Chapter 4: Logistic Regression

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.

Interpreting the Logistic Regression Model:

We will look at the logistic regression model with one explanatory variable. The binary response variable is defined as

$$Y = \begin{cases} 1 & \text{when the actual response is "yes" or success} \\ 0 & \text{when the actual response is "no" or failure} \end{cases}.$$

We want to model

$$\pi(x) = \text{pr}(Y = 1|X = x)$$

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 $\pi(x)$:

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

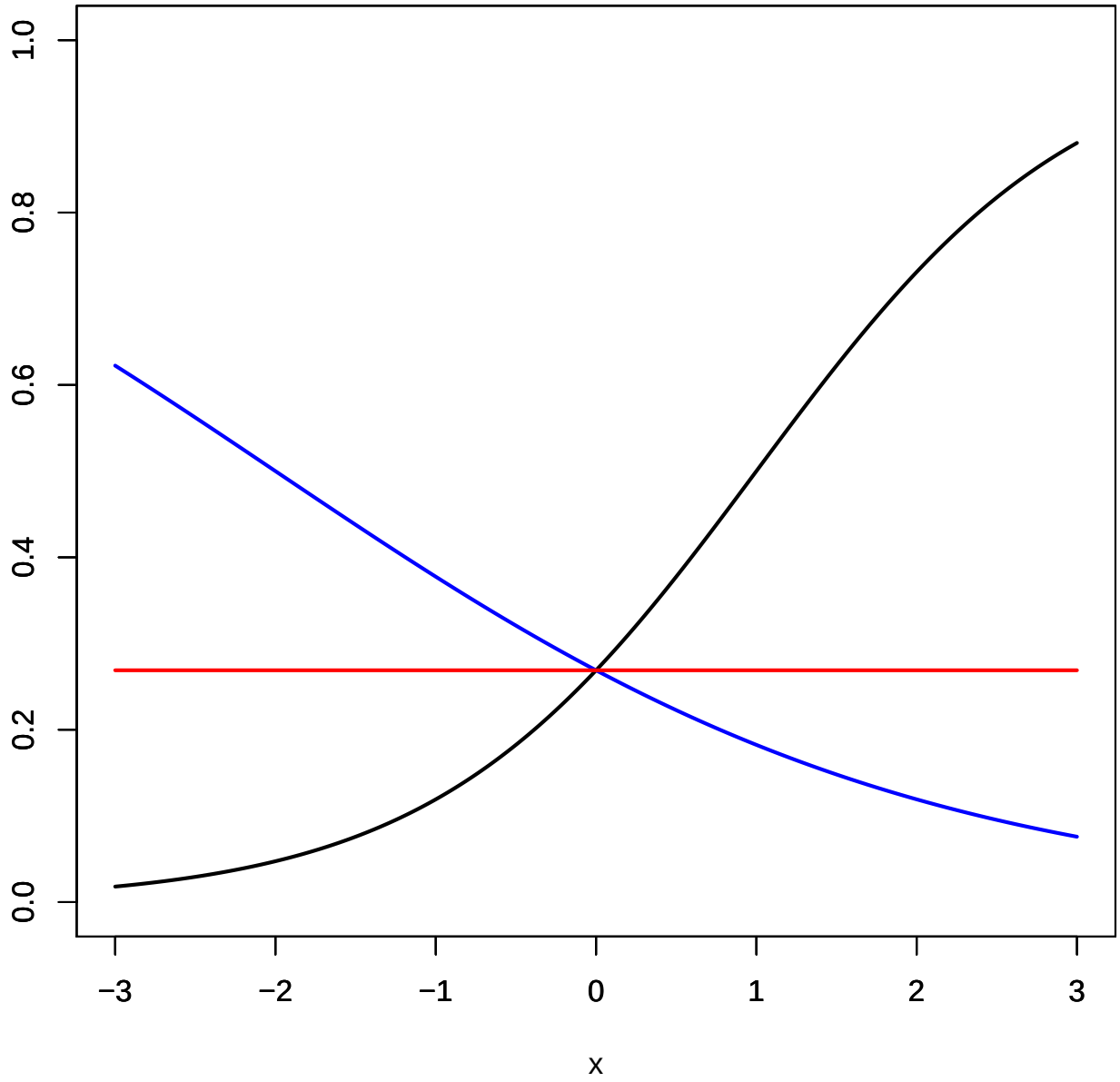
Note: The logistic model is a special case of the generalized linear model with the following components:

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

Fun with R

Code

```
x=seq(-3, 3, 0.001)
prob.pos= exp(-1+1*x)/(1+exp(-1+1*x))
prob.neg= exp(-1-0.5*x)/(1+exp(-1-0.5*x))
prob.0= exp(-1-0*x)/(1+exp(-1-0*x))

plot(x, prob.pos, type="l", lwd=2, ylim=c(0, 1), ylab="")
par(new=T)
plot(x, prob.neg, type="l", lwd=2, ylim=c(0, 1), col="blue",
ylab="", xlab="")
par(new=T)
plot(x, prob.0, type="l", lwd=2, ylim=c(0, 1), col="red",
 ylab="", xlab="")
```


1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
x

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

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

Beetle Mortality -modelling mortality in terms of dose

Code

```
n=c(59, 60, 62, 56, 63, 59, 62, 60)
obs.y=c(6, 13, 18, 28, 52, 53, 61, 59)
obs.x=c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
glm(obs.y/n~obs.x, family=binomial, weight=n)

Call:  glm(formula = obs.y/n ~ obs.x, family = binomial, weights = n)

Coefficients:
(Intercept)        obs.x
     -59.19        33.40

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      274.9
Residual Deviance: 8.64  AIC: 40.82
```

Code

```
glm(obs.y/n~obs.x, family=binomial(link=probit), weight=n)

Call:  glm(formula = obs.y/n ~ obs.x, family = binomial(link = probit),
    weights = n)

Coefficients:
(Intercept)        obs.x
     -33.92        19.15

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     274.9
Residual Deviance: 8.43  AIC: 40.61
```

Code

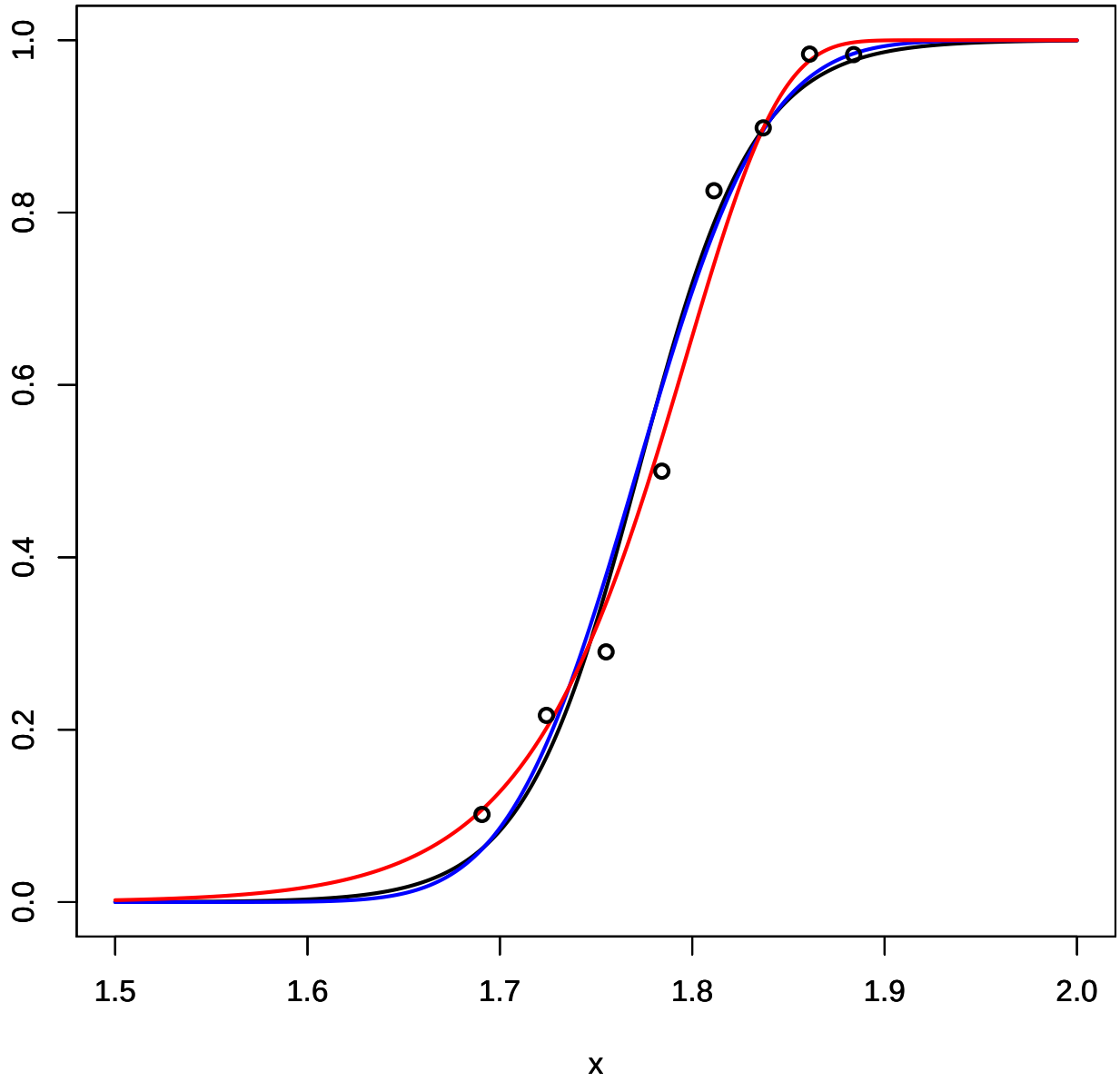
```
glm(obs.y/n~obs.x, family=binomial(link=cloglog), weight=n)

Call:  glm(formula = obs.y/n ~ obs.x, family = binomial(link = cloglog),
    weights = n)

Coefficients:
(Intercept)        obs.x
     -36.90        20.53

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      274.9
Residual Deviance: 6.185  AIC: 38.37
```

Beetle Mortality Data (results with different link functions)

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

- $\text{logit}(\hat{\pi}(x)) = -59.183 + 33.398x$ (black)
- $\text{probit}(\hat{\pi}(x)) = -33.917 + 19.148x$ (blue)
- $\text{cloglog}(\hat{\pi}(x)) = -36.895 + 20.534x$ (red)

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


1.0
0.8
0.6
0.4
0.2
0.0
1.5
1.6
1.7
1.8
1.9
2.0
x

Linear Approximation Interpretation

Suppose that the success probability of a binary outcome Y follows a logistic model in terms of the predictor X, that means

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

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

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

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

This result shows that the rate of change ($\pi'(x)$) is the same at two tails. For instance the rate of change ($\pi(x)$) is the same if $\pi(x)$ is 0.1 or 0.9, similarly it is the same for $\pi(x)$ 0.2 or 0.8, etc. This property is referred to as symmetry property of the link function.

Note that the marginal effect for a continuous predictor x is measured via $\pi'(x)$ (measures the rate of change of the mean response for a unit change in the predictor). So, from the above table we see that the marginal effect of the predictor is the same if the success probability is 0.2 or 0.8, 0.3 or 0.7, 0.1 or 0.9, etc.

- 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 by 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 corresponding slope or the marginal effect of dose when $\pi(x) = 0.5$ is $0.5 \times 0.5 \times 33.255 = 8.313$.
- 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.

Odds ratio interpretation

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

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 $x_1 = x$ and $x_2 = x + 1$, then $\theta_{21} = e^{\beta\{(x+1)-x\}} = e^{\beta}$.
For a unit increase in x, the odds is increased by a factor of e^{β}. If $\beta > 0$, $\exp(\beta) > 1$ otherwise $\exp(\beta) < 1$.

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. The code is given in the next slide.

Code

```
a=read.csv("agresti_crab.csv")
head(a)
  color spine width satell weight y dark
1     2     3  28.3      8   3.05 1    1
2     3     3  22.5      0   1.55 0    1
3     1     1  26.0      9   2.30 1    1
4     3     3  24.8      0   2.10 0    1
5     3     3  26.0      4   2.60 1    1
6     2     3  23.8      0   2.10 0    1
glm(y~width, family=binomial, data=a)
Call:  glm(formula = y ~ width, family = binomial, data = a)

Coefficients:
(Intercept)        width
   -12.3508       0.4972
Degrees of Freedom: 172 Total (i.e. Null);  171 Residual
Null Deviance:     225.8
Residual Deviance: 194.5  AIC: 198.5
```

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 a typical scenario in of some retrospective sampling designs such as case-control studies.

In a case-control study, a random sample of subjects is drawn who we know are controls $Y = 0$, and another random sample of subjects is drawn who we know are cases $Y = 1$. For each of the sampled subjects, we record the explanatory variable X. If the distribution of X differs between the cases and controls, there is evidence of an association between X and Y. When X is binary, we are 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 (association/regression parameter) of X using parameters that refer to odds and odds ratios. However, the estimated intercept term from case-control data is not useful to estimate α of (1) because the estimate is confounded with the relative number of $Y = 1$ and $Y = 0$ in the population..

Logistic Regression and the Normal Distribution

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

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

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

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

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 is a linear function of x (i.e., $\text{logit}\{\pi(x)\} = \beta_0 + \beta_1 x$)
- When $\sigma_1^2 \neq \sigma_0^2$, the log-odds is a quadratic function of x (i.e., $\text{logit}\{\pi(x)\} = \beta_0 + \beta_1 x + \beta_2 x^2$)

Multiple Logistic Regression

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

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

The logistic regression model is

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

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

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

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
- $x_i = (x_{i1}, \ldots, x_{ik}) =$ the values of the k explanatory variables

When there are n_i observations at a fixed x_i value, the number of successes Y_i forms a *sufficient statistic* and has a Binomial (n_i, π_i) distribution where

$$\pi_i = \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}.$$

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

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

The log-likelihood is

$$\begin{aligned}
\ell(\beta) = \log\{\mathcal{L}(\beta)\} &= \sum_{i=1}^{n} [\log \binom{n_i}{y_i} + y_i \log(\pi_i) + (n - y_i)\log(1-\pi_i)] \\
&= \sum_i y_i \beta_0 + \sum_{j=1}^{k} \left(\sum_i y_i x_{ij} \right) \beta_j \\
&\quad - \sum_i n_i \log \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right] + \sum_i \log \binom{n_i}{y_i}.
\end{aligned}$$

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

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

$$U_1(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_0} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} n_i \pi_i = 0$$

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

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

There are $k+1$ equations in $k+1$ unknowns. These equations are solved numerically by the iteratively weighted least square (IWLS) method .

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

$$\widehat{SE}(\hat{\beta}_j) = \sqrt{\widehat{\text{var}}(\hat{\beta}_j)}.$$

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

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_a : \text{At least one } \beta_j \neq 0, j = 1, \ldots, k$$

We typically use the likelihood ratio statistic:

$$G^2 = Q_L = -2 \log \left\{ \frac{\mathcal{L}(\tilde{\beta}_0)}{\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)} \right\} = 2(\ell_{Full} - \ell_{Reduced})$$

Here $\tilde{\beta}_0$ is the MLE of β_0 under the null hypothesis that the model with intercept only holds. When H_0 is true,

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

We reject H_0 for large values of Q_L.

Tests for a sets of coefficients

We can use methods for comparing nested models to test whether a set of coefficients all equal zero. We consider testing the significance of the $m(< k)$ regression coefficients in the logistic regression model. The hypotheses of interest are

$$H_0 : \beta_{i_1} = \cdots = \beta_{i_m} = 0 \quad \text{versus} \quad H_a : \text{At least one } \beta_{i_j} \neq 0$$

We typically use the likelihood ratio statistic:

$$G^2 = Q_L = 2(\ell_{Full} - \ell_{Reduced}) = \text{Deviance(Reduced)} - \text{Deviance(Full)}$$

where the full model contains all k predictors and the reduced model has $k - m$ predictors. When H_0 is true,

$$G^2 \xrightarrow{d} \chi^2_m \text{ as } n \to \infty.$$

We reject H_0 for large values of Q_L.

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 may use the LR statistic for testing these hypotheses.

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

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

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 x:

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

The estimated logit is given by

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

This has estimated asymptotic variance

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

In the above formula, $x_0 = 1$.

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

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

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

Since

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

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

$$\widehat{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sqrt{\widehat{\text{var}}(\hat{\beta}_0 + \hat{\beta}_1 x_0)} = \sqrt{\widehat{\text{var}}(\hat{\beta}_0) + x_0^2 \widehat{\text{var}}(\hat{\beta}_1) + 2x_0 \widehat{\text{cov}}(\hat{\beta}_0, \hat{\beta}_1)}$$

Since

$$\pi(x_0) = \text{pr}(Y = 1 | X = x_0) = \frac{e^{\beta_0 + \beta_1 x_0}}{1 + e^{\beta_0 + \beta_1 x_0}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_0)}}.$$

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

- 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 $\pi(x_0)$.

Examples

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	.9833

The fitted logit model using `proc logistic` is:

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

The observed proportions and the estimated proportion from the fitted model are

Dose	Obs	Est
1.6907	0.102	0.062
1.7242	0.217	0.168
1.7552	0.290	0.363
1.7842	0.500	0.600
1.8113	0.825	0.788
1.8369	0.898	0.897
1.8610	0.984	0.951
1.8839	0.983	0.977

Beetles mortality example

- Test for $H_0 : \beta_1 = 0$. The three tests strongly reject H_0.
- 95% confidence interval for β_1:

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

- Confidence interval for $\text{logit}(\pi(x_0))$ Let $x_0 = 1.8113$. The estimated logit is $-59.18 + 33.40 \times 1.8113 = 1.3111$.

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

The 95% confidence interval for the logit is

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

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.

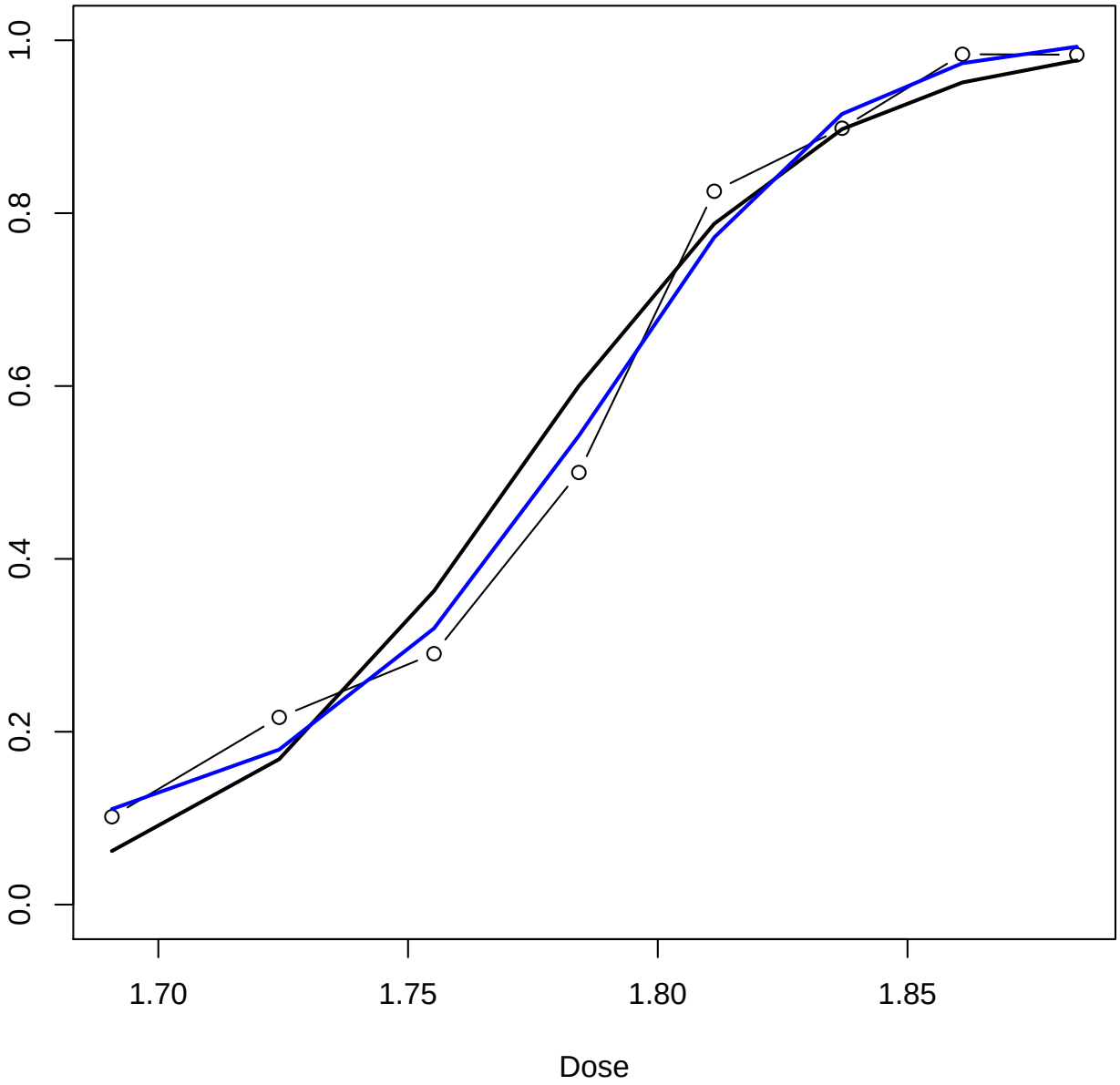
The following table presents the confidence intervals for $\pi(x_0)$ for all the observed values of the covariate:

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:

Code

```
n=c(59, 60, 62, 56, 63, 59, 62, 60)
obs.y=c(6, 13, 18, 28, 52, 53, 61, 59)
obs.x=c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
myx=obs.x-mean(obs.x)
myx2=myx*myx
out=glm(obs.y/n~obs.x, family=binomial(link=logit), weight=n)
out2=glm(obs.y/n~myx+myx2, family=binomial(link=logit), weight=n)

plot(obs.x, obs.y/n, type="b", ylim=c(0, 1), axes=F, ylab="", xlab="")
par(new=T)
plot(obs.x, fitted.values(out), type="l", ylim=c(0, 1), lwd=2, axes=F,
ylab="", xlab="")
par(new=T)
plot(obs.x, fitted.values(out2), type="l", ylim=c(0, 1), lwd=2,
 col="blue", ylab="", xlab="Dose")
```


1.0
0.8
0.6
0.4
0.2
0.0
1.70
1.75
1.80
1.85
Dose

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.

Dummy Variables in Logit Models: As in ordinary regression, dummy variables are used to incorporate qualitative variables into the model.

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 (Y)		
Group (X)	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

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

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 in 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):

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

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

- $\hat{\alpha}$ 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 $X = 0$

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

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

$$\text{var}(\hat{\alpha}) = \frac{n_{2+}}{n_{21}(n_{2+} - n_{21})}$$

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

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

where $\widehat{SE}(\hat{\beta}) = \sqrt{\widehat{\text{var}}(\hat{\beta})}$

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:

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

R code for the analysis of the cancer relapse data

Code

```
x=c(rep(1, 75), rep(0, 103));
y=c(rep(1, 30), rep(0, 45), rep(1, 8), rep(0, 95))
out=glm(y~x, family=binomial)
summary(out)
Call:
glm(formula = y ~ x, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0108  -0.8586  -0.4021  -0.4021   2.2607
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.4744     0.3681  -6.721 1.80e-11 ***
x             2.0690     0.4371   4.733 2.21e-06 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 184.6  on 177  degrees of freedom
Residual deviance: 157.2  on 176  degrees of freedom
AIC: 161.2
Number of Fisher Scoring iterations: 5
```

For this model $\exp(\beta)$ represents the odds ratio of relapse from traces of cancer to cancer free group. The estimate of this odds ratio is $\exp(2.069) = 7.92$. In other words, the odds of relapse in the traces of cancer group is 7.92 times that in the cancer free group.
On the other hand, odds ratio of relapse from the cancer free group to traces of cancer group is $\exp(-2.069) = 1/7.92 = 0.126$. That means, the odds of relapse in the cancer free group is approximately 13% that of the group with traces of cancer.

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 (which is of less important)

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

Noting that $\pi(0) = \text{pr}(\text{relapse}|\text{cancer free}) = \frac{1}{1+e^{-\alpha}}$, we can obtain its estimate

$$\frac{1}{1 + \exp\{-(-2.4744)\}} = 0.0776,$$

and a 95% confidence interval for $\pi(0)$:

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

Alternative coding: $X = 0$ for traces of cancer, $X = 1$ cancer free

Code

```
x=c(rep(0, 75), rep(1, 103));
y=c(rep(1, 30), rep(0, 45), rep(1, 8), rep(0, 95))
out=glm(y~x, family=binomial)
Call:
glm(formula = y ~ x, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0108  -0.8586  -0.4021  -0.4021   2.2607
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4055     0.2357  -1.720   0.0854 .
x            -2.0690     0.4371  -4.733 2.21e-06 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 184.6  on 177  degrees of freedom
Residual deviance: 157.2  on 176  degrees of freedom
AIC: 161.2
Number of Fisher Scoring iterations: 5
```

For this model $\exp(\beta)$ represents the odds ratio of relapse from cancer free group to the group with traces of cancer. The estimate of this odds ratio is $\exp(-2.069) = 0.126$. In other words, the odds of relapse in the cancer free group is approximately 13% that of the group with traces of cancer.
Let us estimate $\mathrm{pr}(\text{relapse}|\text{cancer free}) = 1/\{1 + \exp(-\alpha - \beta)\}$, and the estimate

$$\frac{1}{1 + \exp\{-(-0.4055 - 2.069)\}} = 0.0776.$$

Observe that the estimate of this proability is the same for both approaches. The bottom line is that, not the individual parameter, but the probability estimates will be identical for any definition of the dummy variable.

An Alternative Form of Coding

Another coding method is called *deviation from the mean coding or effects coding*. This method assigns -1 as the score to the first group and 1 to the second group. In this case, the log-odds-ratio for $Y = 1$ from the second to the first group 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

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

Analysis of the data with an alternative coding (effect coding)

Code

```
x.1=x
x.1[x==0]= -1
out.1=glm(y~x.1, family=binomial)
summary(out.1)
Call:
glm(formula = y ~ x.1, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0108  -0.8586  -0.4021  -0.4021   2.2607
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4400     0.2186  -6.588 4.45e-11 ***
x.1          -1.0345     0.2186  -4.733 2.21e-06 ***
---
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 184.6  on 177  degrees of freedom
Residual deviance: 157.2  on 176  degrees of freedom
```

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

Note that when the independent variable takes on category k, we set $x_1 = \cdots = x_{k-1} = 0$.
The resulting logistic regression model is

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

This parameterization 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	$\alpha + \beta_1$
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.
The odds ratio for comparing category j to j' is

$$OR = e^{\beta_j - \beta_{j'}}.$$

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.

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:

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

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

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

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

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. In the main effects model, $H_0 : \beta_1 = 0$ is equivalent to $H_0 : X$ and Y are conditionally independent controlling for Z.

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

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

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

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

- 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 $H_0 : \beta_3 = 0$. That means we check if X-Y association is homogeneous across the levels of Z, or Z-Y association is homogeneous across the levels of X.

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:

- Set the last parameter $\hat{\beta}_I^X$ equal to zero.
- Set the first parameter $\hat{\beta}_1^X$ equal to zero.
- 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 parameter 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=$\geq$ 0.1	0.000	1.055	0.5272

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.

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

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.

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.

Estimation of the Odds Ratio

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

- Write down the expressions for the logits at the two levels of the risk factor being compared.
- Take the difference, simplify and compute.
- 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 $X = x$.

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

More generally, if we wish to compute an odds ratio comparing two settings of the predictors, (x_1, z_1) and (x_0, z_0), we can apply a similar approach to that on the previous slide.

- Write down the expressions for the logits at the two settings of the predictors being compared..
- Take the difference, simplify and compute.
- Exponentiate the value found in step 2.

Let X and Z denote the two predictors and $X \times Z$ their interaction. Suppose we want the OR at levels (x_1, z_1) and (x_0, z_0).

- $g(x, z) = \text{logit}(\pi(x, z)) = \alpha + \beta_1 x + \beta_2 z + \beta_3 x * z.$
- $$\begin{aligned} \log(OR) &= g(x_1, z_1) - g(x_0, z_0) \\ &= \alpha + \beta_1 x_1 + \beta_2 z_1 + \beta_3 x_1 z_1 \\ &\quad -(\alpha + \beta_1 x_0 + \beta_2 z_0 + \beta_3 z_0 * x_0) \\ &= \beta_1(x_1 - x_0) + \beta_2(z_1 - z_0) + \beta_3(x_1 z_1 - x_0 z_0) \end{aligned}$$
- $OR = \exp\left[\beta_1(x_1 - x_0) + \beta_2(z_1 - z_0) + \beta_3(x_1 z_1 - x_0 z_0)\right]$

Summarizing Predictive Power: Classification Tables

A common use for binary regression is classification. One can use a cut-off π_0 as a classification criterion:

- If $\hat{\pi} > \pi_0$, predict $\hat{y} = 1$.
- If $\hat{\pi} \leq \pi_0$, predict $\hat{y} = 0$.

We then form a 2×2 *classification table* to summarize the predictive power of the logistic regression model.

Example: We form the classification table for crab data using the logistic regression model with predictors `width, dark` using cut-off values, $\pi_0 = 0.50$ and $\pi_0 = 0.642$ where $0.642 = 111/173$ is the sample proportion of crabs with satellites.

	Prediction, $\pi_0 = 0.64$		Prediction, $\pi_0 = 0.50$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	78	33	97	14	111
$y = 0$	22	40	34	28	62

We can use the classification table to estimate the sensitivity and specificity of the model:

$$\text{sensitivity} = \text{pr}(\hat{y} = 1 | y = 1), \qquad \text{specificity} = \text{pr}(\hat{y} = 0 | y = 0)$$

Another commonly reported measure is the proportion of correct classifications. This estimates

$$\text{pr}(\text{correct classification}) = \text{pr}(y = 1, \hat{y} = 1) + \text{pr}(y = 0, \hat{y} = 0).$$

Estimates corresponding to the two cut-off values are

π_0	Sensitivity	Specificity	Proportion of Correct Classifications
0.50	$\frac{97}{111} = 0.874$	$\frac{28}{62} = 0.452$	$\frac{97+28}{173} = 0.723$
0.642	$\frac{78}{111} = 0.703$	$\frac{40}{62} = 0.645$	$\frac{78+40}{173} = 0.682$

Remark: The classification table depends on the value of the cut-off. If one makes the cut-off larger, the sensitivity will decrease and the specificity will increase. Also, the results will be sensitive to the relative numbers of times that $y = 1$ and $y = 0$.

Summarizing Predictive Power: ROC Curves

A *receiver operating characteristic* (ROC) curve is a plot of sensitivity as a function of (1 − specificity). Thus, the ROC curve summarizes predictive power for all values of π_0. For a given specificity, better predictive power corresponds to higher sensitivity. A higher ROC curve indicates better predictive power. The area under the ROC curve is used as a measure of predictive ability and is called the *concordance index*.

Example: For the crab data, ROC curves are plotted for the logistic regression models with `width` and `dark`.

Analysis of the crab data

Code

```
library(MASS)
library(e1071)
library(caret)
a=read.csv("agresti_crab.csv")
outnew=glm(y~width+dark, family="binomial", data=a)
out2new=predict(outnew)
confusionMatrix(data=as.factor(as.numeric(out2new>0.6)),
reference=as.factor(a$y))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 43 36
         1 19 75
```

Analysis of the crab data

Code

```
               Accuracy : 0.6821
                 95% CI : (0.6071, 0.7507)
    No Information Rate : 0.6416
    P-Value [Acc > NIR] : 0.15123

                  Kappa : 0.3482

 Mcnemar's Test P-Value : 0.03097

            Sensitivity : 0.6935
            Specificity : 0.6757
         Pos Pred Value : 0.5443
         Neg Pred Value : 0.7979
             Prevalence : 0.3584
         Detection Rate : 0.2486
   Detection Prevalence : 0.4566
      Balanced Accuracy : 0.6846

       'Positive' Class : 0
```

Predictive models

- The performance of a predictive model judged via sensitivity, specificity, accuracy.
- Accuracy = pr(observed response = predicted response) and it is estimated by the

$$\frac{\text{the total of the diagonal entries}}{\text{the total of the confusion matrix}}$$

- Accuracy can be seen as

$$\begin{aligned} \text{accuracy} &= 1 - \text{misclassification probability} \\ &= 1 - \text{pr}(\text{observed response} \neq \text{predicted response}). \end{aligned}$$

Predictive models

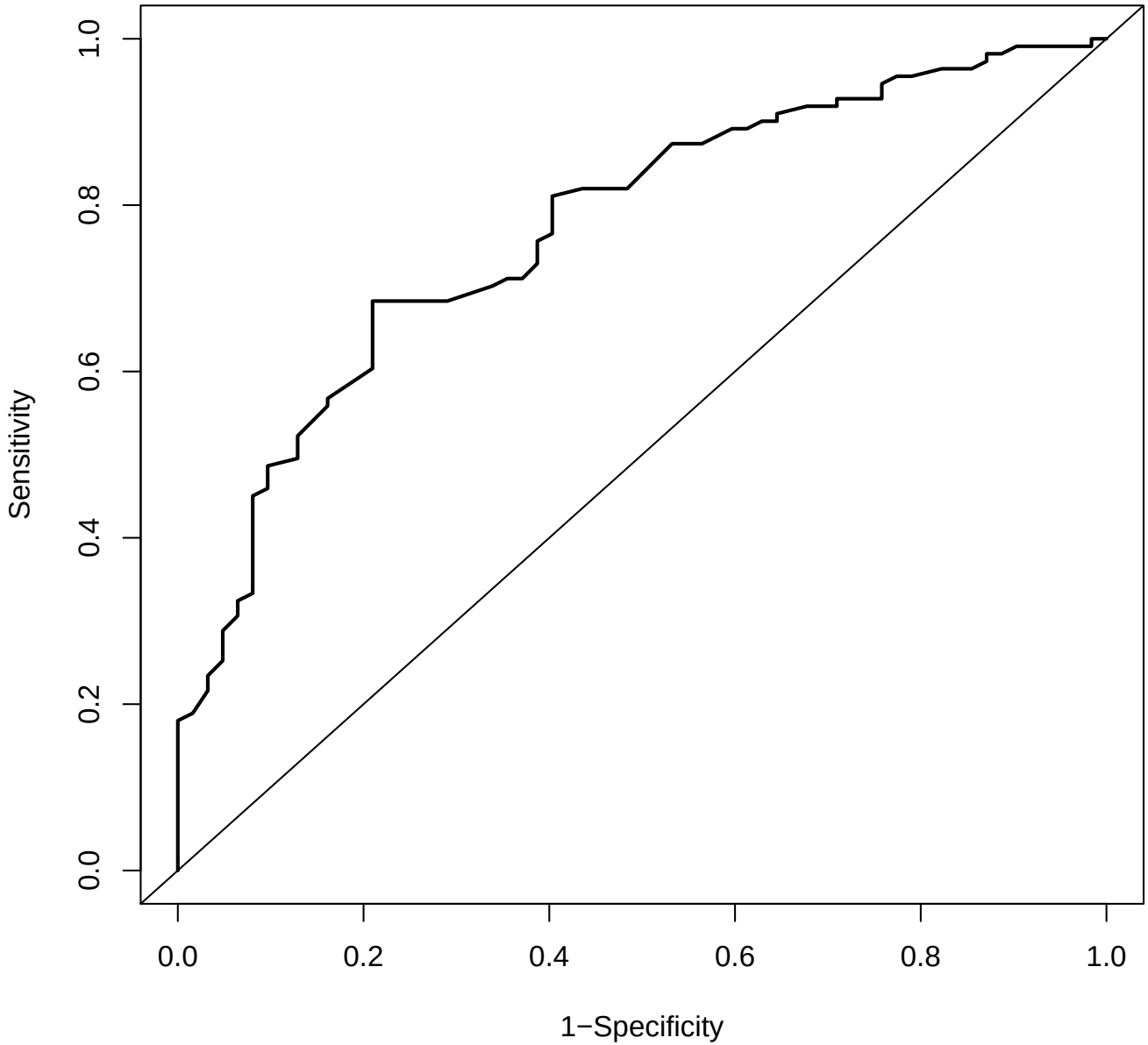
- $$\begin{aligned} & \text{Positive predictive value} \\ = \; & \text{pr}(\text{observed response} = \text{yes} | \text{predicted response} = \text{yes}) \\ = \; & \frac{TP}{TP + FP} \end{aligned}$$

- $$\begin{aligned} & \text{Negative predictive value} \\ = \; & \text{pr}(\text{observed response} = \text{no} | \text{predicted response} = \text{no}) \\ = \; & \frac{TN}{TN + FN} \end{aligned}$$

Analysis of the crab data

Code

```
library(pROC)
outnew=glm(y~width+dark, family="binomial", data=a)
out2new=predict(outnew, type="response")
roccurve= roc(a$y~out2new)
roccurve$auc
Area under the curve: 0.772
plot(1-roccurve$specificities, roccurve$sensitivities, type="l", lwd=2,
xlab="1-Specificity", ylab="Sensitivity")
abline(a=0, b=1)
```


Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Analysis of the crab data

Based on the maximum of (sensitivity+specificity) you may find the optimal threshold to define if a subject is 1 or 0.

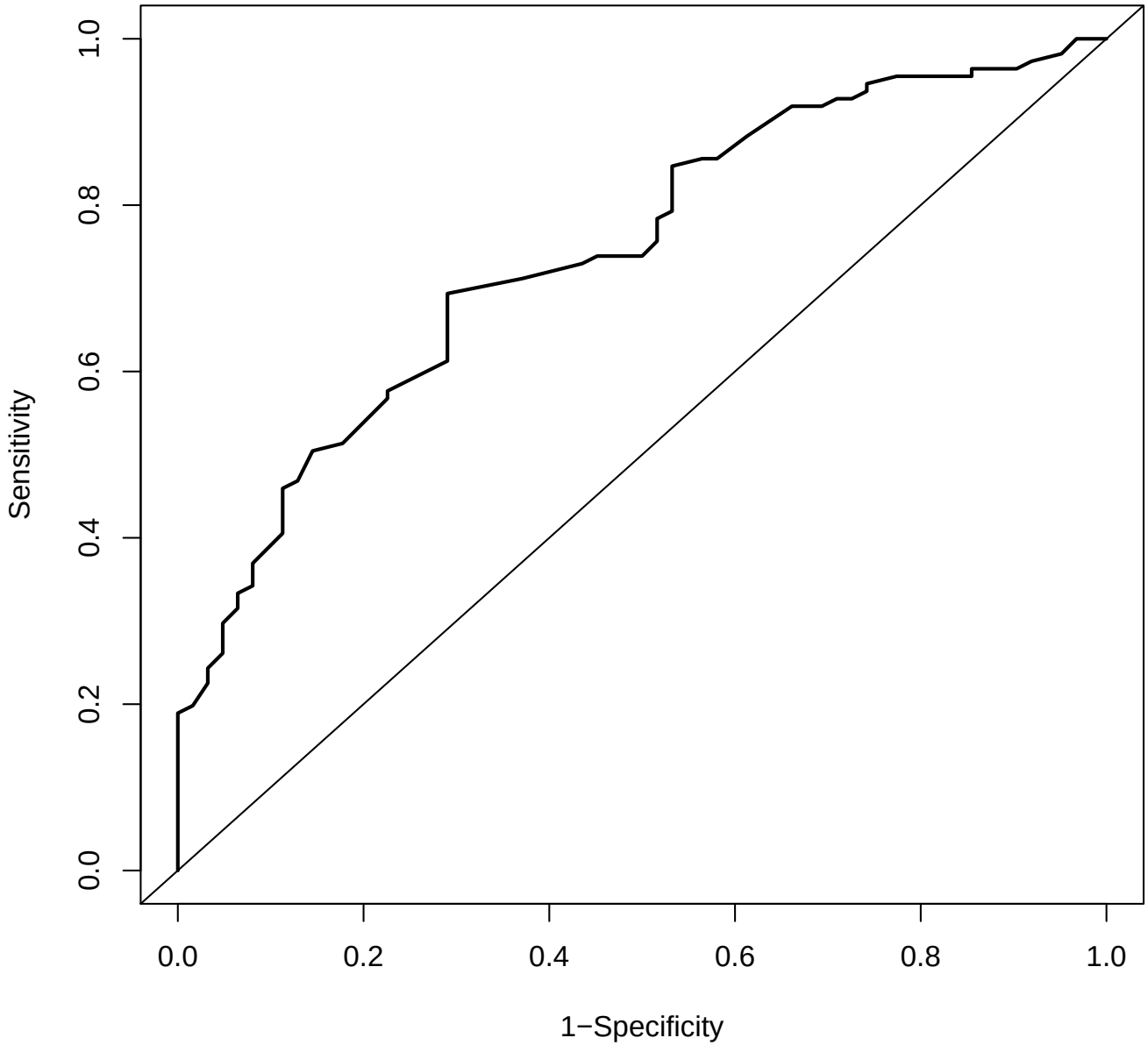
Code

```
total.sen.spe= roccurve$specificities+roccurve$sensitivities
 roccurve$thresholds[total.sen.spe==max(total.sen.spe)]
[1] 0.6643917
# That means if the predicted value is 0.6643917 or more that
will predicted to be Y=1 otherwise 0.
```

Analysis of the crab data

Code

```
library(pROC)
outw=glm(y~width, family="binomial", data=a)
out2=predict(outw, type="response")
roccurve= roc(a$y~out2)
roccurve$auc
Area under the curve: 0.7424
plot(1-roccurve$specificities, roccurve$sensitivities, type="l", lwd=2,
xlab="1-Specificity", ylab="Sensitivity")
abline(a=0, b=1)
```


Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

In this example, the area under the ROC curve is 74%. Mathematically, it is an estimate of $\text{pr}(\widehat{\pi}_i > \widehat{\pi}_j | Y_i = 1 \text{ and } Y_j = 0)$, where $\widehat{\pi}_i$ denotes the estimated success probability for the ith observation. Thus, it is the probability that the estimated success probability for an observed success is larger than that for an observed failure. This is also known as the c-statistic, concordance statistic. This tells us the classification power of the model where a higher value indicates a better classification (prediction) capability. The area under the ROC curve 0.5 indicates the model does not have a better prediction than mere guessing.
The area under the ROC (concordance index) can be estimated by

$$\frac{\sum_{(i,j):Y_i=1,Y_j=0} I(\widehat{\pi}_i > \widehat{\pi}_j)}{\#\{(i,j) : Y_i = 1, Y_j = 0\}}$$