Modeling count data

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

15 May 2020

Goals for today

- Understand the application of Poisson regression to count data
- Understand how to fit Poisson regression models in **R**
- Understand how to evaluate model fits and diagnostics for Poisson regression

Count data

Counts form the basis for much of our data in environmental sciences

- Number of adult salmon returning to spawn in a river
- Number of days of rain in a year
- Number of bees visiting a flower

Counts vs proportions

We have seen how to model proportional data with GLMs

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

Counts vs proportions

With count data, we only know the *frequency of occurrence*

That is, how often something occurred without knowing how often it *did not occur*

Modeling count data

Standard regression models are inappropriate for count data for 4 reasons:

1. linear model might lead to predictions of negative counts
2. variance of the response variable may increase with the mean
3. errors are not normally distributed
4. zeros are difficult to transform

Distribution for discrete counts

The Poisson distribution is perhaps the best known

It gives the probability of a given number of events occurring in a fixed interval of time or space

Poisson distribution

Examples

- the number of Prussian soldiers killed by horse kicks per year from 1868 - 1931
- the number of new COVID-19 infections per day in the US
- the number of email messages I receive per week from students in QERM 514

Poisson distribution

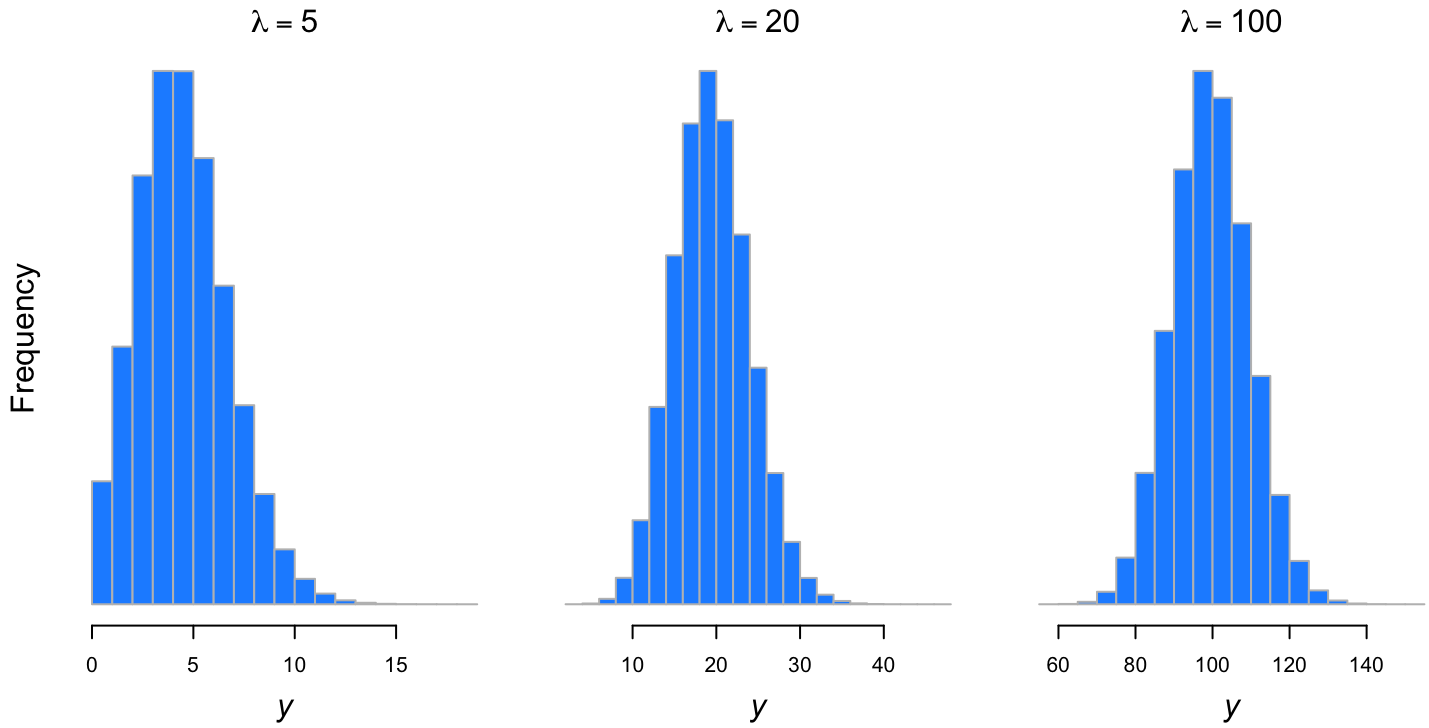
It's unique in that it has one parameter λ to describe both the mean *and* variance

$$y_i \sim \text{Poisson}(\lambda)$$

$$\text{Mean}(y) = \text{Var}(y) = \lambda$$

Poisson distribution

As $\lambda \to \infty$ the Poisson $\to$ Normal



Poisson distribution

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

$$\Downarrow$$

$$f(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

with $\theta = \log(\mu)$ and $\phi = 1$

$a(\phi) = 1 \qquad b(\theta) = \exp(\theta) \qquad c(y,\phi) = -\log(y!)$

Poisson distribution

An interesting property of the Poisson is that

$$y_i \sim \text{Poisson}(\lambda)$$
$$\Downarrow$$
$$\sum_i y_i \sim \text{Poisson}(\sum_i \lambda_i)$$

This means we can use aggregated data if we lack individual-level data

Poisson and binomial

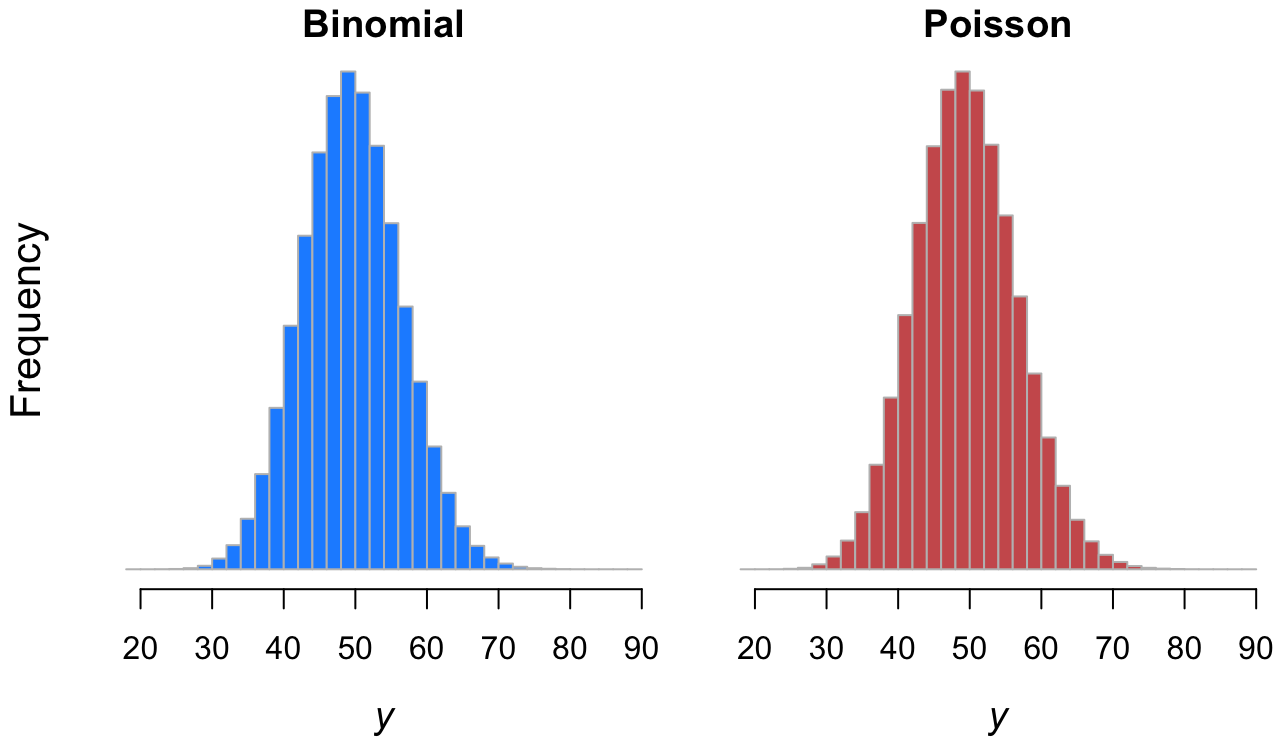
The Poisson distribution can also approximate a binomial distribution if n is large and p is small

As $p \to 0$, $\text{logit}(p) \to \log(p)$

Binomial with logit link $\to$ Poisson with log link

Poisson and binomial

An example with $p = 0.05$ and $n = 1000$



Rethinking density

We have been considering (log) density itself as a response

$$\text{Density}_i = f(\text{Count}_i, \text{Area}_i)$$

$$\Downarrow$$

$$\text{Density}_i = \frac{\text{Count}_i}{\text{Area}_i}$$

Rethinking density

We have been considering (log) density itself as a response

$$\text{Density}_i = f(\text{Count}_i, \text{Area}_i)$$

$$\Downarrow$$

$$\text{Density}_i = \frac{\text{Count}_i}{\text{Area}_i}$$

With GLMs, we can shift our focus to

$$\text{Count}_i = f(\text{Area}_i)$$


VALHALLA
Magellan
MEXICO
SINCE 1534

Example of a Poisson regression

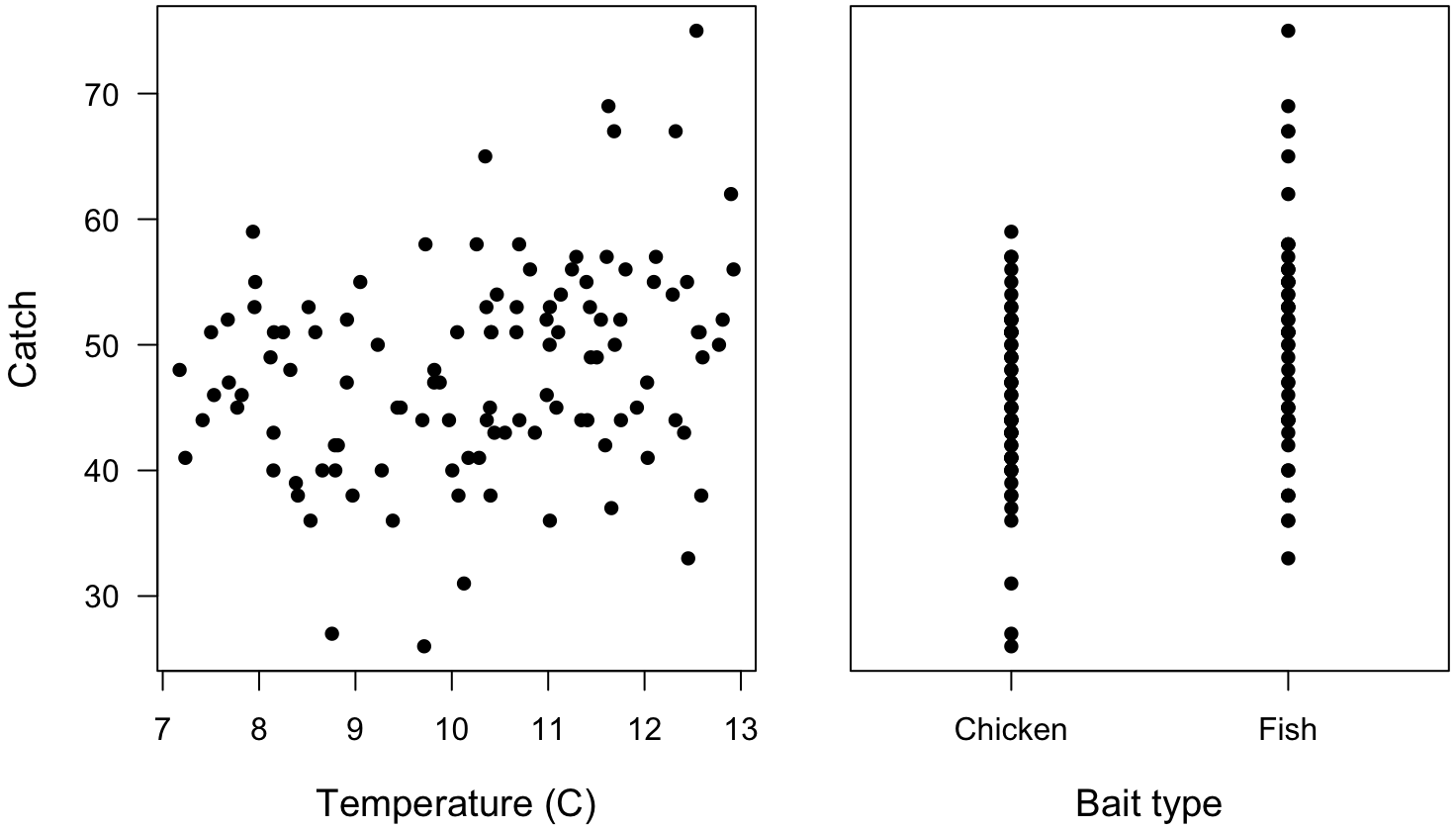
Catches of spot prawns y_i as a function of bait type C_i and water temperature T_i

$$\text{data distribution: } y_i \sim \text{Poisson}(\lambda_i)$$

$$\text{link function: } \log(\lambda_i) = \mu_i$$

$$\text{linear predictor: } \mu_i = \alpha + \beta_1 C_i + \beta_2 T_i$$

Catches of spot prawns



Catches of spot prawns

```
## Poisson regression
cmod <- glm(catch ~ fish + temp, data = prawns,
            family = poisson(link = "log"))
faraway::sumary(cmod)
```

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 3.5644284  0.0906850  39.306 < 2.2e-16
## fish        0.0894061  0.0274085   3.262  0.001106
## temp        0.0256769  0.0087425   2.937  0.003314
##
## n = 113 p = 3
## Deviance = 135.32140 Null Deviance = 157.85016 (Difference = 22.52876)
```

Inference from Poisson regression

We can easily estimate the CI's on the model parameters with `confint()`

```
## CI's for prawn model
ci_prawns <- confint(cmod)
ci_tbl <- cbind(ci_prawns[,1], coef(cmod), ci_prawns[,2])
colnames(ci_tbl) <- c("Lower", "Estimate", "Upper")
signif(ci_tbl, 3)
```

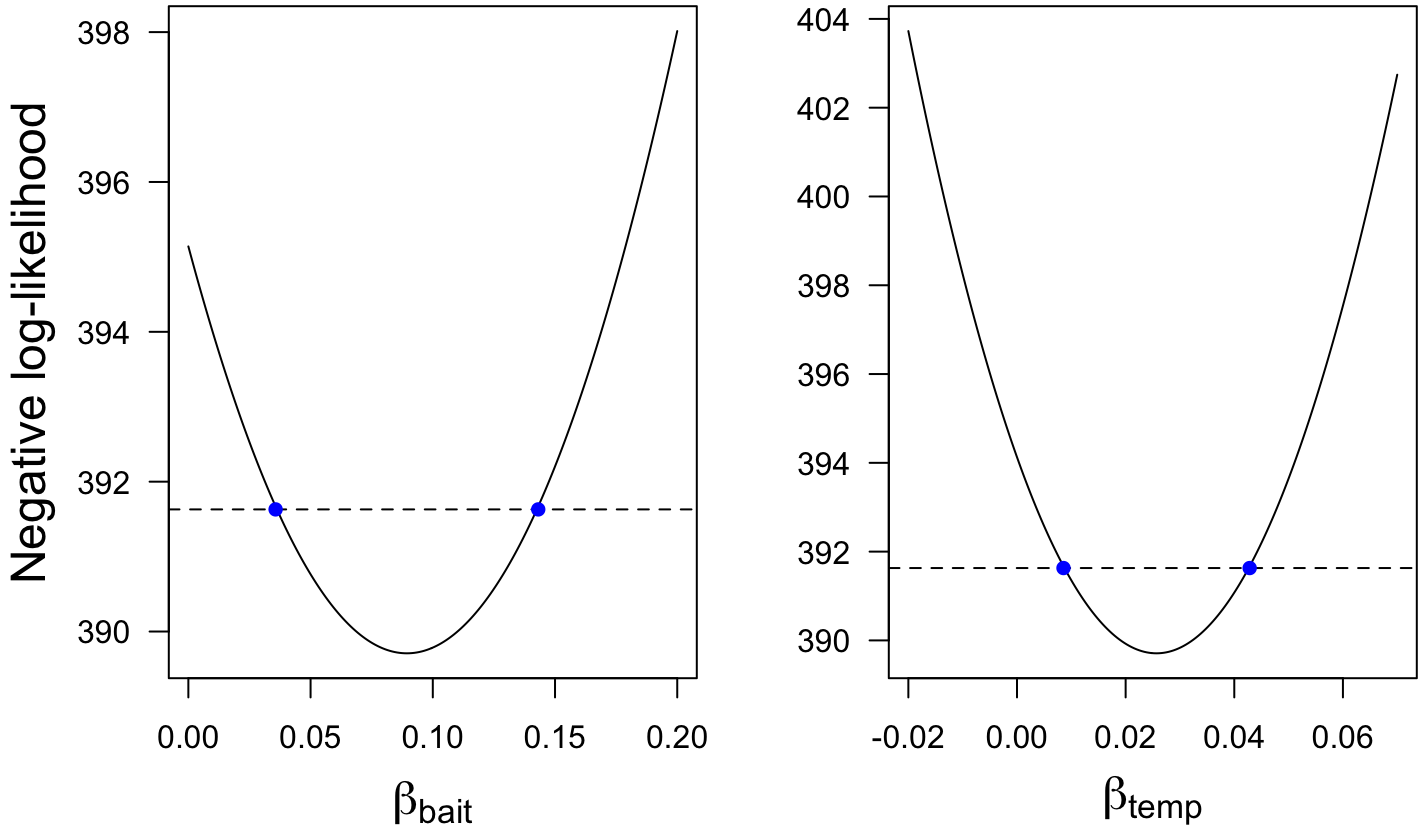
```
##               Lower Estimate  Upper
## (Intercept) 3.39000   3.5600 3.7400
## fish        0.03570   0.0894 0.1430
## temp        0.00856   0.0257 0.0428
```

Profile likelihood

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

```
## number of points to profile
nb <- 200
## possible beta's
beta_bait <- seq(0, 0.2, length = nb)
## calculate neg-LL of possible beta_bait
## fix beta_temp at its MLE
plb <- rep(NA, nb)
for(i in 1:nb) {
  mm <- glm(catch ~ 1 + offset(beta_bait[i] * fish
                               + offset(coef(cmod)[3] * temp)),
            data = prawns,
            family = poisson(link = "log"))
  plb[i] <- -logLik(mm)
}
```

Confidence interval for β_i



Goodness of fit

It's natural to ask how well a model fits the data

As with binomial models, we can check the deviance D against a χ^2 distribution

Deviance for Poisson

Recall that the deviance for any model is

$$D_i = \text{-}2 \left[\log \mathcal{L}(M_i) - \log \mathcal{L}(M_0)\right]$$

where M_i is the model of interest and M_0 is an intercept-only model

Deviance for Poisson

The log-likelihood for a Poisson is

$$\log \mathcal{L}(y; \lambda) = \sum_{i=1}^{n} \left[y_i \log(\lambda) - \lambda - \log(y_i !) \right]$$

The deviance for a Poisson is

$$\log \mathcal{L}(y; \lambda) = \sum_{i=1}^{n} \left[y_i \log(y_i / \hat{\lambda}) - (y_i - \hat{\lambda}) \right]$$

Goodness of fit for prawn model

H_0: Our model is correctly specified

```
## deviance of prawn model
D_full <- summary(cmod)$deviance
## LRT with df = 1
(p_value <- pchisq(D_full, nn - length(coef(cmod)),
                   lower.tail = FALSE))
```

```
## [1] 0.05096932
```

We cannot reject the H_0

Goodness of fit for prawn model

It turns out that the assumption of $D \sim \chi^2_{n-k}$ can be violated with Poisson models unless λ is large

Another option is Pearson's X^2 statistic we saw for binomial models

Pearson's goodness of fit

Recall that Pearson's X^2 is

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-k)}$$

So for our Poisson model

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \sim \chi^2_{n-k}$$

Pearson's goodness of fit

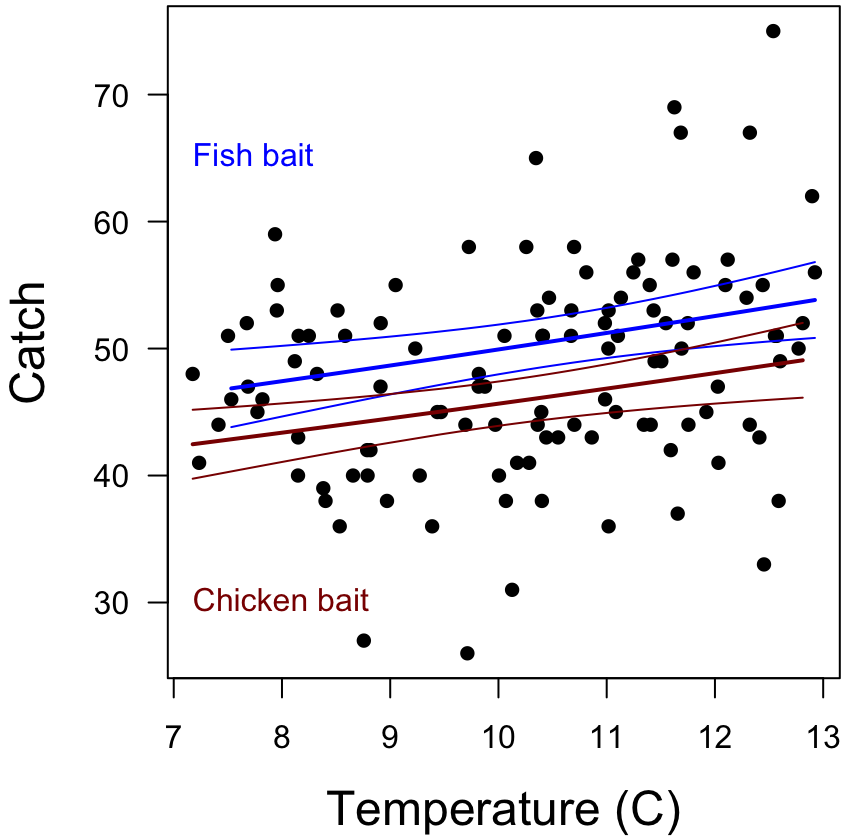
H_0: Our model is correctly specified

```
## numerator
nm <- (prawns$catch - fitted(cmod))^2
## denominator
dm <- fitted(cmod)
## Pearson's
X2 <- sum(nm / dm)
## test
(p_value <- pchisq(X2, nn - length(coef(cmod)), lower.tail = FALSE))
```

```
## [1] 0.07074179
```

We cannot reject the H_0

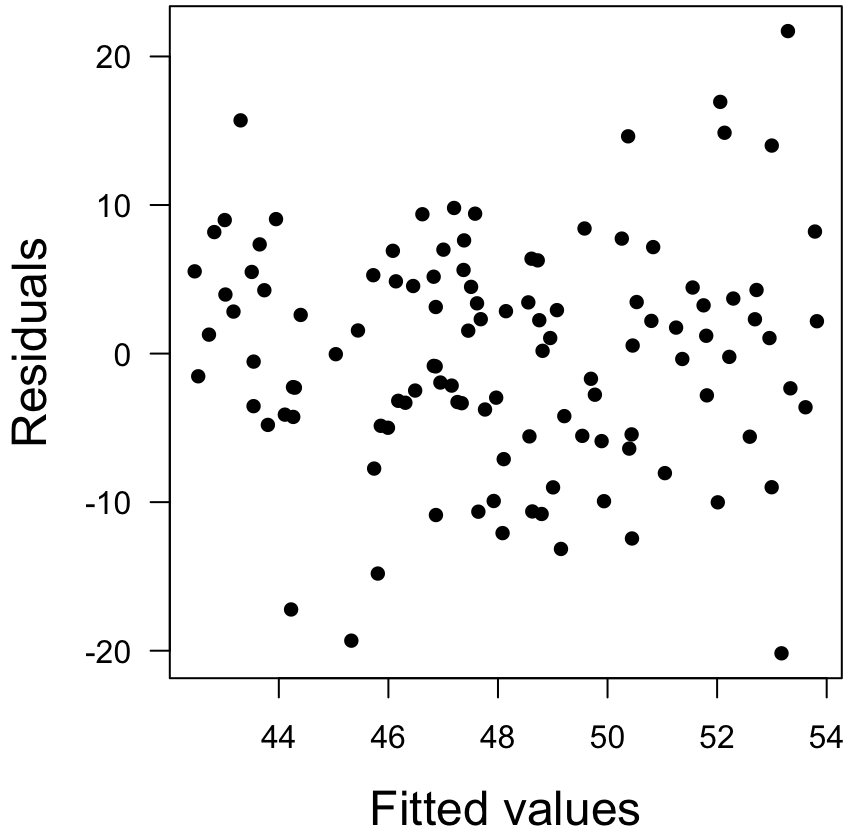
Fitted values & CI's



Model diagnostics

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

Residual plots



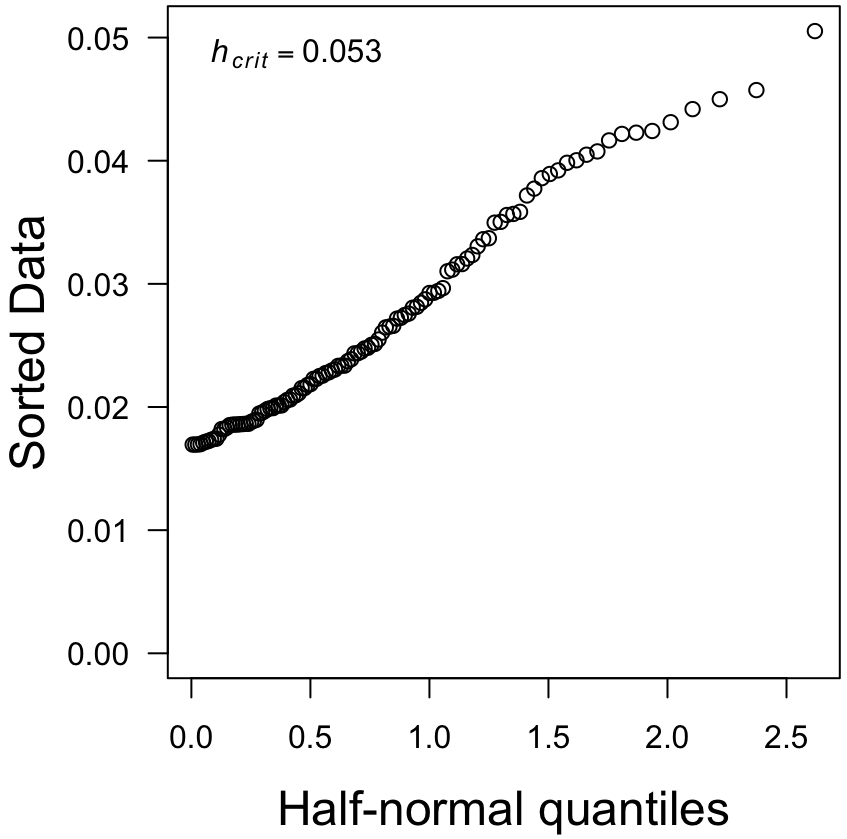
Leverage

We can 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} = 0.053
0.05
0.04
0.03
0.02
0.01
0.00
Sorted Data
0.0
0.5
1.0
1.5
2.0
2.5
Half-normal quantiles

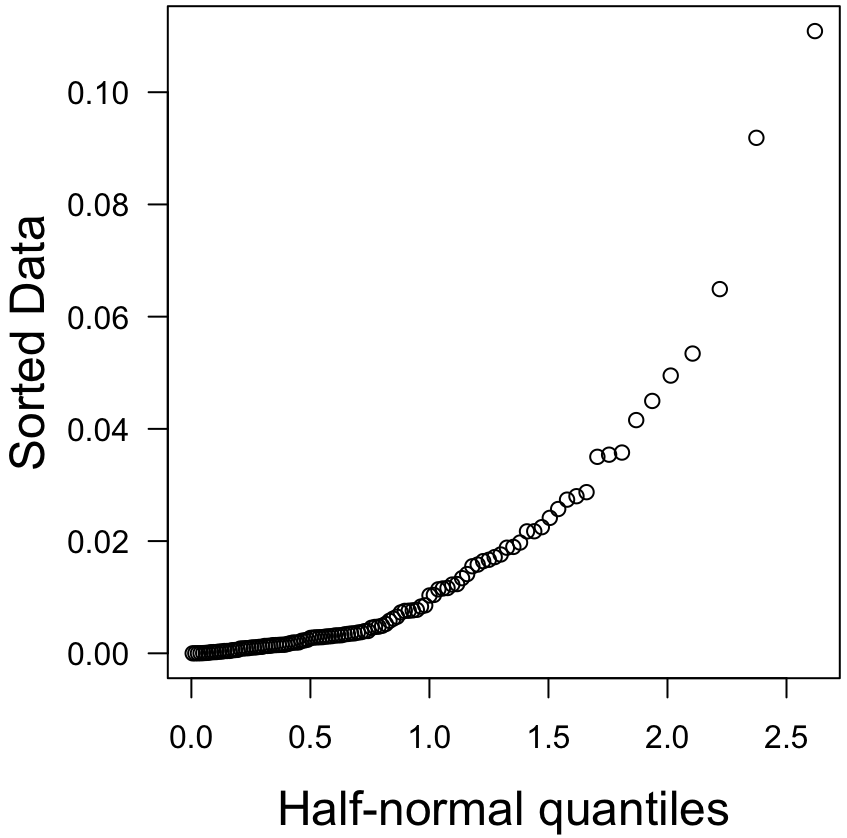
Cook's Distance

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

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

In general we are concerned about $D_i > F_{n,n-k}^{(0.5)} \approx 1$

Cook's Distance



Model selection for prawn model

We can use a likelihood ratio test to compare our model to an intercept-only model

```
## deviance of full model
D_full <- summary(cmod)$deviance
## deviance of null model
D_null <- summary(cmod)$null.deviance
## test statistic
lambda <- D_null - D_full
## LRT with df = 2
(p_value <- pchisq(lambda, 2, lower.tail = FALSE))
```

```
## [1] 1.282157e-05
```

We reject H_0 (that the data come from the null model)

Summary

- Lots of ecological data consists of counts
- We can use Poisson regression for count data instead of a log-transformation
- We can use many of the same goodness-of-fit measures and diagnostics as for other GLMs and LMs