Comparing Pairs to Assess the Fit of a Logistic Regression Model

If a logistic regression model predicts its own data accurately, then we say that the model fits the data well. PROC LOGISTIC calculates several different goodness-of-fit measures, in addition to the AIC, SC, and -2 log L, and displays the resulting statistics in the Association of Predicted Probabilities and Observed Responses table. Here's the table from the last demonstration. Let's take a closer look at how PROC LOGISTIC compares pairs.

We're analyzing the relationship between the continuous predictor Basement_Area and the binary response Bonus. To start, PROC LOGISTIC creates two groups of observations, one for each value of Bonus. One group contains all the observations in which the value of Bonus is 0, meaning the homes that are not bonus eligible. The other group contains all the observations in which the value of Bonus is 1, the homes that are bonus eligible. PROC LOGISTIC then selects pairs of observations, one from each group, until no more pairs can be selected, and then determines whether each pair is concordant, discordant, or tied. Let's look at each type of pair.

Suppose a pair consists of a home with a 1200-square-foot basement that is bonus eligible and a home with an 800-square-foot basement that is not. The home with a smaller basement area has a lower predicted probability of being bonus eligible, .0204, than the home with a larger basement area, .2865. A pair is concordant when the observation with the event, in this case, bonus=1, has a higher predicted probability of having the event than the observation without the event. That is, the model assigns a higher probability of being bonus eligible to the house that is bonus eligible than to the house that isn't. When the model sorts the pair correctly, as in this case, the pair is concordant.

Let's look at another pair. This pair compares a home with a 1400-square-foot basement that is bonus eligible with a home with a 1600-square-foot basement that is not bonus eligible. From our model, we know that larger basement areas have a higher predicted probability, .8855, of being bonus eligible, than smaller basement areas, .6379. However, our model did not sort this pair correctly. A pair is discordant if the observation with the desired outcome has a lower predicted probability than the observation without the outcome.

The last pair compares two homes with 1350-square-foot basements. One home is bonus eligible, and the other is not. According to our model, both have a predicted probability of .5490, so our model cannot distinguish between them. A pair is tied if it is neither concordant nor discordant, that is, the probabilities are the same.

Let's take a look at the Association of Predicted Probabilities and Observed Responses table.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	**Somer's D**	0.791
Percent Discordant	10.4	**Gamma**	0.792
Percent Tied	0.1	**Tau-a**	0.202
Pairs	11475	**c**	0.896

Response Profile		
Ordered Value	**Bonus**	**Total Frequency**
1	0	255
2	1	45

The left column lists the percentage of pairs of each type: concordant, discordant, and tied. Pairs represents the total number of observation pairs on which the percentages are based. The value is calculated as 255, the frequency of values that are not bonus eligible, times 45, the frequency of values that are bonus eligible, for a total of 11475 pairs of observations with different outcome values.

You can use the percentages of concordant, discordant, and tied pairs as goodness-of-fit measures to compare one model to another. In general, higher percentages of concordant pairs and lower percentages of discordant and tied pairs indicate a more desirable model. This table also shows the four rank correlation indices that are computed from the numbers of concordant, discordant, and tied pairs of observations: Somers' D, Gamma, Tau-a, and c. In general, a model with higher values for these indices has better predictive ability than a model with lower values.

The c value is the most commonly used. The c, concordance statistic, estimates the probability of an observation with the event having a higher predicted probability than an observation without the event. The c value is calculated as the number of concordant outcomes plus one half times the number of ties divided by the total number of pairs. The range of possible values is 0.5 to 1.0, where 1.0 is perfect prediction. The value of 0.896 shows a very strong ability of Basement_Area to discriminate between homes that are bonus eligible and homes that are not.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Close