BIOS 6244 Analysis of Categorical Data
November 14, 2005 Lecture

Residuals for Logit Models (Sec. 5.3.3)

Goodness-of-fit tests such as Pearson's χ^2 and the likelihood-ratio test are useful for assessing the overall fit of a GLM to the observed data. However, if lack of fit is detected, additional analyses should be performed in order to gain an understanding of the possible cause(s) of the lack of fit. The *residuals*, i.e., the differences between the observed and fitted values, are useful for this purpose.

Let y_i denote the observed number of "successes" out of n_i trials at the i'th combination of values of the explanatory variables $X_{1,}$ X_2, …, X_K. (Agresti calls each of these combinations a *setting*.) Let $\widehat{\pi}_i$ denote the fitted (predicted) probability of success at this setting. Then $n_i \widehat{\pi}_i$ is the fitted # of successes at the i'th setting. For a GLM with binomial random component, the *Pearson residual* for the fit at the i'th setting is

$$e_i = \frac{y_i - n_i \widehat{\pi}_i}{\sqrt{n_i \widehat{\pi}_i (1 - \widehat{\pi}_i)}} .$$

Note that the denominator is the estimated SE of of the numerator, provided that the binomial is the correct random component. Thus, when n_i is large enough for the normal approximation to the binomial to be valid [usually both $n_i \widehat{\pi}_i$ and $n_i (1 - \widehat{\pi}_i)$ ≥ 5], then each e_i has an approximate standard normal distribution and is interpreted as if it were a z-score, i.e., any $| e_i | > 2$ indicates a possible lack of fit.

Note that for continuous explanatory variables, it is likely that $n_i = 1$ for many settings of the explanatory variables. In that case, $y_i = 0$ or 1 and the corresponding residual e_i has only 2 possible values. Thus, examining these individual residuals as an indicator of lack of fit is usually not very productive. If the data are grouped (using either the values of a single explanatory variable or the fitted values if there is more than 1 explanatory variable), the residuals will typically be much more informative.

Table 5.3 shows the Pearson residuals for the fit of 2 different LR models to the grouped horseshoe crab data: the model in Equation (21) with width as a single explanatory variable and the same model with the coefficient of width, β, set equal to zero. (The latter model is sometimes referred to as the *independence model*, since it is equivalent to saying that having at least 1 satellite is independent of carapace width.) In Table 5.3, we see that some of the residuals for the independence model (denoted with a superscript of *a* in the Table) exceed the 2.0 threshhold in absolute value. Furthermore, they increase monotonically as carapace width increases. On the other hand, the Pearson residuals for the LR model that includes carapace width are all less than 2 in absolute value and there is no obvious trend as width increases.

(Table 5.3 is given on the following page.)

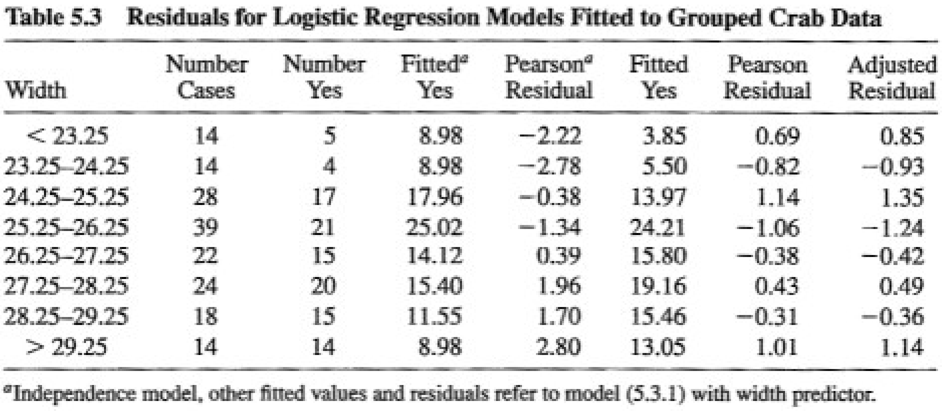

Table 5.3 Residuals for Logistic Regression Models Fitted to Grouped Crab Data

Width	Number Cases	Number Yes	Fitted[a] Yes	Pearson[a] Residual	Fitted Yes	Pearson Residual	Adjusted Residual
< 23.25	14	5	8.98	−2.22	3.85	0.69	0.85
23.25–24.25	14	4	8.98	−2.78	5.50	−0.82	−0.93
24.25–25.25	28	17	17.96	−0.38	13.97	1.14	1.35
25.25–26.25	39	21	25.02	−1.34	24.21	−1.06	−1.24
26.25–27.25	22	15	14.12	0.39	15.80	−0.38	−0.42
27.25–28.25	24	20	15.40	1.96	19.16	0.43	0.49
28.25–29.25	18	15	11.55	1.70	15.46	−0.31	−0.36
> 29.25	14	14	8.98	2.80	13.05	1.01	1.14

[a] Independence model, other fitted values and residuals refer to model (5.3.1) with width predictor.

As we have noted previously, graphical displays are also extremely useful for assessing the GOF of GLM's. One should always plot the fitted values vs. the observed values *and* plot the fitted values vs. each explanatory variable. In Figure 5.3, the observed and fitted proportions for crabs having at least 1 satellite are plotted vs. mean carapace width for each interval in Table 5.3. Note that the raw residuals are represented by the vertical distance between the observed proportion and the fitted proportion in each interval. Based on the graph, the LR model appears to fit the data fairly well, as indicated by all of our previous examinations of GOF.

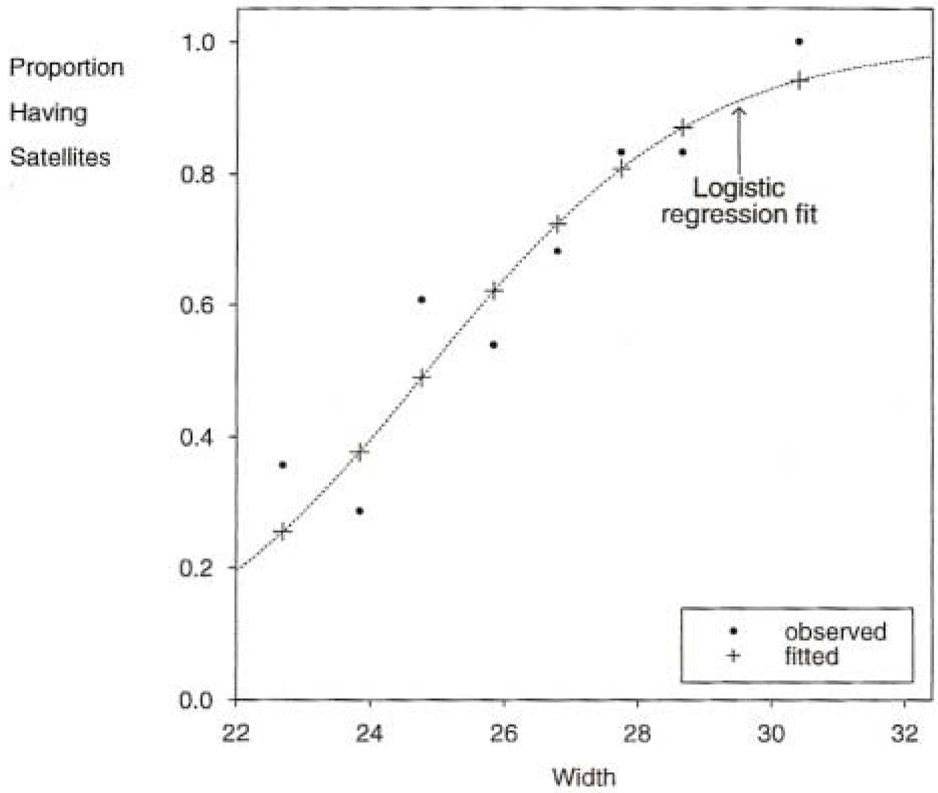

Figure 5.3 Observed and fitted proportions of satellites by width of female crab.

Diagnostic Measures of Influence (Sec. 5.3.4)

As in ordinary regression analysis, some observations may have too much *influence* in determining the fit of a GLM to the observed data. In other words, the fit of the model could be quite different if these observations were deleted from the data set. An observation is more likely to have large influence when it has an extreme value (i.e., it is a possible *outlier*) for one or more of the explanatory variables. It may be informative to report changes in the fit of the model after removing 1 or 2 influential observations if the fitted model when they are included gives unexpected or misleading results.

Several measures have been developed for describing various aspects of influence. Many of them have to do with the effect of removing the observations on certain characteristics of the fitted model. These measures are related to the observation's *leverage*, which is obtained from the diagonal elements of the *hat matrix*. This matrix is used in the matrix formulation of GLM's and can be used to obtain the fitted values from the observed values. The greater an observation's leverage, the greater its potential influence on the fitted model.

Formulas for diagnostic measures of influence are complex and involve matrix calculations, so we will not cover them in this course. Software packages such as SAS and SPSS are capable of calculating these measures.

We illustrate 3 of the most commonly used influence measures using the LR model fit to the grouped carapace width data. These measures include:

(1) For each explanatory variable in the model, we can examine the change in its parameter estimate that results when a given observation is deleted. This change divided by its SE is called *Dfbeta*.

(2) We can examine the change in a joint confidence interval for all parameters in the model that results from deleting a given observation. The diagnostic measure based on this change in the CI is denoted by *c*.

(3) We can also examine the changes in the Pearson χ^2 and likelihood ratio GOF test statistics that result when a given observation is deleted.

Table 5.4 contains the Dfbeta measure for the coefficient of the width variable in the LR model, the confidence interval diagnostic *c*, and the changes in X^2 and G^2 resulting from deleting each width category one-at-a-time. None of the values in Table 5.4 indicate that any of the width categories have undue influence on the fit of the LR model.

By way of contrast, Table 5.4 also contains the changes in X^2 and G^2 that result from deleting each width category when the independence model $\text{logit}[\pi(x)] = \alpha$ is fit to the data. For the intervals corresponding to both smaller and larger horseshoe crabs, the changes in X^2 and G^2 that result from deleting these intervals are quite striking. This tells us that the lack of fit we detected for the independence model is likely due to the crabs with smaller (< 24.25 cm) and larger (> 29.25) carapace widths.

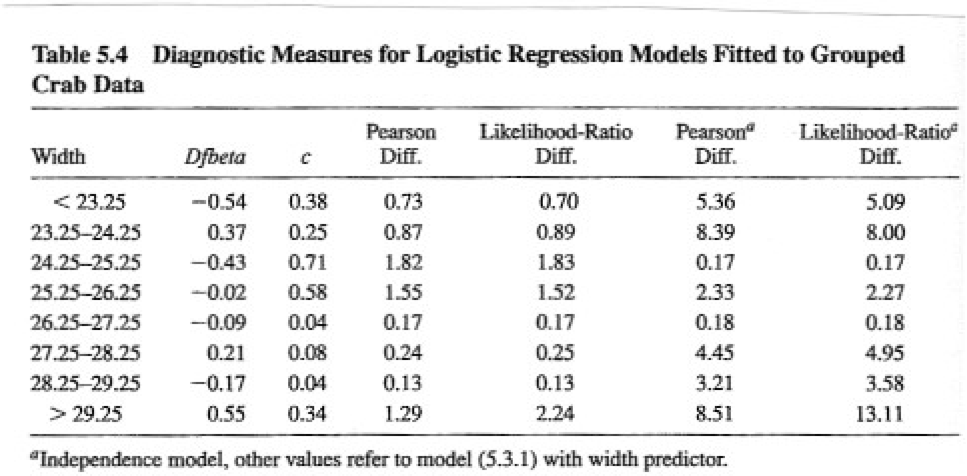

Table 5.4 Diagnostic Measures for Logistic Regression Models Fitted to Grouped Crab Data

Width	*Dfbeta*	*c*	Pearson Diff.	Likelihood-Ratio Diff.	Pearson[a] Diff.	Likelihood-Ratio[a] Diff.
< 23.25	−0.54	0.38	0.73	0.70	5.36	5.09
23.25–24.25	0.37	0.25	0.87	0.89	8.39	8.00
24.25–25.25	−0.43	0.71	1.82	1.83	0.17	0.17
25.25–26.25	−0.02	0.58	1.55	1.52	2.33	2.27
26.25–27.25	−0.09	0.04	0.17	0.17	0.18	0.18
27.25–28.25	0.21	0.08	0.24	0.25	4.45	4.95
28.25–29.25	−0.17	0.04	0.13	0.13	3.21	3.58
> 29.25	0.55	0.34	1.29	2.24	8.51	13.11

[a] Independence model, other values refer to model (5.3.1) with width predictor.

One can also use the leverage values for individual observations to adjust the Pearson residuals so as to improve the standard normal approximation to the distribution of the e_i's. For observation i with leverage h_i, the *adjusted residual* is given by

$$\frac{e_i}{\sqrt{1-h_i}} = \frac{y_i - n_i \widehat{\pi}_i}{\sqrt{n_i \widehat{\pi}_i (1-\widehat{\pi}_i)(1-h_i)}}\,.$$

Table 5.3 contains the adjusted residuals for the LR model fitted to the grouped data with carapace width as the explanatory variable. Note that the adjusted residuals are slightly larger than the Pearson residuals in absolute value, but also do not indicate any lack of fit in the LR model.