Print

# Demo: Fitting a Binary Logistic Regression Model Using PROC LOGISTIC

Filename: **st107d04.sas**

Let's fit a binary logistic regression model in PROC LOGISTIC to characterize the relationship between the continuous variable Basement_Area and our categorical response, Bonus.

**PROC LOGISTIC DATA=***SAS-data-set <options>***;**
**MODEL** *variable <(variable_options)> = <effects> < / options>***;**
**RUN;**

1. Open program st107d04.sas.

   ```
   /*st107d04.sas*/
   ods graphics on;
   proc logistic data=STAT1.ameshousing3 alpha=0.05
                 plots(only)=(effect oddsratio);
       model Bonus(event='1')=Basement_Area / clodds=pl;
       title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
   run;
   ```

   The PROC LOGISTIC statement specifies the amehousing3 data set and has several options. The PLOTS= option requests only the EFFECT and ODDSRATIO plots. The ALPHA= option requests confidence intervals for each parameter estimate.

   The MODEL statement specifies the response variable Bonus, and in parentheses, the variable option EVENT= specifies the event category for the binary response. PROC LOGISTIC then models the probability of the event category you specify. In this example, the event category is the value 1 for Bonus, which indicates a Bonus Eligible home. If you don't include this option, event=0 would be modeled instead, because it's the first level in alphanumeric order. You then specify an equal sign, followed by the predictor variable, Basement_Area.

   After the forward slash, you use the CLODDS= option to compute confidence intervals for the odds ratios of all predictor variables. Following the equal sign, you specify a keyword to indicate the type of confidence interval: PL for profile likelihood, WALD, or BOTH. If you don't specify the CLODDS= option, PROC LOGISTIC computes Wald confidence intervals by default. Wald statistics require fewer computations to perform. Profile-likelihood confidence intervals are desirable for small sample sizes. The CLODDS= option also enables the production of the odds ratio plot that's specified in the PLOTS= option.

2. Submit this program.

3. Review the output.

   The Model Information table describes the data set, the response variable, the number of response levels, the type of model, and the algorithm used to obtain the parameter estimates. The Optimization Technique is the iterative numerical technique that PROC LOGISTIC uses to estimate the model parameters. The model is assumed to be binary logit when there are exactly two response levels.

   In the Observation Summary, the Number of Observations Used is the count of all nonmissing observations. In this case, there were no missing observations for the variables specified in the MODEL statement.

The Response Profile table shows the response variable values listed according to their ordered values. Because we used the EVENT= option in this example, the model is based on the probability of being bonus eligible (Bonus=1). This table also shows frequencies of response values. In this sample of 300 homes, only 45 sold for more than $175,000, and 255 sold for less.

Next, you should always check that the modeled response level is the one you intended. Otherwise your interpretation of the model will be erroneous.

The Model Convergence Status simply indicates that the convergence criterion was met, and there are many options to control the convergence criterion. The optimization technique doesn't always converge to a maximum likelihood solution. When this is the case, the output after this point cannot be trusted. Always check to see that the convergence criterion is satisfied.

The Model Fit Statistics table reports the results of three tests: AIC, SC, which is also known as Schwarzs Bayesian Criterion, or SBC, and -2 Log L, which is -2 times the natural log of the likelihood. The AIC, SC, and -2 Log L are goodness-of-fit measures. These statistics measure relative fit and are used only to compare models. They do not measure absolute fit of any single model. Smaller values for all these measures indicate better fit. However, -2 Log L can be reduced by simply adding more regression parameters to the model. Therefore, it's not used to compare the fit of models that use different numbers of parameters except for comparisons of nested models using likelihood ratio tests. 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.

The Global Tests table, Testing Global Null Hypothesis: BETA=0, provides three statistics to test the null hypothesis that all regression coefficients of the model are 0. A significant pvalue for these tests provides evidence that at least one of the regression coefficients for a predictor variable is significantly different from 0. In this way, they are like the overall F test in linear regression. The Likelihood Ratio Chi-Square value is calculated as the difference between the -2 Log L value of the baseline model (intercept only) and the -2 Log L value of the hypothesized model (intercept and covariates).

The degrees of freedom are equal to the difference in number of parameters between the hypothesized model and the baseline model. In this case, there's only one additional predictor, Basement_Area, compared to the intercept-only model. The Score and Wald tests are also used to test whether all the regression coefficients are 0, and all three tests are asymptotically equivalent and often give very similar values. However, the Likelihood Ratio test is the most reliable, especially for small sample sizes.

The Parameter Estimates table, Analysis of Maximum Likelihood Estimates, lists the estimated model parameters, their standard errors, Wald Chi-Square values, and p-values. The parameter estimates are the estimated coefficients of the fitted logistic regression model. For this example, the logistic regression equation is logit(p-hat) = -9.7854 (0.00739) * Basement_Area.

The Wald Chi-Square and its associated p-value tests whether the parameter estimate is significantly different from 0. The p-value for the variable Basement_Area is significant at the 0.05 alpha level.

The estimated model is displayed on the probability scale in the effect plot. You can see the sigmoidal shape of the estimated probability curve and that the probability of being bonus eligible increases as the basement area increases.

We'll take a closer look at the information in the last two tables and plot after this demonstration.

Close