Summary

Understanding the Logistic Regression Model

The logistic regression model looks for a probability, so the value must be between 0 and 1. In the logistic regression model equation, p_i represents the probability that y equals 1 given the inputs. The model has parameters β_1 through β_k that have to be estimated from the data. The probabilities have a nonlinear relationship with the input variables. Here is the functional form of the logistic regression model:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

The logit transformation linearizes the outcome of the logistic regression model. The logit transformation is the log of the odds of p_i. Here is the functional form of the logit transformation:

$$\text{logit}(p_i) = \ln(\frac{p_i}{1 - p_i}) = n_i$$

The linear relationship between the logit and the predictor variables leads to a linear response surface. However, the nonlinear relationship between the probabilities and the predictor variables leads to a nonlinear response surface.

The logistic regression results are often presented in terms of odds ratios. The odds ratio shows you the strength of the association between the input variable and the target variable. PROC LOGISTIC computes odds ratios for you.

Logistic regression is often used for classification. This type of analysis is more correctly called logistic discrimination. You typically use predictive models to categorize cases using a cutoff.

You use maximum likelihood estimation to estimate the parameters that are most likely to occur given the data and model assumptions. The likelihood function is a mathematical expression that computes the probability that the observed data would occur as a function of the parameters.

In PROC LOGISTIC output, goodness-of-fit measures include the percentage of concordant, discordant, and tied pairs. In general, higher percentages of concordant pairs and lower percentages of discordant and tied pairs indicate a more desirable model.

PROC LOGISTIC has many statements and options that you can use to control your output.

```
PROC LOGISTIC <options>;
     CLASS variable</v-options>;
     EFFECTPLOT <plot-type <(plot-definition-options)>> </options>;
     MODEL response=<effects></options>;
     ODDSRATIO <'label'> variable </options>;
     SCORE <options>;
     CODE <options>;
     UNITS <predictor1=list1> </option>;
RUN;
```

View the "Fitting a Basic Logistic Regression Model" demonstration that shows how to use PROC LOGISTIC to fit a basic logistic regression model.

One of the main purposes of predictive modeling is to score new cases. View the "Scoring New Cases" demonstration that shows three methods of scoring new cases using PROC LOGISTIC.

Correcting for Oversampling

When you fit a predictive model to a biased sample, you get skewed results that need to be corrected. When you correct for oversampling, you make the probabilities much, much smaller.

A response surface based on a biased sample has an intercept that's too high. The corrected plane on the population scale has an intercept that's lower. The difference between the two intercepts is called the offset. In order to get accurate predicted probabilities, you need to adjust for the offset.

View the "Correct for Oversampling" demonstration that shows how to correct for oversampling by specifying the proportion of the target event in the population.