BIOS 6244 Analysis of Categorical Data
November 9, 2005 Lecture

Model Checking (Sec. 5.3)

Just because we can calculate the ML estimates of the logit GLM parameters does not mean that the model accurately represents the relationship between X & Y. Once we have fit the model to the data, we should perform several checks of the model to verify that we have an adequate fit. If we do not have a model that fits the data well, we should continue to examine alternative models until we find one that does fit.

We have already mentioned certain "eyeball" methods that help us determine if we have a good fitting model. One such method involves overlaying a graph of the fitted curve on top of a scatterplot of the observed data (or a scatterplot constructed using grouped data). Another involves an informal comparison of the observed vs. fitted probability values or frequencies within each category of grouped X values. In this section, we describe several objective methods that can be used to test the GOF of a logistic regression (LR) model.

One of the simplest methods for checking *any* theoretical model is to use Pearson's χ^2 or the likelihood ratio test to compare observed & expected frequencies. At each combination of values of the explanatory variables in the LR model, we have an observed frequency of all subjects for which Y = 1 and all subjects for which Y = 0. The model also produces fitted (or expected) frequencies for Y = 1 and Y = 0 at each of these combinations of values of the explanatory variables.

Conditional on the observed combinations of values of the explanatory variables, the X^2 and G^2 test statistics have approximate χ^2 distributions (assuming that the expected frequencies are ≥ 5 for most of the combinations). The degrees of freedom (called the *residual df*) for the model = # of sample logits - # of model parameters. As usual, large values of X^2 and G^2 indicate poor agreement between observed and expected; in our case, this is equivalent to poor GOF of the hypothesized model. We are only interested in an upper-tailed test based on the right-tailed p-value. If we detect a poor fitting model, we can use residuals and other diagnostic measures to examine the *influence* of individual observations on the model fit and to help identify possible reasons for the inadequate fit of the model.

GOF for Models with Continuous Predictors (Sec. 5.3.1)

In this section, we consider a LR model in which X = carapace width is the only explanatory variable:

$$\text{logit}\,[\pi(x)] = \alpha + \beta x. \qquad (21)$$

For the 173 female horseshoe crabs in the dataset listed in Table 4.2 (now available on the course website), there are 66 distinct values for width, with very few crabs having the same carapace width. One could regard this as a 66x2 contingency table in which the 2 columns are defined by

Y = 1 (at least 1 satellite) and Y = 0 (no satellites). As expected, the cell counts in this table tend to be small, as are the fitted counts produced by the LR model.

The large-sample theory for the χ^2 and likelihood ratio tests applies when there is a fixed (i.e., known prior to the time of data collection) number of cells, all of which have "reasonably large" (≥ 5) fitted counts. However, for the 66x2 table based on the horseshoe crab data, there are 2 features that adversely affect the performance of the χ^2 & likelihood ratio tests: (1) the fitted cell counts tend to be small and (2) if more horseshoe crab data were added to the sample, there would likely be additional carapace widths that had not appeared previously, and this would add rows to the 66x2 table. In other words, the # of cells in the contingency table is not fixed. Because of these 2 features (which would be a problem for any LR model with a continuous explanatory variable unless n were huge), the χ^2 approximations for the distributions of X^2 and G^2 tend to be very poor. The χ^2 works best when the explanatory variable is discrete and relatively few of the fitted counts are small.

Instead of basing the X^2 or G^2 test statistics on the observed values of the explanatory variables, it is better to group these values into intervals, as we have already done for the horseshoe crab data in Table 5.1. In Table 5.2, the observed and fitted cell frequencies are presented for the 8x2 contingency table formed from the 8 carapace width intervals originally chosen by Agresti.

Table 5.2 Grouping of Observed and Fitted Values for Fit of Logistic Regression Model to Horseshoe Crab Data

Width	Number Yes	Number No	Fitted Yes	Fitted No
< 23.25	5	9	3.64	10.36
23.25-24.25	4	10	5.31	8.69
24.25-25.25	17	11	13.78	14.22
25.25-26.25	21	18	24.23	14.77
26.25-27.25	15	7	15.94	6.06
27.25-28.25	20	4	19.38	4.62
28.25-29.25	15	3	15.65	2.35
> 29.25	14	0	13.08	0.92

In each row of this table, the fitted value for a "yes" response (i.e., 1 or more satellites) is obtained by summing all of the predicted probabilities $\widehat{\pi(x)}$ for all crabs having a carapace width in that interval. Similarly, the the fitted value for a "no" response (i.e., no satellites) is obtained by summing 1- $\widehat{\pi(x)}$ for those crabs. As expected, we obtain much larger fitted values for the cells of this 8x2 table than we did for the 66x2 table prior to grouping the carapace widths and the χ^2 approximations are therefore expected to be more accurate.

Substituting the observed and fitted counts in Table 5.2 into the usual X^2 and G^2 test statistics [see p. 26 of these lecture notes or Equation (2.4.3) in our textbook], we obtain $X^2 = 5.3$ and $G^2 = 6.2$. To determine the residual df for the χ^2 approximations, we count the number of sample logits [i.e., the number of sample values of $\pi(x)$] and then subtract the # of parameters in

the LR model. For the data in Table 5.2, there are 8 sample logits (1 for each width category) and the LR model has 2 parameters (α & β), so df = 8 – 2 = 6. Using SimCalc, we find the approximate p-value for the χ^2 test to be $Pr(X^2 \geq 5.3 \mid df = 6) = .506$ and for the likelihood ratio test, it is $Pr(G^2 \geq 6.2 \mid df = 6) = .401$. Thus, there is no evidence of lack of fit and we can feel justified in using the LR model for this set of data.

A simpler (but also more approximate) method for testing the GOF is to fit the LR model directly to the observed counts in the contingency table based on the grouped data. For example, for the data in Table 5.2, we would fit the LR model in Equation (21) using the logit of the observed proportion of "success" in each width interval as the outcome variable and a score for each row as the explanatory variable. Agresti recommends using the mean width in each interval as the score for that row. These mean widths are {22.69, 23.84, 24.77, 25.84, 26.79, 27.74, 28.67, 30.41}. The LR model fitted to the grouped horseshoe crab data in this way is given by

$$\text{logit}\,[\widehat{\pi(x)}\,] = -11.51 + .465x. \qquad (22)$$

(Compare this LR model with the one fit to the original data: $\text{logit}\,[\widehat{\pi(x)}\,] = -12.35 + .497x$.)

The χ^2 & likelihood ratio test statistics for the model in Equation (22) are $X^2 = 5.0$ and $G^2 = 6.0$, which are very similar to those obtained for the model that was fit to the original data ($X^2 = 5.3$ and $G^2 = 6.2$).

When explanatory variables in LR models are continuous (as in the horseshoe crab example with carapace width as the predictor), it is difficult to assess lack of fit of the LR model without grouping the values of X in some way. As the # of explanatory variables increases, the cross-classifications of the intervals for the grouped versions of all of the explanatory variables are likely to result in a multi-dimensional contingency table with a large # of cells and hence a strong likelihood of having several expected cell frequencies that are too small.

An alternative method of grouping is based on a *partition* of the fitted (or predicted) probabilities. One can regard the grouping in Table 5.2 as having been done in this way (since there was only a single explanatory variable). For the fitted LR model, the 14 crabs in the 1st width interval are the ones with the smallest fitted probability of at least 1 satellite, the 14 in the 2nd width interval have the 2nd smallest fitted probability, etc. Regardless of how many explanatory variables there are in the model, one can always group the study subjects according to the fitted probabilities. One commonly used approach is to group the fitted probabilities so that all of the intervals have approximately the same # of subjects in them. To form 10 intervals of approximately equal size, for example, one starts with the n/10 subjects having the smallest fitted probabilities; the next interval then consists of the subjects in the 2nd decile of fitted probabilities, etc. For each interval in the partition, the fitted frequency is obtained by summing the fitted probabilities for all study subjects that fall in that interval.

Hosmer & Lemeshow, authors of one of the classic texts on logistic regression, proposed a GOF test for LR models using such a partition of the fitted probabilities. Their Pearson-like test

statistic does not have a χ^2 distribution, but simulations have shown that a χ^2 approximation with df = # of intervals – 2 works reasonably well. Agresti applied the H-L test using 10 intervals to the horseshoe crab data and obtained a test statistic of 3.5. Using a $\chi^2(8)$ distribution, the approximate p-value is .899, indicating no evidence whatsoever of lack of fit of the LR model.

In Section 5.5, we will consider another approach to GOF testing for LR models in which methods are used to compare a simple LR model like the one we are considering for the horseshoe crab data with a more complex one (e.g., one that includes higher order terms for X). If none of the more complex models provide a significantly better fit, then we have additional assurance that our simple fitted model is reasonable. In some ways, this approach is preferable since Pearson-type GOF tests can only tell us that our fitted model appears to be inadequate; they provide no guidance whatsoever about the nature of the lack of fit.

GOF & Likelihood-Ratio Model Comparison Tests (Sec. 5.3.2)

In Section 5.2.2, we briefly discussed the likelihood-ratio test statistic $-2(L_0 - L_1)$ that can be used to test whether certain parameters in a model are equal to 0. The test compares the maximized log-likelihood (L_1) for the fitted model to the maximized log-likelihood (L_0) for a simpler model that omits the parameters in the fitted model (i.e., these parameters are set equal to 0 rather than estimated via ML).

Denote the fitted model by M_1 and the simpler model in which the parameters are set equal to 0 by M_0. The GOF test statistic G^2 for testing the adequacy of a LR model M is a special case of the likelihood-ratio test in which $M_0 = M$ and M_1 is the most complex LR model possible. This complex model has a separate parameter for each logit and provides a perfect fit to the sample logits. It is called the *saturated model* and has residual df = 0. In testing the adequacy of our model M, we test whether *all* parameters that are in the saturated model but not in our model M are equal to 0. Thus, if we fail to reject the null hypothesis, we can conclude that our model M is adequate since none of the additional parameters in M_1 are required.

Denote the test statistic for testing the fit of our LR model M by $G^2(M)$. In GLM terminology, this test statistic is called the *deviance* of the model. Let L_S denote the maximized log-likelihood for the saturated model. Then the deviances for models M_0 and M_1 are $G^2(M_0) = -2(L_0 - L_S)$ & $G^2(M_1) = -2(L_1 - L_S)$.

Denote the likelihood-ratio statistic for testing M_0, given that M_1 holds, by $G^2(M_0|M_1)$. This statistic is equal to $-2(L_0 - L_1)$. By adding and subtracting L_S, we see that

$$G^2(M_0|M_1) = -2(L_0 - L_1) = -2(L_0 - L_S) - [-2(L_1 - L_S)] = G^2(M_0) - G^2(M_1),$$

which is the difference in the G^2 GOF statistics for the 2 models. In other words, the likelihood-ratio statistic for comparing the 2 models is just the difference in the deviances. This statistic is large when model M_0 fits poorly compared to model M_1. The likelihood-ratio test statistic has a χ^2 distribution for large df, with df = the difference between the residual df for the 2 models being compared.

Example (horseshoe crab data, revisited)

M_1: LR model with carapace width as the only explanatory variable:

$$\text{logit}[\pi(x)] = \alpha + \beta x.$$

M_0: simpler LR model in which the parameter we wish to test has been omitted:

$$\text{logit}[\pi(x)] = \alpha.$$

Assuming that model M_0 is correct is equivalent to hypothesizing that having at least 1 satellite is independent of carapace width. Thus, the G^2 GOF test for model M_0 is just the usual likelihood ratio test for testing independence in a 2-way contingency table. For the observed counts presented in the 8x2 contingency table derived from the data in Table 5.2, $G^2(M_0) = 34.0$ (df = 7, $p < .0001$).

For model M_1, we have already seen that $G^2(M_1) = 6.0$ (df = 6).

Thus, the likelihood-ratio test statistic for comparing the models M_1 and M_0 is given by

$$G^2(M_0|M_1) = G^2(M_0) - G^2(M_1) = 34.0 - 6.0 = 28.0 \quad (df = 7 - 6 = 1)$$

Using a χ^2 with 1 df to calculate an approximate p-value for an observed test statistic of 28.0, we see that $p < .0001$. This indicates extremely strong evidence that the parameter β is needed in the LR model. This test is equivalent to performing the likelihood-ratio test of $H_0: \beta = 0$ in the LR model fitted to the grouped data in Table 5.2.