4 Logistic Regression Model

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

4.1 Interpreting the Logistic Regression Model

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

$$Y \quad : \quad \text{binary response variable} \begin{cases} 1 & \text{“yes” or success} \\ 0 & \text{no or failure} \end{cases}$$

$$X \quad : \quad \text{quantitative explanatory variable}$$

We want to model

$$\pi(x) = P(Y = 1|X = x)$$

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

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

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

We can solve for $\pi(x)$:

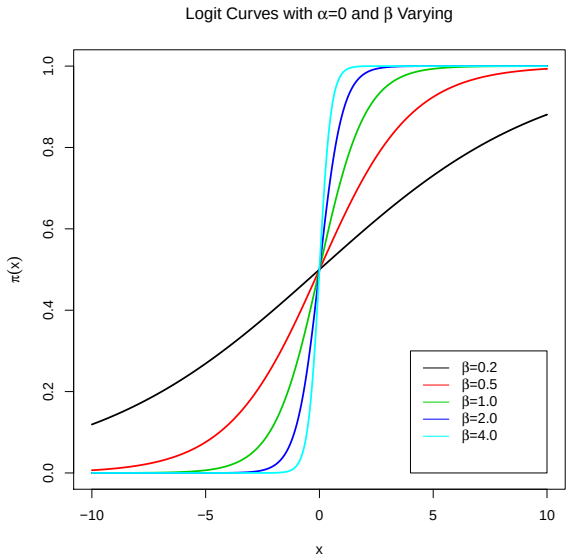
$$\pi(x) = \frac{\exp\{\alpha+\beta x\}}{1+\exp\{\alpha+\beta x\}} = \frac{1}{1+\exp\{-(\alpha+\beta x)\}}.$$

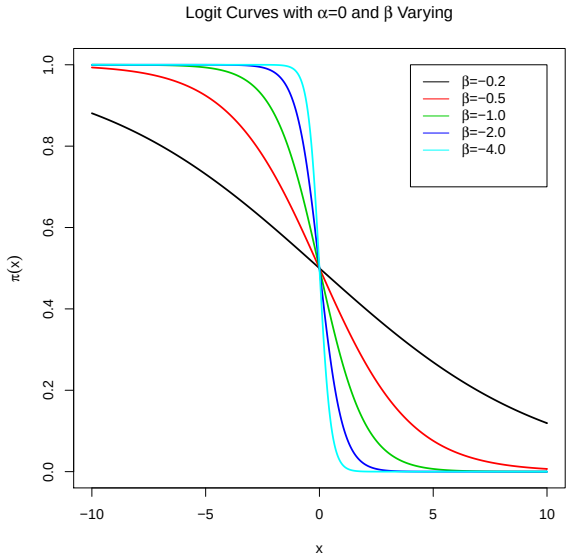
Note: This is a generalized linear model with the following components:

- Link: Logit (log-odds)
- Linear predictor: $\alpha + \beta x$
- Distribution: Binomial

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

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



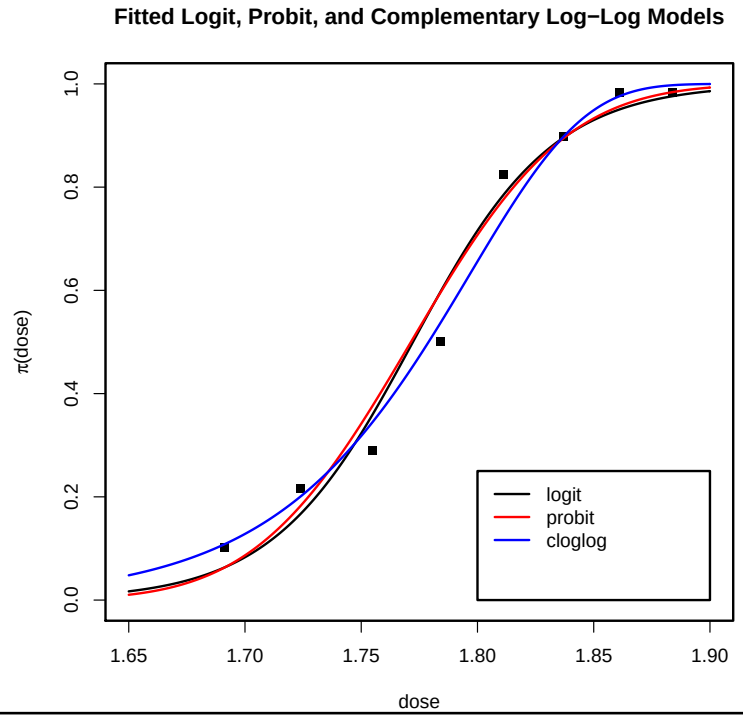


Example: Beetle Mortality Data

Beetles were treated with various concentrations of insecticide for 5 hrs. The fitted models using $x =$dose as a predictor were:

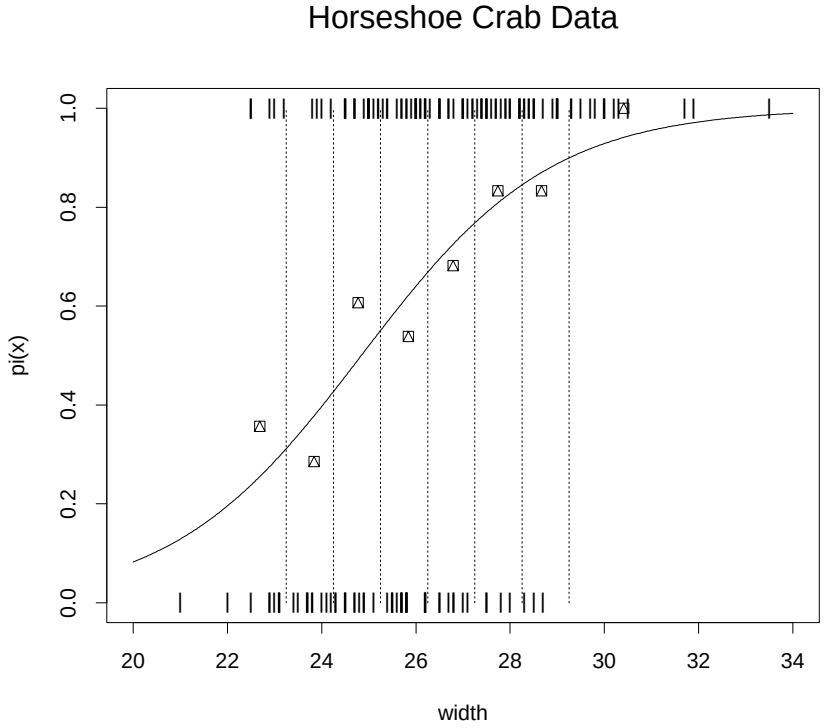
- $\text{logit}(\hat{\pi}(x)) = -58.936 + 33.255x$
- $\text{probit}(\hat{\pi}(x)) = -33.791 + 19.076x$
- $\text{cloglog}(\hat{\pi}(x)) = -36.862 + 20.513x$

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



Example: Agresti's Horseshoe Crab Data

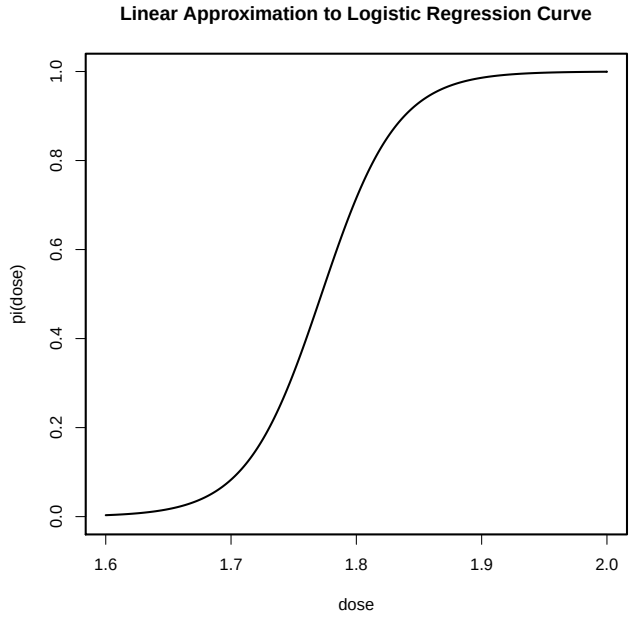
Agresti, Ch. 4, presents a data set from a study of nesting crab. Each female in the study had a male crab accompanying her. Additional male crabs living near her are called *satellites*. The response equals 1 if the female crab has a satellite and 0, otherwise. Predictors include the color, spine condition,width, 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.



4.1.1 Linear Approximation Interpretation

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

$$\text{slope} = \frac{d\pi(x)}{dx} = \beta\pi(x)(1-\pi(x))$$



$\pi(x)$	.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 $\pi(x) = 0.5$ or $x = -\alpha/\beta$. This value is known as *the median effective level* and is denoted EL_{50}

- In the beetle mortality example from Chapter 3, $\text{logit}(\hat{\pi}(x)) = -58.936 + 33.255x$. Thus, $EL_{50} = 1.772$ and the slope is 8.313.

- In the horseshoe crab example, $\text{logit}(\hat{\pi}(x)) = -12.35 + 0.497x$. Thus, $EL_{50} = 24.84$ and the slope is 0.124.

4.1.2 Odds Ratio Interpretation

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

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

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

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

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

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

For an increase of 1 unit in x, the odds are multiplied by a factor of e^{β}.

Example: In the horseshoe crab example, the odds for a female to have a satellite are multiplied by a factor of $e^{0.497} = 1.644$ for each centimeter increase in carapace width.

4.1.3 Logistic Regression and Retrospective Studies

Logistic regression can also be applied in situations where the explanatory variable X rather than the response variable Y is random. This is typical of some retrospective sampling designs such as case-control studies.

In a case-control study, samples of cases ($Y = 1$) and controls ($Y = 0$) are taken, and the value of X is recorded. If the distribution of X differs between the cases and controls, there is evidence of an association between X and Y. When X was binary, we were able to estimate the odds-ratio between X and Y in Chapter 2. We will develop logit models for matched case-control studies in Chapter 8.

For more general distributions of X, we can use logistic regression to estimate the effect of X using parameters that refer to odds and odds ratios. However, the intercept term α is not useful because it is related to the relative number of times $Y = 1$ and $Y = 0$.

4.1.4 Logistic Regression and the Normal Distribution

When Y is a binary response and X is a predictor with a discrete distribution, one can use Bayes theorem to show that

$$\frac{\pi(x)}{1-\pi(x)} = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)}.$$

We can take the logarithm of the corresponding result for a continuous predictor and obtain

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right).$$

Suppose next that the conditional distribution of X given $Y=i$ is $N(\mu_i, \sigma_i^2)$, $i=0,1$. Then substituting the normal density into the above expression yields

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 \text{ where } \beta_1 = \frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2} \text{ and } \beta_2 = \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right).$$

- When $\sigma_1^2 = \sigma_0^2$, the log-odds ratio is a linear function of x.
- When $\sigma_1^2 \neq \sigma_0^2$, the log-odds ratio is a quadratic function of x.

4.2 Multiple Logistic Regression

We consider logistic regression models with one or more explanatory variables.

- Binary response: Y
- k predictors: $\boldsymbol{x} = (x_1, \ldots, x_k)$
- Quantity to estimate: $\pi(\boldsymbol{x}) = P(Y = 1 | x_1, \ldots, x_k)$

The logistic regression model is

$$\text{logit}(\pi(\boldsymbol{x})) = \log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- The parameter β_j reflects the effect of a unit change in x_j on the log-odds that $Y = 1$, keeping the other 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:

$$\begin{aligned}
&\text{logit}(\pi(x_1 + 1, x_2, \ldots, x_k)) - \text{logit}(\pi(x_1, x_2, \ldots, x_k)) \\
&= \alpha + \beta_1(x_1 + 1) + \cdots + \beta_k x_k - (\alpha + \beta_1 x_1 + \cdots + \beta_k x_k) \\
&= \beta_1
\end{aligned}$$

4.2.1 Estimation of Parameters

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

- $Y_i =$ binary response for i^{th} observation
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik}) =$ the values of the k explanatory variables

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

$$\pi_i = \pi(\boldsymbol{x}_i) = \frac{\exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}$$

.

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

$$\ell(\alpha, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

The log-likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}) = \log(\ell(\boldsymbol{\beta})) &= \sum_{i=1}^{n}[\log \binom{n_i}{y_i} + y_i \log(\pi_i) + (n - y_i)\log(1-\pi_i)] \\ &= \textstyle\sum_j \left(\sum_i y_i x_{ij}\right)\beta_j - \sum_i n_i \log\left[1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right] + \sum_i \log\binom{n_i}{y_i}. \end{aligned}$$

We wish to estimate $\alpha, \beta_1, \ldots, \beta_k$ using maximum likelihood.

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

$$U_1(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \alpha} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} n_i \pi_i = 0$$

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

$$j = 1, \ldots, k$$

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:

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

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

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

4.2.2 Overall Test for the Model

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

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

We typically use the likelihood ratio statistic:

$$G^2 = Q_L = -2 \log \left[\frac{\ell(\tilde{\alpha})}{\ell(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k)} \right] = 2[L_{Full} - L_{Reduced}]$$

Here $\tilde{\alpha}$ is the m.l.e. of α under the null hypothesis that the model with intercept only holds. When H_0 is true,

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

We reject H_0 for large values of Q_L.

4.2.3 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 Wald statistic is given by

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

We reject H_0 for large values of $|Z|$. Alternatively, we can use the LR statistic for testing these hypotheses.

4.2.4 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

$$\hat{\beta}_j \pm Z_{\alpha/2}\widehat{SE}(\hat{\beta}_j)$$

4.2.5 Confidence Intervals for the Logit and for the Probability of a Success

We next consider forming a confidence interval for the logit (linear predictor) at a given value of $\boldsymbol{x}$:

$$g(\boldsymbol{x}) = \text{logit}(\pi(\boldsymbol{x})) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k.$$

The estimated logit is given by

$$\hat{g}(\boldsymbol{x}) = \text{logit}(\hat{\pi}(\boldsymbol{x})) = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k.$$

This has estimated asymptotic variance

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

In the above formula, $x_0 = 1$.

A $100(1-\alpha)$% confidence interval for $\text{logit}(\pi(\boldsymbol{x}))$

$$\hat{g}(\boldsymbol{x}) \pm Z_{\alpha/2} \widehat{SE}(\hat{g}(\boldsymbol{x}))$$

where $\widehat{SE}(\hat{g}(\boldsymbol{x})) = \sqrt{\widehat{\text{Var}}(\hat{g}(\boldsymbol{x}))}$

Since

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

we can find a $100(1-\alpha)$% confidence interval for $\pi(\boldsymbol{x})$ 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

$$\hat{\alpha} + \hat{\beta}x_0 \pm Z_{\alpha/2}\widehat{SE}(\hat{\alpha} + \hat{\beta}x_0).$$

The estimated asymptotic standard error is

$$\widehat{SE}(\hat{\alpha} + \hat{\beta}x_0) = \sqrt{\widehat{\mathrm{Var}}(\hat{\alpha} + \hat{\beta}x_0)} = \sqrt{\widehat{\mathrm{Var}}(\hat{\alpha}) + x_0^2\widehat{\mathrm{Var}}(\hat{\beta}) + 2x_0\widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta})}$$

Since

$$\pi(x_0) = \Pr(Y = 1|X = x_0) = \frac{e^{\alpha+\beta_1 x_0}}{1 + e^{\alpha+\beta_1 x_0}} = \frac{1}{1 + e^{-(\alpha+\beta_1 x_0)}}.$$

We substitute the endpoints of the confidence interval for the logit into the above formula to obtain a confidence interval for $\pi(x_0)$.

Remarks

- The fitted value $\hat{\pi}(x_0)$ 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, $\hat{\pi}(x_0)$ 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 $\pi(x_0)$ 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 $\pi(x_0)$.

4.2.6 Examples

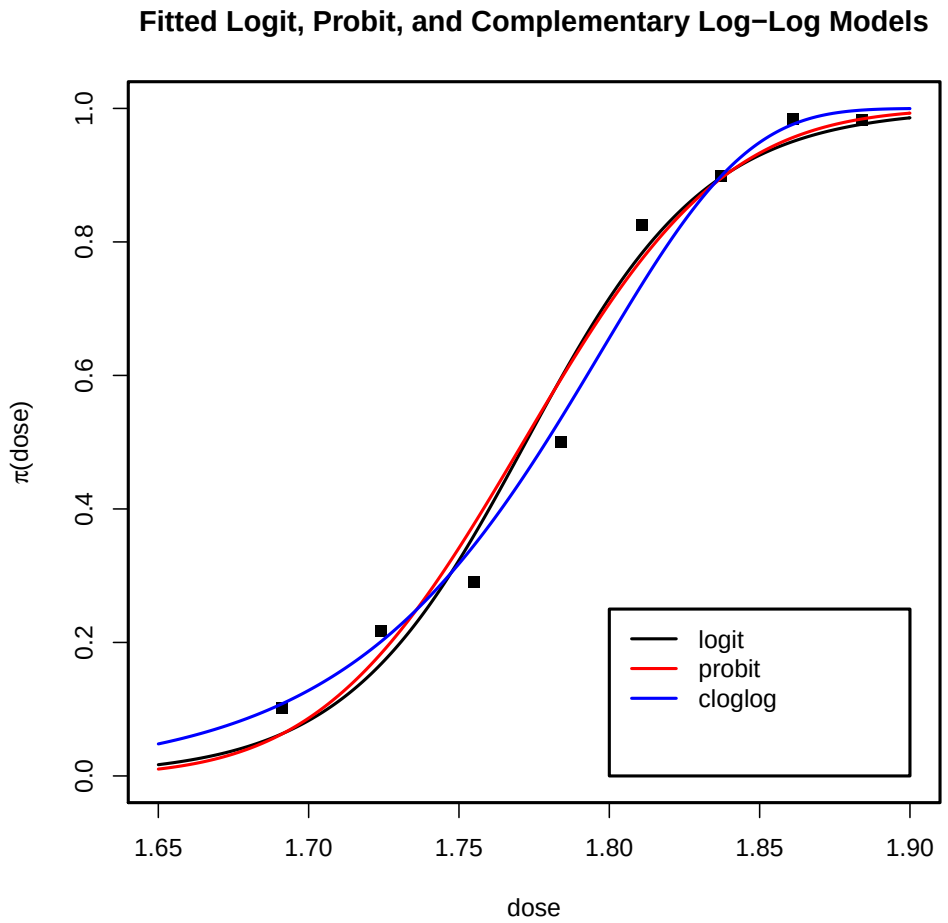
Example: Beetles were treated with various concentrations of insecticide for 5 hrs. The data appear in the following table:

Dose x_i $(\log_{10} \mathrm{CS}_2\mathrm{mgl}^{-1})$	Number of insects, n_i	Number killed, Y_i	Proportion killed, $\frac{y_i}{n_i}$
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	0.9833

The fitted logit model using `proc logistic` is:

$$\text{logit}(\hat{\pi}(x)) = -59.1834 + 33.3984x$$

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



- Test for $H_0 : \beta = 0$.

  The three tests strongly reject H_0.

- 95% confidence interval for β:

$$33.3984 \pm 1.96 \cdot 2.8392 = 33.3984 \pm 5.5648$$

- Confidence interval for $\text{logit}(\pi(x_0))$

  Let $x_0 = 1.8113$. The estimated logit is $-59.18 + 33.40 \cdot 1.8113 = 1.3111$.

$$\begin{aligned} \widehat{SE} &= \sqrt{25.529 + 1.8113^2 \cdot 8.0613 + 2 \cdot 1.8113 \cdot (-14.341)} \\ &= \sqrt{0.02517} = 0.159 \end{aligned}$$

  The 95% confidence interval for the logit is

$$1.311 \pm 1.96 \cdot 0.159 = 1.311 \pm .311 \quad \text{or} \quad (1.000, 1.622)$$

  The 95% confidence interval for $\pi(1.8113)$ is

$$\left(\frac{e^1}{1+e^1}, \frac{e^{1.622}}{1+e^{1.622}} \right) = (0.731, 0.835)$$

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

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

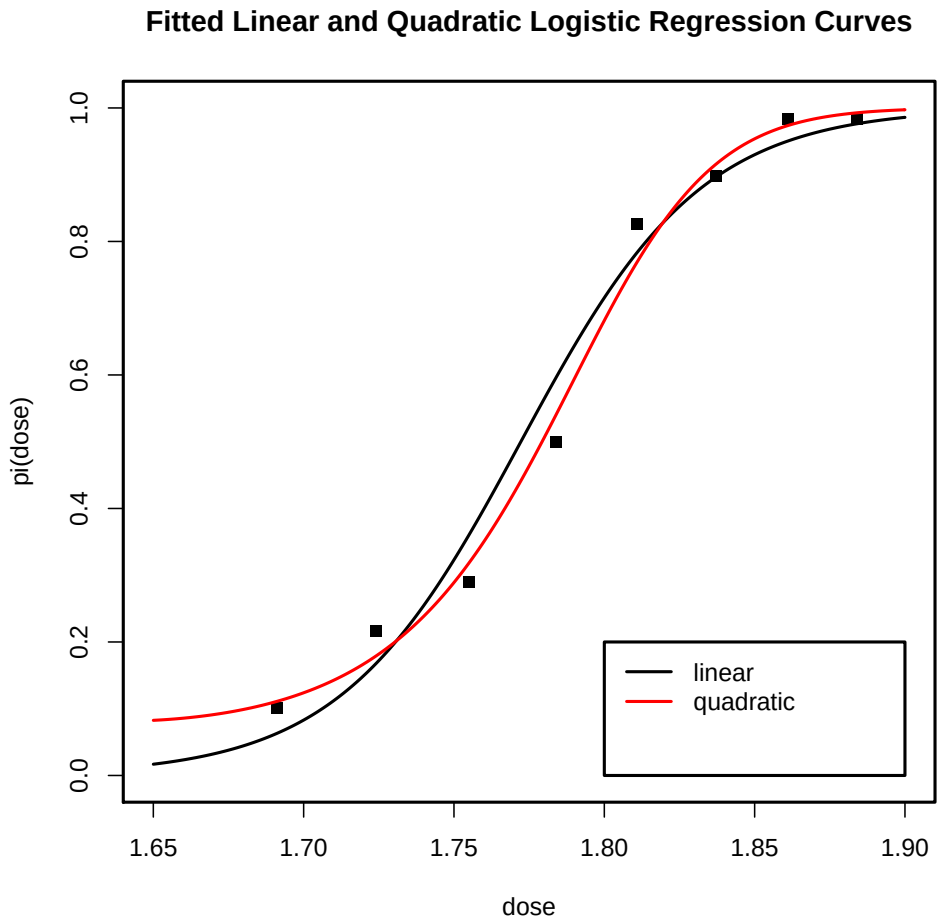
Notice how this interval is wider than the one based on the logistic regression model.

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

Dose x_i	Number of insects, n_i	Number killed, Y_i	Proportion killed, $\frac{y_i}{n_i}$	Predicted	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	.731	.835
1.8369	59	53	.8983	.897	.853	.929
1.8610	62	61	.9839	.951	.920	.970
1.8839	60	59	0.9833	.977	.957	.988

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:



4.3 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.

4.3.1 Dummy Variables in Logit Models

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

Example: Investigators examined a sample of 178 children who appeared to be in remission from leukemia using the standard criterion after undergoing chemotherapy. A new test (PCR) detected traces of cancer in 75 of these children. During 3 years of followup, 30 of these children suffered a relapse. Of the 103 children who did not show traces of cancer, 8 suffered a relapse.

	Relapse		
Group	Yes	No	Total
Traces of Cancer	30	45	75
Cancer Free	8	95	103
Total	38	140	178

Here $Y = 1$ if "yes" and $Y = 0$ if "no". Also, $X = 1$ if "traces" and $X = 0$ if "cancer free".

The logistic regression model is given by

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

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

	Response(Y)		
Explanatory Variable (X)	$y = 1$	$y = 0$	Total
$x = 1$	$\pi_1 = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$	$1 - \pi_1 = \frac{1}{1+e^{\alpha+\beta}}$	1
$x = 0$	$\pi_0 = \frac{e^{\alpha}}{1+e^{\alpha}}$	$1 - \pi_0 = \frac{1}{1+e^{\alpha}}$	1

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

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{\left(\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}\right) / \left(\frac{1}{1+e^{\alpha+\beta}}\right)}{\left(\frac{e^{\alpha}}{1+e^{\alpha}}\right) / \left(\frac{1}{1+e^{\alpha}}\right)} = e^{\beta}.$$

4.3.2 Inference for a Dichotomous Covariate

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

- 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 the 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 2

Setting the likelihood equations equal to zero yields:

$$\frac{e^{\hat{\alpha}+\hat{\beta}}}{1+e^{\hat{\alpha}+\hat{\beta}}} = \frac{n_{11}}{n_{1+}} = \hat{\pi}_1$$

$$\frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}} = \frac{n_{21}}{n_{2+}} = \hat{\pi}_0$$

We solve to obtain the mle for (α, β):

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

$$\hat{\beta} = \log\left[\frac{n_{11}/n_{12}}{n_{21}/n_{22}}\right]$$

- $\hat{\alpha}$ is the log-odds of a "yes" for $X = 0$ (Reference Group)
- $\hat{\beta}$ is the log odds-ratio of a "yes" for $X = 1$ relative to $X = 0$

Note that the estimating equations above imply that

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

Score Test for $H_0 : \beta = 0$

The score statistic for testing $H_0 : \beta = 0$ is Pearson's Chi-squared Statistic for Independence in a 2×2 table. Under H_0 this has approximately a χ^2_1 distribution.

Confidence Intervals for α and β

Using the information matrix, one can show that

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

$$\mathrm{Var}(\hat{\alpha}) = \frac{n_{2+}}{n_{21}(n_{2+}-n_{21})}$$

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

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

where $\widehat{SE}(\hat{\beta}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}$

Confidence Interval for the Odds Ratio

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

$$\exp\left[\hat{\beta} \pm Z_{\alpha/2}\widehat{SE}(\hat{\beta})\right]$$

An Alternative Form of Coding

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

$$\begin{aligned} \log(OR) &= \text{logit}(\pi(1)) - \text{logit}(\pi(-1)) \\ &= (\alpha + \beta \times 1) - (\alpha + \beta \times (-1)) = 2\beta \end{aligned}$$

The endpoints of the 100$(1 - \alpha)$% c.i. for the OR are

$$\exp\left[2\hat{\beta} \pm 2Z_{\alpha/2}\widehat{SE}(\hat{\beta})\right]$$

Example: Calculations for the cancer relapse data

From the output for the logistic regression model, we obtain

$$\hat{\beta} = 2.0690 \qquad \text{and} \qquad \widehat{SE}(\hat{\beta}) = 0.4371.$$

We compute a 95% confidence interval for β:

$$2.0690 \pm (1.96)(0.4371) = 2.0690 \pm 0.8567.$$

The resulting confidence interval is $(1.2123, 2.9257)$. We can exponentiate the endpoints to obtain a 95% confidence interval for the odds ratio:

$$(e^{1.2123}, e^{2.9257}) = (3.361, 18.65)$$

The 95% confidence interval for α is

$$-2.4744 \pm (1.96)(0.3681) = -2.4744 \pm 0.7215 \qquad \text{or} \qquad (-3.196, -1.753).$$

Noting that $\pi(0) = P(\text{Yes}|\text{Normal}) = \frac{1}{1+e^{-\alpha}}$, we can obtain a 95% confidence interval for $\pi(0)$:

$$\left(\frac{1}{1+e^{3.196}}, \frac{1}{1+e^{1.753}}\right) = (0.0393, 0.148).$$

4.3.3 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:

- $x_1 = 1$ if Category 1, $= 0$, otherwise
- $x_2 = 1$ if Category 2, $= 0$, otherwise
- $\vdots$
- $x_{k-1} = 1$ if Category $k - 1$, $= 0$ otherwise

The resulting logistic regression model is

$$\text{logit}(\pi(\boldsymbol{x})) = \alpha + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}.$$

This parameterization treats Category k as a reference category.

Remark: One can also use use effects coding when there are k categories. For $i = 1, \ldots, k - 1$, we define

$$x_i = \begin{cases} 1 & \text{if Category } i \\ -1 & \text{if Category } k \\ 0 & \text{otherwise} \end{cases}$$

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

Category	Logit
1	$\alpha + \beta_1$
2	$\alpha + \beta_2$
$\vdots$	$\vdots$
$k-1$	$\alpha + \beta_{k-1}$
k	α

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

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

We can form Wald confidence intervals 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.

4.3.4 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

$$(X_i, y_i, z_i), i = 1, \ldots, n$$

- $Y_i = 1$ if yes, $= 0$ if no
- $x_i = 1$ if Group 1, $= 0$ if Group 0
- $z_i = 1$ if Layer 1, $= 0$ if Layer 0

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

Define the following probabilities:

$$\pi_{00} = P(Y = 1|X = 0, Z = 0)$$
$$\pi_{10} = P(Y = 1|X = 1, Z = 0)$$
$$\pi_{01} = P(Y = 1|X = 0, Z = 1)$$
$$\pi_{11} = P(Y = 1|X = 1, Z = 1)$$

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

$\text{logit}(\pi_{00})$	α
$\text{logit}(\pi_{10})$	$\alpha + \beta_1$
$\text{logit}(\pi_{01})$	$\alpha + \beta_2$
$\text{logit}(\pi_{11})$	$\alpha + \beta_1 + \beta_2$

$$\begin{array}{rcccccc}
\alpha & = & \log\left(\frac{\pi_{00}}{1-\pi_{00}}\right) & & \log \text{ odds of reference} & & \\
\beta_1 & = & \text{logit}(\pi_{10}) - \text{logit}(\pi_{00}) & = & \log\left(\frac{\text{odds}_{10}}{\text{odds}_{00}}\right) & = & \log\theta_{XY|Z=0} \\
 & = & \text{logit}(\pi_{11}) - \text{logit}(\pi_{01}) & = & \log\left(\frac{\text{odds}_{11}}{\text{odds}_{01}}\right) & = & \log\theta_{XY|Z=1} \\
\beta_2 & = & \text{logit}(\pi_{01}) - \text{logit}(\pi_{00}) & = & \log\left(\frac{\text{odds}_{01}}{\text{odds}_{00}}\right) & = & \log\theta_{ZY|X=0} \\
 & = & \text{logit}(\pi_{11}) - \text{logit}(\pi_{10}) & = & \log\left(\frac{\text{odds}_{11}}{\text{odds}_{10}}\right) & = & \log\theta_{ZY|X=1}
\end{array}$$

Notes:

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

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

$\text{logit}(\pi_{00})$	α
$\text{logit}(\pi_{10})$	$\alpha + \beta_1$
$\text{logit}(\pi_{01})$	$\alpha + \beta_2$
$\text{logit}(\pi_{11})$	$\alpha + \beta_1 + \beta_2 + \beta_3$

$$
\begin{array}{lcccl}
\alpha & = & \log\left(\frac{\pi_{00}}{1-\pi_{00}}\right) & = & \log \text{ odds of the reference} \\
\beta_1 & = & \text{logit}(\pi_{10}) - \text{logit}(\pi_{00}) & = & \log\left(\frac{\text{odds}_{10}}{\text{odds}_{00}}\right) = \log \theta_{XY|Z=0} \\
\beta_1 + \beta_3 & = & \text{logit}(\pi_{11}) - \text{logit}(\pi_{01}) & = & \log\left(\frac{\text{odds}_{11}}{\text{odds}_{01}}\right) = \log \theta_{XY|Z=1} \\
\beta_2 & = & \text{logit}(\pi_{01}) - \text{logit}(\pi_{00}) & = & \log\left(\frac{\text{odds}_{01}}{\text{odds}_{00}}\right) = \log \theta_{ZY|X=0} \\
\beta_2 + \beta_3 & = & \text{logit}(\pi_{11}) - \text{logit}(\pi_{10}) & = & \log\left(\frac{\text{odds}_{11}}{\text{odds}_{10}}\right) = \log \theta_{ZY|X=1}
\end{array}
$$

Note:

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

4.3.5 ANOVA-Type Representation of Factors

We have used $k-1$ dummy variables to model a factor with k levels in logistic regression. An alternative representation of factors in logistic regression resembles ANOVA models:

$$\text{logit}(\pi(x)) = \alpha + \beta_i^X + \beta_k^Z,\ i = i, \ldots, I,\ k = 1, \ldots, K.$$

The parameters $\{\beta_i^X\}$ and $\{\beta_k^Z\}$ represent the effects of X and Z. A test of conditional independence between X and Y conditional on Z corresponds to

$$H_0 : \beta_1^X = \beta_2^X = \cdots = \beta_I^X.$$

This parameterization includes one redundant parameter for each effect. There are several ways of defining parameters to account for the redundancies:

1. Set the last parameter $\hat{\beta}_I^X$ equal to zero.
2. Set the first parameter $\hat{\beta}_1^X$ equal to zero.
3. Set the sum of the parameters equal to zero (effects coding).

We note that each coding scheme:

- The differences $\beta_1^X - \beta_2^X$ and $\beta_1^Z - \beta_2^Z$ are the same.
- The different coding schemes yield the same odds ratios.
- The different coding schemes yield the same probabilities.

The following table gives the estimates corresponding to different coding schemes for the Coronary Artery Disease Data:

	Definition of Parameters		
Parameter	Last $= 0$	First $= 0$	Sum $= 0$
Intercept	1.157	-1.175	-0.0090
Gender=Female	-1.277	0.000	-0.6385
Gender=Male	0.000	1.277	0.6385
ECG=< 0.1	-1.055	0.000	-0.5272
ECG=≥ 0.1	0.000	1.055	0.5272

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.

$$x_2 = \begin{cases} 1 & \text{Lt. Med} \\ 0 & \text{otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{Med} \\ 0 & \text{otherwise} \end{cases} \qquad x_4 = \begin{cases} 1 & \text{DkMed} \\ 0 & \text{otherwise} \end{cases}$$

Consider the two main effect models:

- Ordinal Color (z):

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

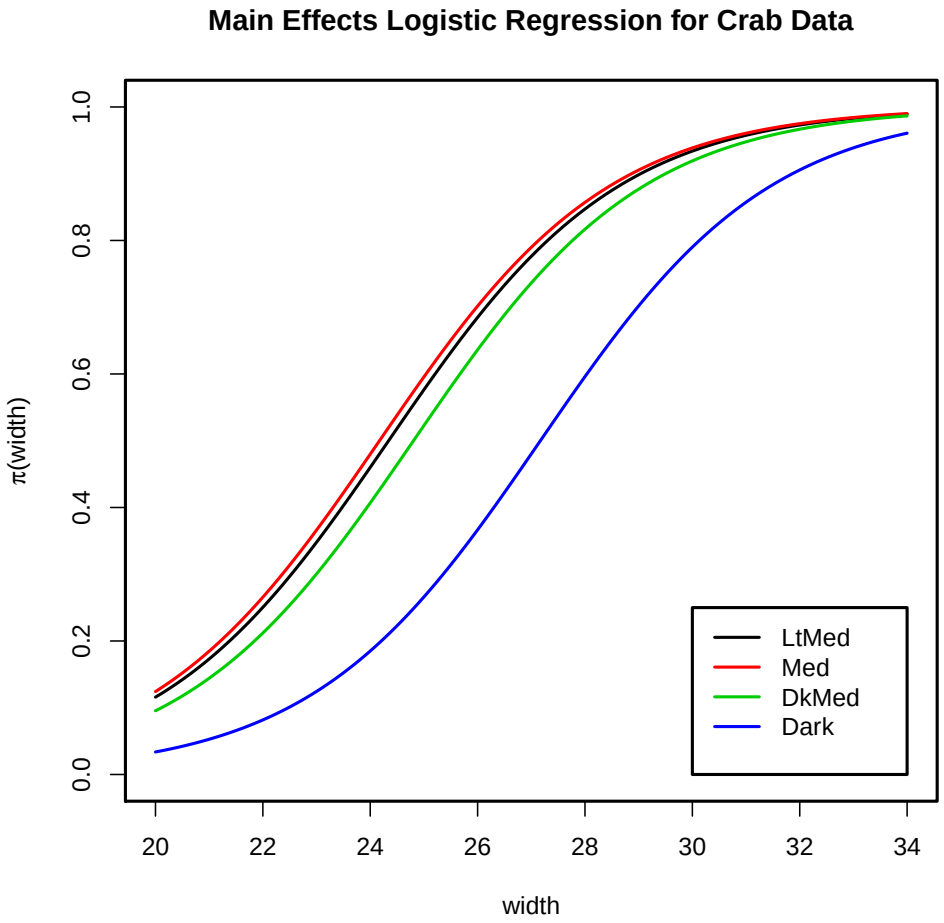
- Nominal color (x_2, x_3, x_4)

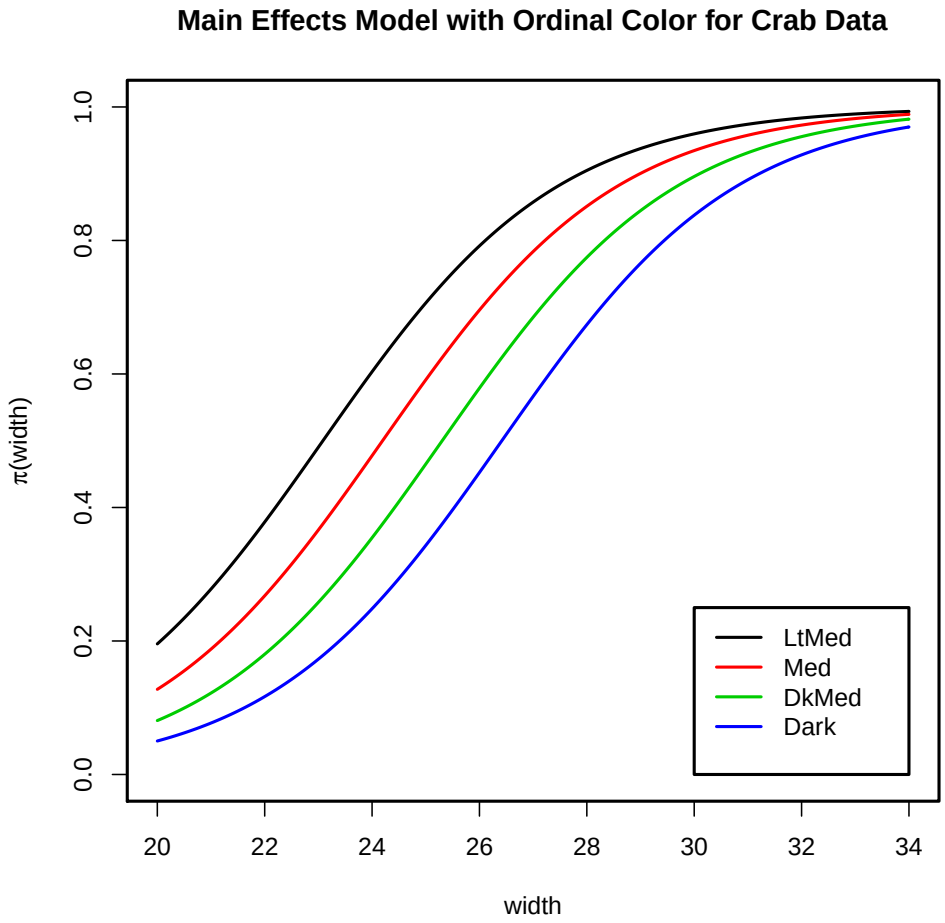
$$\text{logit}(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

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

Color	Ordinal	Nominal
Lt Med	$\alpha + \beta_1 x_1 + \beta_2^*$	$\alpha + \beta_1 x_1 + \beta_2$
Med	$\alpha + \beta_1 x_1 + 2\beta_2^*$	$\alpha + \beta_1 x_1 + \beta_3$
Dk Med	$\alpha + \beta_1 x_1 + 3\beta_2^*$	$\alpha + \beta_1 x_1 + \beta_4$
Dk	$\alpha + \beta_1 x_1 + 4\beta_2^*$	$\alpha + \beta_1 x_1$

- 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.





4.4 Models with Interaction or Confounding

In this section we will consider models where interaction or confounding is present.

- Multivariable logistic regression models enable us to adjust the model relating a response variable (CHD) to an explanatory variable (age) for the presence of other explanatory variables (high blood pressure).

- The variables do not *interact* if their effects on the logit are additive. This implies that the logits all have the same slope for different levels of the second explanatory variable.

- Epidemiologists use the term *confounder* to describe a covariate (Z) that is associated with both another explanatory variable (X) and the outcome variable (Y).

We now look at the situation where the possible confounder is qualitative and the explanatory variable is quantitative.

- Model without interaction:

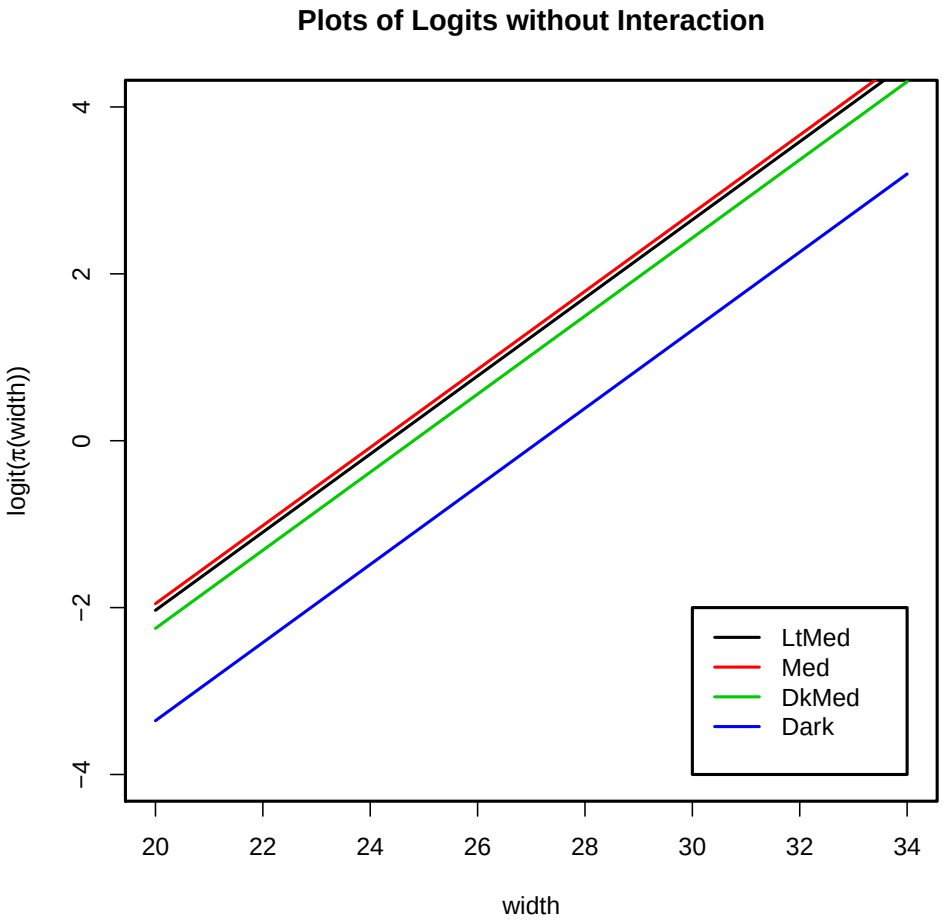
$$\text{logit}(\pi(x, z)) = \alpha + \beta_1 x + \beta_2 z$$

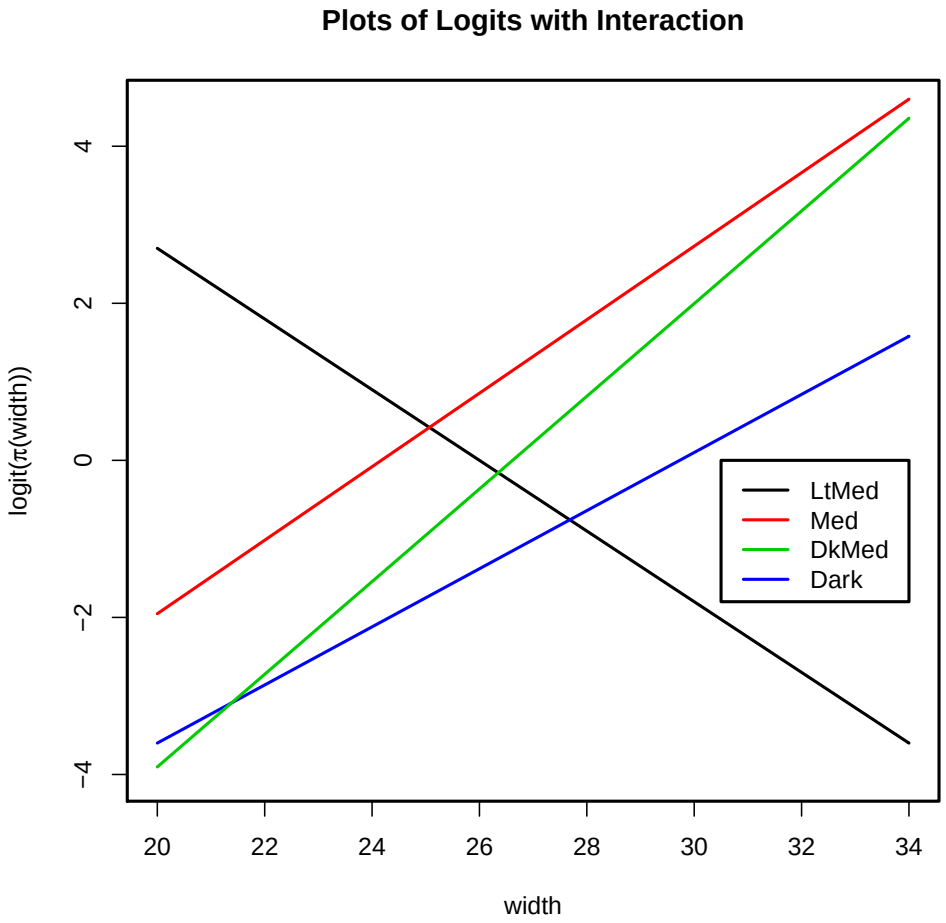
- For the model without interaction, the vertical distance between the logits represents the log odds ratio for comparing the two groups while controlling for the width. This distance is the same for all widths.

- Model with interaction:

$$\text{logit}(\pi(x, z)) = \alpha + \beta_1 x + \beta_2 z + \beta_3 (x \times z)$$

- When interaction is present, the distance between the logits depends on width. The log odds ratio between the groups now depends on the width.





4.4.1 Estimation of the Odds Ratio

When interaction is present, one cannot estimate the odds ratio comparing groups by simply exponentiating a coefficient since the OR depends on the value of the covariate. An approach that can be used is the following:

1. Write down the expressions for the logits at the two levels of the risk factor being compared.
2. Take the difference, simplify and compute.
3. Exponentiate the value found in step 2.

Let Z denote the risk factor, X denote the covariate, and $Z \times X$ their interaction. Suppose we want the OR at levels z_0 and z_1 of Z when $X = x$.

1. $g(x, z) = \text{logit}(\pi(x, z)) = \alpha + \beta_1 x + \beta_2 z + \beta_3 z * x.$
2. $$\begin{aligned} \log(OR) &= g(x, z_1) - g(x, z_0) \\ &= \alpha + \beta_1 x + \beta_2 z_1 + \beta_3 z_1 * x \\ &\quad -(\alpha + \beta_1 x + \beta_2 z_0 + \beta_3 z_0 * x) \\ &= \beta_2(z_1 - z_0) + \beta_3 x(z_1 - z_0) \end{aligned}$$
3. $OR = \exp\left[\beta_2(z_1 - z_0) + \beta_3 x(z_1 - z_0)\right]$