Modeling binary data

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

11 May 2020

Goals for today

- Understand the characteristics of binary data and the Bernoulli distribution
- Understand how to model binary data with logistic regression
- Understand approaches to inference in logistic regression
- Understand diagnostic measures for logistic regression

Bernoulli distribution

The Bernoulli distribution describes the probability of a single “event” y_i occurring

- present (1/1) or absent (0/1)
- alive (1/1) or dead (0/1)
- mature (1/1) or immature (0/1)

Binomial distribution

The binomial distribution is closely related to the Bernoulli

It describes the number of k “successes” in a sequence of n independent Bernoulli “trials”

For example, the number of heads in 4 coin tosses

Binomial distribution

For a population, these could be

- k survivors out of n tagged individuals
- k infected individuals out of n susceptible individuals
- k counts of allele A in n total chromosomes

Binomial distribution

The probability mass function

$$\Pr(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Bernoulli distribution

Special case of binomial with $n = 1$

$$\Pr(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\Downarrow$$

$$\Pr(k; p) = p^k (1-p)^{(1-k)}$$

$$\Downarrow$$

$$k = \begin{cases} 1 \text{ if success (T, Y) with probability } p \\ 0 \text{ if failure (F, N) with probability } (1-p) \end{cases}$$

Bernoulli distribution

$$\Pr(k;p) = p^k(1-p)^{(1-k)}$$

$$\Downarrow$$

$$k = \begin{cases} 1 \text{ if success (T, Y) with probability } p \\ 0 \text{ if failure (F, N) with probability } (1-p) \end{cases}$$

where

$$\text{Mean}(k) = p \qquad \text{Var}(k) = p(1-p)$$

Bernoulli distribution

Likelihood

$$\mathcal{L}(k;p) = \prod_{i=1}^{n} p^{k_i}(1-p)^{(1-k_i)}$$

$$\Downarrow$$

$$\log \mathcal{L}(k;p) = \log p \sum_{i=1}^{n} k_i + \log(1-p) \sum_{i=1}^{n} (1-k_i)$$

Bernoulli distribution

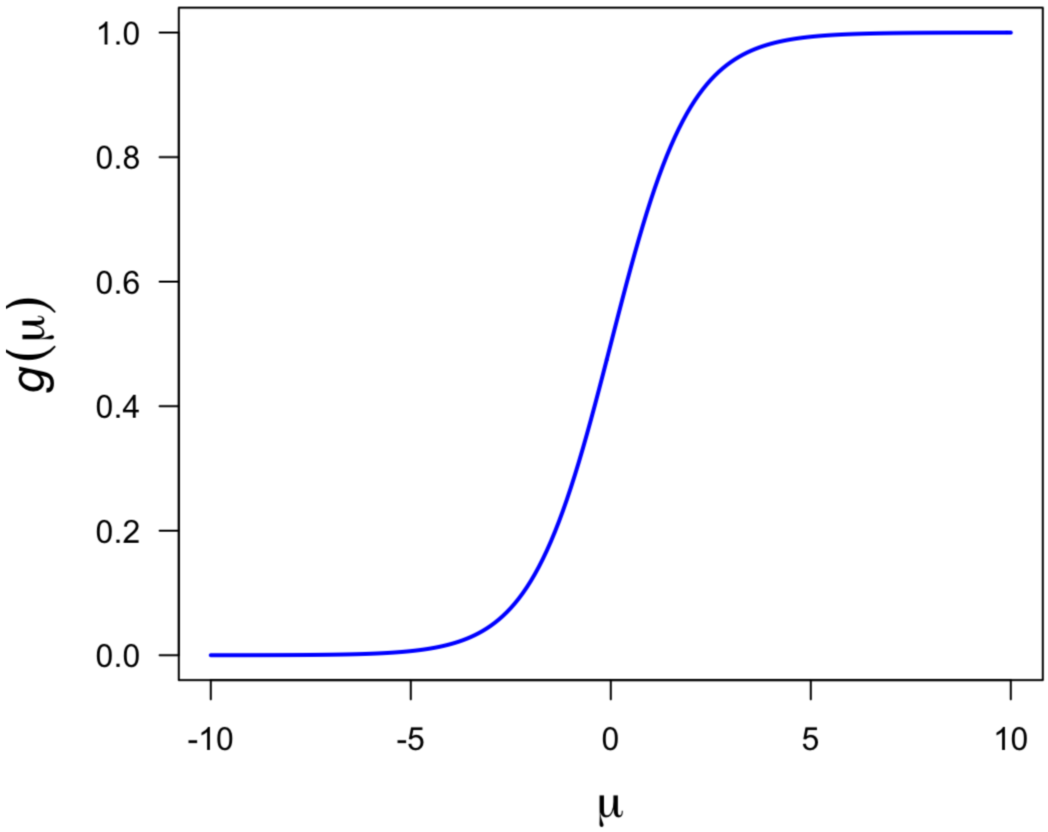
Canonical link is the logit

$$\log\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\boldsymbol{\beta}$$

$$\Downarrow$$

$$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$$

Logit link



Logistic regression

Similar to other regression in that we assume

- the predictors are linear
- the observations are independent of one another
- no(ish) multicollinearity among predictors

Logistic regression

Different from other regression in that

- the response is binary
- the relationship between response and predictors is often non-linear
- the errors can be non-normal
- the errors can be heteroscedastic

Logistic regression is a GLM

We need 3 things to specify our GLM

1. Distribution of the data: $y \sim \text{Bernoulli}(p)$
2. Link function: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta$
3. Linear predictor: $\eta = \mathbf{X}\boldsymbol{\beta}$

Logistic regression

The probability of a success is given by

$$p = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}\boldsymbol{\beta})}$$

$$= \frac{1}{1 + \exp(\text{-}\mathbf{X}\boldsymbol{\beta})}$$

Logistic regression

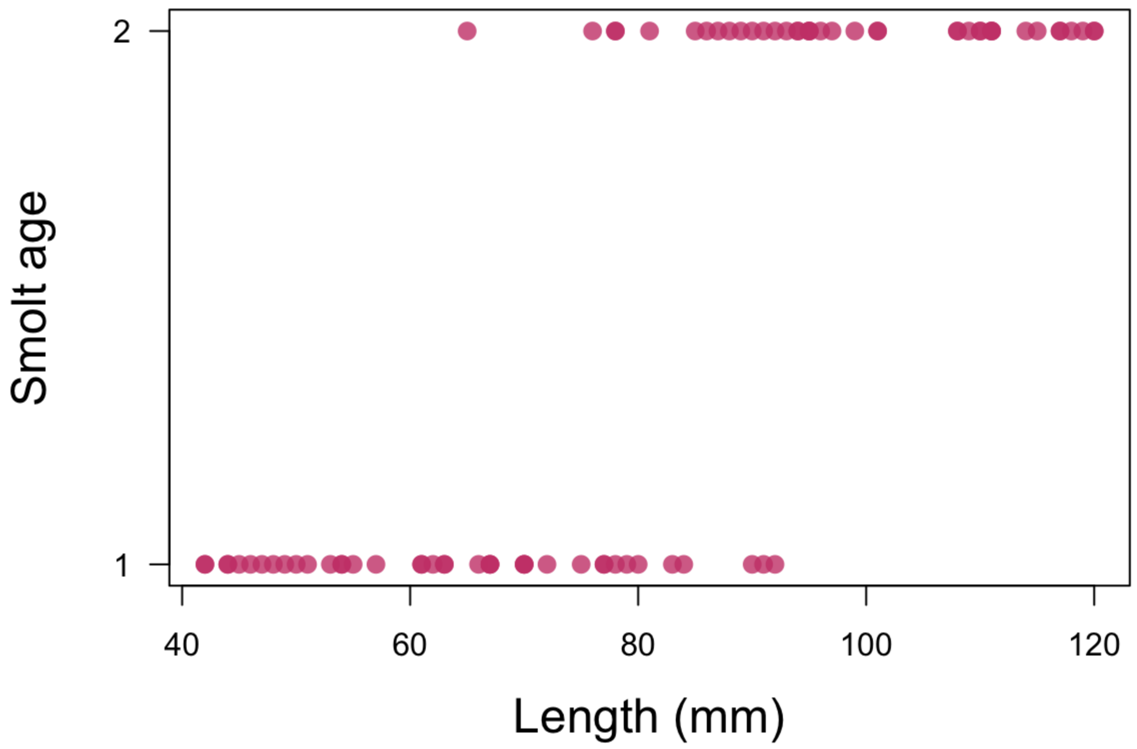
Example

Sockeye salmon are born in freshwater and rear there for some time before migrating to the ocean as *smolts*

The age at which sockeye smolt can vary from 1 to 2 years, which is thought to depend on their body size

Let’s examine the relationship between fish length and its probability of smolting at age-2 instead of age-1

Smolt age versus length



Smolt age versus length

In **R** we use `glm()` to fit logistic regression models (and other GLMs)

```
## fit model with glm
fit_mod <- glm(age ~ length, data = df,
               family = binomial(link = "logit"))
faraway::sumary(fit_mod)
```

```
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -13.982707   3.308236 -4.2266 2.372e-05
## length        0.170646   0.039786  4.2891 1.794e-05
##
## n = 80 p = 2
## Deviance = 42.05294 Null Deviance = 110.90355 (Difference = 68.85061)
```

Smolt age versus length

The probability of smolting at age-2 is given by

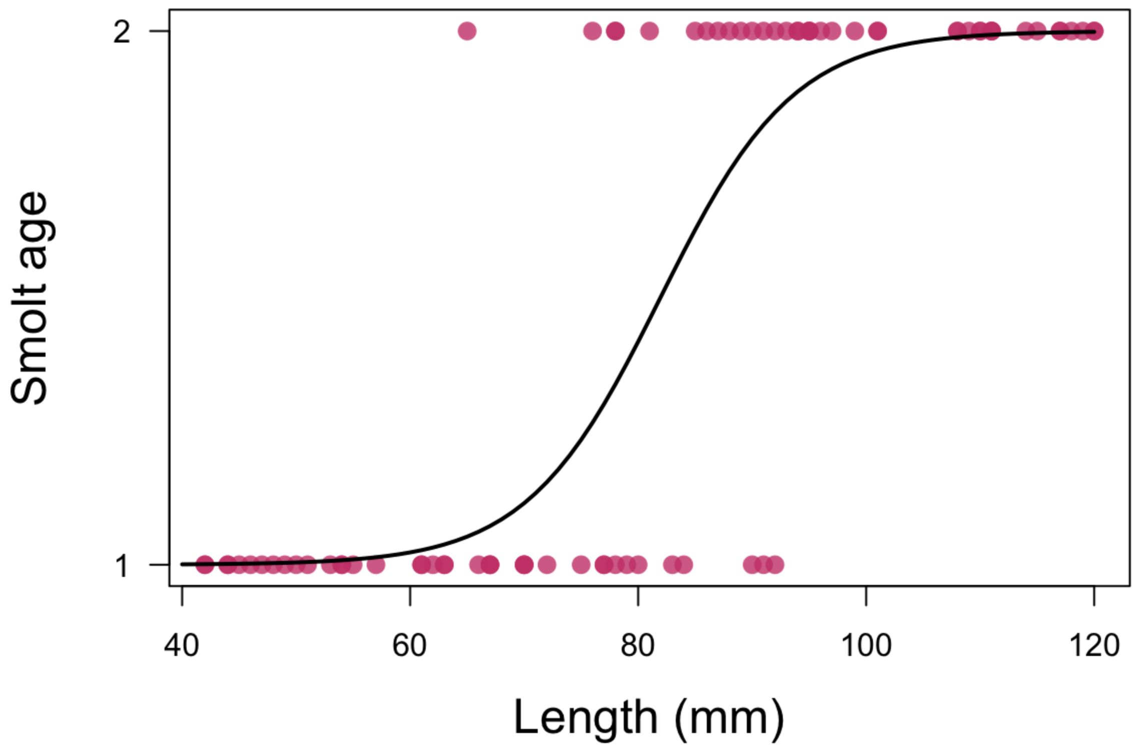
$$\begin{aligned} p_i &= \frac{1}{1 + \exp(\text{-}\mathbf{X}_i \boldsymbol{\beta})} \\ &\approx \frac{1}{1 + \exp(14 - 0.17 L_i)} \end{aligned}$$

Smolt age versus length

We can get the fitted values with `predict()`

```r
## get fitted values
newdata <- data.frame(length = seq(40, 120))
p_hat <- 1 / (1 + exp(-predict(fit_mod, newdata)))
```

Smolt age versus length



Smolt age versus length

What is the length at which the probability of smolting at age-2 is 0.5?

$$p_i = \frac{1}{1 + \exp(\text{-}\mathbf{X}_i \boldsymbol{\beta})}$$

$$\Downarrow$$

$$0.5 = \frac{1}{1 + \exp(14 - 0.17 L_{0.5})}$$

$$\Downarrow$$

$$L_{0.5} \approx 82 \text{ mm}$$

Logistic regression and odds

We have talked a bit about odds with respect to evidence ratios

Odds o are an unbounded alternative to probability p

If we represent the k-to-1 odds against something as $1/k$, then the following holds

$$o = \frac{1}{1-p} \;\Rightarrow\; p = \frac{o}{1+o}$$

For example, if p = 0.8, then $o = \frac{1}{1-0.8} = 5$

Logistic regression and odds

$$\mathrm{logit}(p) = \mathbf{X}\boldsymbol{\beta}$$

$$\Downarrow$$

$$\log\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta}$$

$$\Downarrow$$

$$\log(\mathrm{odds}) = \mathbf{X}\boldsymbol{\beta}$$

$$\Downarrow$$

$$\mathrm{odds} = \exp(\mathbf{X}\boldsymbol{\beta})$$

Smolt age versus length

```
## our fitted model
faraway::sumary(fit_mod)
```

```
##                Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -13.982707   3.308236 -4.2266 2.372e-05
## length        0.170646   0.039786  4.2891 1.794e-05
##
## n = 80 p = 2
## Deviance = 42.05294 Null Deviance = 110.90355 (Difference = 68.85061)
```

Smolt age versus length

$$\log\left(\frac{p}{1-p}\right) = \text{-}14 + 0.17L$$

$$\Downarrow$$

$$\log(\text{odds}) = \text{-}14 + 0.17L$$

A unit increase in L increases the log-odds by 0.17

Smolt age versus length

$$\log\left(\frac{p}{1-p}\right) = \text{-}14 + 0.17L$$

$$\Downarrow$$

$$\log(\text{odds}) = \text{-}14 + 0.17L$$

$$\Downarrow$$

$$\text{odds} = \exp(\text{-}14 + 0.17L)$$

A unit increase in L increases odds by exp(0.17) $\approx$ 1.19 = 19%

QUESTIONS?

Inference

Consider 2 models, A & B, such that B is a subset of A

A = $f(x_1, x_2)$

B = $g(x_1)$

We have seen that we can compare A & B via a likelihood ratio test

$$\lambda = \text{-}2 \log \frac{\mathcal{L}_A}{\mathcal{L}_B} \sim \chi^2_{df = k_A - k_B}$$

Log-likelihood

The log-likelihood using a logit link is

$$\log \mathcal{L}(k; p) = \log p \sum_{i=1}^{n} k_i + \log(1 - p) \sum_{i=1}^{n} (1 - k_i)$$

Deviance

Deviance D is a goodness-of-fit statistic

It's a generalization of using the sum-of-squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood

$$D = \text{-}2 \log \mathcal{L}$$

Deviance for logistic regression

$$
\begin{aligned}
D &= \text{-}2\left[\log p\sum_{i=1}^{n} k_i + \log(1-p)\sum_{i=1}^{n}(1-k_i)\right] \\
&= \text{-}2\sum_{i=1}^{n}\left[p_i \operatorname{logit}(p_i) + \log(1-p_i)\right]
\end{aligned}
$$

Likelihood ratio test

$$\lambda = \text{-}2 \log \frac{\mathcal{L}_A}{\mathcal{L}_B} \sim \chi^2_{df=k_A-k_B}$$

$$\Downarrow$$

$$\lambda = \text{-}2(\log \mathcal{L}_A - \log \mathcal{L}_B) \sim \chi^2_{df=k_A-k_B}$$

$$\Downarrow$$

$$\lambda = D(B) - D(A) \sim \chi^2_{df=k_A-k_B}$$

Smolt age versus length

The output from `glm()` includes the deviances for the full model and a null model with no predictors

```
## our fitted model
faraway::sumary(fit_mod)
```

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -13.982707   3.308236 -4.2266 2.372e-05
## length        0.170646   0.039786  4.2891 1.794e-05
##
## n = 80 p = 2
## Deviance = 42.05294 Null Deviance = 110.90355 (Difference = 68.85061)
```

Smolt age versus length

Likelihood ratio test for $H_0 : \beta_1 = 0$

```r
## deviance of full model
D_full <- summary(fit_mod)$deviance
## deviance of null model
D_null <- summary(fit_mod)$null.deviance
## test statistic
lambda <- D_null - D_full
## LRT with df = 1
(p_value <- pchisq(lambda, 1, lower.tail = FALSE))
```

```
## [1] 1.062116e-16
```

Model selection via AIC

$$\begin{aligned} AIC &= 2k - 2\log\mathcal{L} \\ &= 2k + D \end{aligned}$$

```r
## AIC
AIC(fit_mod)
## AIC via likelihood
(2 * 2) - 2 * logLik(fit_mod)
## AIC via deviance
(2 * 2) + summary(fit_mod)$deviance
```

```
## [1] 46.05294
## 'log Lik.' 46.05294 (df=2)
## [1] 46.05294
```

Smolt age versus length

Compare to a null model with no predictors

```
## fit null model
fit_null <- glm(age ~ 1, data = df,
                family = binomial(link = "logit"))
faraway::sumary(fit_null)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.00000    0.22361       0        1
##
## n = 80 p = 1
## Deviance = 110.90355 Null Deviance = 110.90355 (Difference = 0.00000)
```

Model selection via AIC

```
## difference in AIC
AIC(fit_null) - AIC(fit_mod)
```

```
## [1] 66.85061
```

Significance test for β_i

An alternative to the χ^2 test is a z test

$$z = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)} \sim z_{\alpha/2}$$

Significance test for β_i

```
## summary table
faraway::sumary(fit_mod)
```

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -13.982707   3.308236 -4.2266 2.372e-05
## length        0.170646   0.039786  4.2891 1.794e-05
##
## n = 80 p = 2
## Deviance = 42.05294 Null Deviance = 110.90355 (Difference = 68.85061)
```

Confidence interval for β_i

We can also compute a 100(1 - α)% confidence interval

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

Confidence interval for β_i

```
## beta
beta_1 <- coef(fit_mod)[2]
## SE of beta
se_beta_1 <- sqrt(diag(vcov(fit_mod)))[2]
## 95% CI
beta_1 + c(-1,1) * 1.96 * se_beta_1
```

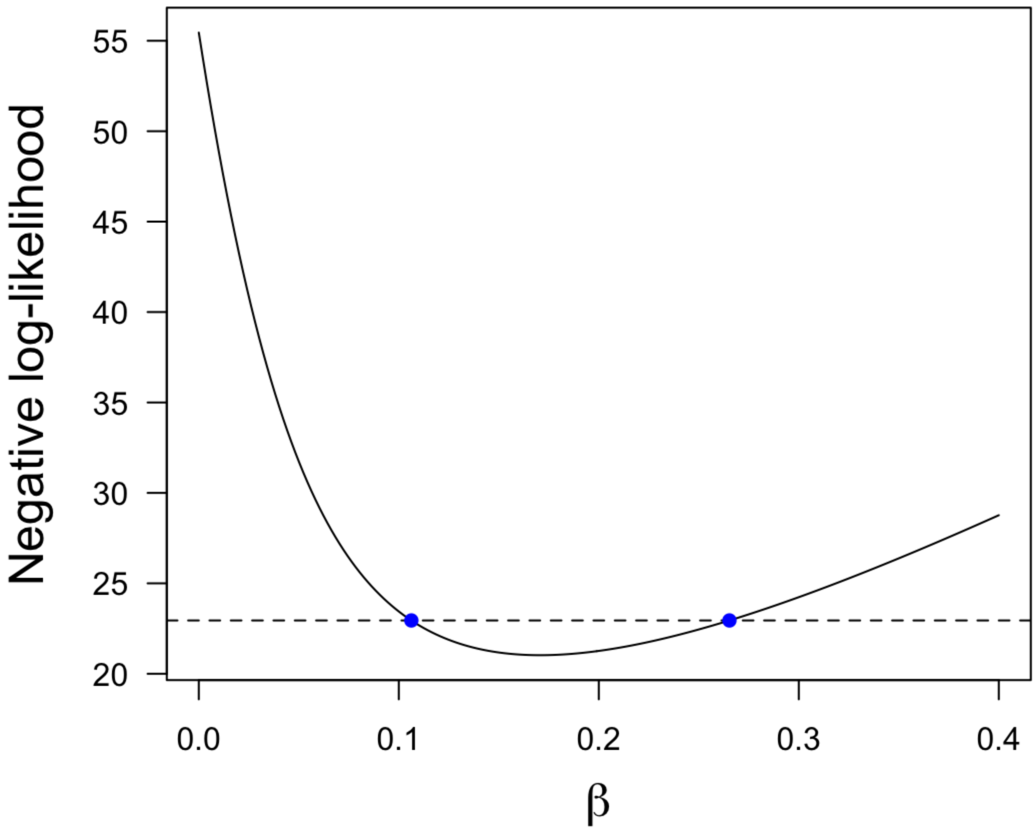
```
## [1] 0.09266613 0.24862556
```

Confidence interval for β_i

Due to possible bias in $\mathrm{SE}(\beta)$, we can compute CI's based on the *profile likelihood*

```r
## number of points to profile
nb <- 200
## possible beta's
beta_hat <- seq(0, 0.4, length = nb)
## calculate neg-LL of possible beta's
pl <- rep(NA, nb)
for(i in 1:nb) {
  mm <- glm(age ~ 1 + offset(beta_hat[i] * length), data = df,
            family = binomial(link = "logit"))
  pl[i] <- -logLik(mm)
}
```

Confidence interval for β_i



Confidence interval for β_i

We can compute CI’s based on the profile likelihood with `confint()`

```
## 95% CI via profile likelihood
confint(fit_mod)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) -21.8553251 -8.6351047
## length        0.1062832  0.2653229
```

Model diagnostics

As with other models, it’s important to examine diagnostic checks for our fitted models

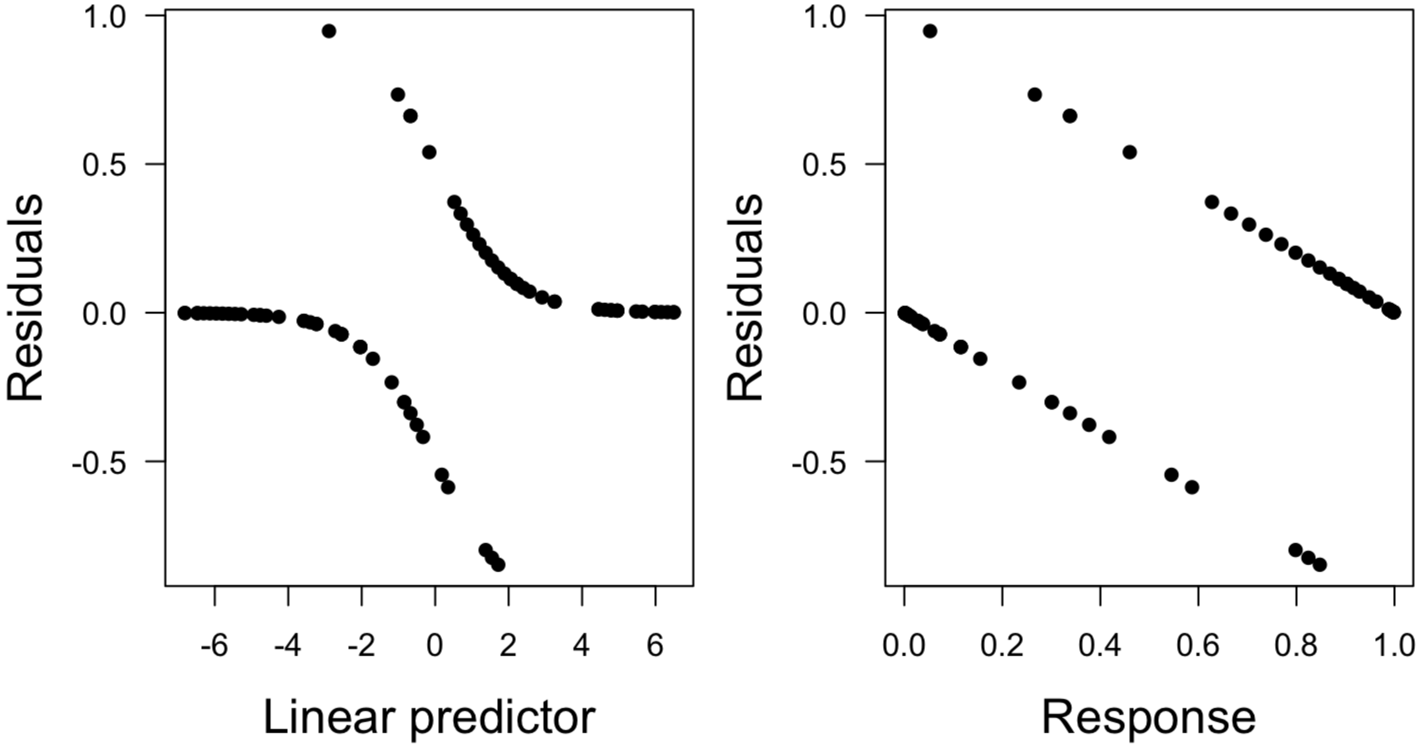
Residuals

We usually think about residuals e as

$$e = y - \hat{y}$$

With logistic regression, the response can take 1 of 2 possible values

Residuals


1.0
0.5
0.0
-0.5
Residuals
-6
-4
-2
0
2
4
6
Linear predictor
1.0
0.5
0.0
-0.5
Residuals
0.0
0.2
0.4
0.6
0.8
1.0
Response

Deviance residuals

We can instead use the *deviance residuals*

$$e_i = (2y_i - 1)D_i$$

$2y - 1$ is 1 (-1) if y is 1 (0)

This is the default for `residuals()`

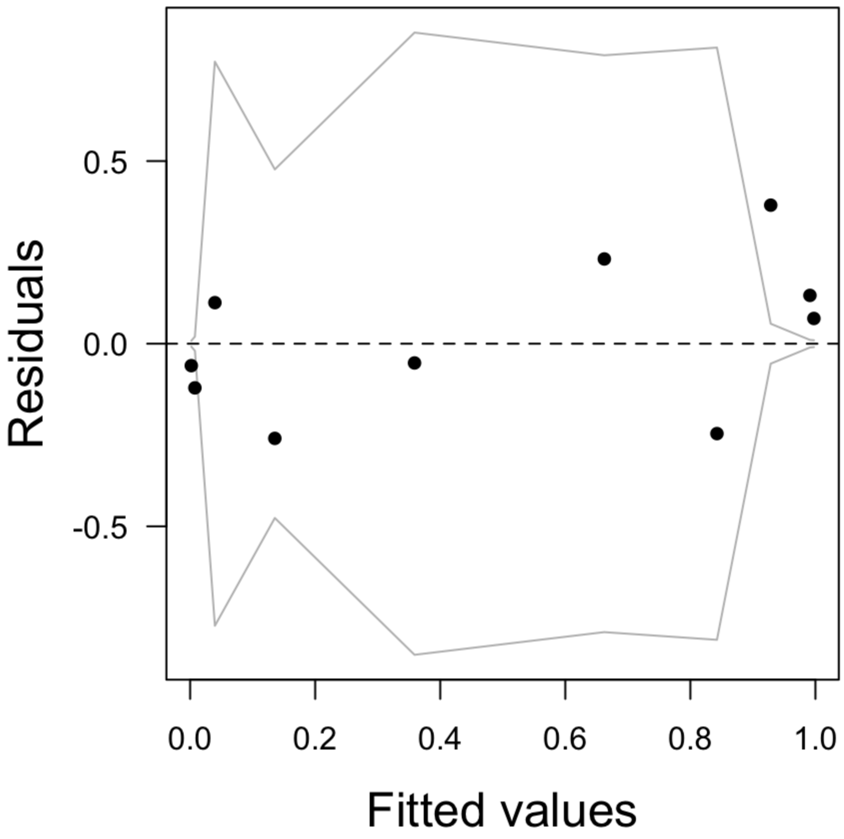
Deviance residuals

We then place the deviance residuals into bins for easier inspection

- Sensitive to the number of bins (~30/bin is good)
- Mean of e not constrained to 0
- Check to see that ~95% of points fall within the CI

Can use `binnedplot()` from the **arm** package to do this

Deviance residuals

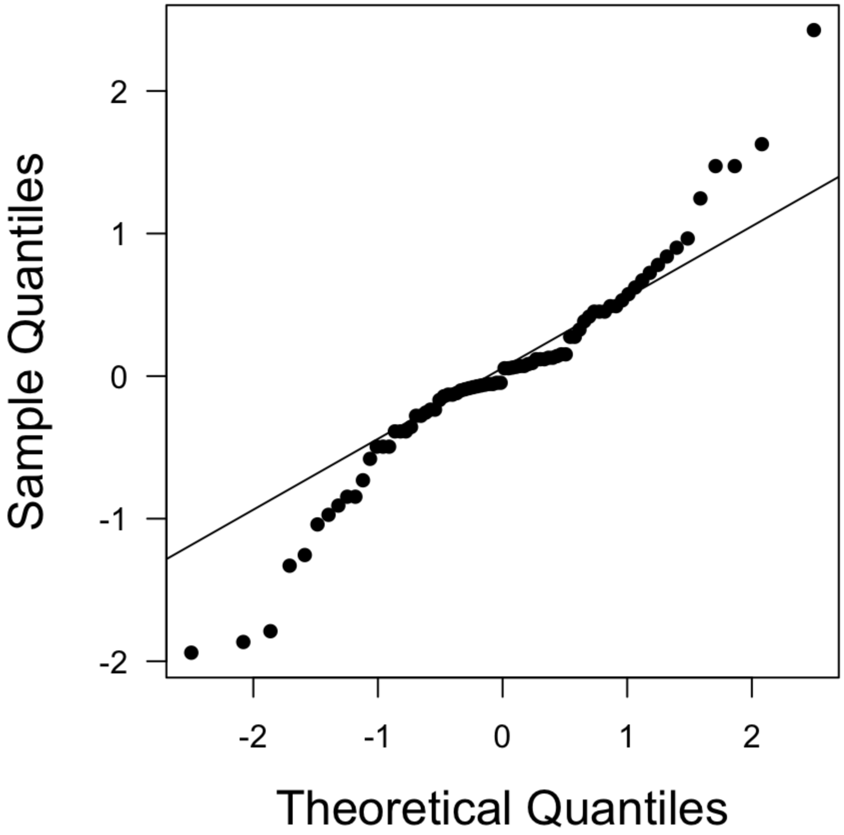


Q-Q plots

We can examine a Q-Q plot, but there is no assumption that the e are normal

It can help to identify unusual points

Q-Q plots



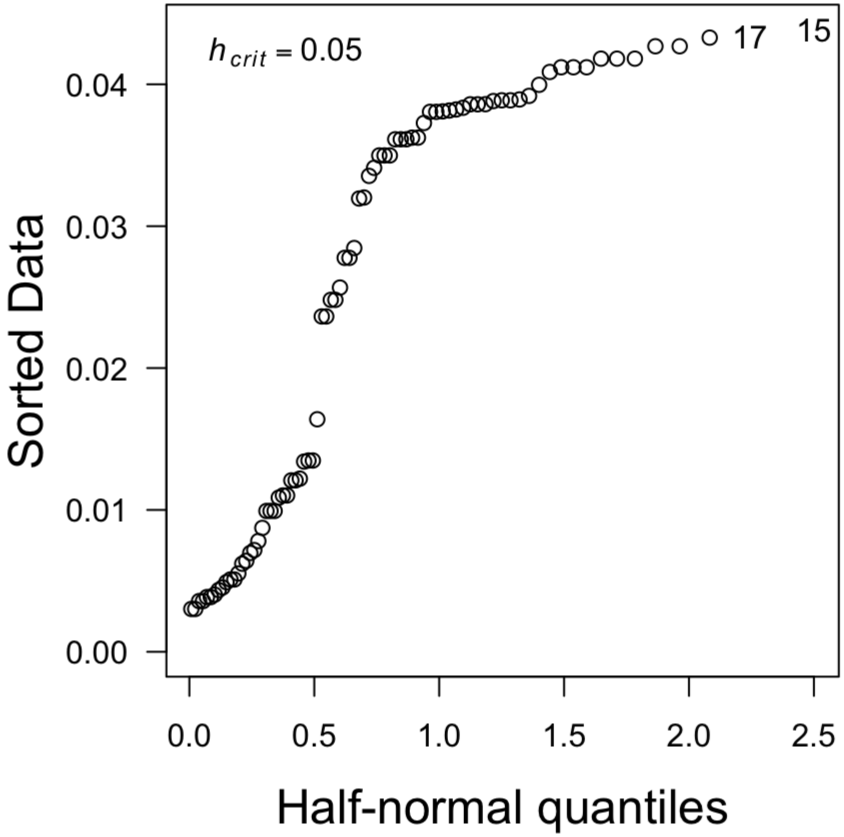
Leverage

We can also calculate the leverages h to look for unusual observation in *predictor space*

Recall we are potentially concerned about $h > 2\frac{k}{n}$

We can use `faraway::halfnorm()`

Leverage



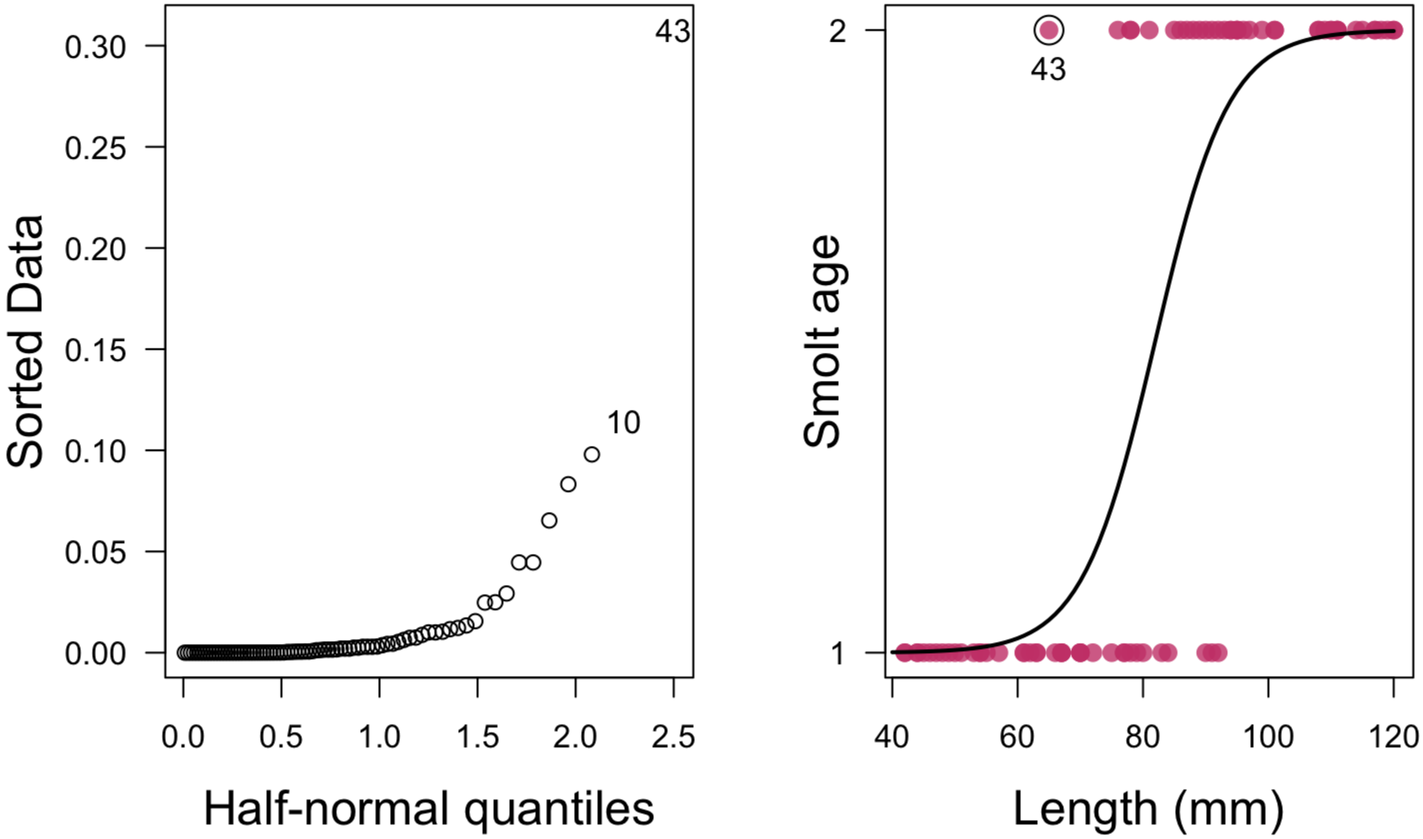
h_{crit} = 0.05
17
15
Sorted Data
0.00
0.01
0.02
0.03
0.04
0.0
0.5
1.0
1.5
2.0
2.5
Half-normal quantiles

Cook’s Distance

Recall that we can use Cook’s D to identify potentially influential points

$$D_i = e_i^2 \frac{1}{k} \left(\frac{h_i}{1 - h_i} \right)$$

Cook's Distance


Sorted Data
Half-normal quantiles
Smolt age
Length (mm)