Chapter 4 Logistic Regression

Outline

1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

Outline

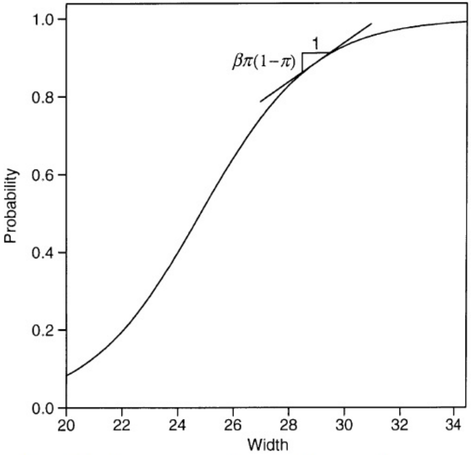
1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

Logistic Regression

Suppose there is a single explanatory variable X, which is quantitative. For a binary response variable Y , recall that $\pi(x)$ denotes the "success" probability at value x.

$$
\begin{aligned}
\operatorname{logit}[\pi(x)] &= \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \\
\pi(x) &= \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.
\end{aligned}
$$

4.1.1 Linear Approximation Interpretations



4.1.1 Linear Approximation Interpretations

- A straight line drawn tangent to the curve at a particular x value,such as shown in Figure 4.1, describes the rate of change at that point. For logistic regression parameter β, that line has slope $\beta\pi(x)[1-\pi(x)]$.
- The steepest slope occurs at x for which $\pi(x) = 0.50$. That x value relates to the logistic regression parameters by $x = -\alpha/\beta$. This x value is sometimes called the *median effective level* and is denoted EL_{50}. It represents the level at which each outcome has a 50% chance.

4.1.2 Horseshoe Crabs: Viewing and Smoothing a Binary Outcome

To illustrate these interpretations, we re-analyze the horseshoe crab data introduced in Section 3.3.2 (Table 3.2). Let Y indicate whether a female crab has any satellites (other males who could mate with her). $Y = 1$ if a female crab has at least one satellite, and $Y = 0$ if she has no satellite. Use the female crab's width (in cm) as the sole predictor.

4.1.2 Horseshoe Crabs: Viewing and Smoothing a Binary Outcome

Better information results from grouping the width values into categories and calculating a sample proportion of crabs having satellites for each category. This reveals whether the true proportions follow approximately the trend required by this model. Consider the grouping shown in Table 4.1. In each of the eight width categories, we computed the sample proportion of crabs having satellites and the mean width for the crabs in that category.

4.1.2 Horseshoe Crabs: Viewing and Smoothing a Binary Outcome

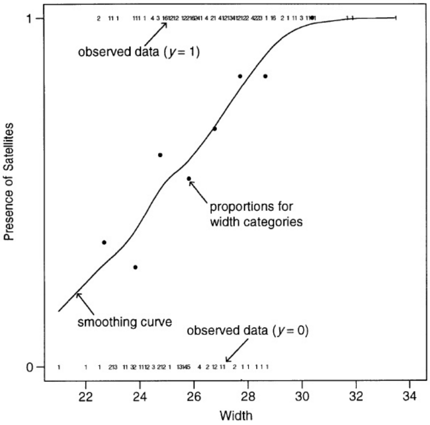


Figure 4.2. Whether satellites are present ($Y = 1$, yes; $Y = 0$, no), by width of female crab.

4.1.2 Horseshoe Crabs: Viewing and Smoothing a Binary Outcome

Table 4.1 Relation Between Width of Female Crab and
Existence of Satellites, and Predicted Values for Logistic Regression Model

Width	Number of Cases	Number Satellites	Sample Proportion	Estimated Probability	Predicted No. Crabs with Stellites
< 23.25	14	5	0.36	0.26	3.63
23.25-24.25	14	4	0.29	0.38	5.31
24.25-25.25	28	17	0.61	0.49	13.78
25.25-26.25	39	21	0.54	0.62	24.23
26.25-27.25	22	15	0.68	0.72	15.94
27.25-28.25	24	20	0.83	0.81	19.38
28.25-29.25	18	15	0.83	0.87	15.65
> 29.25	14	14	1.00	0.93	13.08

4.1.3 Horseshoe Crabs: Interpreting the Logistic Regression Fit

For the ungrouped data in Table 3.2, let $\pi(x)$ denote the probability that a female horseshoe crab of width x has a satellite. The simplest model to interpret is the linear probability model, $\pi(x) = \alpha + \beta x$. During the ML fitting process, some predicted values for this GLM fall outside the legitimate 0-1 range for a binomial parameter, so ML fitting fails. Ordinary least squares fitting yields $\hat{\pi}(x) = -1.766 + 0.092x$. It is inadequate for extreme values.

4.1.3 Horseshoe Crabs: Interpreting the Logistic Regression Fit

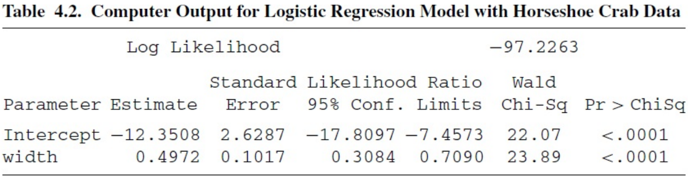


Table 4.2. Computer Output for Logistic Regression Model with Horseshoe Crab Data

Log Likelihood −97.2263

Parameter	Estimate	Standard Error	Likelihood Ratio 95% Conf. Limits		Wald Chi-Sq	Pr > ChiSq
Intercept	−12.3508	2.6287	−17.8097	−7.4573	22.07	<.0001
width	0.4972	0.1017	0.3084	0.7090	23.89	<.0001

$\hat{\pi}(x) = \frac{\exp(-12.351+0.497x)}{1+\exp(-12.351+0.497x)}$. Since $\beta > 0$, the estimated probability $\hat{\pi}$ is larger at larger width values. The median effective level is the width at which $\hat{\pi}(x) = 0.50$. This is $x = EL_{50} = -\hat{\alpha}/\hat{\beta} = 24.8$. Figure 4.1 plots the estimated probabilities as a function of width.

4.1.3 Horseshoe Crabs: Interpreting the Logistic Regression Fit

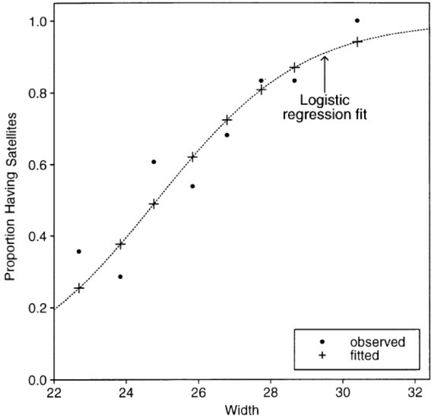


Figure 4.3. Observed and fitted proportions of satellites, by width of female crab.

4.1.3 Horseshoe Crabs: Interpreting the Logistic Regression Fit

Table 4.1 reports the fitted values and the average estimated probabilities of a satellite, in grouped fashion. Figure 4.3 plots the sample proportions and the estimated probabilities against width. These comparisons suggest that the model fits decently.

4.1.4 Odds Ratio Interpretation

An important interpretion of the logistic regression model uses the *odds* and the *odds ratio*. For model(4.1), the odds of response 1 are

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

This exponential relationship provides an interpretation for β: The odds multiply by e^{β} for every 1-unit increase in x. That is, the odds at level $x+1$ equal the odds at x multiplied by e^{β}. When $\beta=0, e^{\beta}=1$, and the odds do not change as x changes.

4.1.5 Logistic Regression with Retrospective Studies

Another property of logistic regression relates to situations in which the explanatory variable X rather than the response variable Y is random. This occurs with retrospective sampling designs. Sometimes such designs are used because one of the response categories occurs rarely, and a prospective study might have too few cases to enable one to estimate effects of predictors well. For a given sample size, effect estimates have smaller SEs when the number of outcomes of the two types are similar than when they are very different.

4.1.5 Logistic Regression with Retrospective Studies

Logistic regression parameters refer to odds and odds ratios. One can fit logistic regression models with data from case-control studies and estimate effects of explanatory variables. The intercept term α in the model is not meaningful, because it relates to the relative numbers of outcomes of $y = 1$ and $y = 0$. We do not estimate this, because the sample frequencies for $y = 1$ and $y = 0$ are fixed by the nature of the case-control study.

4.1.5 Logistic Regression with Retrospective Studies

With case-control studies, it is not possible to estimate effects in binary models with link functions other than the logit. Unlike the odds ratio, the effect for the conditional distribution of X given Y does not then equal that for Y given X. This provides an important advantage of the logit link over links such as the probit. It is a major reason why logistic regression surpasses other models in popularity for biomedical studies.

4.1.6 Normally Distributed X Implies Logistic Regression for Y

Suppose the distribution of X for subjects having $Y = 1$ is normal $N(\mu_1, \sigma^2)$, and suppose the distribution of X for subjects having $Y = 0$ is normal $N(\mu_0, \sigma^2)$; that is, with different mean but the same standard deviation. Then, a Bayes theorem calculation converting from the distribution of X given $Y = y$ to the distribution of Y given $X = x$ shows that $P(Y = 1|x)$ satisfies the logistic regression curve. For that curve, the effect of x is $\beta = (\mu_1 - \mu_0)/\sigma^2$. In particular, β has the same sign as $\mu_1 - \mu_0$. For example, if those with $y = 1$ tend to have higher values of x, then $\beta > 0$.

4.1.6 Normally Distributed X Implies Logistic Regression for Y

Example

Consider Y = heart disease (1 = yes, 0 = no) and X = cholesterol level. Suppose cholesterol levels have approximately a $N(\mu_0 = 160, \sigma^2 = 50^2)$ distribution for those without heart disease and a $N(\mu_1 = 260, \sigma^2 = 50^2)$ distribution for those with heart disease. Then, the probability of having heart disease satisfies the logistic regression function with predictor x and $\beta = (260 - 160)/50^2 = 0.04$.

Outline

1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

4.2.1 Binary Data can be Grouped or Ungrouped

Ungrouped binary data

Widely available software reports the ML estimates of parameters and their standard errors. Sometimes sets of observations have the same values of predictor variables, such as when explanatory variables are discrete. Then, ML model fitting can treat the observations as the binomial counts of successes out of certain sample sizes, at the various combinations of values of the predictors.We will refer to this case as *grouped binary data* and the case in which each observation is a single binary outcome as *ungrouped binary data*.

When at least one explanatory variable is continuous, binary data are naturally ungrouped.

4.2.2 Confidence Intervals for Effects

A large-sample Wald confidence interval for the parameter β in the logistic regression model, $\text{logit}[\pi(x)] = \alpha + \beta x$, is

$$\hat{\beta} \pm z_{\alpha/2}(SE)$$

Exponentiating the endpoints yields an interval for e^{β}, the multiplicative effect on the odds of a 1-unit increase in x.

4.2.2 Confidence Intervals for Effects

From Section 4.1.1, a simpler interpretation uses a straight-line approximation to the logistic regression curve. The term $\beta\pi(x)[1-\pi(x)]$ approximates the change in the probability per 1-unit increase in x. For instance, at $\pi(x) = 0.50$, the estimated rate of change is $0.25\hat{\beta} = 0.124$. A 95% confidence interval for 0.25β equals 0.25 times the endpoints of the interval for β. For the likelihood-ratio interval, this is $[0.25(0.308), 0.25(0.709)] = (0.077, 0.177)$. So, if the logistic regression model holds, then for values of x near the width value at which $\pi(x) = 0.50$, we infer that the rate of increase in the probability of a satellite per centimeter increase in width falls between about 0.08 and 0.18.

4.2.3 Significance Testing

For the logistic regression model, $H_0 : \beta = 0$ states that the probability of success is independent of X. Wald test statistics (Section 1.4.1) are simple. For large samples,

$$z = \hat{\beta}/SE$$

has a standard normal distribution when $\beta = 0$. Refer z to the standard normal table to get a one-sided or two-sided P-value. Equivalently, for the two-sided $H_a : \beta \neq 0$, $z^2 = (\hat{\beta}/SE)^2$ has a large-sample chi-squared null distribution with $df = 1$.

4.2.3 Significance Testing

Although the Wald test is adequate for large samples, the likelihood-ratio test is more powerful and more reliable for sample sizes often used in practice. The test statistic compares the maximum L_0 of the log-likelihood function when $\beta = 0$ to the maximum L_1 of the log-likelihood function for unrestricted β. The test statistic, $-2(L_0 - L_1)$, also has a large-sample chi-squared null distribution with $df = 1$.

4.2.3 Significance Testing

For the horseshoe crab data, the Wald statistic $z = \hat{\beta}/SE = 0.497/0.102 = 4.9$. This shows strong evidence of a positive effect of width on the presence of satellites ($P < 0.0001$). The equivalent chi-squared statistic, $z^2 = 23.9$, has $df = 1$. Software reports that the maximized log likelihoods equal $L_0 = -112.88$ under $H_0 : \beta = 0$ and $L_1 = -97.23$ for the full model. The likelihood-ratio statistic equals $-2(L_0 - L_1) = 31.3$, with $df = 1$. This also provides extremely strong evidence of a width effect ($P < 0.0001$).

4.2.4 Confidence Intervals for Probabilities

Recall that the logistic regression estimate of $P(Y = 1)$ at a fixed setting x is

$$\hat{\pi}(x) = \exp(\hat{\alpha} + \hat{\beta}x)/[1 + \exp(\hat{\alpha} + \hat{\beta}x)].$$

Most software for logistic regression can report this estimate as well as a confidence interval for the true probability $\pi(x)$.

4.2.4 Confidence Intervals for Probabilities

We illustrate by estimating the probability of a satellite for female crabs of width $x = 26.5$, which is near the mean width. The logistic regression fit yields

$$\hat{\pi} = \exp(-12.351 + 0.497(26.5))/[1 + \exp(-12.351 + 0.497(26.5))] = 0.695.$$

From software, a 95% confidence interval for the true probability is (0.61, 0.77).

4.2.6 Confidence Intervals for Probabilities: Details

- The term $\hat{\alpha} + \hat{\beta}x$ in the exponents of the prediction equation is the estimated linear predictor in the logit transform of $\pi(x)$.

$$Var(\hat{\alpha} + \hat{\beta}x) = Var(\hat{\alpha}) + x^2 Var(\hat{\beta}) + 2xCov(\hat{\alpha}, \hat{\beta}).$$

- A 95% confidence interval for the true logit is $(\hat{\alpha} + \hat{\beta}x) \pm 1.96(SE)$.
- For example, for the horseshoe crab data, $\hat{Var}(\hat{\alpha}) = 6.9102$, $\hat{Var}(\hat{\beta}) = 0.0103$, $\hat{Cov}(\hat{\alpha}, \hat{\beta}) = -0.2668$. Therefore, the estimated variance of this estimated logit equals $(x = 26.5)$

$$6.1902 + (26.5)^2(0.0103) + 2(26.5)(-0.2668).$$

4.2.7 Standard Errors of Model Parameter Estimates

- The estimated covariance matrix for the ML parameter estimates is the inverse of the information matrix.
- Let n_i denote the number of observations at setting i of the explanatory variables. (Note $n_i = 1$ when the binary data are ungrouped.) At setting i, let x_{ij} denote the value of explanatory variable j , and let $\hat{\pi}_i$ denote the estimated "success" probability based on the model fit. The element in row a and column b of the information matrix is

$$\sum_i x_{ia} x_{ib} n_i \hat{\pi}(1 - \hat{\pi})$$

4.2.5 Why Use a Model to Estimate Probabilities?

When the logistic regression model holds, the model-based estimator of $\pi(x)$ is much better than the sample proportion. It uses all the data rather than only the data at the fixed x value. The result is a more precise estimate.

Reality is more complicated. In practice, any model will not exactly represent the true relationship between $\pi(x)$ and x. If the model approximates the true probabilities reasonably well, however, it performs well. The model-based estimator tends to be much closer than the sample proportion to the true value, unless the sample size on which that sample proportion is based is extremely large. The model smooths the sample data, somewhat dampening the observed variability.

Outline

1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

Logistic regression, like ordinary regression, can have multiple explanatory variables. Some or all of those predictors can be categorical, rather than quantitative. This section shows how to include categorical predictors, often called *factors*, and Section 4.4 presents the general form of multiple logistic regression models.

4.3.1 Indicator Variables Represent Categories of Predictors

Let x and z each take values 0 and 1 to represent the two categories of each explanatory variable. The model for $P(Y = 1)$,

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z.$$

has main effects for x and z. The variables x and z are called *indicator variables*. They indicate categories for the predictors. Indicator variables are also called *dummy variables*. For this coding, Table 4.3 shows the logit values at the four combinations of values of the two predictors.

4.3.1 Indicator Variables Represent Categories of Predictors

Table 4.3 Logits Implied by Indicator Variables in Model, $\text{logit}[P(Y=1)] = \alpha + \beta_1 x + \beta_2 z$

x	z	Logit
0	0	α
1	0	$\alpha + \beta_1$
0	1	$\alpha + \beta_2$
1	1	$\alpha + \beta_1 + \beta_2$

4.3.1 Indicator Variables Represent Categories of Predictors

- This model assumes an absence of interaction. The effect of one factor is the same at each category of the other factor. At a fixed category z of Z, the effect on the logit of changing from $x = 0$ to $x = 1$ is $= [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1$. This difference between two logits equals the difference of log odds. Equivalently, that difference equals the log of the odds ratio between X and Y , at that category of Z. Thus, $\exp(\beta_1)$ equals the conditional odds ratio between X and Y.

4.3.1 Indicator Variables Represent Categories of Predictors

- Conditional independence exists between X and Y , controlling for Z, if $\beta_1 = 0$. In that case the common odds ratio equals 1. The simpler model,

  $$\text{logit}[P(Y = 1)] = \alpha + \beta_2 z$$

  then applies to the three-way table.

4.3.2 Example: AZT Use and AIDS

Table 4.4. Development of AIDS Symptoms by AZT Use and Race

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

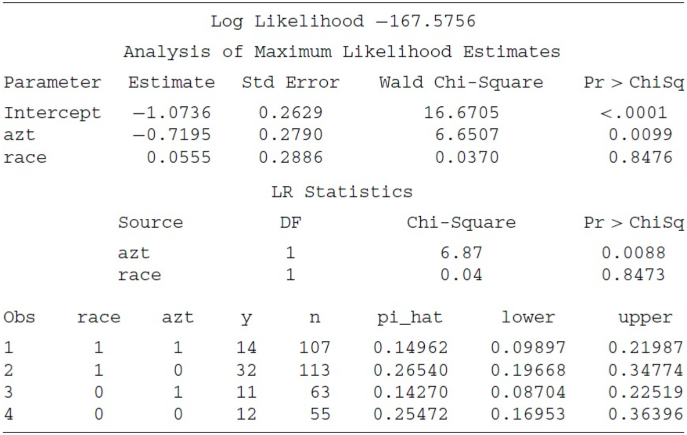
4.3.2 Example: AZT Use and AIDS

Example

Table 4.4, is based on a study described in the New York Times (February 15, 1991) on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. Table 4.4 is a $2 \times 2 \times 2$ cross classification of veterans race, whether AZT was given immediately, and whether AIDS symptoms developed during the 3 year study. Let $X = AZT$ treatment, $Z = race$, and $Y =$ whether AIDS symptoms developed ($1 =$ yes, $0 =$ no).

4.3.2 Example: AZT Use and AIDS

Table 4.5. Computer Output for Logit Model with AIDS Symptoms Data



```
                       Log Likelihood −167.5756

            Analysis of Maximum Likelihood Estimates

Parameter   Estimate   Std Error    Wald Chi-Square    Pr > ChiSq

Intercept   −1.0736     0.2629         16.6705           <.0001
azt         −0.7195     0.2790          6.6507            0.0099
race         0.0555     0.2886          0.0370            0.8476

                          LR Statistics

            Source          DF          Chi-Square       Pr > ChiSq

            azt              1            6.87            0.0088
            race             1            0.04            0.8473

Obs     race     azt     y      n      pi_hat      lower      upper

1        1        1      14    107    0.14962    0.09897    0.21987
2        1        0      32    113    0.26540    0.19668    0.34774
3        0        1      11     63    0.14270    0.08704    0.22519
4        0        0      12     55    0.25472    0.16953    0.36396
```

4.3.3 ANOVA-Type Model Representation of Factors

An alternative representation of factors in logistic regression uses the way ANOVA models often express factors. The model formula

$$\operatorname{logit}[P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z$$

represents the effects of X through parameters $\{\beta_i^X\}$ and the effects of Z through parameters $\{\beta_k^Z\}$. (The X and Z superscripts are merely labels and do not represent powers.) The term β_i^X denotes the effect on the logit of classification in category i of X. Conditional independence between X and Y, given Z, corresponds to $\beta_1^X = \cdots = \beta_I^X$.

4.3.3 ANOVA-Type Model Representation of Factors

It applies for any numbers of categories for X and Z. Each factor has as many parameters as it has categories, but one is redundant. To account for redundancies, most software sets the parameter for the last category equal to zero. The term β_i^X in this model then is a simple way of representing

$$\beta_1^X x_1 + \beta_2^X x_2 + \cdots + \beta_{I-1}^X x_{I-1}.$$

Category I does not need an indicator.

4.3.3 ANOVA-Type Model Representation of Factors

By itself, the parameter estimate for a single category of a factor is irrelevant. Different ways of handling parameter redundancies result in different values for that estimate. An estimate makes sense only by comparison with one for another category. Exponentiating a difference between estimates for two categories determines the odds ratio relating to the effect of classification in one category rather than the other.

4.3.4 The Cochran-Mantel-Haenszel Test for $2 \times 2 \times K$ Contingency Tables

With K categories for Z, model can then be expressed as $logit[P(Y = 1)] = \alpha + \beta x + \beta_k^Z$, where x is an indicator variable for the two categories of X. Then, $exp(\beta)$ is the common XY odds ratio for each of the K partial tables for categories of Z. This is the homogeneous association structure for multiple 2×2 tables. One can test conditional independence by the Wald test or the likelihood-ratio test of $H_0 : \beta = 0$.

4.3.4 The Cochran-Mantel-Haenszel Test for $2 \times 2 \times K$ Contingency Tables

- The Cochran-Mantel-Haenszel test is an alternative test of XY conditional independence in $2 \times 2 \times K$ contingency tables. This test conditions on the row totals and the column totals in each partial table. Like Fishers exact test, the test statistic utilizes this cell in each partial table.
- In partial table k, the row totals are $\{n_{1+k}, n_{2+k}\}$, and the column totals are $\{n_{+1k}, n_{+2k}\}$. Given these totals, under H_0,

$$\mu_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}$$

$$Var(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k} - 1)$$

4.3.4 The Cochran-Mantel-Haenszel Test for $2 \times 2 \times K$ Contingency Tables

The Cochran-Mantel-Haenszel(CMH) test statistic summarizes the information from the K partial tables using

$$CMH = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k Var(n_{11k})}$$

This statistic has a large-sample chi-squared null distribution with $df = 1$. The approximation improves as the total sample size increases, regardless of whether the number of strata K is small or large.

4.3.5 Testing the Homogeneity of Odds Ratios

Sometimes it is of interest to test the hypothesis of homogeneous association (although it is not necessary to do so to justify using the CMH test). A test of homogeneity of the odds ratios is, equivalently, a test of the goodness of fit of model. Section 5.2.2 will show how to do this.

Some software reports a test, called the Breslow-Day test, that is a chi-squared test specifically designed to test homogeneity of odds ratios. It has the form of a Pearson chi-squared statistic, comparing the observed cell counts to estimated expected frequencies that have a common odds ratio. This test is an alternative to the goodness-of-fit tests.

Outline

1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

We consider the general logistic regression model with multiple explanatory variables. Denote the k predictors for a binary response Y by $x_1, x_2, ..., x_k$. The model for the log odds is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

The parameter β_i refers to the effect of x_i on the log odds that Y $= 1$, controlling the other x.

4.4.1 Example: Horseshoe Crabs with Color and Width Predictors

Example

We continue the analysis of the horseshoe crab data (Sections 3.3.2 and 4.1.3) by using both the female crab's shell width and color as predictors. Color has five categories: light, medium light, medium, medium dark, dark. Color is a surrogate for age, older crabs tending to have darker shells. The sample contained no light crabs, so we use only the other four categories.

4.4.1 Example: Horseshoe Crabs with Color and Width Predictors

The Model is $\text{logit}[P(Y = 1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x$, where x denotes width and $c_1 = 1$ for color = medium light, 0 otherwise; $c_2 = 1$ for color = medium, 0 otherwise; $c_3 = 1$ for color = medium dark, 0 otherwise. The crab color is dark (category 4) when $c_1 = c_2 = c_3 = 0$. Table 4.6 shows the ML parameter estimates.

4.4.1 Example: Horseshoe Crabs with Color and Width Predictors

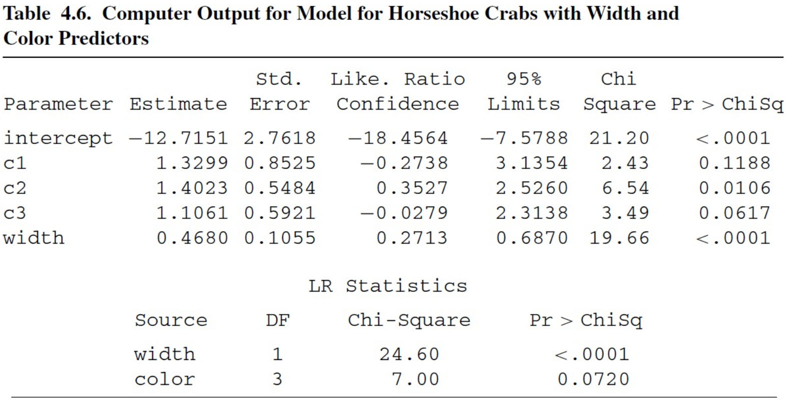


Table 4.6. Computer Output for Model for Horseshoe Crabs with Width and Color Predictors

		Std.	Like. Ratio	95%	Chi	
Parameter	Estimate	Error	Confidence	Limits	Square	Pr > ChiSq
intercept	−12.7151	2.7618	−18.4564	−7.5788	21.20	<.0001
c1	1.3299	0.8525	−0.2738	3.1354	2.43	0.1188
c2	1.4023	0.5484	0.3527	2.5260	6.54	0.0106
c3	1.1061	0.5921	−0.0279	2.3138	3.49	0.0617
width	0.4680	0.1055	0.2713	0.6870	19.66	<.0001

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
width	1	24.60	<.0001
color	3	7.00	0.0720

4.4.1 Example: Horseshoe Crabs with Color and Width Predictors

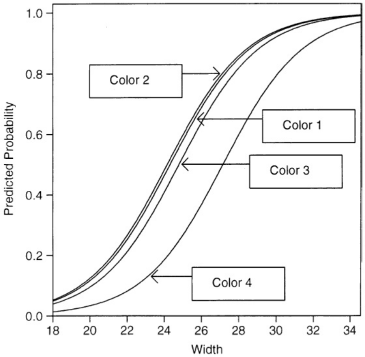


Figure 4.4. Logistic regression model using width and color predictors.

4.4.2 Model Comparison to Check Whether a Term is Needed

- Are certain terms needed in a model? To test this, we can compare the maximized log-likelihood values for that model and the simpler model without those terms.
- To test whether color contributes to model, we test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. This hypothesis states that, controlling for width, the probability of a satellite is independent of color. The likelihood-ratio test compares the maximized log-likelihood L1 for the full model to the maximized log-likelihood L_0 for the simpler model in which those parameters equal 0.

4.4.3 Quantitative Treatment of Ordinal Predictor

A simpler model treats color in a quantitative manner. It supposes a linear effect, on the logit scale, for a set of scores assigned to its categories. To illustrate, we use scores $c = \{1, 2, 3, 4\}$ for the color categories and fit the model $\text{logit}[P(Y = 1)] = \alpha + \beta_1 c + \beta_2 x$. The prediction equation is

$$\text{logit}[\hat{P}(Y = 1)] = -10.071 - 0.509c + 0.458x.$$

4.4.3 Quantitative Treatment of Ordinal Predictor

In summary, the nominal-scale model, the quantitative model with color scores $\{1, 2, 3, 4\}$, and the model with binary color scores $\{1, 1, 1, 0\}$ all suggest that dark crabs are least likely to have satellites. When the sample size is not very large, it is not unusual that several models fit adequately.

It is advantageous to treat ordinal predictors in a quantitative manner, when such models fit well. The model is simpler and easier to interpret, and tests of the effect of the ordinal predictor are generally more powerful when it has a single parameter rather than several parameters.

4.4.4 Allowing Interaction

The models we have considered so far assume a lack of interaction between width and color. Let us check now whether this is sensible.We can allow interaction by adding cross products of terms for width and color. Each color then has a different-shaped curve relating width to the probability of a satellite, so a comparison of two colors varies according to the value of width.

$$\text{logit}[\hat{P}(Y = 1)] = -5.854 - 6.958c + 0.200x + 0.322(c \times x).$$

$$\text{logit}[\hat{P}(Y = 1)] = -5.854 + 0.200x.$$

$$\text{logit}[\hat{P}(Y = 1)] = -12.812 + 0.522x.$$

Outline

1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

4.5.1 Probability-Based Interpretations

Consider a setting of predictors at which $\hat{P}(Y=1)=\hat{\pi}$. Then, controlling for the other predictors, a 1-unit increase in x_j corresponds approximately to a $\hat{\beta}_j\hat{\pi}(1-\hat{\pi})$ change in $\hat{\pi}$.

4.5.1 Probability-Based Interpretations

This straight-line approximation deteriorates as the change in the predictor values increases. More precise interpretations use the probability formula directly. One way to describe the effect of a predictor x_j sets the other predictors at their sample means and finds $\hat{\pi}$ at the smallest and largest x_j values. The effect is summarized by reporting those $\hat{\pi}$ values or their difference. However, such summaries are sensitive to outliers on x_j . To obtain a more robust summary, it is more sensible to use the quartiles of the x_j values.

4.5.1 Probability-Based Interpretations

To summarize the effect of an indicator explanatory variable, it makes sense to report the estimated probabilities at its two values rather than at quartiles, which could be identical.

Table 4.7 summarizes effects using estimated probabilities. It also shows results for the extension of the model permitting interaction. The estimated width effect is then greater for the lighter colored crabs. However, the interaction is not significant.

4.5.1 Probability-Based Interpretations

Table 4.7. Summary of Effects in Model with Crab Width and Whether Color is Dark as Predictors of Presence of Satellites

Variable	Estimate	SE	Comparison	Change in Probability
No interaction model				
Intercept	−12.980	2.727		
Color(0=dark, 1= other)	1.300	0.526	$\bar{x}$ at $(1, 0)$	$0.31 = 0.71 - 0.40$
Width(x)	0.478	0.104	$\bar{c}$ at (UQ, LQ)	$0.29 = 0.80 - 0.51$
Interaction model				
Intercept	−5.854	6.694		
Color(0=dark, 1=other)	−6.958	7.318		
Width(x)	0.200	0.262	$c = 0$ at (UQ, LQ)	$0.13 = 0.43 - 0.30$
Width*color	0.322	0.286	$c = 1$ at (UQ, LQ)	$0.29 = 0.84 - 0.55$

4.5.2 Standardized Interpretations

- With multiple predictors, it is tempting to compare magnitudes of $\{\hat{\beta}_j\}$ to compare effects of predictors. For binary predictors, this gives a comparison of conditional log odds ratios, given the other predictors in the model. For quantitative predictors, this is relevant if the predictors have the same units, so a 1-unit change means the same thing for each. Otherwise, it is not meaningful.

4.5.2 Standardized Interpretations

- An alternative comparison of effects of quantitative predictors having different units uses *standardized* coefficients. The model is fitted to standardized predictors, replacing each x_j by $(x_j - \bar{x}_j)/s_{x_j}$. A 1-unit change in the standardized predictor is a standard deviation change in the original predictor. Then, each regression coefficient represents the effect of a standard deviation change in a predictor, controlling for the other variables.

Outline

1. 4.1 Interpreting the Logistic Regression Model
2. 4.2 Inference for Logistic Regression
3. 4.3 Logistic Regression with Categorical Predictors
4. 4.4 Multiple Logistic Regression
5. 4.5 Summarizing Effects in Logistic Regression
6. Homework 4

Homework 4

1. Analysis the GDS5037 data. Suppose the samples are randomly chosen. Let Y denote patient's status, $Y = 1$ for mild asthma (MMA) and severe asthma (SA), $Y = 0$ for control.

   (1) Choose 4 identifiers by yourself as independent variables to built a logistic regression. Explain your result.

   (2) In (1) do you have any reasons or evidence to choose those 4 identifiers? If yes, please explain. If no, please try other 4 identifiers, report and explain the result, and compare the result with that in (1).

2. Problems in textbook 4.5, 4.7, 4.10, 4.12, 4.15, 4.16, 4.35.