Categorical Data Analysis
Lecture Note 8

Instructor: Seokho Lee

Hankuk University of Foreign Studies

8. Logistic Regression Model

Logistic regression model is a technique for relating a binary response variable to explanatory variables. The explanatory variables may be categorical, continuous, or both.

8.1. Interpreting the Logistic Regression Model

We will look at the logistic model with one explanatory variable:

$$\begin{aligned} Y &: \text{binary response variable} \begin{cases} 1 & \text{yes or success} \\ 0 & \text{no or failure} \end{cases} \\ X &: \text{quantitative explanatory variable} \end{aligned}$$

We want to model

$$p(x) = \Pr(Y = 1 | X = x)$$

This is the probability of a success when $X = x$.

The *logistic regression model* has a linear form for logarithm of the odds, or *logit function*,

$$\text{logit}[p(x)] = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

We can solve for $p(x)$:

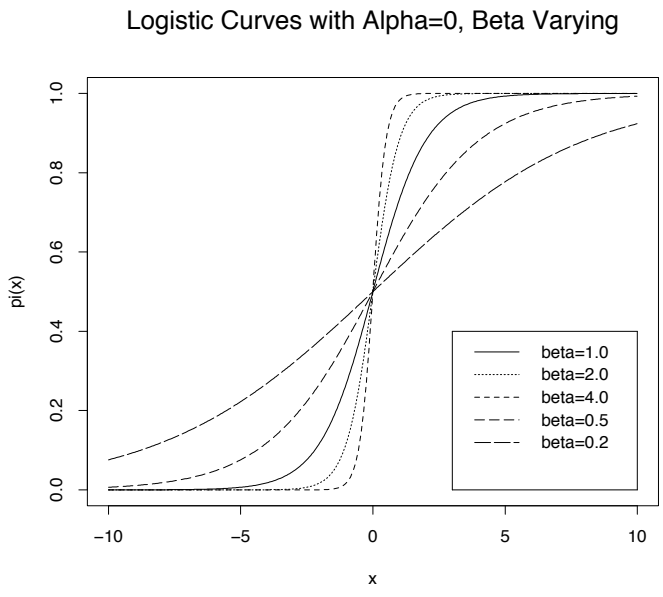
$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x)\}}$$

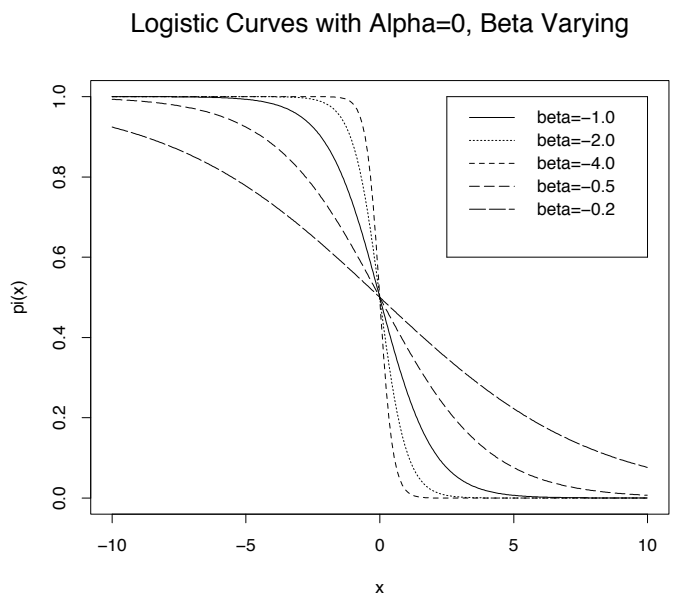
Note: This is a generalized linear model with the following components:

- Link: Logit (log-odds)
- Linear predictor: $\beta_0 + \beta_1 x$
- Error distribution: Binomial

From the figures on the next page, we see that $p(x)$ is a monotone function of x.

- If $\beta > 0$, $p(x)$ is an increasing function of x
- If $\beta < 0$, $p(x)$ is a decreasing function of x
- If $\beta = 0$, $p(x)$ is constant and the probability of a success does not depend on x.

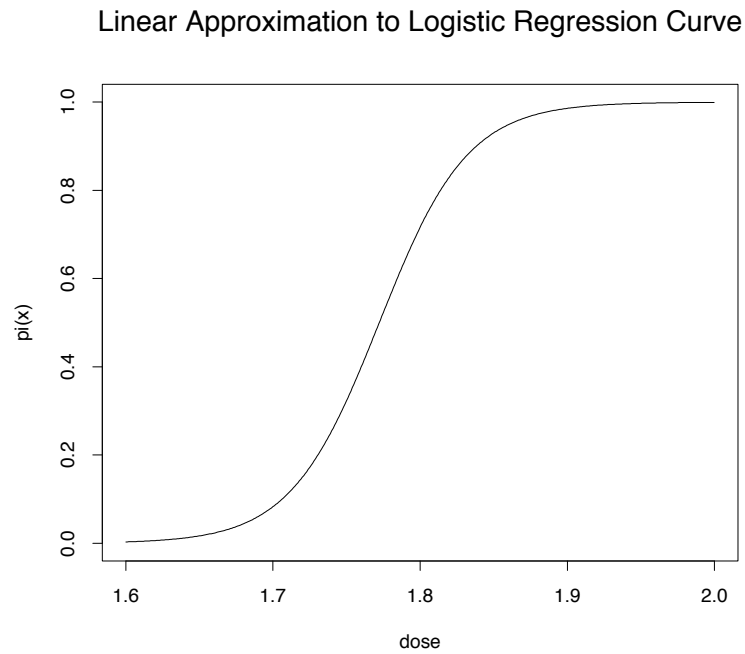

Logistic Curves with Alpha=0, Beta Varying
pi(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
beta=1.0
beta=2.0
beta=4.0
beta=0.5
beta=0.2


Logistic Curves with Alpha=0, Beta Varying
pi(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
beta=−1.0
beta=−2.0
beta=−4.0
beta=−0.5
beta=−0.2

8.1.1. Linear Approximation Interpretation

The parameter β determines the rate of increase or decrease of the S-shaped curve. The rate of change in $p(x)$ per unit change in x is the slope. This can be found by taking the derivative:

$$\text{slope} = \frac{dp(x)}{dx} = \beta p(x)(1 - p(x))$$



$p(x)$	.5	.4 or .6	.3 or .7	.2 or .8	.1 or .9
slope	$.25\beta$	$.24\beta$	$.21\beta$	$.16\beta$	$.09\beta$

- The steepest slope occurs at $p(x) = 0.5$ or $x = -\beta_0/\beta$. This value is known as *the median effective level* and is denoted EL_{50}.
- In the example from Chapter 6, $\text{logit}(\hat{p}(x)) = -59.188 + 33.401x$. Thus, $EL_{50} = 1.772$ and the slope is 8.350.

8.1.2. Odds Ratio Interpretation

$$\text{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

For an increase of 1 unit in x, the logit increases by β.

The odds for the logistic regression model when $X = x$ is given by

$$\frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x) = e^{\beta_0}(e^{\beta_1})^x.$$

Consider two values of x_1 and x_2 of the explanatory variable. The odds ratio comparing x_2 to x_1 is

$$\theta_{21} = OR(x_2 : x_1) = \frac{odds(x_2)}{odds(x_1)} = e^{\beta_1(x_2 - x_1)}.$$

Let $x_2 = x + 1$, $x_1 = x$, and $\beta_1 > 0$, then $\theta_{21} = e^{\beta_1((x+1)-x)} = e^{\beta_1}$.

For an increase of 1 unit in x, the odds increase multiplicatively by a factor of e^{β_1}.

Example: In the example, $\hat{\beta}_1 = 33.401$, so the multiplicative increase in odds per unit increase in x is

$$e^{\hat{\beta}_1} = e^{33.401}.$$

Examining the example more closely, we see that

$$x = \log_{10} CS_2 mgl^{-1}$$

or the logarithm of a concentration. So letting c_1 and c_2 be the concentrations corresponding to x_1 and x_2, we get

$$\theta_{21} = \left(\frac{c_2}{c_1}\right)^{\hat{\beta}_1 / \log 10} = \left(\frac{c_2}{c_1}\right)^{14.51}$$

The odds ratio equals the ratio of concentrations raised to a power depending on β_1.

Remark: Logistic regression has parameters that refer to odds ratios. Because one can estimate odd ratios in case-control studies, logistic regression can be used for such retrospective studies.

8.2. Multiple Logistic Regression

We consider logistic regression models with more than one explanatory variable.

- Binary response: Y
- k predictors: $\mathbf{x} = (x_1, \cdots, x_k)$
- Quantity to estimate: $p(\mathbf{x}) = \Pr(Y = 1 | x_1, \cdots, x_k)$

The logistic regression model is

$$\text{logit}(p(\mathbf{x})) = \log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- The parameter β_j reflects the effect of a unit change in x_j on the log-odds that $Y = 1$, keeping the others x_is constant.
- Here, e^{β_j} is the multiplicative effect on the odds that $Y = 1$ of a one-unit increase in x_j, keeping the other x_is constant:

$$\begin{aligned} & \text{logit}(p(x_1 + 1, x_2, \cdots, x_k)) - \text{logit}(p(x_1, x_2, \cdots, x_k)) \\ = \quad & \beta_0 + \beta_1(x_1 + 1) + \cdots + \beta_k x_k - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \\ = \quad & \beta_1 \end{aligned}$$

8.2.1. Estimation of Parameters

Suppose that we have n independent observations, $(x_{i1}, \cdots, x_{ik}, Y_i)$, $i = 1, \cdots, n$.

- $Y_i =$ binary response for ith observation
- $\mathbf{x}_i = (x_{i1}, \cdots, x_{ik}) =$ the values of the k explanatory variables

When there are n_i observations at a fixed $\mathbf{x}_i$ value, the number of successes Y_i forms a *sufficient statistics* and has a Binomial(n_i, p_i) distribution where

$$p_i = p(\mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ki})}.$$

Suppose that there are N distinct settings on $\mathbf{x}$. The responses $(Y_1, \cdots, Y_N)$ are independent binomial random variables with joint likelihood equal to

$$\mathcal{L}(\beta_0, \beta_1, \cdots, \beta_k) = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}.$$

The log-likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}) &= \log(\mathcal{L}(\boldsymbol{\beta})) = \sum_{i=1}^{N} \left[\log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i) \log(1 - p_i) \right] \\ &= \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) + \sum_{i=1}^{N} n_i \log\{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\} \\ &\quad + \sum_{i=1}^{N} \log \binom{n_i}{y_i}. \end{aligned}$$

We wish to estimate $\beta_0, \beta_1, \cdots, \beta_k$ using maximum likelihood.

Setting the scores equal to zero gives us the estimating equations:

$$U_0(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} n_i p_i = 0$$

$$U_j(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} x_{ji} y_i - \sum_{i=1}^{N} n_i x_{ji} p_i = 0$$

for $j = 1, \cdots, k$.

There are $k + 1$ equations in $k + 1$ unknowns. These equations are solved numerically by `SAS`.

To obtain the asymptotic variances and covariances of the estimators, we obtain Fisher's information matrix:

$$\begin{pmatrix} \sum_{i=1}^{N} n_i p_i(1-p_i) & \sum_{i=1}^{N} n_i x_{i1} p_i(1-p_i) & \cdots & \sum_{i=1}^{N} n_i x_{ik} p_i(1-p_i) \\ \sum_{i=1}^{N} n_i x_{i1} p_i(1-p_i) & \sum_{i=1}^{N} n_i x_{i1}^2 p_i(1-p_i) & \cdots & \sum_{i=1}^{N} n_i x_{i1} x_{ik} p_i(1-p_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} n_i x_{ik} p_i(1-p_i) & \sum_{i=1}^{N} n_i x_{i1} x_{ik} p_i(1-p_i) & \cdots & \sum_{i=1}^{N} n_i x_{ik}^2 p_i(1-p_i) \end{pmatrix}$$

The asymptotic variance-covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ is the inverse of the information matrix. The estimated asymptotic variances of the estimators $\widehat{Var}(\hat{\beta}_j)$ are the diagonal entries of this matrix. The asymptotic standard error of $\hat{\beta}_j$ is given by

$$\hat{se}(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)}.$$

8.2.2. Overall Tests for the Model

We consider testing the overall significance of the k regression coefficients in the logistic regression model. The hypotheses of interest are

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \qquad \text{versus} \qquad H_a : \text{At least one } \beta_j \neq 0.$$

We typically use the likelihood ratio statistic:

$$G^2 = -2 \log \left[\frac{\mathcal{L}(\tilde{\beta}_0)}{\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)} \right] = 2 \left[L_{Full} - L_{Reduced} \right].$$

Here $\tilde{\beta}_0$ is the m.l.e. of β_0 under the null hypothesis. When H_0 is true,

$$G^2 \xrightarrow{d} \chi^2_k \quad \text{as} \quad n \longrightarrow \infty.$$

We reject H_0 for large values of G^2.

8.2.3. Tests on Individual Coefficients

To help determine which explanatory variables are useful, it is convenient to examine the Wald test statistics for the individual coefficients. To determine whether x_j is useful in the model given that the other $k-1$ explanatory variables are in the model, we will test the hypotheses:

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_a : \beta_j \neq 0$$

The test statistic is given by

$$Z = \frac{\hat{\beta}_j}{\widehat{se}(\hat{\beta}_j)}.$$

We reject H_0 for large values of $|Z|$.

8.2.4. Confidence Intervals for Coefficients

Confidence intervals for coefficients in multiple logistic regression are formed in essentially the same way as they were for a single explanatory variable.

A $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\hat{\beta}_j \pm z_{\alpha/2}\widehat{se}(\hat{\beta}_j)$$

8.2.5. Confidence Intervals for the Logit and for the Probability of a Success

We next consider forming a confidence interval for the logit at a given value of $\mathbf{x}$:

$$g(\mathbf{x}) = \text{logit}(p(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

The estimated logit is given by

$$\hat{g}(\mathbf{x}) = \text{logit}(\hat{p}(\mathbf{x})) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k.$$

This has estimated asymptotic variance

$$\widehat{Var}(\hat{g}(\mathbf{x})) = \sum_{j=0}^{k} \widehat{Var}(\hat{\beta}_j) + \sum_{j=0}^{k} \sum_{\ell=j+1}^{k} 2x_j x_\ell \widehat{Cov}(\hat{\beta}_j, \hat{\beta}_\ell).$$

In the above formula, $x_0 = 1$.

A $100(1-\alpha)\%$ confidence interval for $\text{logit}(p(\mathbf{x}))$ is

$$\hat{g}(\mathbf{x}) \pm z_{\alpha/2} \widehat{se}(\hat{g}(\mathbf{x}))$$

where $\widehat{se}(\hat{g}(\mathbf{x})) = \sqrt{\widehat{Var}(\hat{g}(\mathbf{x}))}$.

Since

$$p(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} = \frac{1}{1 + e^{-g(\mathbf{x})}},$$

we can find a 100(1-α)% confidence interval for $p(\mathbf{x})$ by substituting the endpoints of the confidence interval for the logit into this formula.

In the case of one predictor $x = x_0$, the confidence interval for the logit is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm z_{\alpha/2} \widehat{se}(\hat{\beta}_0 + \hat{\beta}_1 x_0).$$

The estimated asymptotic standard error is

$$\widehat{se}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)} = \sqrt{\widehat{Var}(\hat{\beta}_0) + x_0^2 \widehat{Var}(\hat{\beta}_1) + 2x_0 \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1)}.$$

Since

$$p(x_0) = \Pr(Y = 1 | X = x_0) = \frac{e^{\beta_0 + \beta_1 x_0}}{1 + e^{\beta_0 + \beta_1 x_0}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_0)}},$$

we substitute the endpoints of the confidence interval for the logit into the above formula to obtain a confidence interval for $p(x_0)$.

Remarks

- The fitted value $\hat{p}(x_0)$ is analogous to a particular point on the line in simple linear regression. This is the estimated mean response for individuals with covariate x_0. In our case $\hat{p}(x_0)$ is an estimate of the proportion of all individuals with covariate x_0 that result in a success. Any particular individual is either a success or a failure.

- An alternative method of estimating $p(x_0)$ is to compute the sample proportion of successes among all individuals with covariate x_0. When the logistic model truly holds, the model-based estimate can be considerably better than the sample proportion. Instead using just a few observations, the model uses all the data to estimate $p(x_0)$.

8.2.6. Examples

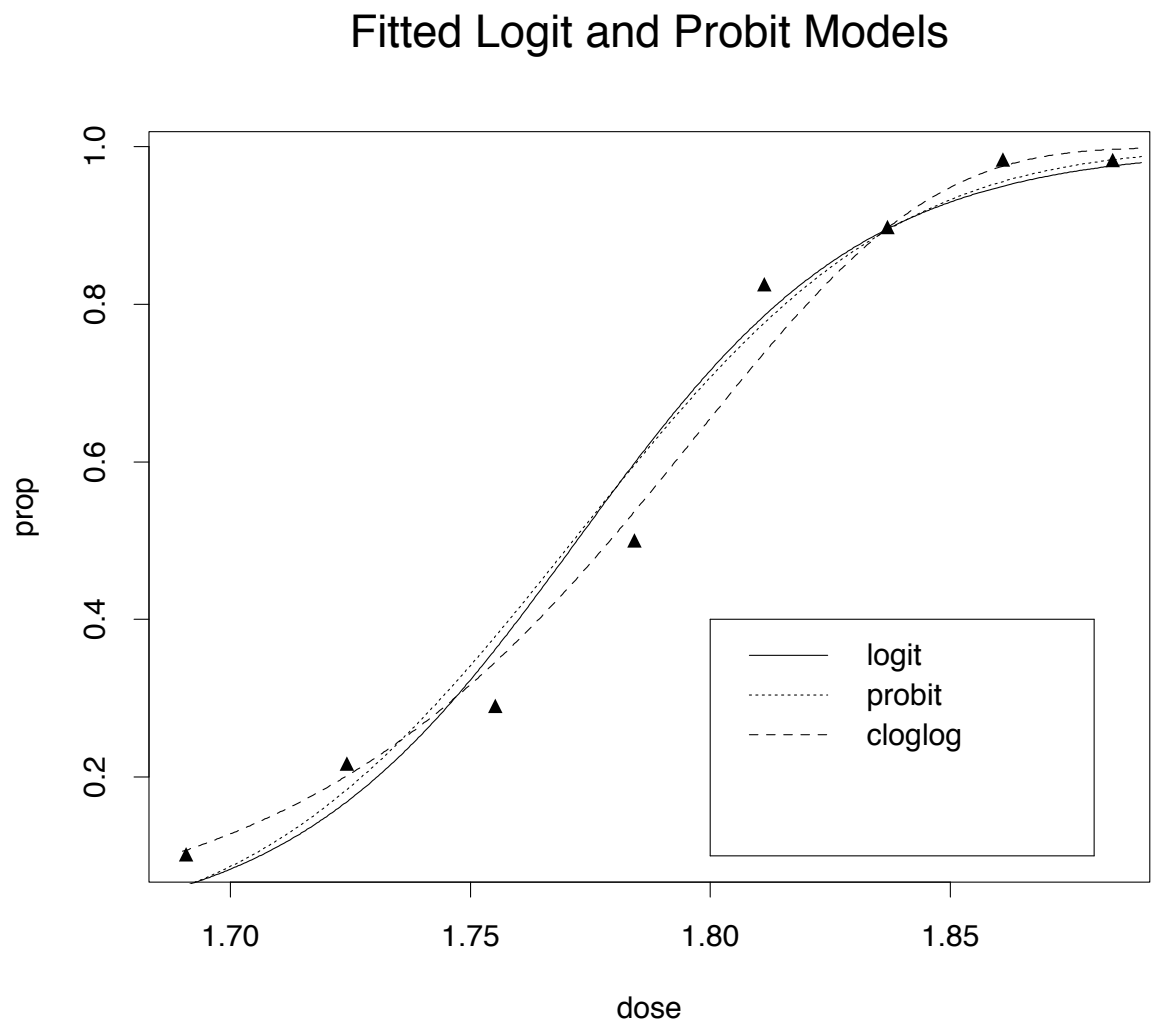
Example: Beetles were treated with various concentrations of insecticide for 5 hours. The data appear in the following table:

Dose x_i ($\log_{10} CS_2 mgl^{-1}$)	Number of insects, n_i	Number killed, y_i	Proportion killed, y_i/n_i
1.6907	59	6	.1017
1.7242	60	13	.2167
1.7552	62	18	.2903
1.7842	56	28	.5000
1.8113	63	52	.8254
1.8369	59	53	.8983
1.8610	62	61	.9839
1.8839	60	59	.9833

The fitted logit model is:

$$\text{logit}(\hat{p}(x)) = -59.1875 + 33.4007x$$

The observed proportions and the fitted model appear in the following graph:



- Test for $H_0 : \beta_1 = 0$

  The test strongly reject H_0.

- 95% confidence interval for β_1:

$$33.4007 \pm (1.96)(2.8394) = 33.4007 \pm 5.565$$

- Confidence interval for logit($p(x_0)$)

  Let $x_0 = 1.8113$. The estimated logit is $-59.18 + (33.40)(1.8113) = 1.3111$.

$$\begin{aligned} \widehat{se} &= \sqrt{(5.0530)^2 + (1.8113)^2(2.8394)^2 + 2(1.8113)(-14.341)} \\ &= \sqrt{0.0316} = 0.1778 \end{aligned}$$

  The 95% confidence interval for the logit is

$$1.3111 \pm (1.96)(0.1778) = 1.311 \pm 0.3485 \quad \text{or} \quad (0.9625, 1.6595)$$

  The 95% confidence interval for $p(1.8113)$ is

$$\left(\frac{e^{0.9625}}{1 + e^{0.9625}}, \frac{e^{1.6595}}{1 + e^{1.6595}} \right) = (0.7236, 0.8402)$$

We can also find the 95% confidence interval for $p(1.8113)$ based on the 63 insects that received this dose:

$$\frac{52}{63} \pm 1.96\sqrt{\frac{\frac{52}{63}(1-\frac{52}{63})}{63}} = 0.825 \pm 0.094 \quad \text{or} \quad (0.731, 0.919)$$

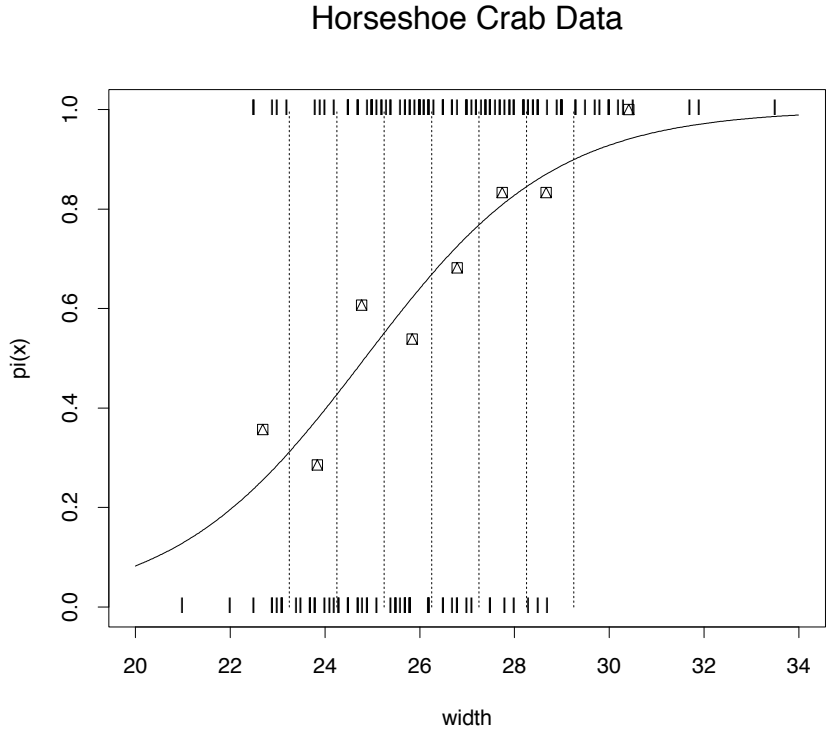
Notice how this interval is wider than the one based on the logistic regression model.

The following table presents the confidence intervals for $p(x_0)$ for all the observed values of the covariate:

Dose x_i	Number of insects, n_i	Number killed, y_i	Proportion killed, y_i/n_i	Predicted Proportion	Lower Bound	Upper Bound
1.6907	59	6	.1017	.062	.037	.103
1.7242	60	13	.2167	.168	.120	.231
1.7552	62	18	.2903	.363	.300	.431
1.7842	56	28	.5000	.600	.538	.659
1.8113	63	52	.8254	.788	.724	.840
1.8369	59	53	.8983	.897	.853	.929
1.8610	62	61	.9839	.951	.920	.970
1.8839	60	59	.9833	.977	.957	.988

Example: Agresti's Horseshoe Crab Data

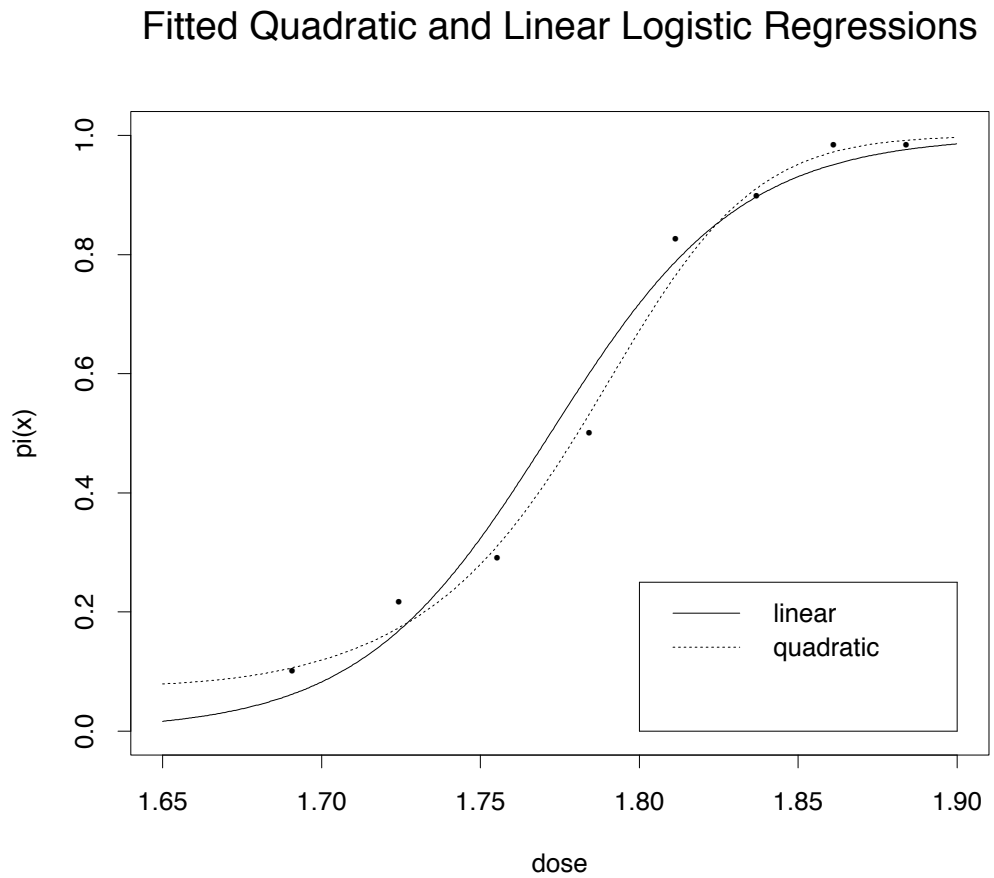
Agresti, Ch. 4, presents a data set from a study of nesting crab. Each female in the study had a male crab accompanying her. Additional male crabs living near her are called *satellites*. The response equals 1 if the female crab has a satellite and 0, otherwise. Predictors include the color, spine condition, width, and weight of the female crab. The plot below depicts the response as a function of carapace width. The observed data are plotted using the symbol "|". The groups used later for model checking are separated using the vertical lines. The plotted line is the logistic regression for the entire data set.



Example: Beetle Mortality Data

On the output, we also fit a quadratic logistic regression function to the beetle mortality data. We use this to illustrate the comparison of models using the deviances.

The fitted linear and quadratic regressions are in the following graph:



8.3. Logit Models for Qualitative Predictors

We have looked at logistic regression models for quantitative predictors. Similar to ordinary regression, we can have qualitative explanatory variables.

8.3.1. Dummy Variables in Logit Models

As in ordinary regression, dummy variables are used to incorporate qualitative variables into the model.

Example: An antibiotic for pneumonia was injected into 100 patients with kidney malfunctions and into 100 normal patients. Some allergic reactions developed in 38 uremic patients and 21 normal patients.

	Allergic Reaction		
Treatment	Yes	No	Total
Uremic	38	62	100
Normal	21	79	100
Total	59	141	200

Here $Y = 1$ if "yes" and $Y = 0$ if "no".

Also, $X = 1$ if uremic and $X = 0$ if normal.

The logistic regression model is given by

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x$$

We can obtain a table for the values of the logistic regression model:

	Response (Y)		
Explanatory Variable (X)	$y = 1$	$y = 0$	Total
$x = 1$	$p_1 = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$1 - p_1 = \frac{1}{1+e^{\beta_0+\beta_1}}$	100
$x = 0$	$p_0 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$	$1 - p_0 = \frac{1}{1+e^{\beta_0}}$	100

The odds ratio for a 2×2 table can be expressed by

$$\theta = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\frac{1}{1+e^{\beta_0}}\right)} = e^{\beta_1}.$$

8.3.2. Inference for a Dichotomous Covariate

Data: (x_i, Y_i), $i = 1, \cdots, n$

- The response Y_i equals 1 if "yes" and 0 if "no".
- The explanatory variable x_i equals 1 if Group 1 or 0 if Group 0.

We summarize the data are following 2×2 table:

	Response (Y)		
Explanatory Variable (X)	$y = 1$ (yes)	$y = 0$ (no)	Total
$x = 1$	n_{11}	n_{12}	n_{1+}
$x = 0$	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- $n_{+1} = \sum_{i=1}^{n} Y_i =$ Total # of "yes" responses
- $n_{11} = \sum_{i=1}^{n} x_i Y_i =$ Total # of "yes" responses in Group 1
- $n_{21} = \sum_{i=1}^{n} (1 - x_i) Y_i =$ Total # of "yes" responses in Group 0

Setting the likelihood equations equal to zero yields:

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{n_{11}}{n_{1+}} = \hat{p}_1$$

$$\frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \frac{n_{21}}{n_{2+}} = \hat{p}_0$$

We solve to obtain the mle for (β_0, β_1):

$$\hat{\beta}_0 = \log\left(\frac{n_{21}}{n_{22}}\right)$$

$$\hat{\beta}_1 = \log\left(\frac{n_{11}/n_{12}}{n_{21}/n_{22}}\right)$$

- $\hat{\beta}_0$ is the log odds for $X = 0$ (Reference Group)
- $\hat{\beta}_1$ is the log odds ratio for $X = 1$ relative to $X = 0$

Note that the estimating equations above imply that

$$\hat{p}_1 = \frac{n_{11}}{n_{1+}} \quad \text{and} \quad \hat{p}_0 = \frac{n_{21}}{n_{2+}}$$

Confidence Intervals for β_0 and β_1

Using the information matrix, one can show that

$$\begin{aligned}
Var(\hat{\beta}_1) &= \frac{n_{2+}}{n_{21}(n_{2+}-n_{21})} + \frac{n_{1+}}{n_{11}(n_{1+}-n_{11})} \\
&= \frac{1}{n_{11}} + \frac{1}{n_{1+}-n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{2+}-n_{21}} \\
&= \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \\
Var(\hat{\beta}_0) &= \frac{n_{2+}}{n_{21}(n_{2+}-n_{21})} = \frac{n_{2+}}{n_{21}n_{22}}
\end{aligned}$$

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm z_{\alpha/2}\widehat{se}(\hat{\beta}_1)$$

where $\widehat{se}(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)}$.

Confidence Interval for the Odds Ratio

- Recall that the odds ratio for $X = 1$ relative to $X = 0$ is e^{β_1}.
- The logarithm of the odds ratio is simply the logistic regression coefficient β_1.
- The confidence interval for β_1 can be exponentiated to form a 100(1-α)% confidence interval for the odds ratio:

$$\exp\left[\hat{\beta}_1 \pm z_{\alpha/2}\widehat{se}(\hat{\beta}_1)\right]$$

An Alternative Form of Coding

Another coding method is called *deviation from the mean coding*. This method assigns -1 to the lower code and 1 to the higher code. In this case, the log odds ratio becomes

$$\begin{aligned} \log\theta &= \text{logit}(p(1)) - \text{logit}(p(-1)) \\ &= (\beta_0 + \beta_1 \times 1) - (\beta_0 + \beta_1 \times (-1)) = 2\beta_1. \end{aligned}$$

The endpoints of the $100(1-\alpha)\%$ confidence interval for the odds ratio are

$$\exp\left[2\hat{\beta}_1 \pm 2z_{\alpha/2}\widehat{se}(\hat{\beta}_1)\right]$$

Example: Calculation for the antibiotic data

From the output for the logistic regression model, we obtain

$$\hat{\beta}_1 = 0.8354 \qquad \text{and} \qquad \widehat{se}(\hat{\beta}_1) = 0.3205.$$

We compute a 95% confidence interval for β_1:

$$.8354 \pm (1.96)(.3205) = .8354 \pm .6272.$$

The resulting confidence interval is (.2072, 1.4636). We can exponentiate the endpoints to obtain a 95% confidence interval for the odds ratio:

$$(e^{.2072}, e^{1.4636}) = (1.2302, 4.3215)$$

The 95% confidence interval for β_0 is

$$-1.3249 \pm (1.96)(.2455) = -1.3249 \pm .4812 \quad \text{or} \quad (-1.8016, -0.8437).$$

Noting that $p(0) = \Pr(\text{yes}|\text{Normal}) = \frac{1}{1+e^{-\beta_0}}$, we can obtain a 95% confidence interval for $p(0)$:

$$\left(\frac{1}{1 + e^{1.8016}}, \frac{1}{1 + e^{.8437}} \right) = (.1411, .3007)$$

8.3.3. Polytomous Independent Variables

We now suppose that instead of two categories the independent variable can take on $k > 2$ distinct values. We define $k - 1$ dummy variables to form a logistic regression model:

- $x_1 = \begin{cases} 1 & \text{if category 1} \\ 0 & \text{otherwise} \end{cases}$
- $x_2 = \begin{cases} 1 & \text{if category 2} \\ 0 & \text{otherwise} \end{cases}$
- $\vdots$
- $x_{k-1} = \begin{cases} 1 & \text{if category k-1} \\ 0 & \text{otherwise} \end{cases}$

The resulting logistic regression model is

$$\text{logit}(p(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}.$$

This parametrization treats Category k as a reference category.

Implication of Model: We can form a table of the logits corresponding to the different categories:

Category	Logit
1	$\beta_0 + \beta_1$
2	$\beta_0 + \beta_2$
$\vdots$	$\vdots$
$k-1$	$\beta_0 + \beta_{k-1}$
k	β_0

The odds ratio for comparing category j to the reference category k is

$$\theta = e^{\beta_j}.$$

We can form confidence interval for the β_js and then exponentiate the endpoints to obtain the confidence intervals for the odds ratios.

Remark: When the categories are ordinal, one can use scores in fitting a linear logistic regression model. To test for an effect due to categories, one can test $H_0 : \beta_1 = 0$. An alternative analysis to test for a linear trend for category probabilities uses the Cochran-Armitage statistic. These approaches yield equivalent results with the score statistic from logistic regression being equivalent to the Cochran-Armitage statistic.

Example: Horseshoe Crab Data Continued

Earlier we used width (x_1) as a predictor of the presence of satellites. We now include color as a second explanatory variable. Color is a surrogate for age with older crabs tending to be darker.

- We can treat color as an *ordinal* variable by assigning scores to the levels: (1) Light Medium, (2) Medium, (3) Dark Medium, (4) Dark
- We can treat color as a *nominal* variable by defining dummy variables or by using a "class" statement in `proc logistic`.

$$x_2 = \begin{cases} 1 & \text{light medium} \\ 0 & \text{otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{medium} \\ 0 & \text{otherwise} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{dark medium} \\ 0 & \text{otherwise} \end{cases}$$

Consider the two main effect models:

- Ordinal color (z):

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \beta_2^* z$$

- Nominal color (x_2, x_3, x_4):

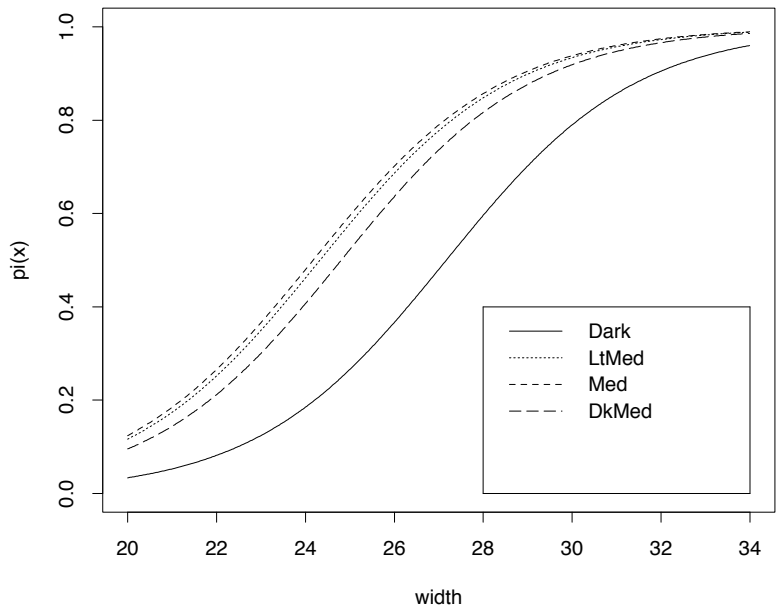
$$\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

The logits for the two models are in the following table:

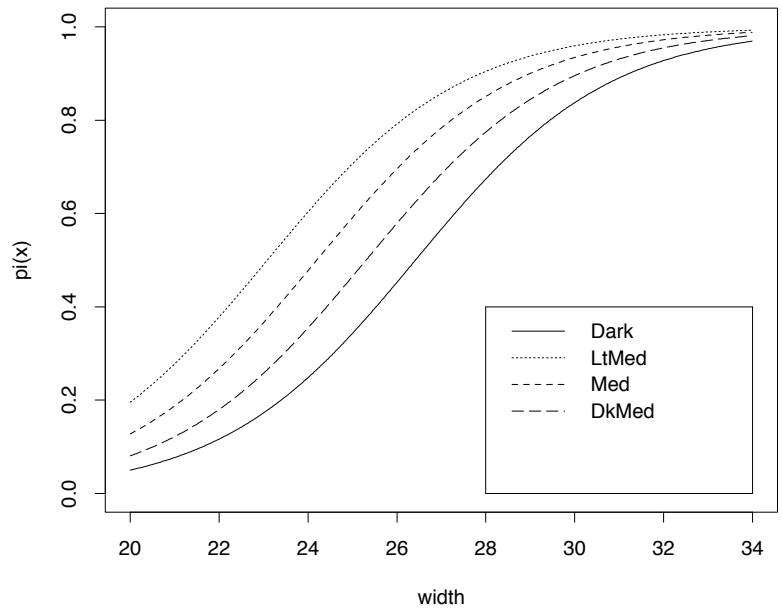
Color	Ordinal	Nominal
Light Medium	$\beta_0 + \beta_1 x_1 + \beta_2^*$	$\beta_0 + \beta_1 x_1 + \beta_2$
Medium	$\beta_0 + \beta_1 x_1 + 2\beta_2^*$	$\beta_0 + \beta_1 x_1 + \beta_3$
Dark Medium	$\beta_0 + \beta_1 x_1 + 3\beta_2^*$	$\beta_0 + \beta_1 x_1 + \beta_4$
Dark	$\beta_0 + \beta_1 x_1 + 4\beta_2^*$	$\beta_0 + \beta_1 x_1$

- These models assume no interaction between width and color
- Width has the same effect for all four colors-the slope β_1 is the same.
- Thus, the shapes of the curves are the same. Any curve can be obtained from the others by shifting either to the left or to the right.
- The curves are “parallel” in that they never cross.

Main effects model logistic regression



Main effects model with ordinal color



8.4. Models with Two Qualitative Predictors

Suppose that there are two qualitative predictors, X and Z, each with two levels. We then have a $2 \times 2 \times 2$ table. The data are

$$(x_i, Y_i, z_i), \quad i = 1, \cdots, n$$

- $Y_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$
- $x_i = \begin{cases} 1 & \text{if Group 1} \\ 0 & \text{if Group 0} \end{cases}$
- $z_i = \begin{cases} 1 & \text{if Layer 1} \\ 0 & \text{if Layer 0} \end{cases}$

We will consider two logistic regression models, a main effects model and a model with interaction.

Define the following probabilities:

$$p_{00} = \Pr(Y = 1 | X = 0, Z = 0)$$

$$p_{01} = \Pr(Y = 1 | X = 1, Z = 0)$$

$$p_{10} = \Pr(Y = 1 | X = 0, Z = 1)$$

$$p_{11} = \Pr(Y = 1 | X = 1, Z = 1)$$

- **Main Effects Model:** $\text{logit}(p(x,z)) = \beta_0 + \beta_1 x + \beta_2 z$

$\text{logit}(p_{00})$	β_0
$\text{logit}(p_{10})$	$\beta_0 + \beta_1$
$\text{logit}(p_{01})$	$\beta_0 + \beta_2$
$\text{logit}(p_{11})$	$\beta_0 + \beta_1 + \beta_2$

$$\beta_0 = \log\left(\frac{p_{00}}{1-p_{00}}\right) \qquad \text{log odds of reference}$$

$$\begin{aligned}
\beta_1 &= \text{logit}(p_{01}) - \text{logit}(p_{00}) = \log\left(\frac{odds_{01}}{odds_{00}}\right) = \log\theta_{XY|Z=0} \\
&= \text{logit}(p_{11}) - \text{logit}(p_{10}) = \log\left(\frac{odds_{11}}{odds_{10}}\right) = \log\theta_{XY|Z=1}
\end{aligned}$$

$$\begin{aligned}
\beta_2 &= \text{logit}(p_{10}) - \text{logit}(p_{00}) = \log\left(\frac{odds_{10}}{odds_{00}}\right) = \log\theta_{ZY|X=0} \\
&= \text{logit}(p_{11}) - \text{logit}(p_{01}) = \log\left(\frac{odds_{11}}{odds_{01}}\right) = \log\theta_{ZY|X=1}
\end{aligned}$$

Note:

1. This main effects model assumes that the XY association is homogeneous across levels of Z and that the ZY association is homogeneous across levels of X.
2. $H_0 : \beta_1 = 0$ is equivalent to $H_0 :$ X and Y are conditionally independent controlling for Z.

- **Interaction Model:** $\text{logit}(p(x,z)) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 (x \times z)$

This model adds an interaction term $x \times z$ to the main effects model.

$\text{logit}(p_{00})$	β_0
$\text{logit}(p_{10})$	$\beta_0 + \beta_1$
$\text{logit}(p_{01})$	$\beta_0 + \beta_2$
$\text{logit}(p_{11})$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

$$\beta_0 = \log\left(\frac{p_{00}}{1-p_{00}}\right) \qquad \text{log odds of reference}$$

$$\beta_1 = \text{logit}(p_{01}) - \text{logit}(p_{00}) = \log\left(\frac{odds_{01}}{odds_{00}}\right) = \log\theta_{XY|Z=0}$$

$$\beta_1 + \beta_3 = \text{logit}(p_{11}) - \text{logit}(p_{10}) = \log\left(\frac{odds_{11}}{odds_{10}}\right) = \log\theta_{XY|Z=1}$$

$$\beta_2 = \text{logit}(p_{10}) - \text{logit}(p_{00}) = \log\left(\frac{odds_{10}}{odds_{00}}\right) = \log\theta_{ZY|X=0}$$

$$\beta_2 + \beta_3 = \text{logit}(p_{11}) - \text{logit}(p_{01}) = \log\left(\frac{odds_{11}}{odds_{01}}\right) = \log\theta_{ZY|X=1}$$

Note:

1. This model does not assume that XY association is homogeneous across levels of Z and that the ZY association is homogeneous across levels of X.
2. We test homogeneity across layers by testing $H_0 : \beta_3 = 0$.

8.4.1. The Loglinear-Logit Connection

Loglinear models for contingency tables have all variables as response variables whereas logit models have a binary response variable which depends on a set of explanatory variables. To help interpret a loglinear model, it is sometimes useful to construct an equivalent logit model.

Consider the homogeneous association model:

$$\log e_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

We will suppose that Y is a binary variable and treat it as a response. We let X and Z be considered as explanatory variables. Let p be the probability that $Y = 1$. The logit for Y is

$$\begin{aligned}
\text{logit}(p) &= \log\left(\frac{p}{1-p}\right) = \log\left(\frac{\Pr(Y=1|X=i,Z=k)}{\Pr(Y=0|X=i,Z=k)}\right) \\
&= \log\left(\frac{e_{i1k}}{e_{i2k}}\right) = \log(e_{i1k}) - \log(e_{i2k}) \\
&= \lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ} \\
&\quad -(\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}) \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}).
\end{aligned}$$

The first term is a constant and does not depend on i or k. The second term depends on the level i of X. The third term depends on the level k of Z. Thus, the logit can be written

$$\text{logit}(p) = \alpha + \beta_i^X + \beta_k^Z.$$

- When Y is binary, the loglinear model with homogeneous association is equivalent to the above logit model.

- When X is also binary, this logit model and the loglinear model (XY, XZ, YZ) have equal odds ratios at each of the K levels of Z. The G^2 and X^2 goodness-of-fit statistics are an alternative way to test for a common odds ratio.

- When we derived the logit model corresponding to the (XY, XZ, YZ) loglinear model, the λ_{ik}^{XZ} terms cancelled out. Thus, the same derivation for the (XY, YZ) model would also lead the same logit model. However, the loglinear model that has the same fit as the logit model is the one containing a general interaction term for relationships among the explanatory variables. The logit model does not describe relationships among the explanatory variables, so it allows a general interaction pattern.

- Some equivalent loglinear and logit models when Y is a binary response variable:

Loglinear	Logit
(Y, XZ)	α
(XY, XZ)	$\alpha + \beta_i^X$
(YZ, XZ)	$\alpha + \beta_k^Z$
(XY, XZ, YZ)	$\alpha + \beta_i^X + \beta_k^Z$
(XYZ)	$\alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$

8.5. Model Checking

We have assumed that the logistic regression model is the appropriate model for a set of data. In this section we will look at ways of assessing the fit of the model. We first discuss various goodness-of-fit tests. We will also discuss the use of residuals in assessing the fit of the model.

8.5.1. Goodness-of-Fit Tests

A fitted logistic regression model provides estimated probabilities that $Y = 1$ and $Y = 0$ at each setting of the explanatory variables. We can then calculate the predicted number of subjects at each setting by multiplying the estimated probability times the number of subjects at the setting. We can then compare the observed and estimated frequencies using Pearson's X^2 or the likelihood ratio G^2 statistic.

Suppose that there are N settings of the explanatory variables. We define the following variables:

- n_i = number of trials at the ith setting
- y_i = number of successes at the ith setting
- $\hat{p}_i$ = predicted probability of success at the ith setting
- $\hat{y}_i = n\hat{p}_i$ = predicted number of successes at the ith setting
- The Pearson's residual is defined as

$$r_i^P = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}$$

Pearson's goodness-of-fit statistic can be written as

$$X^2 = \sum_{i=1}^{N}(r_i^P)^2 = \sum_{i=1}^{N}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

Large values of X^2 will cause us to conclude that the proposed model does not fit the data.

Remarks:

- Each squared Pearson residual is a component of X^2.
- For large n, r_i^P is approximately $N(0,1)$ when the model holds.
- For a given n_i, $y_i - n_i\hat{p}_i = y_i - \hat{y}_i$ tends to be smaller than $y_i - n_i p_i$ and so the actual variance of the Pearson residual is less than 1. The standardized Pearson residual is defined as

$$\tilde{r}_i^P = \frac{r_i^P}{\sqrt{1-\hat{h}_i}} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)(1-\hat{h}_i)}}$$

  where $\hat{h}_i$ is the leverage associated with observation i.
- Absolute values of $\tilde{r}_i^P$ or r_i^P larger than 2 or 3 provide some evidence of lack of fit.
- Residual plots against explanatory variables or linear predictor values can help detect a lack of fit.