BIOS 6244 Analysis of Categorical Data
November 7, 2005 Lecture

Logistic Regression (Chap. 5)

Interpreting the Logistic Regression Model (Sec. 5.1)

Recall the logit GLM (same as the logistic regression model):

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x, \qquad (17)$$

where $\pi(x)$ denotes the probability of "success" for the binary response variable Y when the explanatory variable X = x.

This model implies that $\pi(x)$ increases or decreases as an "S-shaped" function of x. (Recall Figure 4.2.) By exponentiating both sides of Equation (17), we can derive an equivalent form of logistic regression that has the success probability $\pi(x)$ on the LHS of the model equation:

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \qquad (18)$$

Linear Approximation Interpretations (Sec. 5.1.1)

Recall that the parameter β determines the rate of increase or decrease of the S-shaped curve defined by the logistic function. In Figure 5.1, the logistic model fitted to the horseshoe crab data (Sec. 5.1.2) is graphed. Here, "Probability" on the y-axis refers to the probability that a female crab has at least 1 satellite.

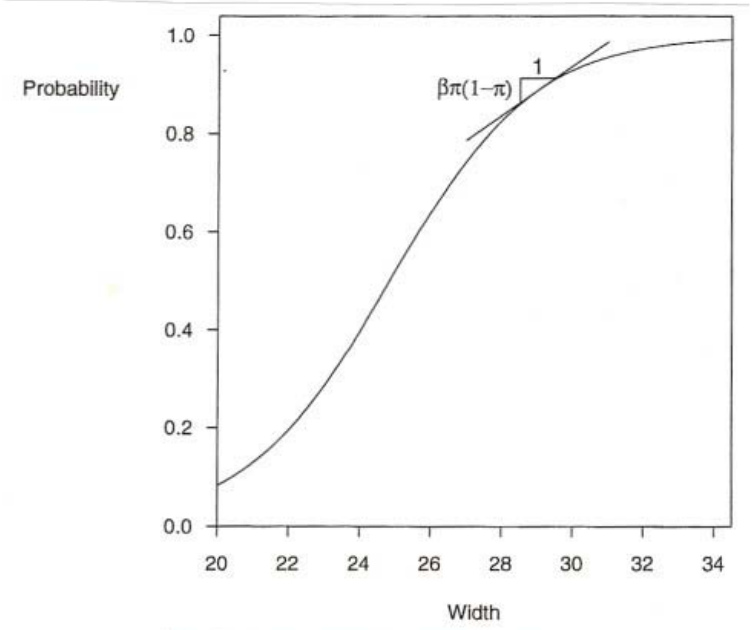

Figure 5.1 Linear approximation to logistic regression curve.

Since the graph is curved, the rate of change in $\pi(x)$ per unit change in x varies according to the value of x (unlike in the linear probability model). A straight line drawn at a tangent to the curve at a particular x value, such as that shown in Figure 5.1, describes the rate of change at that point. For a logistic regression parameter β, the tangent line has slope equal to $\beta\pi(x)[1-\pi(x)]$. For example, the line tangent to the curve at the value x where $\pi(x) = .5$ (in this case, $x = 24.8$) has slope equal to $\beta(.5)(.5) = .25\,\beta$ (= .041 for the curve in Fig. 5.1). When $\pi(x) = .1$ or .9, the tangent line has slope $\beta(.9)(.1) = .09\beta$ (= .015 for the curve in Fig. 5.1). The slope of the tangent line $\rightarrow$ zero as $\pi(x) \rightarrow 0$ or 1. The steepest slope of the curve occurs at the value x for which $\pi(x) = .5$. This value is given by $x = -\alpha/\beta$, which is obtained by solving Equation (17) for x when $\pi(x) = .5$. This x value is called the *median effective level* and is denoted by EL_{50}. In *dose-response studies*, which attempt to establish the relationship between a binary outcome (e.g., lung cancer) and the "dose" of an exposure (e.g., cigarette smoking) , this value of x is called the *median effective dose* and is denoted by ED_{50}. If the binary outcome is death/no death, then this value of x is called the *median lethal dose* and is denoted by LD_{50}.

Horseshoe Crab Example, revisited (Sec. 5.1.2)

Figure 5.2 contains a scatterplot of the horsehoe crab data in which Y is an indicator variable for whether the crab had at least one satellite and X is the width of the crab's carapace. The plotting symbol used is the # of observations having that combination of (x,y) values (as in Figure 4.3). Note that for "large" X (> 29 cm), there are all 1's and no 0's for Y. Beyond that, the scatterplot is not very informative, and it is difficult to determine if a LR model is reasonable.

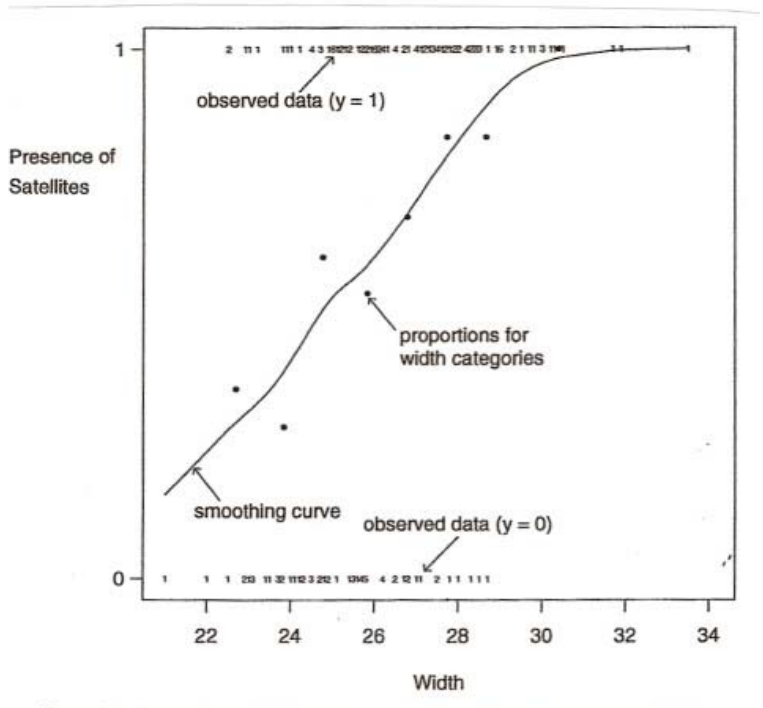

Figure 5.2 Whether satellites are present ($Y = 1$, yes; $Y = 0$, no), by width of female crab.

A more informative plot can be constructed in a manner similar to what was done in Sec. 4.3.2; that is, we group the width values into intervals and then calculate the sample proportion of crabs

having at least 1 satellite within each interval. The data grouped in this way are presented in Table 5.1, showing for each interval the # of female crabs, the # having at least 1 satellite, and the resulting sample proportion.

Table 5.1 Relation Between Width of Female Crab and Existence of Satellites, and Predicted Values for Logistic Regression Model

Width	Number Cases	Number Having Satellites	Sample Proportion	Predicted Probability	Predicted Number Crabs with Satellites
< 23.25	14	5	.36	.26	3.64
23.25–24.25	14	4	.29	.38	5.31
24.25–25.25	28	17	.61	.49	13.78
25.25–26.25	39	21	.54	.62	24.23
26.25–27.25	22	15	.68	.72	15.94
27.25–28.25	24	20	.83	.81	19.38
28.25–29.25	18	15	.83	.87	15.65
> 29.25	14	14	1.00	.93	13.08

Figure 5.2 contains 8 dots showing the sample proportion in each interval (Y) plotted vs. the mean carapace width in that interval (X). As in Fig. 4.4, Agresti also considers a "smoother" for the data, in which one fits a curve to the points in the scatterplot without specifying any particular functional form for the relationship between X & Y. Figure 5.2 also contain a graph of the smoothing curve for these horseshoe crab data.

Both the plotted points from the grouped data and the smoothed curve indicate a general increasing trend, so we are justified in trying to fit a GLM that allows for monotonic trends (for example, a logit or logistic regression model).

Agresti first considers the linear probability model for these data: $\pi(x) = \alpha + \beta x$. However, because of the "major structural defect" pointed out by Agresti in Sec. 4.2.1, ML estimation of the model parameters fails. Agresti considers the least squares fit of the linear probability model instead (i.e., he fits a simple linear regression model) and obtains $\hat{\alpha} = -1.766$ and $\hat{\beta} = .092$. However, this model produces some fitted values that are outside (0, 1) for extreme values of x. For example, when x = 33.5 (the sample max), $\hat{\pi} = -1.766 + .092(33.5) = 1.3$.

The ML estimates for the parameters α and β in the logistic regression model in Equation (17) are difficult to compute by hand, but are easily available in SAS or SPSS. Using PROC LOGISTIC or PROC GENMOD in SAS, we obtain $\hat{\alpha} = -12.351$ and $\hat{\beta} = .497$. At the sample min (x = 21.0 cm), the fitted probability value is

$$\hat{\pi} = \frac{e^{-12.351+.497(21)}}{1+e^{-12.351+.497(21)}} = .129,$$

and at the sample max (x = 33.5 cm), the fitted probability value is

$$\hat{\pi} = \frac{e^{-12.351+.497(33.5)}}{1+e^{-12.351+.497(33.5)}} = .987 .$$

As we saw before, the median effective level is $x = EL_{50} = -\frac{\hat{\alpha}}{\hat{\beta}} = \frac{12.351}{.497} = 24.8$; that is, at a carapace width of 24.8, there is an estimated 50% chance that the female crab will have at least one satellite.

The fitted logistic GLM is graphed in Figure 5.1. An "eyeball" comparison of the fitted curve with the 8 plotted points in Fig. 5.2 indicates that the logistic model appears to provide a good fit to the data. Objective methods for assessing the GOF of a logistic model will be covered in Sec. 5.3.

At the sample mean width of x = 26.3 cm, the fitted value is $\hat{\pi} = .674$. Thus, the incremental rate of change in the fitted probability at x = 26.3 is $\hat{\beta}\hat{\pi}(1-\hat{\pi}) = .497(.674)(.326) = .11$. Thus, for female crabs with carapaces near the mean width, the estimated probability of having at least 1 satellite increases at the rate of .11 per 1 cm increase in width. The estimated rate of change is greatest at the x value (24.8) at which $\hat{\pi}$ = .5: (.497)(.5)(.5) = .12. This illustrates the fact that, unlike the linear probability model, the logistic GLM allows for the estimated rate of change in $\pi(x)$ to vary as x varies.

To further describe the fit of the logistic model, we can compare the fitted value for the # of crabs having at least 1 satellite with the observed number in each width category. Table 5.1 also contains these values. To obtain the fitted values for the # of crabs, we simply add the fitted probability values for all crabs in each width interval. For example, the fitted probabilities for all 14 crabs with carapace widths less than 23.25 cm is 3.64 (compare with the observed value of 5 crabs in this category who had at least 1 satellite). We can also calculate a fitted (or predicted) probability for each width interval by dividing the fitted value for the # of crabs having at least 1 satellite by the total number of crabs in that interval. So, for example, in the 1st width category (< 23.25 cm), the fitted probability is $\frac{3.64}{14} = .26$.

An eyeball comparison of the 3rd & 6th columns in Table 5.1 (or of the 4th & 5th columns) suggests that the model fits the data "decently," especially for the larger carapace widths (> 26.25 cm). Objective criterial for measuring model GOF will be presented in Sec. 5.3.

<u>Odds Ratio Interpretation of LR Models (Sec.5.1.3)</u>

Logistic regression model parameters can also be interpreted in terms of estimated OR's. In the model represented by Equation (17), the odds of success are modeled as

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x} .$$

This exponential relationship provides an interpretation for β: the odds of success increase multiplicatively by e^{β} for every 1 unit increase in x. That is, the odds at level x + 1 are equal to the odds at level x multiplied by e^{β}. If x is binary with values 1 and 0, then e^{β} = the odds ratio for "success" for subjects in the category corresponding to X = 1 relative to those in the category with X = 0. In health sciences research, the explanatory variable of primary interest in the logistic regression will be "exposure," typically coded as a binary variable with values 1 for "exposed" and 0 for "unexposed." Thus, $e^{\hat{\beta}}$ provides an estimate of the OR for the disease given the exposure. This will be illustrated further in Sec. 5.4. Note that when β (or $\hat{\beta}$) = 0, this corresponds to an OR of 1.

For the horseshoe crab data, the estimated odds of a crab having at least 1 satellite are multiplied by $e^{\hat{\beta}} = e^{.497}$=1.64 for every 1 cm increase in carapace width. (As an exercise, verify this for an increase in width from 26.3 to 27.3 cm.) Thus, the odds ratio for having at least 1 satellite, as a function of carapace width measured in cm, is 1.64.

Using Logistic Regression with Underlying Normal Populations

Suppose that, for all study subjects having Y=1, the distribution of X is $N(\mu_1, \sigma^2)$ and that , for all study subjects having Y=0, the distribution of X is $N(\mu_2, \sigma^2)$. Then it can be shown that $\pi(x)$ *must* satisfy the logistic regression model in Equation (18) for some α & β, with β having the same sign as $\mu_1-\mu_2$. Even if the variances of the two normal populations differ, the logistic regression model should approximate the true relationship fairly well, as long as the heteroscedasticiy is not too great. If the variances differ substantially, then a logistic model with a quadratic term (i.e., one that contains both x and x^2) will often fit the data well. (See Exercise 5.5, p. 136).

Inference for Logistic Regression (Sec. 5.2)

Confidence Intervals for Effects (Sec. 5.2.1)

As we have seen before, a general formula for finding an approximate 95% CI for any parameter θ is

$$\hat{\theta} \pm 1.96\,\mathrm{ASE}(\hat{\theta})$$

where $\hat{\theta}$ is the estimate of the parameter and $\mathrm{ASE}(\hat{\theta})$ is the asymptotic standard error.

So, for the parameter β in the logistic regression model given by Equation (17), we can use

$$\hat{\beta} \pm 1.96\,\mathrm{ASE}(\hat{\beta}).$$

Exponentiating the endpoints of this interval yields an approximate 95% CI(OR) since $\widehat{\mathrm{OR}} = e^{\hat{\beta}}$.

Horseshoe Crab Example, revisited

PROC LOGISTIC (or PROC GENMOD) in SAS yields $\hat{\beta}$ = .497 and ASE($\hat{\beta}$) = .102. Therefore, an approximate 95% CI(β) is .497 ± 1.96(.102) = (.298, .697) and an approximate 95% CI(OR) is given by $(e^{.298}, e^{.697}) = (1.35, 2.01)$. The interpretation of this CI is that a 1 cm increase in carapace width corresponds to at least a 35% increase and at most a doubling of the odds that a female crab will have at least one satellite.

Recall from the discussion of the linear approximation to the logistic curve in Sec. 5.1.1 that $\hat{\beta}\hat{\pi}(1-\hat{\pi})$ approximates the change in the probability of "success" π per 1 unit increase in X. For example, at π = .5, the estimated rate of change is $.25\hat{\beta}$ = .25(.497) = .124 for the horseshoe crab data. An approximate 95% CI for .25β can be found by multiplying .25 times the endpoints of the approximate 95% CI(β), yielding [(.25)(.298), (.25)(.697)] =(.074, .174). Thus, for values of X near the carapace width at which π = .5 (namely, 24.8 cm), the rate of increase in the probability that a female crab has at least 1 satellite per 1 cm increase in carapace width is likely to fall somewhere between .07 and .17.

Significance Testing (Sec. 5.2.2)

In addition to finding a CI(β) [or CI(OR)], we are also usually interested in testing the null hypothesis H_0: β = 0 (which is equivalent to testing H_0: OR = 1). Using our general method for testing hypotheses about a single parameter θ, we can use the test statistic

$$z = \frac{\hat{\beta}}{\text{ASE}(\hat{\beta})}$$

in a test of H_0: β = 0 and calculate an approximate p-value using the standard normal distribution. (This approach is called the *Wald test* if one calculates z^2 and then uses $\chi^2(1)$ to calculate an approximate p-value for z^2.)

Another approach that is generally preferable for testing H_0: β = 0 is the *likelihood ratio test*. (See Section 4.4.1 in our text.) The test statistic here is based on the difference between the max of the log-likelihood function when β is assumed to be 0 (denoted L_0) and the max of the log-likelihood function under the "full model" when β is unrestricted (denoted L_1). Multiplying this difference by -2 yields a test statistic that also has an approximate $\chi^2(1)$ distribution.

Statistical packages such as SAS & SPSS report the values of the maximized log-likelihood functions and the resulting likelihood ratio test statistic & p-value.

For the horshoeshoe crab data, the Wald test yields

$$z^2 = \left(\frac{\hat{\beta}}{\text{ASE}(\hat{\beta})}\right)^2 = \left(\frac{.497}{.102}\right)^2 = 23.9$$

and $p < .0001$ using $\chi^2(1)$.

For the likelihood ratio test, the maximized log-likelihood under H_0 is $L_0 = -112.88$ and for the full model it is $L_1 = -97.23$. Therefore, the likelihood ratio test statistic is

$$-2(L_0 - L_1) = 31.3.$$

Again, we use $\chi^2(1)$ to calculate an approximate p-value, so the evidence here is even stronger that $\beta \neq 0$ and hence that there is an association between carapace width and having at least 1 satellite.

Distribution of Probability Estimates (Sec. 5.2.3)

In addition to performing inference for the model parameter β, we may also be interested in CI's & hypothesis tests for the true probability $\pi(x)$ at some given value of X.

We know that the estimated probability (i.e., fitted value) at X = x is given by

$$\widehat{\pi(x)} = \frac{e^{\hat{\alpha}+\hat{\beta}x}}{1+e^{\hat{\alpha}+\hat{\beta}x}}. \qquad (19)$$

Thus, in the horseshoe crab example, we estimate the probability of having at least 1 satellite for a female crab with carapace width = 26.5 cm to be

$$\widehat{\pi(26.5)} = \frac{e^{-12.351+.497(26.5)}}{1+e^{-12.351+.497(26.5)}} = .695.$$

To find a 95% CI[$\widehat{\pi(x)}$], we need the ASE for $\widehat{\pi(x)}$. This can be obtained from the estimated variances for $\hat{\alpha}$ and $\hat{\beta}$ and estimated Cov($\hat{\alpha}$, $\hat{\beta}$) as follows.

We start with estimating Var($\hat{\alpha}+\hat{\beta}x$):

$$\text{Var}(\hat{\alpha}+\hat{\beta}x) = \text{Var}(\hat{\alpha}) + x^2\text{Var}(\hat{\beta}) + 2x\text{Cov}(\hat{\alpha},\hat{\beta}). \qquad (20)$$

After substituting sample estimates of the variances and covariance into Equation (20), we take the square root to obtain ASE ($\hat{\alpha}+\hat{\beta}x$). We then construct an approximate 95% CI for logit[$\widehat{\pi(x)}$] and then back transform to find an approximate 95% CI[$\pi(x)$].

For the horseshoe crab data, PROC LOGISTIC (or PROC GENMOD) in SAS yields $\widehat{\text{Var}(\hat{\alpha})} = 6.910$, $\widehat{\text{Var}(\hat{\beta})} = .01035$, and $\widehat{\text{Cov}(\hat{\alpha},\hat{\beta})} = -.2668$. Thus, the ASE for $\hat{\alpha}+\hat{\beta}x$ when x = 26.5 cm is

$$\sqrt{6.910 + (26.5)^2(.01035) - 2(26.5)(.2668)} = .195.$$

The estimated logit at x = 26.5 is given by $\hat{\alpha} + \hat{\beta}(26.5) = -12.351 + .497(26.5) = .820$ and an approximate 95% CI for logit[$\pi(26.5)$] is given by .820 ± 1.96(.195) = (.438, 1.202). Back-transforming using Equation (19) yields

$$\left(\frac{e^{.438}}{1+e^{.438}}, \frac{e^{1.202}}{1+e^{1.202}} \right) = (.61, .77)$$

as an approximate 95% CI for the probability that crabs with a carapace width equal to 26.5 cm will have at least one satellite.

What do we gain by using the logistic regression model to estimate such probabilities? The naïve approach to obtaining a CI for the probability that a female crab with width 26.5 would have at least one satellite would be to treat the 6 crabs in the data set that had a carapace of this width as a random sample from a binomial distribution and then find an exact 95% CI(π). Of the 6 crabs with width 26.5, 4 had at least 1 satellite. This yields a sample estimate of $\frac{4}{6} = .667$, which is very close to the logistic model-based estimate of .695. For these data, SimCalc yields an exact 95% CI(π) of (.22, .96). The width of this CI is much, much larger than that of the approximate 95% CI we obtained using the logistic regression model: .74 vs. .16. Assuming that the logistic model is an accurate representation of reality and that n is sufficiently large, we see that using the logistic model can provide much more precise estimates of the probabilities. The main reason for this is that the logistic model uses information from all 173 female crabs in the sample, not just the 6 with carapace width = 26.5 cm. For small n, when the approximate CI's & hypothesis tests may not be valid, there are exact methods that can be used to perform inference for the logistic regression model parameters. In Section 5.3, we present objective methods for determining if, in fact, the LR model provides an adequate fit to the observed data. If not, then alternative GLM's or other statistical methods should be considered.