Print

Demo: Fitting a Multiple Logistic Regression Model with Categorical Predictors Using PROC LOGISTIC

Filename: **st107d05.sas**

In this demonstration, we'll fit a multiple logistic regression model that characterizes the relationship between the response Bonus, and the variables Basement_Area, Lot_Shape_2, and Fireplaces.

PROC LOGISTIC DATA=*SAS-data-set <options>*;
    CLASS *variable <(options)> ... < / options>*;
    MODEL *variable <(variable_options)> = <effects> < / options>*;
    UNITS *<independent1=list1> ... < / options>*;
RUN;

1. Open program st107d05.sas.

```
/*st107d05.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3 plots(only)=(effect oddsratio);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2 / clodds=pl;
    units Basement_Area=100;
    title 'LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces Lot_Shape_2';
run;
```

   In the PROC LOGISTIC statement, we specify the ameshousing3 data, and the PLOTS= option specifies an effect plot and an odds ratio plot. The CLASS statement specifies the categorical predictors Fireplaces and Lot_Shape_2 and the parameterization method. By default, PROC LOGISTIC uses effect coding. To specify reference cell coding, you use the PARAM= option.

   As the default reference level for either effect coding or reference cell coding, PROG LOGISTIC uses the last level in alphanumeric order. To specify a reference level, you use the REF= variable option. You can specify the actual value, or level, in quotation marks, or the keyword FIRST or LAST. In this example, Fireplaces has a reference level of 0, and Lot_Shape_2 has a reference level of Regular.

   The MODEL statement is the same as the previous demonstration, but it now specifies all three predictor variables.

   The UNITS statement enables you to obtain customized odds ratio estimates for a specified unit of change in one or more continuous predictor variables. By default, PROC LOGISTIC finds odds ratios for a one-unit change in the continuous predictors. However, for a variable like Basement_Area, it doesn't make sense to find the odds ratios between two homes where one has only a single square foot larger basement area. Instead, we might want to find the odds ratios for homes with a difference of 100 square feet in basement area, because it's more interpretable. For each continuous predictor, you specify the variable name, an equal sign, and a list of one or more units of change, separated by spaces. In this example, we specify the number 100 as the unit of change for the continuous predictor Basement_Area.

2. Submit the code.

3. Review the output.

   The Model Information and Response Profile are the same as for the binary logistic regression model that we ran in the previous demonstration. It's useful to look at this information to make sure that your model is set up as you intended. The statement below the table indicates that, once again, we're modeling the probability of

being bonus eligible (Bonus=1).

The Class Level Information table includes the predictor variables in the CLASS statement. Because we used the PARAM=REF and REF='Regular' options, this table reflects our choice of Lot_Shape_2='Regular' as the reference level. The design variable is 1 when Lot_Shape_2='Irregular' and 0 when Lot_Shape_2='Regular'. The reference level for Fireplaces is 0, so there are two design variables, each coded 0 for observations where Fireplaces=0.

The SC value in the Basement_Area only model was 169.246. Here in the Fit Statistics table, it's 159.001. Recalling that smaller values imply better fit, you can conclude that this model fits better.

In the Global Tests table, Testing Global Null Hypothesis: Beta=0, we see that this model is statistically significant, indicating at least one of the predictors in the model is significantly different from zero.

The Type 3 Analysis of Effects table is generated when you use the CLASS statement. This table displays the significance of each of the effects individually, adjusting for other predictors included in the model. Fireplaces is not statistically significant at the 0.05 level, but the other remaining predictors are statistically significant. Notice for Fireplaces, there are two degrees of freedom. This is because Fireplaces has three levels and two dummy variables, or effects in the model. This table tests for the significance of Fireplaces, and therefore, tests for the significance of both of its effects jointly. The continuous predictor Basement_Area only has one degree of freedom, and Lot_Shape_2 has two levels, and therefore, only 1 dummy variable and degree of freedom.

Next let's look at the Parameter Estimates table, Analysis of Maximum Likelihood Estimates. For CLASS variables, effects are displayed for each of the design variables. Because reference cell coding was used, each effect is measured against the reference level. For example, the estimate for Lot_Shape_2 | Irregular shows the difference in logits between houses with Irregular and Regular lot shapes, which is 1.9025. Fireplaces | 1 shows the logit difference between houses with 1 fireplace and 0 fireplaces, while Fireplaces | 2 shows the difference in logits between houses with 2 fireplaces and 0 fireplaces. Not all of these effects are statistically significant. Notice that the Wald Chi-Square values and p-values are the same in both tables for Basement_Area and Lot_Shape_2. This is because they only have one degree of freedom. Whereas each effect for Fireplaces is tested separately in this table, so the chi-square values are different than the type 3 analysis of effects.

In the Association of Predicted Probabilities and Observed Responses table, the c statistic value is 0.930, and the percent concordant is approximately 93% for this model, indicating that 93% of the positive and negative response pairs are correctly sorted using Basement_Area, Fireplaces, and Lot_Shape_2.

The Odds Ratio Estimates table shows that, adjusting for other predictor variables, homes with Irregular lots had 6.703 times the odds of being bonus eligible than homes with Regular lots. Homes with 1 fireplace had nearly 2.5 times the odds of being bonus eligible than homes with 0 fireplaces, and homes with 2 fireplaces had less than half the odds than the homes with 0 fireplaces. The table shows that for each 100-square-foot increase in basement area, the bonus eligibility odds increase by 110.5%. Notice that the confidence intervals for the Basement_Area and Lot_Shape_2 odds ratios do not cover a value of 1, indicating the odds ratios are statistically significant for these two variables. However, the odds ratios for Fireplaces both cover 1, and therefore, they are not statistically significant.

The ODDSRATIO plot displays these confidence intervals graphically, and you can see which intervals cover the vertical line at a value of 1.

The final plot is an effect plot of Predicted Probablities for Bonus=1. It shows the estimated probability of a bonus eligible home across different basement areas, given the different combinations of the categorical variables Lot_Shape_2 and Fireplaces. For example, we can see that the left most probability curve corresponds to homes with 1 fireplace and an irregular lot shape. This means that as the values of Basement_Area increase, the probability of being bonus eligible increases faster than any other combination of the categorical variables. However, as the Basement_Area increases past 1500 square feet, each curve shows a high probability of the home selling for more than $175,000.

Close