Categorical Data Analysis

Chapter 5

Deyuan Li
School of Management
Fudan University

Fall 2014

Outline

5.1 Interpreting Parameters in Logistic Regression

For a binary response variable Y and an explanatory variable X, let $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$.

The logistic regression model is

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

Equivalently, the log odds, called the *logit*, has the linear relationship

This equates the logit link function to the linear predictor.

5.1.1 Interpreting β

The sign of β determines whether $\pi(x)$ is increasing or decreasing as x increases.

The $|\beta|$ determines the rate of climb or descent.
$\beta \to 0$: the curve flattens to a horizontal straight line.
$\beta = 0$: Y is independent of X.

For quantitative x with $\beta > 0$, the curve for $\pi(x)$ has the shape of the cdf of the logistic distribution.

5.1.1 Interpreting β

β and odds

$$\text{odds} = \pi(x)/[1-\pi(x)] = \exp(\alpha+\beta x) = e^{\alpha} e^{\beta x}.$$

- the odds increase multiplicatively by e^{β} for every 1-unit increase in x;
- e^{β} is an odds ratio, i.e.,

$$\frac{\text{odds at } X=x+1}{\text{odds at } X=x} = \frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = \frac{e^{\alpha} e^{\beta(x+1)}}{e^{\alpha} e^{\beta x}} = e^{\beta}.$$

β and linear approximation

Since $\pi(x)$ changes with x and $\partial\pi(x)/\partial x = \beta\,\pi(x)[1-\pi(x)]$, the rate of change in $\pi(x)$ per unit change in x varies.

5.1.1 Interpreting β

$\pi(x)$	slope $= \beta\,\pi(x)[1-\pi(x)]$
1/2	$\beta/4$
0.9 or 0.1	0.09β
$\to$ 1 or 0	$\to$ 0

The steepest slope occurs when

$$\begin{aligned} & \pi(x) = 1/2 \\ \Rightarrow \quad & \text{odds} = 1 \quad \Rightarrow \text{logit} = 0 \\ \Rightarrow \quad & \alpha + \beta x = 0 \quad \Rightarrow x = -\alpha/\beta. \end{aligned}$$

$x = (-\alpha/\beta)$ is sometimes called the *median effective level* and denoted by EL_{50}.

In toxicology studies it is called LD_{50} (LD = lethal dose, 致命剂量), i.e., the dose with a 50% chance of a lethal result.

5.1.1 Interpreting β

β and probability

An alternative way to interpret the effect reports the values of $\pi(x)$ at certain x values, such as their quartiles (e.g., $x_{0.25}$, $x_{0.75}$).

The change in $\pi(x)$ over the middle half of x values, from the lower quartile to the upper quartile of x, then describes the effect, i.e., $\Delta\pi(x) = \pi(x_{0.75}) - \pi(x_{0.25})$.

5.1.2 Looking at the data

Aim: to check if the logistic regression model is appropriate.

Tool: to plot sample proportions or logits against x.

Let

$n_i =$ number of observations at setting i of x,

$y_i =$ number of "1" outcomes at setting i,

$p_i =$ y_i/n_i, the sample proportion at setting i (MLE for π_i).

The sample logit at setting i is

$$\log \frac{p_i}{1-p_i} = \log \frac{y_i/n_i}{1-y_i/n_i} = \log \frac{y_i}{n_i - y_i}.$$

This is not finite when $y_i = 0$ or n_i.

5.1.2 Looking at the data

The adjustment

is the least-biased estimator of this form of the true logit.

Plot $(x_i, \log \frac{y_i+1/2}{n_i-y_i+1/2})$ and see if it is roughly a straight line.

Remark:

(1) When X is not continuous and n_i are not too small, the plot of sample logits against x should be roughly linear.

(2) When X is continuous or all n_i are small, the plot of sample logits against x is unsatisfactory.

5.1.2 Looking at the data

When X is continuous, two approaches.

1. Grouping data

   Group the data with nearby x values into categories.

   The plot of sample logits against x category should be roughly linear.

2. Use a smoothing mechanism to reveal trends

   For example, fit a generalized additive model (Section 4.8).

   This approach is better because it does not require choosing arbitrary categories.

5.1.3 Horseshoe crabs revisited

Recall data in Section 4.3.2.

Response: $Y = 1$ if a female crab has at least one satellite,
$Y = 0$ if she has none.

Predictor: $X =$ width of the crab.

$n = 173$ female crabs

Figure 4.7 plotted the data and showed the smoothed prediction of the mean provided by a generalized additive model (GAM), assuming a binomial response and logit link.

$\Rightarrow$ The logistic regression model appears to be adequate (since the curve of GAM is S-shape).

5.1.3 Horseshoe crabs revisited

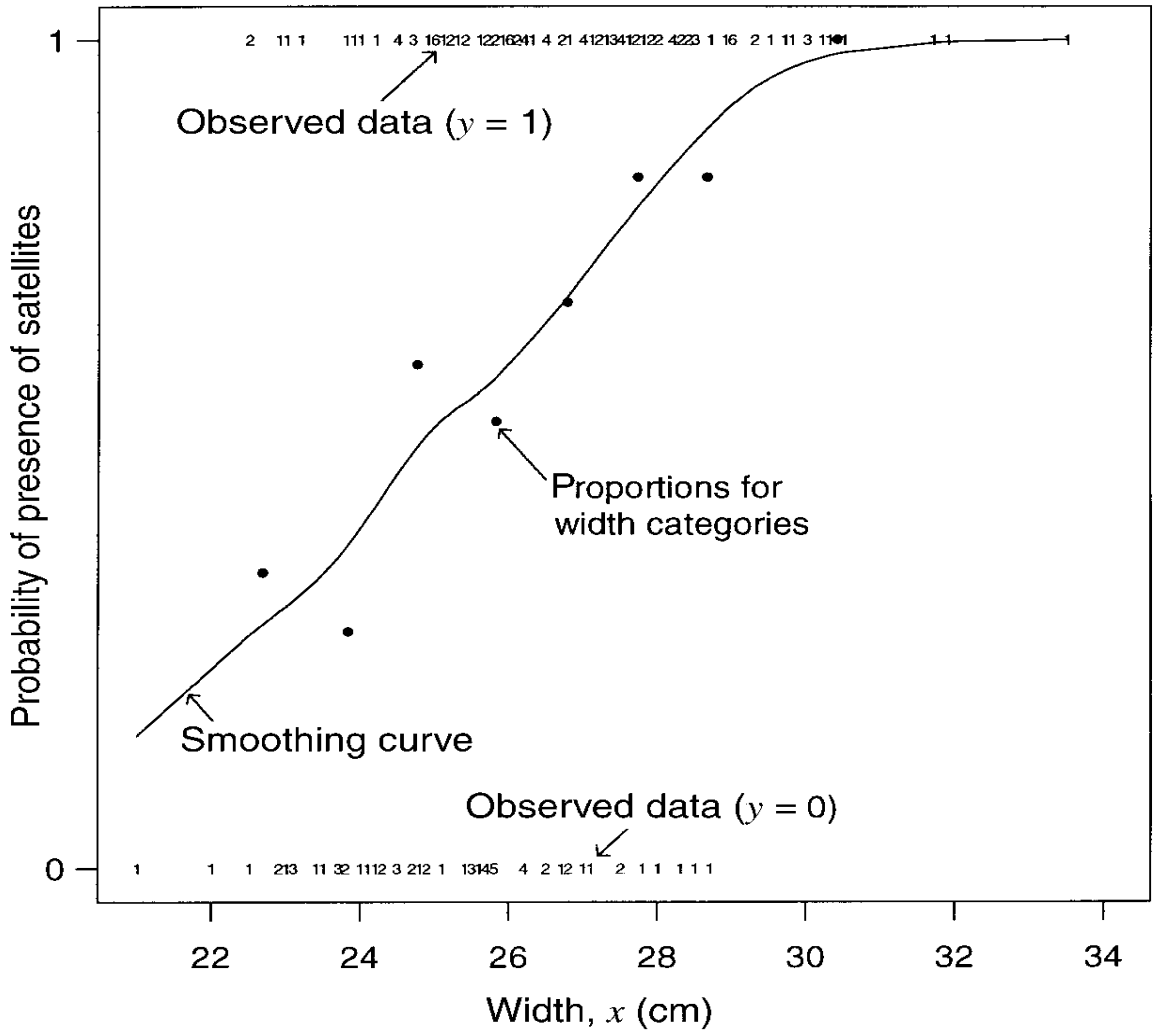

FIGURE 4.7 Whether satellites are present (1, yes; 0, no), by width of female crab, with smoothing fit of generalized additive model.

5.1.3 Horseshoe crabs revisited

Table 4.4 Sample Mean and Variance of Number of Satellites

Width(cm)	Number of Cases	Number of Satellites	Sample Mean	Sample Variance
$<$23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
$>$29.25	14	72	5.14	8.29

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.

5.1.3 Horseshoe crabs revisited

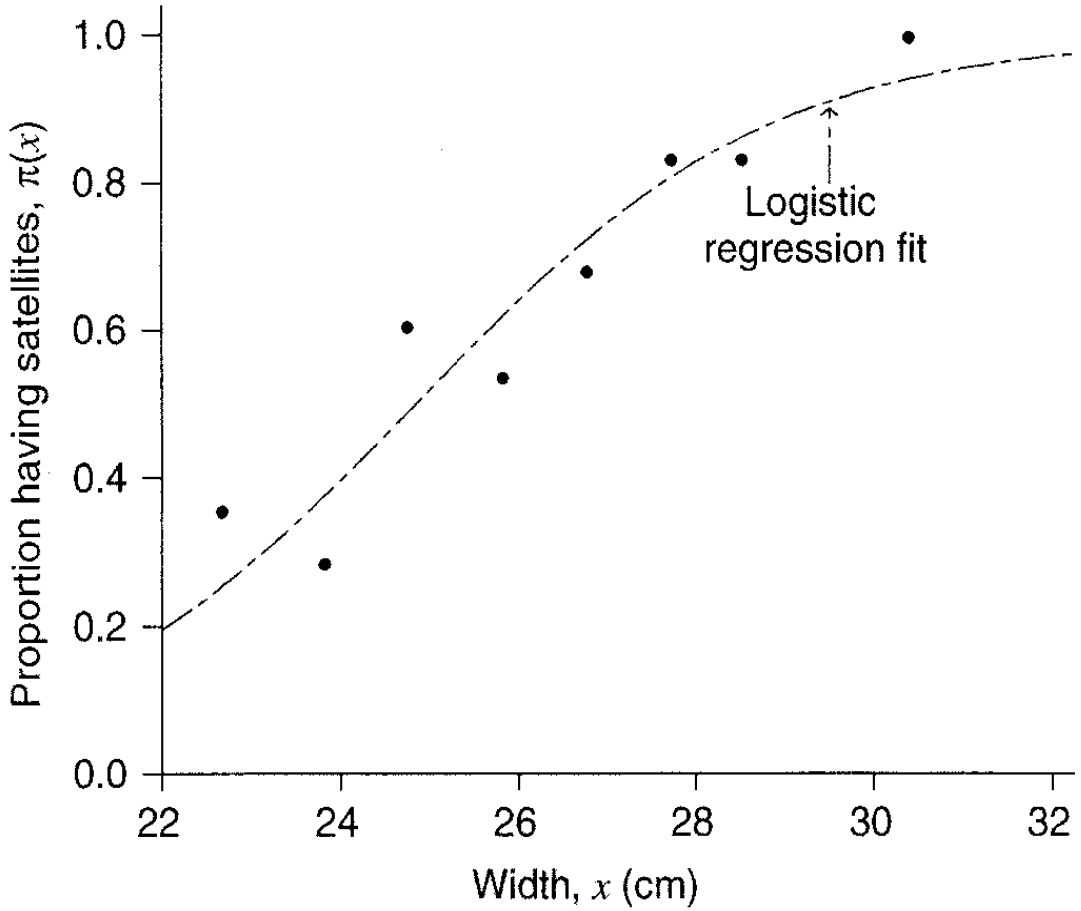

FIGURE 5.2 Observed and fitted proportions of satellites by width of female crab.

5.1.3 Horseshoe crabs revisited

Both the sample proportions plotted in Figure 5.2 and the GAM curve in Figure 4.7 show a roughly increasing trend.

$\Rightarrow$ fitting the logistic regression model with linear width predictor.

Table 5.1: SAS output of logistic regression model for ungrouped data (Table 4.3).

TABLE 5.1 Computer Output for Logistic Regression Model with Horseshoe Crab Data

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	171	194.4527
Pearson Chi-Square	171	165.1434
Log Likelihood		−97.2263

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf Limits		Wald Chi-Sq	P>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

5.1.3 Horseshoe crabs revisited

$\hat{\alpha} = -12.3508$ and $\hat{\beta} = 0.4972$, so

$$\hat{\pi}(x) = \frac{\exp(-12.3508 + 0.4972\,x)}{1 + \exp(-12.3508 + 0.4972\,x)}.$$

At $x = 26.3$ cm (the overall mean width), $\hat{\pi}(x) = 0.674$.
At $x = -\hat{\alpha}/\hat{\beta} = 12.3508/0.4972 = 24.8$, $\hat{\pi}(x) = 1/2$.

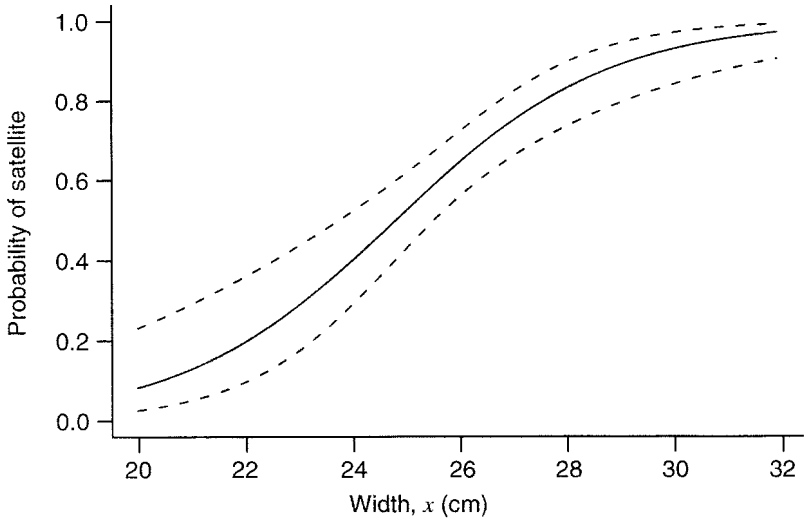

FIGURE 5.3 Prediction equation and 95% confidence bands for probability of satellite as a function of width.

5.1.3 Horseshoe crabs revisited

Odds ratio: $\exp(\hat{\beta}) = \exp(0.4972) = 1.64$.
$\Rightarrow$ For each 1-cm increase in width, there is a 64% increase in odds of having at least one satellite.

Linear approximation: At the overall mean width, slope
$= \hat{\beta}\,\hat{\pi}(x)[1 - \hat{\pi}(x)] = 0.4971 \times 0.674 \times 0.326 = 0.11$;
$\Rightarrow$ For each 1-cm increase in width, $\hat{\pi}(x)$ increases by about 0.11.

Probability: The quartiles of x are 24.9, 26.1 and 27.7. The corresponding $\hat{\pi}(x)$ are 0.51, 0.65 and 0.81.
$\Rightarrow$ The $\hat{\pi}(x)$ increases by $\Delta\hat{\pi}(x) = 0.81 - 0.51 = 0.30$ over the x values for the middle half of the sampled widths.

5.1.3 Horseshoe crabs revisited

The summary based on quartiles is useful for comparing the effects of predictors having different units.

For instance, with crab weight as the predictor, $\text{logit}[\hat{\pi}(x)] = -3.695 + 1.815\,x$.

- Weight is not comparable to width, so $\hat{\beta}_{\text{width}} = 0.497$ is not comparable to $\hat{\beta}_{\text{weight}} = 1.815$.
- The quartiles for weight are 2.00, 2.35 and 2.85. The corresponding $\hat{\pi}(x)$ are 0.48, 0.64 and 0.81.
- The $\hat{\pi}(x)$ increases by $\Delta\hat{\pi}(x) = 0.81 - 0.48 = 0.33$ over the x values for the middle half of the sampled weights.
- The effect of weight ($\Delta\hat{\pi}(x) = 0.33$) is similar to that of width ($\Delta\hat{\pi}(x) = 0.30$).

Outline

5.2.1 Types of inference

For the model with a single predictor,

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

significance tests focus on

$$H_0 : \beta = 0, \text{ the hypothesis of independence.}$$

Hypothesis testing

Wald test: uses the log likelihood at $\hat{\beta}$.
Test statistic $z = \hat{\beta}/\text{SE}(\hat{\beta})$ or z^2.
Under H_0, z^2 is asymptotically χ^2_1.

5.2.1 Types of inference

Likelihood-ratio test (LRT): uses twice the difference between the maximized log likelihood at $\hat{\beta}$ and at $\beta = 0$.
Under H_0, it is also asymptotically χ^2_1.

Score test: uses the log likelihood at $\beta = 0$ through the derivative of the log likelihood at that point.

Remarks:

(1) For large samples, three tests usually give similar results.

(2) The LRT is preferred over the Wald, since it uses more information.

5.2.1 Types of inference

Confidence intervals

CI for β

An interval for β results from inverting a test of $H_0 : \beta = \beta_0$.

For the Wald approach, this means $[(\hat{\beta} - \beta_0)/\text{SE}(\hat{\beta})]^2 \leq z^2_{\alpha/2}$; so the interval is $\hat{\beta} \pm z_{\alpha/2} \times \text{SE}(\hat{\beta})$.

CI for $\pi(x)$

For $x = x_0$, $\text{logit}[\hat{\pi}(x_0)] = \hat{\alpha} + \hat{\beta}\, x_0$ has a large-sample SE:

$$\text{SE}(\hat{\alpha} + \hat{\beta}\, x_0) = [\text{var}(\hat{\alpha}) + x_0^2\, \text{var}(\hat{\beta}) + 2\, x_0\, \text{cov}(\hat{\alpha}, \hat{\beta})]^{1/2}.$$

A 95% CI for $\text{logit}[\hat{\pi}(x_0)]$ is

$$(\hat{\alpha} + \hat{\beta}\, x_0) \pm 1.96 \times \text{SE}(\hat{\alpha} + \hat{\beta}\, x_0) = (L_{\text{logit}}, U_{\text{logit}}).$$

5.2.1 Types of inference

The corresponding 95% CI for $\pi(x_0)$ is

$$\left(\frac{\exp(L_{\text{logit}})}{1+\exp(L_{\text{logit}})}, \frac{\exp(U_{\text{logit}})}{1+\exp(U_{\text{logit}})}\right).$$

5.2.2 Inference for horseshoe crab data

TABLE 5.1 Computer Output for Logistic Regression Model with Horseshoe Crab Data

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	171	194.4527
Pearson Chi-Square	171	165.1434
Log Likelihood		−97.2263

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf Limits		Wald Chi-Sq	P>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

5.2.2 Inference for horseshoe crab data

The Wald statistic $z = \hat{\beta}/\text{SE}(\hat{\beta}) = 0.4972/0.1017 = 4.8889$ or $z^2 = 4.8889^2 = 23.89 \sim \chi^2_1$, with $P < 0.0001$.
$\Rightarrow$ Strong evidence of a positive (because $\hat{\beta} > 0$) width effect.

The maximized log likelihood equal -112.88 under $H_0 : \beta = 0$, and -97.23 for the full model.

The likelihood-ratio statistic equals $-2[-112.88 - (-97.23)] = 31.3$ with $df = 1$.

This provides even stronger (because $31.3 > 23.89$) evidence than the Wald test.

5.2.2 Inference for horseshoe crab data

The Wald 95% CI for β is
$0.4972 \pm 1.96 \times 0.1017 = (0.298, 0.697)$.

Likelihood-ratio CI for β is $(0.308, 0.709)$. The CI for the effect on the odds per 1-cm increase in width equals
$(e^{0.308}, e^{0.709}) = (1.36, 2.03)$.
$\Rightarrow$ A 1-cm increase in width has at least a 36% increase and at most a doubling in the odds of having at least one satellite.

Most software for logistic regression also reports estimates and CIs for $\pi(x)$ (e.g., PROC GENMOD in SAS with the OBSTATS option).

5.2.2 Inference for horseshoe crab data

For $x = 26.5$ (near the mean width), the estimated logit is $-12.3508 + 0.4972 \times 26.5 = 0.825$ and $\hat{\pi}(26.5) = e^{0.825}/(1 + e^{0.825}) = 0.695$.

Software reports: $\widehat{\text{var}}(\hat{\alpha}) = 6.910$, $\widehat{\text{var}}(\hat{\beta}) = 0.01035$, and $\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) = -0.2668$.
$\Rightarrow \widehat{\text{var}}\{\text{logit}[\hat{\pi}(x)]\} = 6.910 + x^2(0.01035) + 2x(-0.2668)$.

At $x = 26.5$ this is 0.038, so the 95% CI for logit$[\pi(26.5)]$ equals $0.825 \pm 1.96 \times \sqrt{0.038} = (0.44, 1.21)$.
$\Rightarrow$ The CI for $\pi(26.5)$ equals $(e^{0.44}/(1 + e^{0.44}),\ e^{1.21}/(1 + e^{1.21})) = (0.61, 0.77)$.

5.2.2 Inference for horseshoe crab data

Use sample proportions (the saturated model) to estimate $\pi(x)$.

Six female crabs in the sample had width $x = 26.5$ and four of them had satellites.
$\Rightarrow \hat{\pi}(26.5) = 4/6 = 0.67$, similar to the model-based estimate 0.695.
$\Rightarrow$ The 95% score CI based on these six observations equals $(0.30, 0.90)$, wider than the model-based CI $(0.61, 0.77)$.

When the logistic regression model truly holds, the model-based estimator of a probability is considerably better than the sample proportion.

5.2.3 Checking goodness of fit: ungrouped and grouped data

Compared with more complex models:

Likelihood-ratio test compares the model to more complex ones, which:

- might contain a nonlinear effect, such as a quadratic term;
- might consider interaction for models with multiple predictors.

If more complex models do not fit better, this provides some assurance that the model chosen is reasonable.

Overall goodness of fit:

The test of the model compares the observed counts and fitted values using a Pearson X^2 or likelihood-ratio G^2 statistic.

5.2.3 Checking goodness of fit: ungrouped and grouped data

Grouped data:

- counts of success/failure at each setting for the predictors.
- the saturated model has a parameter at each **setting**;
- number of subjects can be increased without increasing number of settings (parameters).

Ungrouped data:

- individual 0-1 observations at the subject level;
- the saturated model has a parameter for each **subject**;
- increase in number of subjects also increases number of parameters.

An asymptotic chi-square distribution can be applied to grouped data, but not ungrouped data.

5.2.4 Goodness of fit of model for horseshoe crabs

Compare the model containing a single predictor $x =$ width with a more complex model containing a quadratic term.

With width centered at 0 by subtracting its mean of 26.3, the more complex model has fit

$$\text{logit}[\hat{\pi}(x)] = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = 0.618 + 0.533 x + 0.040 x^2.$$

The quadratic estimate has $\text{SE}(\hat{\beta}_2) = 0.046$.

The LR statistic for testing $\beta_2 = 0$ equals 0.83 (df$=1$).
$\Rightarrow$ No much evidence to support adding quadratic term.

5.2.4 Goodness of fit of model for horseshoe crabs

For overall goodness of fit, one may take 66 distinct widths as the number of settings and view the data as a 66×2 contingency table.

However, the chi-square theory for X^2 and G^2 does not apply to this situation because:

(1) most fitted counts are very small due to few observations at most widths;

(2) when more data are collected, additional distinct width values would occur, so the contingency table would contain more cells rather than a fixed number.

5.2.4 Goodness of fit of model for horseshoe crabs

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

Width (cm)	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

5.2.4 Goodness of fit of model for horseshoe crabs

TABLE A.6 SAS Code for Modeling Grouped Crab Data in Table 5.2

```
data crab;
input width y n satell;  logcases=log(n);
datalines;
22.69  5 14 14
...
30.41 14 14 72
;
proc genmod;
  model y/n=width/dist=bin link=logit 1rci alpha=.01 type3;
proc logistic;
  model y/n=width/influence stb;
  output out=predict p=pi_hat lower=LCL upper=UCL;
proc print data=predict;
proc genmod;
  model satell=width/dist=poi link=log offset=logcases residuals;
```

5.2.4 Goodness of fit of model for horseshoe crabs

With such grouped data, the fitted values are much larger than those from the 66×2 contingency table.

Then, X^2 and G^2 have better validity, although the chi-squared theory still is not perfect since $\pi(x)$ is not constant in each category. We obtain $X^2 = 5.3$ and $G^2 = 6.2$.

Table 5.2 has 8 binomial samples and the model has two parameters (α and β), so df=$8 - 2 = 6$.
$\Rightarrow$ Neither X^2 nor G^2 shows evidence of lack of fit ($P > 0.4$).
$\Rightarrow$ We can feel more comfortable about using the model for the original ungrouped data.

5.2.5 Checking goodness of fit with ungrouped data by grouping (estimated probabilities of success)

To group data is a good way. But, as the number of predictors increases, simultaneous grouping of values for each predictor can produce a contingency table with a large number of cells, most of which have small counts.

Regardless of the number of predictors, one can partition observed and fitted values according to the estimated probabilities of success ($\hat{\pi}$) using the original ungrouped data.

One common approach forms the groups in the partition so they have approximately equal size.

5.2.5 Checking goodness of fit with ungrouped data by grouping (estimated probabilities of success)

Illustration: Consider a model

$$\text{logit}[\pi(\mathbf{x}_k)] = \beta_0 + \beta_1 x_{1,k} + \beta_2 x_{2,k} + \cdots + \beta_p x_{p,k},$$

where $\mathbf{x}_k = (1, x_{1,k}, \ldots, x_{p,k})'$, $k = 1, \ldots, n$.

Fit the model for the ungrouped data and obtain $\hat{\pi}_k$.

Sort the $\hat{\pi}_k$ in descending manner and let $\hat{\pi}_{(k)}$ denote the ordered value such that $\hat{\pi}_{(1)} \geq \hat{\pi}_{(2)} \geq \cdots \geq \hat{\pi}_{(n)}$.

Partition the observations into g groups with equal size.
$\Rightarrow$ Each group has n/g observations.

$\hat{\pi}_{(1)}, \ldots, \hat{\pi}_{(n/g)}$	$\hat{\pi}_{([n/g]+1)}, \ldots, \hat{\pi}_{(2n/g)}$	$\cdots$	$\hat{\pi}_{([(g-1)n/g]+1)}, \ldots, \hat{\pi}_{(n)}$
Group 1	Group 2	$\cdots$	Group g

5.2.5 Checking goodness of fit with ungrouped data by grouping (estimated probabilities of success)

Let y_{ij} denote the binary outcome (0 or 1) for observation j in group i of the partition, $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. Let $\hat{\pi}_{ij}$ denote the corresponding fitted probability.

Then, for group i,

$$\begin{aligned}
\sum_j y_{ij} &= \text{observed number of successes,} \\
n_i - \sum_j y_{ij} &= \text{observed number of failures,} \\
\sum_j \hat{\pi}_{ij} &= \text{fitted number of successes,} \\
n_i - \sum_j \hat{\pi}_{ij} &= \text{fitted number of failures.}
\end{aligned}$$

5.2.5 Checking goodness of fit with ungrouped data by grouping (estimated probabilities of success)

Hosmer and Lemeshow (1980) proposed a Pearson type statistic comparing the observed and fitted counts for such partition:

This statistic does not have a limiting chi-squared distribution, because the observations in a group are not identical trials.

However, when the *number of distinct patterns of covariate values* (协变量不同取值的组合数) equals the sample size, the null distribution is approximated by chi-squared with $df = g - 2$.

5.2.5 Checking goodness of fit with ungrouped data by grouping (estimated probabilities of success)

For the logistic regression fit to the horseshoe crab data with continuous width predictor, the Hosmer-Lemeshow statistic with $g = 10$ groups equals 3.5, with df$= 8$.

$\Rightarrow$ It also indicates a decent fit.

Outline

5.3.1 ANOVA-type representation of factors

We first consider a single factor X with I categories.

In row i of the $I \times 2$ table, y_i is the number of successes out of n_i trials. We treat y_i as binomial with parameter π_i.

The logit model with a factor is

The higher β_i is, the higher the value of π_i.

The right-hand side resembles the model formula for cell means in one-way ANOVA.

$\Rightarrow$ One parameter can be set to 0, say $\beta_I = 0$.

5.3.1 ANOVA-type representation of factors

If the values do not satisfy this, we can recode so that it is true. For instance, set $\tilde{\beta}_i = \beta_i - \beta_I$ and $\tilde{\alpha} = \alpha + \beta_I$, then

$$\text{logit}(\pi_i) = \alpha + \beta_i = (\alpha + \beta_I) + (\beta_i - \beta_I) =: \tilde{\alpha} + \tilde{\beta}_i,$$

where the new parameters $\{\tilde{\beta}_i\}$ satisfy the constraint $\tilde{\beta}_I = 0$.

If we set $\beta_I = 0$,
$\Rightarrow$ $\text{logit}(\pi_I) = \alpha$, i.e., α equals the logit in row *I*;
$\Rightarrow$ $\text{logit}(\pi_i) - \text{logit}(\pi_I) = (\alpha + \beta_i) - \alpha = \beta_i$,
i.e., β_i is the difference between the logits in rows *i* and *I*.
$\Rightarrow$ $\text{logit}(\pi_i) - \text{logit}(\pi_I) = \log\left[\left(\frac{\pi_i}{1-\pi_i}\right) / \left(\frac{\pi_I}{1-\pi_I}\right)\right] = \beta_i$,
i.e., β_i is also the log odds ratio for rows *i* and *I*.

5.3.1 ANOVA-type representation of factors

For any $\{\pi_i > 0\}$, $\{\beta_i\}$ exist such that the above logit model holds.

The model has as many parameters (I) as binomial observations and is *saturated*.

When a factor X has no effect,
$\Rightarrow \beta_1 = \beta_2 = \cdots = \beta_I = 0$,
$\Rightarrow$ equivalent to $\pi_1 = \pi_2 = \cdots = \pi_I$,
$\Rightarrow$ the model has only an intercept term and specifies statistical independence of X and Y.

5.3.2 Dummy variables in logit models

An equivalent expression of the above model in ANOVA-type representation uses *dummy variables*（哑变量）.

0/1 coding

If we take category I as reference, the model is

where the dummy variables

$$x_i^* = \begin{cases} 1 & \text{for observations in category } i, \\ 0 & \text{for observations in other categories.} \end{cases}$$

Having no dummy variable for category I corresponds to the constraint $\beta_I = 0$ in the ANOVA-type representation.

5.3.2 Dummy variables in logit models

-1/1 effect coding

Set $\sum_{j=1}^{I} \beta_j^{\#} = 0. \Rightarrow \beta_i^{\#} = -\sum_{j \neq i} \beta_j^{\#}, \quad i, j = 1, \ldots, I.$

If we take category I as reference, the model is

where the dummy variables

$$x_i^{\#} = \begin{cases} 1 & \text{for observations in category } i, \\ -1 & \text{for observations in category } I, \\ 0 & \text{for observations in other categories.} \end{cases}$$

$\beta_I^{\#}$ is not included in the model, and $\beta_I^{\#} = -\sum_{j=1}^{I-1} \beta_j^{\#}$.

Example: Suppose X has $I = 2$ categories, so $\beta_2^{\#} = -\beta_1^{\#}$. Then the single dummy variable $x^{\#} = 1$ in category 1 and $x^{\#} = -1$ in category 2.

5.3.2 Dummy variables in logit models

Comparisons

Taking category I as reference, the logits expressed by different coding schemes are as follows:

	Logit		
Category	ANOVA	0/1 Coding	± 1 Coding
$i = 1, \ldots, I-1$	$\alpha + \beta_i$	$\alpha^* + \beta_i^*$	$\alpha^\# + \beta_i^\#$
I	α	α^*	$\alpha^\# - \sum_{i=1}^{I-1} \beta_i^\#$

5.3.2 Dummy variables in logit models

So,

- ANOVA vs. 0/1 coding:
  $\alpha = \alpha^*$ and $\beta_i = \beta_i^*$ for all $i = 1, \ldots, I-1$ and $\beta_I = \beta_I^* = 0$.
  $\Rightarrow$ The parameters of these two approaches are equivalent.

- ANOVA vs. ± 1 coding:
  For category I, $\alpha = \alpha^\# - \sum_{j=1}^{I-1} \beta_j^\#$.
  For other categories, $\alpha + \beta_i = \alpha^\# + \beta_i^\#$. $\Rightarrow \beta_i =$
  $\alpha^\# + \beta_i^\# - \alpha = \alpha^\# + \beta_i^\# - (\alpha^\# - \sum_{j=1}^{I-1} \beta_j^\#) = \beta_i^\# + \sum_{j=1}^{I-1} \beta_j^\#$.

5.3.2 Dummy variables in logit models

The log odds ratio for categories a and b ($a, b = 1, \ldots, I-1$) is

$$\begin{aligned}
&\text{logit}(\pi_a) - \text{logit}(\pi_b) \\
&= \begin{cases}
\text{ANOVA} & (\alpha + \beta_a) - (\alpha + \beta_b) = \beta_a - \beta_b\,, \\
\text{0/1 coding} & (\alpha^* + \beta_a^*) - (\alpha^* + \beta_b^*) = \beta_a^* - \beta_b^* = \beta_a - \beta_b\,, \\
\pm 1 \text{ coding} & (\alpha^\# + \beta_a^\#) - (\alpha^\# + \beta_b^\#) = \beta_a^\# - \beta_b^\# \\
& = (\beta_a - \sum_{j=1}^{I-1} \beta_j^\#) - (\beta_b - \sum_{j=1}^{I-1} \beta_j^\#) = \beta_a - \beta_b\,;
\end{cases}
\end{aligned}$$

and for categories a ($a = 1, \ldots, I-1$) and I (reference) is

$$\begin{aligned}
&\text{logit}(\pi_a) - \text{logit}(\pi_I) \\
&= \begin{cases}
\text{ANOVA} & (\alpha + \beta_a) - \alpha = \beta_a\,, \\
\text{0/1 coding} & (\alpha^* + \beta_a^*) - \alpha^* = \beta_a^* = \beta_a\,, \\
\pm 1 \text{ coding} & (\alpha^\# + \beta_a^\#) - (\alpha^\# - \sum_{j=1}^{I-1} \beta_j^\#) = \beta_a^\# + \sum_{j=1}^{I-1} \beta_j^\# = \beta_a\,.
\end{cases}
\end{aligned}$$

5.3.2 Dummy variables in logit models

Reparameterizing a model may change parameter estimates but does not change the model fit $\{\hat{\pi}_i\}$ or the effects of interest.

The values β_i or $\hat{\beta}_i$ for a single category is irrelevant.

A parameter or its estimate makes sense only by comparison with another category.

5.3.3 Alcohol and infant malformation example revisited

TABLE 5.3 Logits and Proportion of Malformation for Table 3.7

Alcohol Consumption	Present	Absent	Logit	Proportion Malformed	
				Observed	Fitted
0	48	17,066	−5.87	0.0028	0.0026
< 1	38	14,464	−5.94	0.0026	0.0030
1–2	5	788	−5.06	0.0063	0.0041
3–5	1	126	−4.84	0.0079	0.0091
≥ 6	1	37	−3.61	0.0263	0.0231

For the ANOVA-type logit model, we treat malformations (present vs absent) as the response and alcohol consumption as an explanatory factor with 5 categories.

5.3.3 Alcohol and infant malformation example

Regardless of the constraint for $\{\beta_i\}$, $\{\hat{\alpha} + \hat{\beta}_i\}$ are the sample logits and reported in Table 5.3. For instance,

$$\begin{aligned} \text{logit}(\hat{\pi}_1) &= \hat{\alpha} + \hat{\beta}_1 \\ &= \log(\text{Nr. Present}/\text{Nr. Absent}) = \log(48/17066) \\ &= -5.87. \end{aligned}$$

If we set $\beta_5 = 0$,
$\Rightarrow \hat{\alpha} = \text{logit}(\hat{\pi}_5) = -3.61$;
$\Rightarrow \hat{\beta}_1 = \text{logit}(\hat{\pi}_1) - \hat{\alpha} = -5.87 - (-3.61) = -2.26.$

If we set $\beta_1 = 0$,
$\Rightarrow \hat{\alpha} = \text{logit}(\hat{\pi}_1) = -5.87$;
$\Rightarrow \hat{\beta}_5 = \text{logit}(\hat{\pi}_5) - \hat{\alpha} = -3.61 - (-5.87) = 2.26.$

5.3.3 Alcohol and infant malformation example

Model with all $\beta_i = 0$ specifies independence. In this case,

$$\hat{\alpha} = \log\left(\frac{48+38+5+1+1}{17066+14464+788+126+37}\right) = \log\left(\frac{93}{32481}\right) = -5.86,$$

i.e., $\hat{\alpha}$ equals the logit for the overall sample proportion of malformations.

To test H_0 : independence (df$= I - 1 = 4$),

- the Pearson statistic is $X^2 = 12.1$ $(P = 0.02)$,
- the likelihood-ratio statistic is $G^2 = 6.2$ $(P = 0.19)$.

These provide mixed signals.

5.3.3 Alcohol and infant malformation example

Reason: Table 5.3 has a mixture of very small, moderate and extremely large counts.
$\Rightarrow$ Even though the total count $n = 32574$ is large, the null sampling distributions of X^2 or G^2 may not be close to χ^2_{df}.

The P-values using the exact conditional distributions of X^2 and G^2 are 0.03 and 0.13.

In any case, these statistics ignore the order of alcohol consumption.

5.3.4 Linear logit model for $I \times 2$ tables

The model $\text{logit}(\pi_i) = \alpha + \beta_i$ treats the explanatory factor as nominal. For ordered factor categories, one can assign scores.

Assign **scores** $\{x_1, x_2, \ldots, x_I\}$ to the categories of factor X.

When one expects a monotone effect of X on Y, it is naturally to fit the *linear logit model* (linear in logit)

$$\text{logit}(\pi_i) = \alpha + \beta x_i.$$

The independence model is the special case with $\beta = 0$.

5.3.4 Linear logit model for $I \times 2$ tables

The factor of alcohol consumption was based on categorizing a naturally continuous variable.

With scores $\{x_1 = 0, x_2 = 0.5, x_3 = 1.5, x_4 = 4.0, x_5 = 7.0\}$ (the last score is somewhat arbitrary), Table 5.4 shows results of the linear logit model.

TABLE 5.4 Computer Output for Logistic Regression Model with Infant Malformation Data

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	3	1.9487
Pearson Chi-Square	3	2.0523
Log Likelihood		−635.5968

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf	Limits	Wald Chi-Sq	Pr>ChiSq
Intercept	−5.9605	0.1154	−6.1930	−5.7397	2666.41	<.0001
alcohol	0.3166	0.1254	0.0187	0.5236	6.37	0.0116

5.3.4 Linear logit model for $I \times 2$ tables

The estimated odds ratio $\exp(0.317) = 1.37$ represents the estimated multiplicative effect of a unit increase in daily alcohol consumption on the odds of malformation.

The observed and fitted proportions of malformation are presented in Table 5.3.

Comparing observed and fitted counts, the goodness of fit test gives $G^2 = 1.95$ and $X^2 = 2.05$ with
df$=$ number of categories $-$ number of parameters $= 5 - 2 = 3$.
$\Rightarrow$ The linear logit model seems to fit well.

Outline

5.4 Multiple logistic regression

Logistic regression model with multiple explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)$:

$$\text{logit}[P(Y = 1)] = \text{logit}[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p .$$

The alternative formula is

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

5.4 Multiple logistic regression

More specifically, taking β_1 for illustration, consider the logits at $x_1 = x_1^*$ (a constant) and at $x_1 = x_1^* + 1$, with other $x_i = x_i^*$ being the same for both situations.

The difference in logit (equivalent to log odds ratio) equals

$$\begin{aligned}
&\text{logit}[\pi(x_1 = x_1^* + 1,\, x_2 = x_2^*,\, \cdots,\, x_p = x_p^*)] \\
&-\text{logit}[\pi(x_1 = x_1^*,\, x_2 = x_2^*,\, \cdots,\, x_p = x_p^*)] \\
&= [\alpha + \beta_1(x_1^* + 1) + \beta_2 x_2^* + \cdots + \beta_p x_p^*] \\
&\quad -[\alpha + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*] \\
&= \beta_1\,,
\end{aligned}$$

so the odds ratio $\exp(\beta_1)$ is the multiplicative effect on the odds of a 1-unit increase in x_1, at fixed levels of other x_i.

5.4.1 Logit models for multiway contingency tables

Model with two binary factors

For two binary predictors X and Z and one binary outcome Y.

X i	Z k	Sample size n_{ik}	Parameter π_{ik}	Count $Y=1$	Count $Y=0$
1	1	n_{11}	π_{11}	y_{11}	$n_{11}-y_{11}$
1	2	n_{12}	π_{12}	y_{12}	$n_{12}-y_{12}$
2	1	n_{21}	π_{21}	y_{21}	$n_{21}-y_{21}$
2	2	n_{22}	π_{22}	y_{22}	$n_{22}-y_{22}$

We treat the sample size n_{ik} at given combinations (i,k) of X and Z as fixed, for $i,k=1,2$; and regard the two counts on Y at each (i,k) setting as binomial, with different binomials treated as independent.

5.4.1 Logit models for multiway contingency tables

Taking category 2 as reference, the dummy variables x and z respectively for X and Z are as follows:

$$x = \begin{cases} 1 \equiv x_1, & \text{if } i = 1; \\ 0 \equiv x_2, & \text{if } i = 2. \end{cases} \qquad z = \begin{cases} 1 \equiv z_1, & \text{if } k = 1; \\ 0 \equiv z_2, & \text{if } k = 2. \end{cases}$$

The model

$$\text{logit}[\pi(X = i, Z = k)] = \text{logit}(\pi_{ik}) = \alpha + \beta_1 x_i + \beta_2 z_k$$

has main effects for X and Z, but no interaction effects.

$\Rightarrow$ The effect of one factor is the same at each level of the other.

For instance, at category k of Z, the effect on the logit of changing categories of X is

$$\begin{aligned} \text{logit}(\pi_{1k}) - \text{logit}(\pi_{2k}) &= (\alpha + \beta_1 x_1 + \beta_2 z_k) - (\alpha + \beta_1 x_2 + \beta_2 z_k) \\ &= \beta_1 x_1 - \beta_1 x_2 = \beta_1 \times 1 - \beta_1 \times 0 = \beta_1. \end{aligned}$$

5.4.1 Logit models for multiway contingency tables

The result is the same at each level of Z, no matter $k = 1$ or $k = 2$. $\Rightarrow$ There is *homogenous XY association*.

$\exp(\beta_1)$ is the **conditional** odds ratio between X and Y.

Controlling for Z, the odds of success when $x = 1$ equal $\exp(\beta_1)$ times the odds when $x = 0$.

When $\beta_1 = 0$, the common odds ratio (for all partial tables XY given Z) equals 1. $\Rightarrow$ X and Y are independent in each partial table, or *conditionally independent, given Z*.

5.4.1 Logit models for multiway contingency tables

General model

For X with I categories and Z with K categories, the model

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

represents effects of X with parameters $\{\beta_i^X\}$ and effects of Z with parameters $\{\beta_k^Z\}$.

Conditional independence between X and Y, given Z, corresponds to $\beta_1^X = \beta_2^X = \cdots = \beta_I^X$.

For each factor (X or Z), one parameter in the above model is redundant. Fixing one at 0, such as $\beta_I^X = \beta_K^Z = 0$.

If X and Z have two categories, then $\beta_1^X = \beta_1$ and $\beta_2^X = 0$, and with $\beta_1^Z = \beta_2$ and $\beta_2^Z = 0$.

5.4.2 AIDS and AZT example

TABLE 5.5 Development of AIDS Symptoms by AZT Use and Race

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

Source: New York Times, Feb. 15, 1991.

$X =$ immediate AZT (zidovudine，齐多夫定（艾滋病防护药）) treatment or not;
yes $\Rightarrow x_1 = 1$, no $\Rightarrow x_2 = 0$.

$Z =$ race; white $\Rightarrow z_1 = 1$, black $\Rightarrow z_2 = 0$.

$Y =$ develop AIDS symptoms in 3 years;
yes $\Rightarrow Y = 1$, no $\Rightarrow Y = 0$.

5.4.2 AIDS and AZT example

TABLE A.7 SAS Code for Logit Modeling of AIDS Data in Table 5.5

```
data aids;
input race $ azt $ y n @@;
datalines;
  White Yes 14 107    White No 32 113    Black Yes 11 63    Black No 12 55
;
proc genmod;  class race azt;
  model y/n = azt race / dist = bin type3 lrci residuals obstats;
proc logistic;  class race azt / param = reference;
  model y/n = azt race / aggregate scale = none clparm = both clodds = both;
  output out = predict p = pi_hat lower = lower upper = upper;
proc print data = predict;
proc logistic;  class race azt (ref = first) / param = ref;
  model y/n = azt / aggregate= (azt race) scale = none;
```

5.4.2 AIDS and AZT example

Follow the model with two binary factors:

$$\text{logit}(\pi_{ik}) = \alpha + \beta_1 x_i + \beta_2 z_k.$$

Then

- α is the log odds of developing AIDS symptoms for black subjects without immediate AZT use;
- β_1 is the increment to the log odds for those with immediate AZT use;
- β_2 is the increment to the log odds for white subjects.

Table 5.6 shows output.

5.4.2 AIDS and AZT example

TABLE 5.6 Computer Output for Logit Model with AIDS Symptoms Data

Goodness-of-Fit Statistics

Criterion	DF	Value	Pr > ChiSq
Deviance	1	1.3835	0.2395
Pearson	1	1.3910	0.2382

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

Odds Ratio Estimates

Effect	Estimate	95% Wald Confidence Limits	
azt	0.487	0.282	0.841
race	1.057	0.600	1.861

Profile Likelihood Confidence Interval for Odds Ratios

Effect	Estimate	95% Confidence Limits	
azt	0.487	0.279	0.835
race	1.057	0.605	1.884

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

5.4.2 AIDS and AZT example

The estimated odds ratio between immediate AZT use and development of AIDS symptoms equals $\exp(-0.7195) = 0.487$.

$\Rightarrow$ For each race, the estimated odds of symptoms are half as high for those who took AZT immediately.

- The Wald CI for this effect is $\exp[-0.7195 \pm 1.96 \times 0.2790] = (0.282, 0.841)$.
- The likelihood-based CI is (0.279, 0.835), very similar.

5.4.2 AIDS and AZT example

The hypothesis of conditional independence of AZT treatment and development of AIDS symptoms, controlling for race, is $H_0 : \beta_1 = 0$.

- The likelihood-ratio statistic comparing models with and without β_1 equals 6.9, with df$= 1$.
  $\Rightarrow$ Evidence of association ($P = 0.01$).
- The Wald statistic provides similar results:
  $[\hat{\beta}_1/\text{SE}(\hat{\beta}_1)]^2 = [-0.7195/0.2790]^2 = 6.6507$ ($P = 0.0099$).

5.4.2 AIDS and AZT example

Table 5.7 shows parameter estimates for three ways of defining factor parameters in the general models:

1) setting parameter of the last category equal to 0;
2) setting parameter of the first category equal to 0;
3) having parameters of a factor sum to zero.

TABLE 5.7 Parameter Estimates for Logit Model Fitted to Table 5.5

	Definition of Parameters		
Parameter	Last = Zero	First = Zero	Sum = Zero
Intercept	−1.074	−1.738	−1.406
AZT Yes	−0.720	0.000	−0.360
No	0.000	0.720	0.360
Race White	0.055	0.000	0.028
Black	0.000	−0.055	−0.028

5.4.2 AIDS and AZT example

For each coding scheme, at a given combination of AZT use and race, the estimated probability of developing AIDS symptoms is the same (because there is no interaction term).

For instance, the intercept estimate plus the estimate for immediate AZT use plus the estimate for being white is -1.738 for each scheme:

$$\begin{array}{llll} 1) & -1.074 & -0.720 & +0.055 \quad = -1.738; \\ 2) & -1.738 & +0 & +0 \quad = -1.738; \\ 3) & -1.406 & -0.360 & +0.028 \quad = -1.738. \end{array}$$

$\Rightarrow$ The estimated probability that white subjects with immediate AZT use develop AIDS symptoms equals $\exp(-1.738)/[1+\exp(-1.738)] = 0.15$.

5.4.2 AIDS and AZT example

Similarly, for each coding scheme, $(\beta_1^X - \beta_2^X)$ is identical and represents the conditional log odds ratio of X with the response, given Z.

Here, $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X) = \exp(-0.720) = 0.49$ estimate the common odds ratio between immediate AZT use and AIDS symptoms, for each race.

5.4.3 Goodness of fit as a likelihood-ratio test

To test whether certain model parameters are zero by comparing the maximized log likelihood L_1 for the fitted model M_1 with L_0 for a simpler model M_0, the likelihood-ratio statistic is

The goodness-of-fit statistic $G^2(M)$ is a special case of LRT: $M_0 = M$ and M_1 is the saturated model.

$L =$ the maximized log likelihood for the fitted model M;
$L_S =$ the maximized log likelihood for the saturated model. Then

$$G^2(M) = -2(L - L_S).$$

5.4.3 Goodness of fit as a likelihood-ratio test

In testing whether M fits, we test whether all parameters in the saturated model but not in M equal zero.

More specifically, if model M has parameters $\{\beta_1, \ldots, \beta_p\}$ and the saturated model has parameters $\{\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_K\}$, then we test $H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_K = 0$.

The asymptotic df is the difference in the number of parameters in the two models, i.e., $K - p$, which is the number of binomials modeled minus the number of parameters in M.

5.4.3 Goodness of fit as a likelihood-ratio test

Illustration: Check the fit of the model in §5.4.2 for AIDS data.

Table 5.5 shows that the goodness-of-fit statistics are $G^2 = 1.38$ and $X^2 = 1.39$.

The model has four binomials, i.e., one at each combination of AZT use and race. Since the model has three parameters, residual df$= 4 - 3 = 1$ for the goodness-of-fit test. Small G^2 and X^2 suggest that the model fits decently ($P > 0.2$).

5.4.3 Goodness of fit as a likelihood-ratio test

LRT for comparing models M_1 and M_0 is identical to the difference in goodness-of-fit G^2 statistics (deviances):

$$\begin{aligned} G^2(M_0|M_1) &= -2(L_0 - L_1) \\ &= -2(L_0 - L_S) - [-2(L_1 - L_S)] \\ &= G^2(M_0) - G^2(M_1). \end{aligned}$$

Test $H_0 : \beta_2 = 0$ (race effect with the AIDS data in Table 5.5):

- The deviance of the fitted model including race effect is $G^2(M_1) = 1.38$ (Table 5.6).
- The deviance of the simpler model without race effect is $G^2(M_0) = 1.42$.
- The likelihood-ratio statistic equals $G^2(M_0) - G^2(M_1) = 0.04$.
  $\Rightarrow$ The simpler model is adequate.

5.4.3 Goodness of fit as a likelihood-ratio test

Remark.

The model comparison likelihood-ratio statistic often has an approximate chi-squared null distribution even when separate $G^2(M_i)$ do not.

5.4.4 Horseshoe crab example revisited

Like ordinary regression, logistic regression can have a mixture of quantitative and qualitative predictors.

We revisit the horseshoe crab data (Section 5.1.3), using the female crab's width (x) and color as predictors.

Color has 4 categories: medium light, medium, medium dark, dark.

5.4.4 Horseshoe crab example revisited

We first treat color as qualitative (unordered categorical), using 3 dummy variables:

$$c_1 = \begin{cases} 1 & \text{for medium-light,} \\ 0 & \text{otherwise;} \end{cases} \qquad c_2 = \begin{cases} 1 & \text{for medium,} \\ 0 & \text{otherwise;} \end{cases}$$

$$c_3 = \begin{cases} 1 & \text{for medium-dark,} \\ 0 & \text{otherwise.} \end{cases}$$

For dark color (category 4), $c_1 = c_2 = c_3 = 0$.
The model is

$$\text{logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x.$$

Table 5.8 shows the ML parameter estimates.

5.4.4 Horseshoe crab example revisited

TABLE 5.8 Computer Output for Model with Width and Color Predictors

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	168	187.4570
Pearson Chi-Square	168	168.6590
Log Likelihood		−93.7285

Parameter	Estimate	Standard Error	Likelihood-Ratio 95% Confidence	Limits	Chi-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

For dark crabs, $\text{logit}(\hat{\pi}) = -12.7151 + 0.4680\,x$.
At the average width $x = 26.3$ cm,

$$\hat{\pi} = \frac{\exp(-12.7151 + 0.4680 \times 26.3)}{1 + \exp(-12.7151 + 0.4680 \times 26.3)} = 0.399.$$

5.4.4 Horseshoe crab example revisited

For medium-light crabs,

$$\text{logit}(\hat{\pi}) = (-12.7151 + 1.3299) + 0.4680\,x = -11.3852 + 0.4680\,x.$$

At the average width of 26.3 cm,

$$\hat{\pi} = \frac{\exp(-11.3852 + 0.4680 \times 26.3)}{1 + \exp(-11.3852 + 0.4680 \times 26.3)} = 0.715.$$

The model assumes a lack of interaction between color and width in their effects.

$\Rightarrow$ Width has the same coefficient 0.4680 for all colors.

$\Rightarrow$ The lines relating width to $\text{logit}(\hat{\pi})$ are parallel with different intercepts.

$\Rightarrow$ The shapes of curves relating width to $\hat{\pi}$ are identical.

For each color, a 1-cm increase in width has a multiplicative effect of $\exp(0.4680) = 1.60$ on the odds that $Y = 1$.

5.4.4 Horseshoe crab example revisited

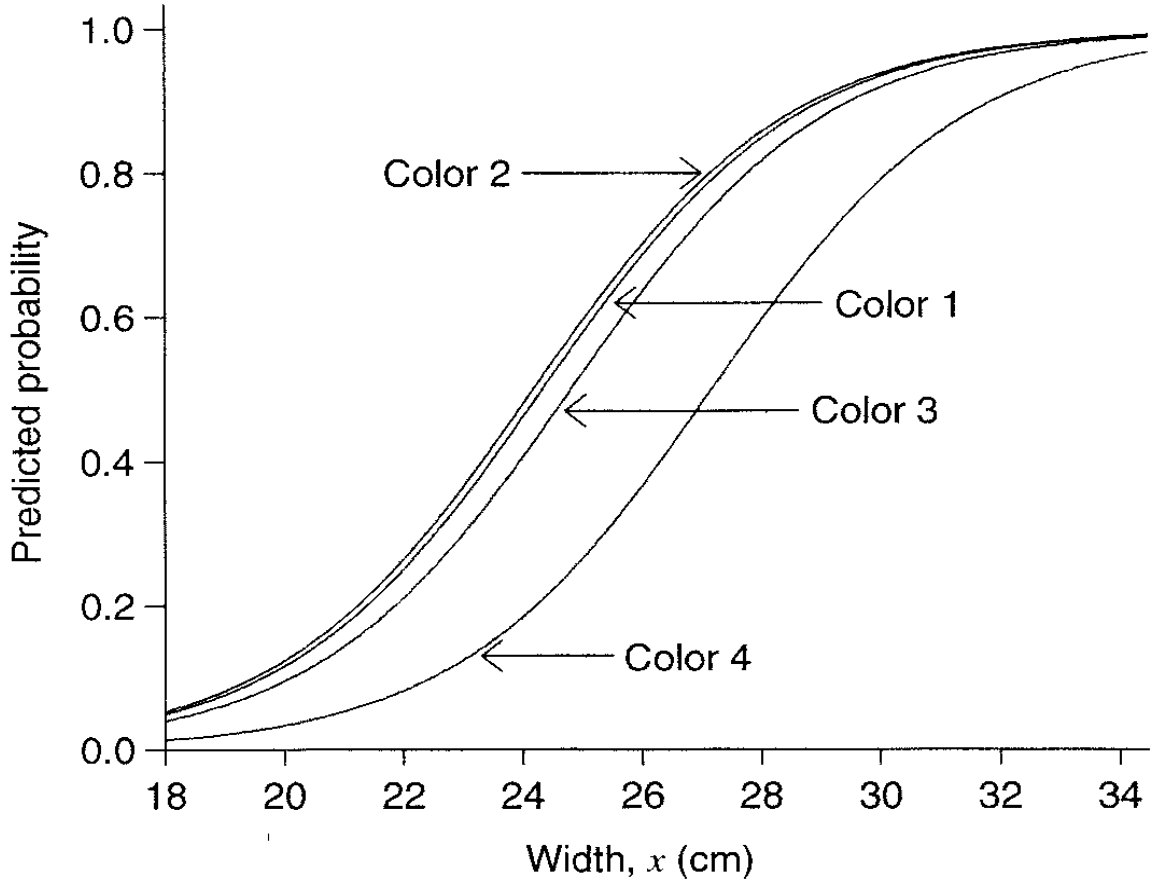

FIGURE 5.5 Logistic regression model using width and color predictors of satellite presence for horseshoe crabs.

5.4.5 Model comparison

Color effect

To test whether color contributes significantly to the model above, we test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, i.e., controlling for width, the probability of a satellite is independent of color.

We compare the maximized log-likelihood for the full model (L_1) to that for the simpler model (L_0).

The test statistic $-2(L_0 - L_1) = 7.0$ has $df = 3$, with $P = 0.07$.

$\Rightarrow$ Slight evidence of a color effect.

5.4.5 Model comparison

Color$\times$width interaction

The more complex model allowing color $\times$ width interaction has three additional terms, i.e., the cross-products of width with the color dummy variables.

The likelihood-ratio statistic comparing the models with and without the interaction terms equals 4.4, with $df = 3$ and $P = 0.22$.

$\Rightarrow$ The evidence of interaction is weak.

5.4.6 Quantitative treatment of ordinal predictor

Color has in fact ordered categories, from lightest to darkest. Assign scores to color.

Monotone scoring: $c = \{1, 2, 3, 4\}$.

Using this scoring, the model

$$\text{logit}(\pi) = \alpha + \beta_1 c + \beta_2 x$$

has $\hat{\beta}_1 = -0.509$ with $\text{SE}(\hat{\beta}_1) = 0.224$
and $\hat{\beta}_2 = 0.458$ with $\text{SE}(\hat{\beta}_2) = 0.104$.
$\Rightarrow$ Strong evidence of color and width effects.

At a given width, for every one-category increase in color darkness, the estimated odds of a satellite multiply by $\exp(-0.509) = 0.60$.

5.4.6 Quantitative treatment of ordinal predictor

LRT comparing this fit to the more complex model with 3 dummy variables equals 1.7 with $df = 2$ and $P = 0.44$.

$\Rightarrow$ Given the more complex model holds, this simpler model is adequate, i.e., the simplification seems permissible.

Binary scoring: $c = \{1, 1, 1, 0\}$

The estimates of color parameters in the more complex model are $(1.33, 1.40, 1.11, 0)$, the 0 value for the dark category reflecting its lack of a dummy variable.

Although these values do not depart significantly from a linear trend (as tested above), the first three are quite similar compared to the last one.

5.4.6 Quantitative treatment of ordinal predictor

LRT comparing this fit with a binary color score to the more complex model with 3 dummy variables equals 0.5 with $df = 2$.
$\Rightarrow$ This simpler model is also adequate.

Its fit is

$$\text{logit}(\hat{\pi}) = -12.980 + 1.300\, c + 0.478\, x,$$

with $\text{SE}(\hat{\beta}_1) = 0.526$ and $\text{SE}(\hat{\beta}_2) = 0.104$.

At a given width, the estimated odds that a lighter-colored crab has a satellite are $\exp(1.300) = 3.7$ times the odds for a dark crab.

5.4.7 Standardized and probability-based interpretations

Standardized coefficients

To compare effects of quantitative predictors having different units, it can be helpful to report standardized coefficients, which can be obtained by

either fitting the model to standardized predictors, i.e., replacing each x_j by $(x_j - \bar{x}_j)/s_{x_j}$, where s_{x_j} is the standard deviation of x_j.

or fitting the model to unstandardized predictors and then multiplying each β_j by its s_{x_j}.

Each standardized coefficient represents the effect of a standard deviation change in a predictor, controlling for the other variables.

5.4.7 Standardized and probability-based interpretations

Probability

Use the quartiles to describe the effect of x_j while set the other predictors at their sample means.

For the model with binary color scores above, the sample mean of x is $\bar{x} = 26.3$ and of c is $\bar{c} = 0.873$.

The lower quartile of x is $x_{0.25} = 24.9$ and the upper quartile is $x_{0.75} = 27.7$. At $c = \bar{c}$, $\hat{\pi}(x_{0.25}, \bar{c}) = 0.51$ and $\hat{\pi}(x_{0.75}, \bar{c}) = 0.80$.
$\Rightarrow \Delta\hat{\pi}(x, \bar{c}) = \hat{\pi}(x_{0.75}, \bar{c}) - \hat{\pi}(x_{0.25}, \bar{c}) = 0.80 - 0.51 = 0.29$.
$\Rightarrow$ A strong width effect.

5.4.7 Standardized and probability-based interpretations

Since c takes only values 0 and 1, one could instead report width effect separately for each value.

Also, when an explanatory variable is a dummy, it makes sense to report the estimated probabilities at its two values rather than at quartiles, which could be identical.

At $x = \bar{x} = 26.3$, $\hat{\pi}(\bar{x}, c = 0) = 0.40$ and $\hat{\pi}(\bar{x}, c = 1) = 0.71$.

The difference is $0.71 - 0.40 = 0.31$.
$\Rightarrow$ This color effect, differentiating dark crabs from others, is also substantial.

5.4.7 Standardized and probability-based interpretations

TABLE 5.9 Summary of Effects in Model (5.14) with Crab Width and Color 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	(1, 0) at $\bar{x}$	0.31 = 0.71 − 0.40
Width, x (cm)	0.478	0.104	(UQ, LQ) at $\bar{c}$	0.29 = 0.80 − 0.51
Interaction model				
Intercept	−5.854	6.694		
Color (0 = dark, 1 = other)	−6.958	7.318		
Width, x (cm)	0.200	0.262	(UQ, LQ) at $c = 0$	0.13 = 0.43 − 0.30
Width × color	0.322	0.286	(UQ, LQ) at $c = 1$	0.29 = 0.84 − 0.55

Since the coefficient of the interaction is positive, the estimated width effect is greater for the light-colored crabs.

However, the interaction is not significant.

Outline

5.5 Fitting logistic regression models

y_i: binary outcome for $i = 1, 2, \ldots, N$.
$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$: explanatory variable for observation i.

The multiple logistic regression model:

$$\text{logit}[\pi(\mathbf{x}_i)] = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij}$$

with $x_{i1} = 1$ for all i (so β_1 is the intercept). This model gives

$$\pi_i = \pi(\mathbf{x}_i) = \frac{\exp(\sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij})}.$$

5.5.1 Likelihood equations

$$
\begin{aligned}
&\prod_{i=1}^{N}(\pi_i)^{y_i}(1-\pi_i)^{n_i-y_i} \\
&= \left\{\prod_{i=1}^{N}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}\right\}\left\{\prod_{i=1}^{N}(1-\pi_i)^{n_i}\right\} \\
&= \left\{\prod_{i=1}^{N}\exp\left[\log\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}\right]\right\}\left\{\prod_{i=1}^{N}(1-\pi_i)^{n_i}\right\} \\
&= \left\{\exp\left[\sum_{i=1}^{N} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right]\right\}\left\{\prod_{i=1}^{N}(1-\pi_i)^{n_i}\right\} \\
&= \left\{\exp\left[\sum_{i=1}^{N} y_i \,\mathrm{logit}(\pi_i)\right]\right\}\left\{\prod_{i=1}^{N}\left[1-\frac{\exp(\sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\sum_{j=1}^{p}\beta_j x_{ij})}\right]^{n_i}\right\} \\
&= \left\{\exp\left[\sum_{i=1}^{N} y_i \sum_{j=1}^{p}\beta_j x_{ij}\right]\right\}\left\{\prod_{i=1}^{N}\left[\frac{1}{1+\exp(\sum_{j=1}^{p}\beta_j x_{ij})}\right]^{n_i}\right\} \\
&= \left\{\exp\left[\sum_{j=1}^{p}\left(\sum_{i=1}^{N} y_i x_{ij}\right)\beta_j\right]\right\}\left\{\prod_{i=1}^{N}\left[1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right]^{-n_i}\right\}.
\end{aligned}
$$

5.5.1 Likelihood equations

Then, the log likelihood is

$$L(\beta) \propto \log\left[\left\{\exp\left[\sum_{j=1}^{p}\left(\sum_{i=1}^{N} y_i\, x_{ij}\right)\beta_j\right]\right\}\left\{\prod_{i=1}^{N}\left[1+\exp\left(\sum_{j=1}^{p}\beta_j\, x_{ij}\right)\right]^{-n_i}\right\}\right]$$

$$= \sum_{j=1}^{p}\left(\sum_{i=1}^{N} y_i\, x_{ij}\right)\beta_j - \sum_{i=1}^{N} n_i \log\left[1+\exp\left(\sum_{j=1}^{p}\beta_j\, x_{ij}\right)\right].$$

Since

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i y_i\, x_{ij} - \sum_i n_i\, x_{ij}\left[\frac{\exp(\sum_{k=1}^{p}\beta_k\, x_{ik})}{1+\exp(\sum_{k=1}^{p}\beta_k\, x_{ik})}\right]$$

$$= \sum_i y_i\, x_{ij} - \sum_i n_i\, x_{ij}\,\pi_i\,,$$

the likelihood equations are

$$\sum_i y_i\, x_{ij} - \sum_i n_i\,\hat{\pi}_i\, x_{ij} = 0, \quad j = 1,\ldots,p,$$

5.5.1 Likelihood equations

where

$$\hat{\pi}_i = \exp(\sum_k \hat{\beta}_k x_{ik})/[1 + \exp(\sum_k \hat{\beta}_k x_{ik})]$$

is the ML estimate of π_i.

The equations are nonlinear and require iterative solution.

Let $\mathbf{X}$ denote the $N \times p$ matrix of values of $\{x_{ij}\}$. The likelihood equations have the form $\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}$, where $\hat{\mu}_i = n_i \hat{\pi}_i$.

This equation illustrates a fundamental results: for GLMs with canonical link, the likelihood equations equate the sufficient statistics to the estimates of their expected values.

5.5.2 Asymptotic covariance matrix

Covariance matrix equal to the inverse of the information matrix.

The observed information matrix has elements

$$
\begin{aligned}
-\frac{\partial^2 L(\beta)}{\partial\beta_a\,\partial\beta_b} &= -\frac{\partial[\sum_i y_i\,x_{ia} - \sum_i n_i\,x_{ia}\,\pi_i]}{\partial\beta_b} = \sum_i n_i\,x_{ia}\,\frac{\partial\pi_i}{\partial\beta_b} \\
&= \sum_i n_i\,x_{ia}\,\frac{\partial\{\exp(\sum_j \beta_j\,x_{ij})/[1+\exp(\sum_j \beta_j\,x_{ij})]\}}{\partial\beta_b} \\
&= \sum_i n_i\,x_{ia}\,\frac{x_{ib}\exp(\sum_j \beta_j\,x_{ij})}{[1+\exp(\sum_j \beta_j\,x_{ij})]^2} \\
&= \sum_i n_i\,x_{ia}\,x_{ib}\left[\frac{\exp(\sum_j \beta_j\,x_{ij})}{1+\exp(\sum_j \beta_j\,x_{ij})}\right]\left[\frac{1}{1+\exp(\sum_j \beta_j\,x_{ij})}\right] \\
&= \sum_i x_{ia}\,x_{ib}\,n_i\,\pi_i\,(1-\pi_i).
\end{aligned}
$$

5.5.2 Asymptotic covariance matrix

The expression above is not a function of $\{y_i\}$, so the observed and expected information matrices are identical. This happens for all GLMs that use canonical links (Section 4.6.4).

The estimated covariance matrix is the inverse of the matrix having the above elements, substituting $\hat{\beta}$:

$$\widehat{\text{cov}}(\hat{\beta}) = \{\sum_i x_{ia}\, x_{ib}\, n_i\, \hat{\pi}_i\,(1-\hat{\pi}_i)\}^{-1} = \{\mathbf{X}'\,\mathbf{diag}[n_i\,\hat{\pi}_i\,(1-\hat{\pi}_i)]\,\mathbf{X}\}^{-1},$$

where $\mathbf{diag}[n_i\,\hat{\pi}_i\,(1-\hat{\pi}_i)]$ denotes the $N \times N$ diagonal matrix having $\{n_i\,\hat{\pi}_i\,(1-\hat{\pi}_i)\}$ on the main diagonal. It is a special case of $\widehat{\text{cov}}(\hat{\beta}) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$ with $\hat{\mathbf{W}}$ having elements $\hat{w}_i = n_i\,\hat{\pi}_i\,(1-\hat{\pi}_i)$.

The square roots of the main diagonal elements of $\widehat{\text{cov}}(\hat{\beta})$ are estimated standard errors of $\hat{\beta}$, denoted by $\text{SE}(\hat{\beta})$.

5.5.3 Distribution of probability estimators

For β_j: 95% CI is $\hat{\beta}_j \pm z_{\alpha/2} \times \text{SE}(\hat{\beta}_j) = (\hat{\beta}_{jL},\ \hat{\beta}_{jU})$.

For logit: at particular settings $\mathbf{x}$, $\text{logit}[\hat{\pi}(\mathbf{x})] = \mathbf{x}\hat{\boldsymbol{\beta}}$.

So the estimated variance of $\text{logit}[\hat{\pi}(\mathbf{x})]$ is $\mathbf{x}\,\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\,\mathbf{x}'$.

For large sample, the 95% CI for the true logit is

$$\text{logit}[\hat{\pi}(\mathbf{x})] \pm z_{\alpha/2} \times \sqrt{\mathbf{x}\,\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\,\mathbf{x}'} = (\text{logit}_L(\mathbf{x}),\ \text{logit}_U(\mathbf{x})).$$

For odds ratio $\pi(\mathbf{x})/[1-\pi(\mathbf{x})]$:

95% CI is $(\exp[\text{logit}_L(\mathbf{x})],\ \exp[\text{logit}_U(\mathbf{x})])$.

For π:

Since $\pi = \exp(\text{logit})/[1+\exp(\text{logit})]$, the 95% CI for $\pi(\mathbf{x})$ is

$$\left(\frac{\exp[\text{logit}_L(\mathbf{x})]}{1+\exp[\text{logit}_L(\mathbf{x})]},\ \frac{\exp[\text{logit}_U(\mathbf{x})]}{1+\exp[\text{logit}_U(\mathbf{x})]}\right).$$

5.5.3 Distribution of probability estimators

TABLE A.8 SAS Code for Logistic Regression Models with Horseshoe Crab Data in Table 4.3

```
data crab;
input color spine width satell weight;
if satell>0 then y=1; if satell=0 then y=0;
if color=4 then light=0; if color<4 then light=1;
datalines;
2 3 28.3 8 3.05
...
2 2 24.5 0 2.00
;
proc genmod descending; class color;
  model y=width color/dist=bin link=logit lrci type3 obstats;
  contrast 'a-d' color 1 0 0 -1;
proc genmod descending;
  model y=width color/dist=bin link=logit;
proc genmod descending;
  model y=width light/dist=bin link=logit;
proc genmod descending; class color spine;
  model y=width weight color spine/dist=bin link=logit type3;
proc logistic descending; class color spine/param=ref;
  model y=width weight color spine/selection=backward lackfit
     outroc=classif1;
proc plot data=classif1; plot _sensit_*_lmspec_ ;
```