probability scale to the real line. Therefore, modeling the logit with a linear combination gives estimated probabilities that are constrained to be between zero and one.

Odds Ration from a Logistic Regression Model

Estimated logistic regression model:

$$\mathrm{logit(p)} = -0.7567 + 0.4373 * (\mathrm{gender})$$

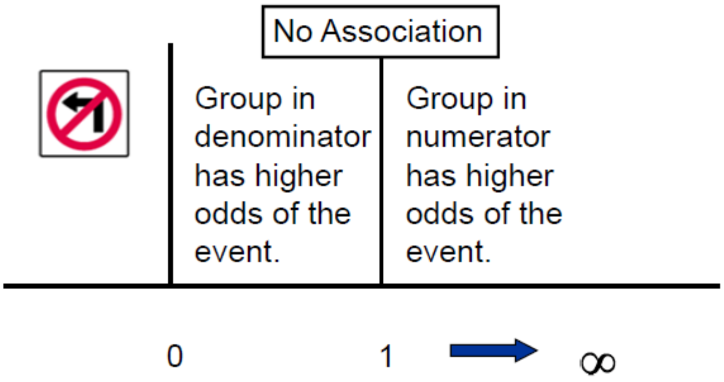
where females are coded 1 and males are coded 0

Estimated odds ratio (Females to Males):

$$\text{odds ratio} = \frac{\mathrm{e}^{-0.7567+0.4373}}{e^{-0.7567}} = 1.55$$

An odds ratio of 1.55 means that females have 1.55 times the odds of having the outcome compared to males.

Properties of the Odds Ratio


No Association
Group in denominator has higher odds of the event.
Group in numerator has higher odds of the event.
0
1
∞

The odds ratio shows the strength of the association between the predictor variable and the response variable. The odds ratio represents the multiplicative effect of each input variable. Moreover, the effect of each input variable does not depend on the values of the other inputs (additivity). However, this simple interpretation depends on the model being correctly specified. In predictive modeling, you should not presume that the true posterior probability has such a simple form.

Logistic Discrimination

In supervised classification, the ultimate use of logistic regression is to allocate cases to classes (termed *logistic discrimination*). An allocation rule is merely an assignment of a cutoff probability, where cases above the cutoff are allocated to class 1 and cases below the cutoff are allocated to class 0. The decision boundary is always linear. Determining the best cutoff is a fundamental concern in logistic discrimination.

Concordant versus Discordant

In logistic regression, several measures that assess the predictive accuracy of the model are reported. These measures are calculated from the percent concordant, discordant, and tied pairs.

concordant If the observation with the outcome has a **higher** predicted outcome probability compared to an observation without the outcome

discordant If the observation with the outcome has a **lower** predicted outcome probability compared to the predicted outcome probability of an observation without the outcome

tie If the predicted outcome probabilities are the same

LOGISTIC Procedure

```
PROC LOGISTIC <options>;
    CLASS variable </v-options>;
    MODEL response=<effects></options>;
    ODDSRATIO <'label'> variable </options>;
    ROC <'label'> <specification> </options>;
    ROCCONTRAST <'label'> <contrast> </options>;
    SCORE <options>;
    UNITS predictor1=list1 </options>;
    OUTPUT <OUT=SAS-data-set> keyword=name... keyword=name> </option>;
RUN;
```

The **LOGISTIC** procedure fits logistic regression models for binary, ordinal, or nominal response data.

CLASS specifies the classification variables to be used in the analysis. The **CLASS** statement must precede the **MODEL** statement.

MODEL specifies the response variable and the predictor variables. The **MODEL** statement is required, and only one is allowed with each invocation of **PROC LOGISTIC**.

ODDSRATIO produces odds ratios for variables even when the variables are involved in interactions with other covariates, and for classification variables that use any parameterization. You can specify several **ODDSRATIO** statements.

ROC specifies models to be used in the **ROC** comparisons. You can specify more than one **ROC** statement. **ROC** statements are identified by their label.

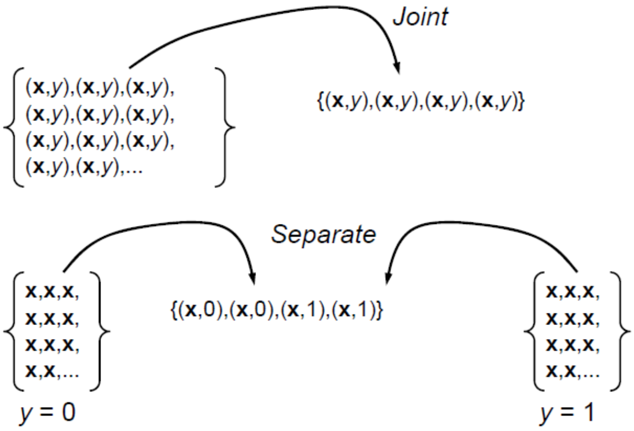
ROCCONTRAST compares the different **ROC** models. You can specify more than one **ROCCONTRAST** statement.

SCORE creates a data set that contains all the data in the **DATA=** data set together with posterior probabilities and, optionally, prediction confidence intervals. You can specify several **SCORE** statements.

UNITS enables you to obtain an odds ratio estimate for a specified change in a predictor variable. The unit of change can be a number, standard deviation or a number of times the standard deviation.

2.2 Adjustments for Oversampling

Sampling Designs



In *joint (mixture) sampling*, the input-target pairs are randomly selected from their joint distribution. In *separate sampling*, the inputs are randomly selected from their distributions within each target class.

Separate sampling is standard practice in supervised classification. When the target event is rare, it is common to oversample the rare event, that is, take a disproportionately large number of event cases. Oversampling rare events is generally done for data efficiency purposes.

The Effect of Oversampling

The maximum likelihood estimates were derived under the assumption that y_i have independent Bernoulli distributions. This assumption is appropriate for joint sampling but not for separate sampling. However, the effects of violating this assumption can be easily corrected, as in logistic regression only the estimate of the intercept (β_0) is affected by using a separate sampling design.

Consequently, the effect of oversampling is to shift the logits by a constant amount – the *offset*

$$\ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right)$$

When rare events have been oversampled $\pi_0 > \rho_0$ and $\pi_1 < \rho_1$, the offset is positive.

Offset

$$\mathrm{logit}(\mathrm{p}_i^*) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Because only the intercept is affected, the adjustments might not be necessary. If the goal of the analysis is to understand the relationships between the inputs and the target, or to rank order the population, then the adjustment is not critical. If the predicted probabilities are important, and not just necessary for rank ordering or classification, then the correction for oversampling is necessary.