Text Version

Expand Print

Fitting a Binary Logistic Regression Model

We already know that there is a significant association between the categorical predictor variable, Gender, and the categorical outcome variable, Purchase, in the Sales_Inc data set. This PROC LOGISTIC program helps us to characterize the relationship between Gender and Purchase.

```
ods graphics / width=700;

proc logistic data=statdata.sales_inc
              plots(only)=(effect);
   class Gender (param=ref ref='Male');
   model Purchase(event='1')=Gender;
   title1 'LOGISTIC MODEL (1):Purchase=Gender';
run;

title;
```

The PLOTS= option in the PROC LOGISTIC statement will display only the effect plot. The keyword ONLY suppresses the default plots. The CLASS statement specifies the categorical predictor variable Gender. Instead of the default parameterization method, which is effect coding, the CLASS statement specifies reference cell coding and Male as the reference level. In the MODEL statement, the event of interest is a value of 1 for Purchase, indicating a customer who spent at least $100 on merchandise.

We submit the program. The log shows that the code ran without errors. A note indicates the probability that PROC LOGISTIC is modeling. Let's look at the results. We look at the first few tables to make sure that the model is set up the way we want.

The Model Information table describes the data set, the response variable, the number of response levels, the type of model, the algorithm used to obtain the parameter estimates, and the number of observations read and used.

The Response Profile table shows the values of the response variable, listed according to their ordered value and frequency. By default, PROC LOGISTIC orders the values of the response variable alphanumerically and bases the logistic regression model on the probability of the lowest value. However, we set the EVENT= option to 1, the highest value, so this model is based on the probability that Purchase=1 (the probability that the person spent at least $100). Below this table, we see the probability that PROC LOGISTIC is modeling, as shown in the log.

The Class Level Information table displays the predictor variable in the CLASS statement: Gender. Our program specifies the PARAM=REF and REF='Male' options, so this table indicates that Male is the reference level. The design variable has the value 1 when Gender=Female and 0 when Gender=Male.

The Model Convergence Status simply indicates that the convergence criterion was met. There are a number of options to control the convergence criterion, but the default is the gradient convergence criterion with a default value of 1E-8.

The Model Fit Statistics table reports the results of three tests. AIC is Akaike's Information Criterion. SC is the Schwarz Criterion, which is also known as Schwarz's Bayesian Criterion. -2Log L is -2 times the log likelihood. AIC and SC are goodness-of-fit measures that you can use to compare one model to another. AIC and SC are not dependent on the number of terms in the model. For both AIC and SC, lower values indicate a more desirable model. AIC adjusts for the number of predictor variables, and SC adjusts for the number of predictor variables and the number of observations. SC uses a bigger penalty for extra variables and therefore favors more parsimonious models. If you are trying to come up with the best explanatory model, AIC is the best of the three statistics to use. If you are trying to come up with the best predictive model, SC is the best statistic to use. You can compare the AIC and SC columns for the model with the intercept only and the model with the intercept and the predictor variables. Remember that lower values indicate a more desirable model. If you are only considering these two models and if you want a better explanatory model, the AIC is lower for the second model, so the second model is better. If you want a better predictive model, the SC is lower for the first model, so the first model is better. In a few minutes, we'll use the second column to compare against other models with different variables to see which model is more desirable.

The table called Testing Global Null Hypothesis: BETA=0 provides three statistics to test the null hypothesis that all regression coefficients in the model are 0. We can look at the chi-square statistics and p-values to determine whether any of the parameter estimates (or betas) are significantly different from 0. The Likelihood Ratio test is the most reliable of the three tests, especially for small sample sizes. The Likelihood Ratio test statistic is similar to the overall F test in linear

regression. The Score and Wald tests are also used to test whether all the regression coefficients are 0. Here, all of the p-values are less than 0.05, so we would reject our null hypothesis and say that at least one of our regression coefficients is statistically different from 0.

The Type 3 Analysis of Effects table is generated when the CLASS statement specifies a categorical predictor variable. The Wald chi-square statistic tests the listed effect. Because Gender is the only predictor variable in the model, the value listed in the table will be identical to the Wald test in the Testing Global Null Hypothesis table.

Here's a question: Is the parameter estimate of Gender statistically different from 0? The parameter estimate is statistically different because the p-value is less than 0.05.

The Analysis of Maximum Likelihood Estimates table lists the estimated model parameters, their standard errors, Wald test statistics, and corresponding p-values. The parameter estimates are the estimated coefficients of the fitted logistic regression model. The regression coefficient (or parameter estimate) for β_0 (the intercept) is -0.7566. This is the logit of the probability that males will spend at least $100 because males are the reference level. The regression coefficient for β_1 is 0.4373. This is the difference in the logit of the probability of females and males spending at least $100. We can use the parameter estimates to construct the logistic regression equation, which is shown here. The Wald chi-square statistics and corresponding p-values test whether the parameter estimates are significantly different from 0. For this example, the p-values for both the intercept and the variable Gender are significant at the 0.05 significance level. The p-value for Gender is 0.0312, which is the same as the p-value in the Type 3 Analysis of Effects table. The significant p-value means that there is an association between Gender and Purchase, so females and males are statistically different from one another in terms of purchasing $100 or more.

The next table shows the odds ratio of females to males for the modeled event. Notice that PROC LOGISTIC calculates Wald confidence limits by default.

The Association of Predicted Probabilities and Observed Responses table lists several goodness-of-fit measures. We'll take a closer look at the information in these two tables after this demonstration.

Let's finish up by looking at the effect plot, which shows the levels of the CLASS predictor variable versus the probability of the desired outcome. In other words, this plot shows the probability of males versus females spending at least $100, as predicted by our model. The vertical axis represents the predicted probability of a customer spending at least $100. Gender is on the horizontal axis. Females have a higher predicted probability of spending at least $100 than males.

Close