Generalised Linear Models (GLMs)

http://bradduthie.github.io/talks/GLMs.pdf

Brad Duthie

25 November 2019

Key Skills Test 3 (KST3)

- Critical analysis of statistical results
- Material from whole course (excluding UG Open Science Session)
- Two previews: Sample KST3 and challenge at start of practical on interactions and fit (reading tables of coefficients)
- Will be provided results as tables and/or figs and asked to interpret and criticize; no coding required
- KST3 assessment Friday 29 NOV

Topics concerning generalised linear models

- Common problems of general linear models
  1. Non-homogenous variance of residuals
  2. Non-normal distribution of residuals
- Generalising the linear model[1]
- Linear predictors and link functions
- Overdispersion (more variance in response variable than expected)

[1] Nelder, JA, & Wedderburn, RW. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370-384.

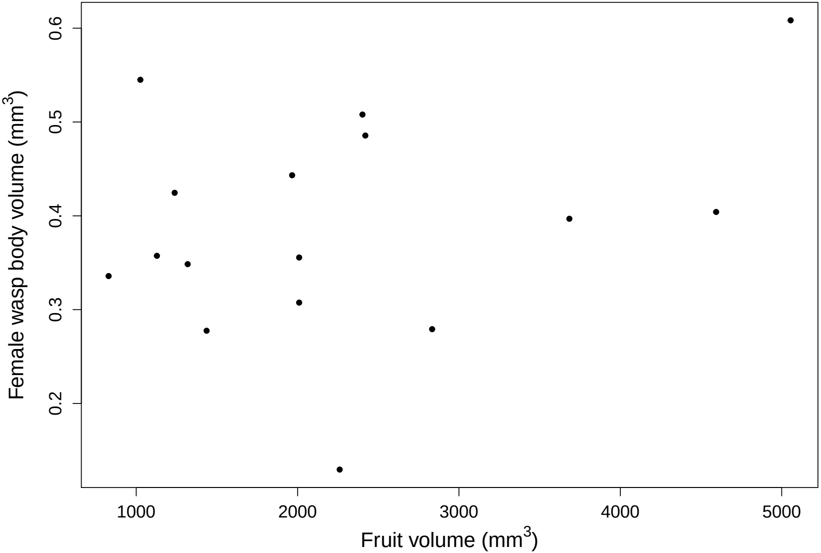
Common problems of general linear models

Common problems of general linear models

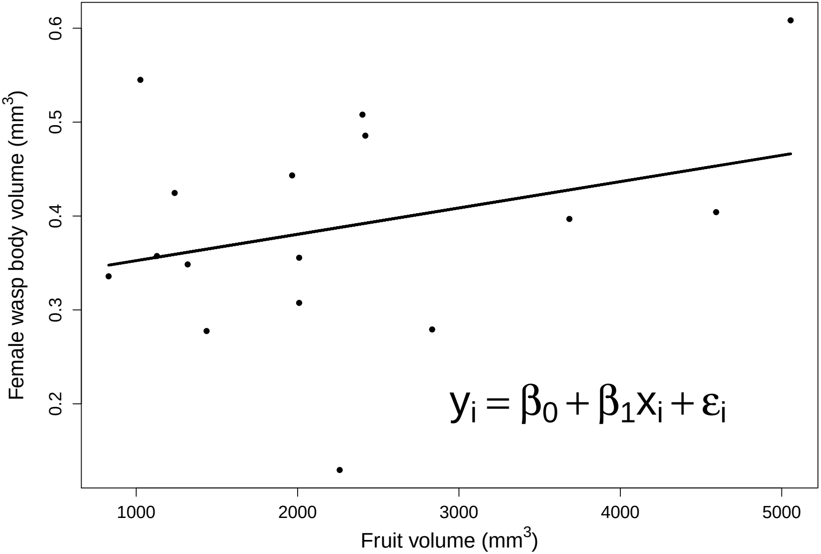
Fig trees in Baja, Mexico are visted by several species of wasps

Wasps use their ovipositors to drill into the side of the enclosed inflorescence (syconia, or colloquially “fruit”)

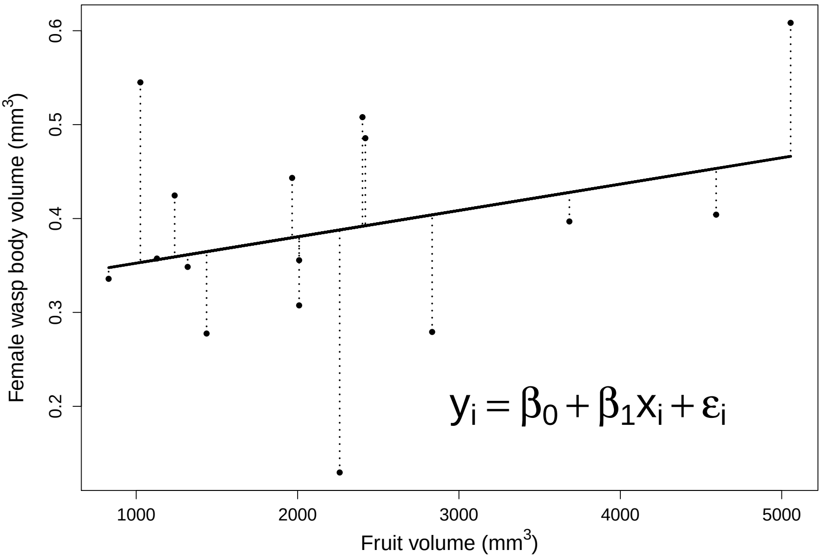
Common problems of general linear models



Common problems of general linear models



Common problems of general linear models



Common problems of general linear models

What if the response (y) variable residuals do not fit general linear model assumptions? This can happen under the following conditions:

- Residuals (ϵ) do not have a constant variance across x values (heteroscadisticity)
- Residuals (ϵ) are not normally distributed

[1] Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons.

Common problems of general linear models

What if the response (y) variable residuals do not fit general linear model assumptions? This can happen under the following conditions:

- Residuals (ϵ) do not have a constant variance across x values (heteroscadisticity)
- Residuals (ϵ) are not normally distributed

The Logan text book[1] (Ch. 17) discusses four situations:

1. Count data
2. Proportion data
3. Binary responses
4. "Time to event" data

[1] Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons.

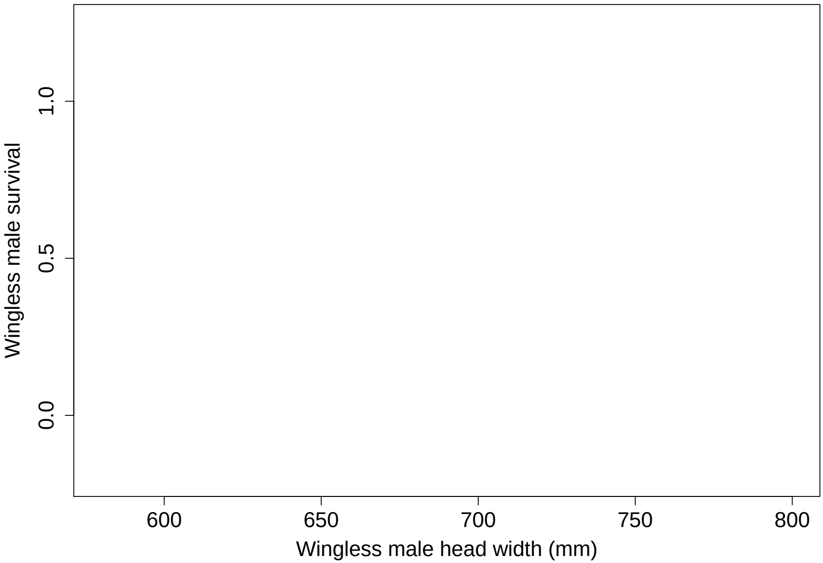
Common problems of general linear models

- Female *Heterandrium* wasps can produce two types of males
- Winged males disperse from their natal fruit to mate
- Wingless males engage in combat within fruit for access to females

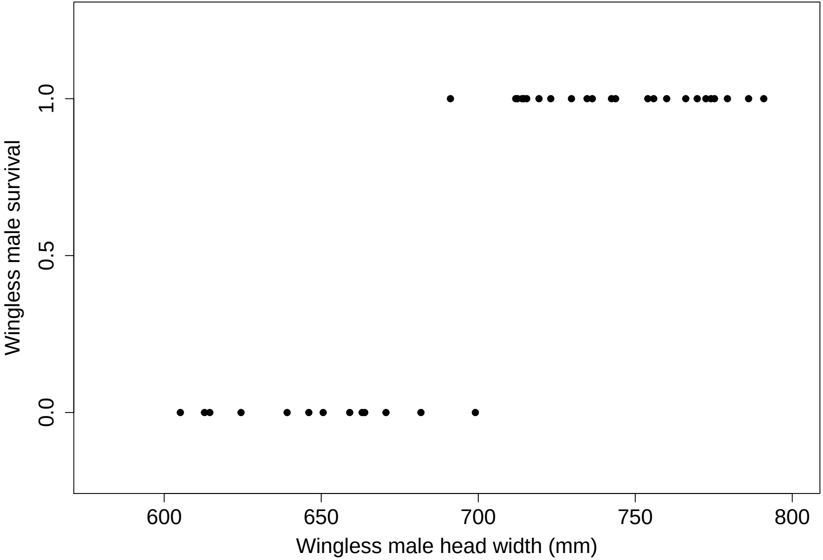
Common problems of general linear models

- Female *Heterandrium* wasps can produce two types of males
- Winged males disperse from their natal fruit to mate
- Wingless males engage in combat within fruit for access to females
- **Do bigger wingless males have a higher probability of survival?**

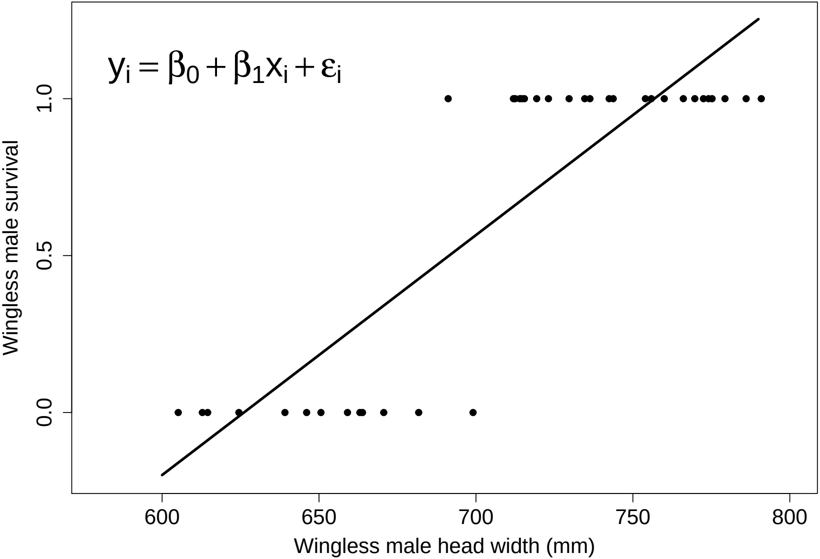
Common problems of general linear models



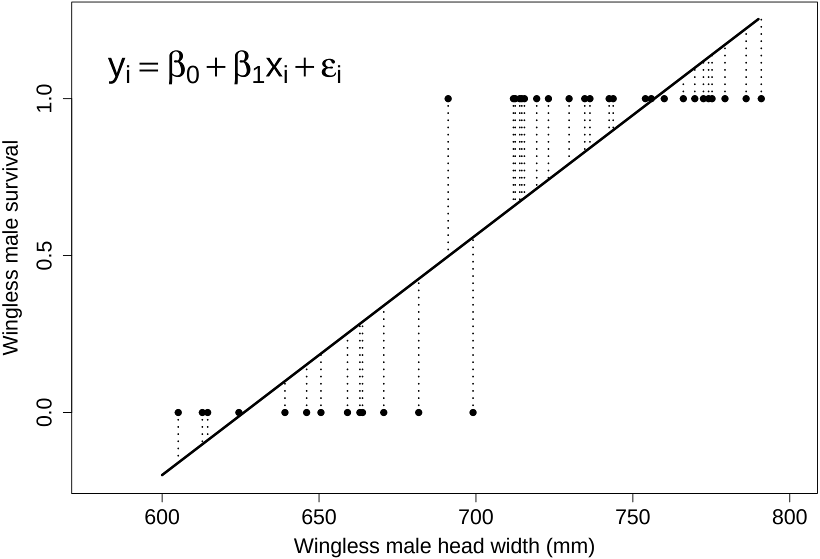
Common problems of general linear models



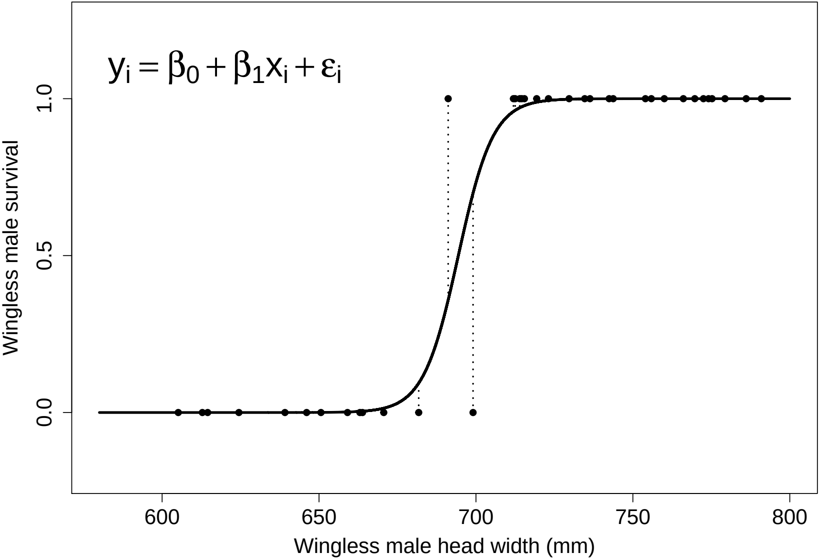
Common problems of general linear models



Common problems of general linear models



Common problems of general linear models



Common problems of general linear models

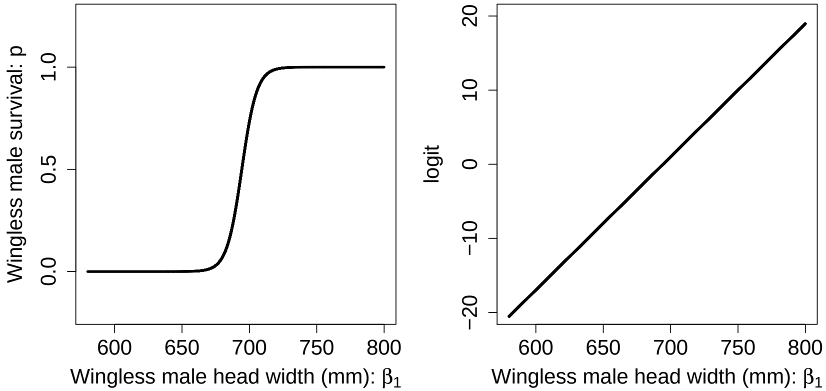
The logit link function linearises the binomial probability function

$$\ln\left(\frac{p}{q}\right) = \beta_0 + \beta_1 x$$

Common problems of general linear models

The logit link function linearises the binomial probability function

$$\ln\left(\frac{p}{q}\right) = \beta_0 + \beta_1 x$$



Generalising the linear model

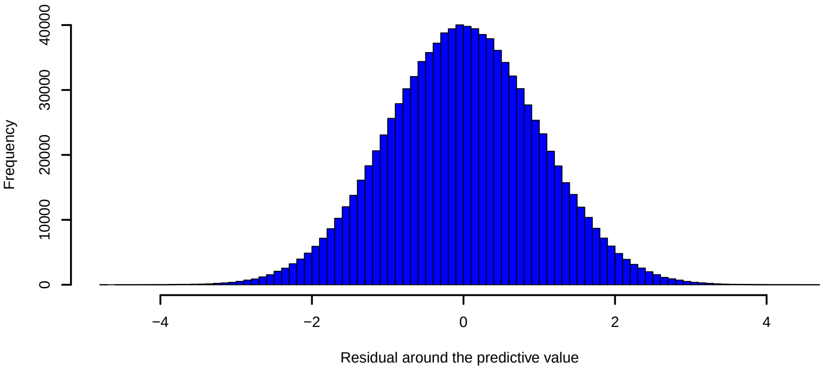
GLMs (pronounced "glims"): Generalised linear models

- Not to be confused with **general** linear models (also sometimes called GLMs)
- have three properties
  1. Error structure
  2. Linear predictor
  3. Link function

Generalising the linear model: error (i.e., residual) structure

General linear models assume normally distributed errors

Actual errors can violate the asumption of normality in several ways



Strong skew, kurtosis, Strict bounds (e.g., values between 0 and 1 as shown earlier, predicted values never below zero as with counts)

Generalising the linear model: error (i.e., residual) structure

Generalised linear models are characterised by independent random variables (i.e., y_1, y_2, $\ldots$, y_n) with an expected value $E(y_i) = \mu_i$, and a density function (error) **from the exponential family**.

[1] Rencher, AC, & Schaalje, GB. 2008. *Linear models in statistics*. John Wiley & Sons, 446-448.

[2] Nelder, JA, & Wedderburn, RW. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370-384.

Generalising the linear model: error (i.e., residual) structure

Generalised linear models are characterised by independent random variables (i.e., y_1, y_2, $\ldots$, y_n) with an expected value $E(y_i) = \mu_i$, and a density function (error) **from the exponential family**.

A density function $f(y_i; \theta_i)$ is in the exponential family if it can be expressed as follows,

$$f(y_i; \theta_i) = e^{y_i \theta_i + b(\theta_i) + c(y_i)}.$$

In the above, θ_i is a parameter of the family.

Statistical distributions in the exponential family include the Poisson, binomial, exponential, and gamma (also the normal).

[1] Rencher, AC, & Schaalje, GB. 2008. *Linear models in statistics*. John Wiley & Sons, 446-448.

[2] Nelder, JA, & Wedderburn, RW. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370-384.

Generalising the linear model: error (i.e., residual) structure

There are four common error structures:

1. Poisson errors (for count data)
2. Binomial errors (for proportion data)
3. Exponential errors (for time to event)
4. Gamma errors (for data with constant coefficient of variation)

Generalising the linear model: linear predictor

The linear predictor (η) is the sum of linear effects of 1 or more explanatory variables (β),

$$\eta = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}.$$

GLMs compare a *transformed* value from η to observations:

- The transformation is specified by the link function (see next)
- The fitted value is the predicted value multiplied by the reciprocal of the link function

Generalising the linear model: link function

The link function describes how the expected value of the response variable (μ_i) relates to η,

$$g(\mu_i) = \eta.$$

- Note that this relates the **mean** of a response variable (i.e., $E(y_i) = \mu_i$) to the linear predictor; it is not transforming individual values of y_i.

Generalising the linear model: link function

The link function describes how the expected value of the response variable (μ_i) relates to η,

$$g(\mu_i) = \eta.$$

- Note that this relates the **mean** of a response variable (i.e., $E(y_i) = \mu_i$) to the linear predictor; it is not transforming individual values of y_i.
- The model prediction is not $E(y_i)$, except in the special case that we have been using up until now (normally distributed residuals), called the *identity link* (i.e., $g(\mu_i) = \mu_i = \eta$).

Linear predictors and link functions

Recreated Table 17.1: Common GLMs and associated canonical link-distribution pairs.

Model	Response variable	Predictor variable(s)	Residual dist.	Link
Linear regression[a]	Continuous	Continuous/ categorical	Gaussian (normal)	Identity $g(\mu) = \mu$
Logistic regression	Binary	Continuous/ categorical	Binomial	Logit $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$
Log-linear models	Counts	Categorical	Poisson	Log $g(\mu) = \ln(\mu)$

[a] Includes the standarad ANOVA and ANCOVA designs.

Applications

Linear predictors and link functions: Count data

Fruit	Females	Wingless_Males	Winged_Males
1	2	0	1
2	0	0	0
3	0	0	0
4	0	0	0
5	2	8	1
6	0	0	1

- Count data are *frequencies* rather than proportions
- Count data have a lower bound (cannot be below zero)
- Variance increases with the mean (heteroscadisticity)
- Errors are not normally distributed
- Data are integers (affects error distribution)

Linear predictors and link functions: Count data

Fruit	Females	Wingless_Males	Winged_Males
1	2	0	1
2	0	0	0
3	0	0	0
4	0	0	0
5	2	8	1
6	0	0	1

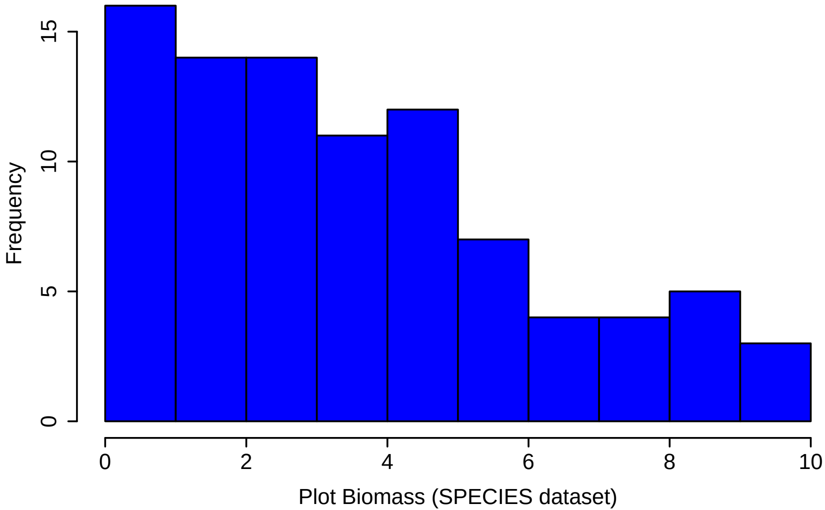
- Use a log link function (ensures fitted values bounded below)
- Use `family = poisson` to specify appropriate error variance
- Can use `family = quasipoisson` if the data are **overdispersed** (this is conservative)
- Alternaive error distributions are available (e.g., negative binomial)

Linear predictors and link functions: Count data

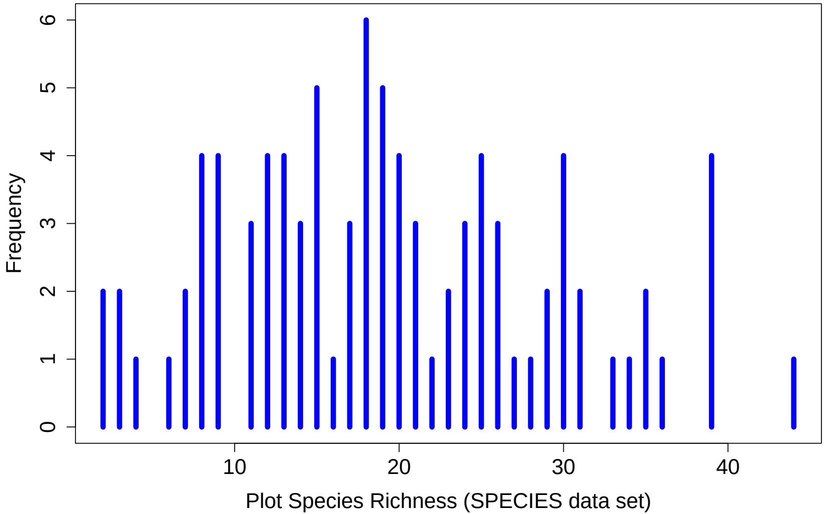
Species in a plot as a function of Biomass and pH

```
## # A tibble: 90 x 3
##    PH    BIOMASS RICHNESS
##    <chr>   <dbl>    <dbl>
##  1 high    0.469       30
##  2 high    1.73        39
##  3 high    2.09        44
##  4 high    3.93        35
##  5 high    4.37        25
##  6 high    5.48        29
##  7 high    6.68        23
##  8 high    7.51        18
##  9 high    8.13        19
## 10 high    9.57        12
## # ... with 80 more rows
```

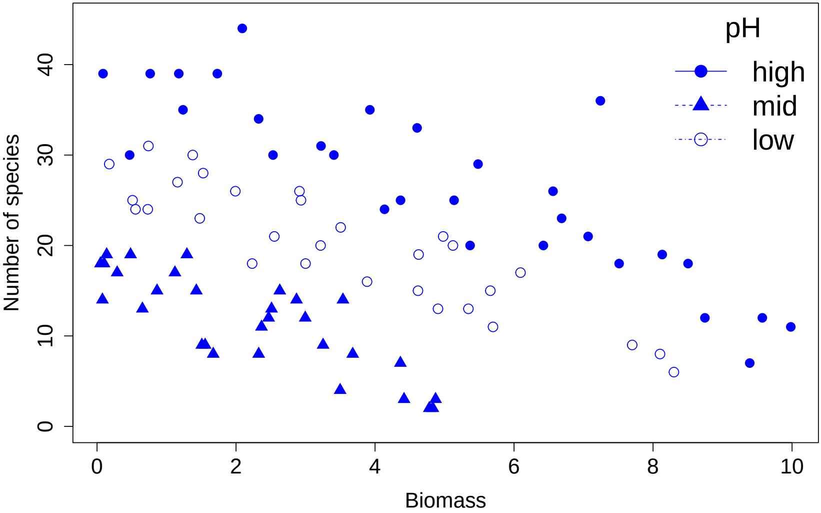
Linear predictors and link functions: Count data



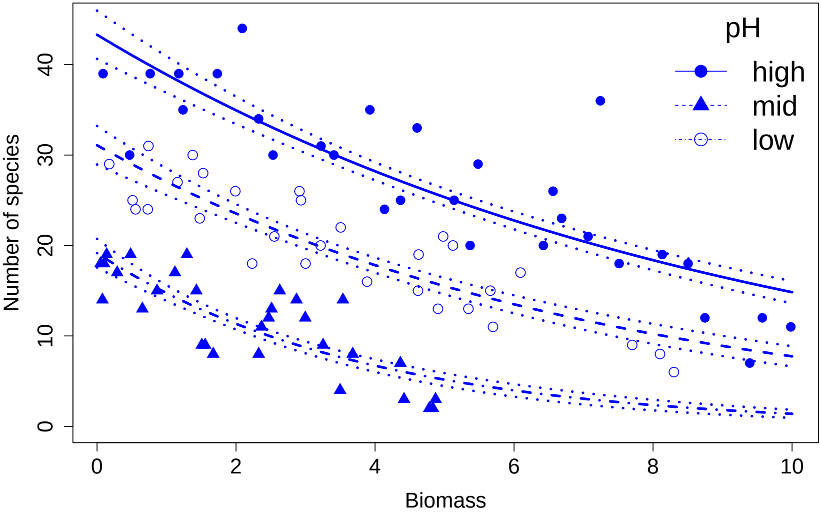
Linear predictors and link functions: Count data



Linear predictors and link functions: Count data



Linear predictors and link functions: Count data



Linear predictors and link functions: Count data

Output from `summary(the_model)`:

```
##
## Call:
## glm(formula = RICHNESS ~ BIOMASS * PH, family = poisson, data = SPECIES)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4978  -0.7485  -0.0402   0.5575   3.2297
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
## BIOMASS       -0.10713    0.01249  -8.577  < 2e-16 ***
## PHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
## PHmid         -0.33146    0.09217  -3.596 0.000323 ***
## BIOMASS:PHlow -0.15503    0.04003  -3.873 0.000108 ***
## BIOMASS:PHmid -0.03189    0.02308  -1.382 0.166954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 452.346  on 89  degrees of freedom
## Residual deviance:  83.201  on 84  degrees of freedom
## AIC: 514.39
##
## Number of Fisher Scoring iterations: 4
```

Linear predictors and link functions: Overdispersion

- Variance of Poisson or binomial models is assumed to relate to the mean or sample size, respectively
- Dispersion (variance) parameter is set to 1.
- But we often get more (or less) variance than expected
- If residual deviance divided by degrees of freedom is less than 0.5 or more than 2, a quasibinomial or quasipoisson can be used to model the dispersion (but this will be conservative)
- We can also try other error distributions (e.g., negative binomial, bevabinomial)
- We can also consider other models (e.g., hurdle models such as zero-altered models)

Linear predictors and link functions: Overdispersion

Output from `summary(the_model)`:

```
##
## Call:
## glm(formula = RICHNESS ~ BIOMASS * PH, family = poisson, data = SPECIES)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4978  -0.7485  -0.0402   0.5575   3.2297
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
## BIOMASS       -0.10713    0.01249  -8.577  < 2e-16 ***
## PHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
## PHmid         -0.33146    0.09217  -3.596 0.000323 ***
## BIOMASS:PHlow -0.15503    0.04003  -3.873 0.000108 ***
## BIOMASS:PHmid -0.03189    0.02308  -1.382 0.166954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 452.346  on 89  degrees of freedom
## Residual deviance:  83.201  on 84  degrees of freedom
## AIC: 514.39
##
## Number of Fisher Scoring iterations: 4
```

Linear predictors and link functions: Overdispersion

We cannot do an F-test to compare generalised models

```r
model1 <- glm(RICHNESS ~ BIOMASS * PH, family = poisson,
              data = SPECIES);
model2 <- glm(RICHNESS ~ BIOMASS + PH, family = poisson,
              data = SPECIES);
anova(model1, model2, test = "Chi");
```

```
## Analysis of Deviance Table
##
## Model 1: RICHNESS ~ BIOMASS * PH
## Model 2: RICHNESS ~ BIOMASS + PH
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        84     83.201
## 2        86     99.242 -2   -16.04 0.0003288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear predictors and link functions: Overdispersion

Output from `summary(the_model)`:

```
##
## Call:
## glm(formula = RICHNESS ~ BIOMASS * PH, family = poisson, data = SPECIES)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4978  -0.7485  -0.0402   0.5575   3.2297
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
## BIOMASS       -0.10713    0.01249  -8.577  < 2e-16 ***
## PHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
## PHmid         -0.33146    0.09217  -3.596 0.000323 ***
## BIOMASS:PHlow -0.15503    0.04003  -3.873 0.000108 ***
## BIOMASS:PHmid -0.03189    0.02308  -1.382 0.166954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 452.346  on 89  degrees of freedom
## Residual deviance:  83.201  on 84  degrees of freedom
## AIC: 514.39
##
## Number of Fisher Scoring iterations: 4
```

Linear predictors and link functions: Binomial data

Properties of binomial data

- Binomial data include proportions or binary outcomes
- Errors (i.e., residuals) are not normally distributed
- Variance of the response variable is not constant
- Response variable is bounded between 0 and 1
- Calculating a percentage and transformation loses information of the size of the sample from which the population was estimated

Linear predictors and link functions: Binomial data

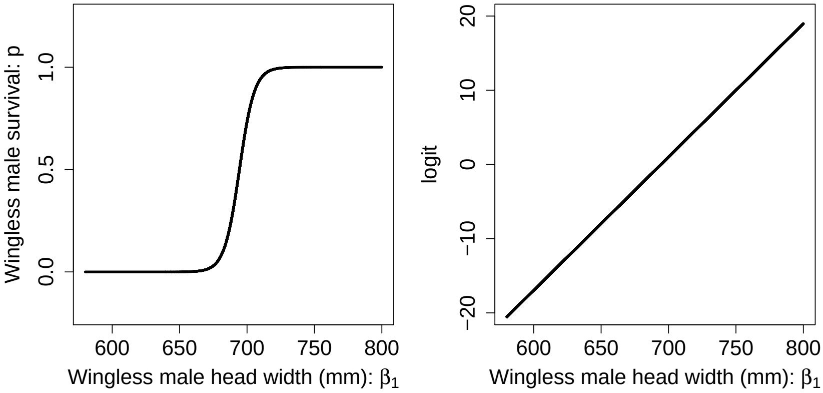
Generalised linear model with binomial data

- Use a **logit link** to ensure fitted values are bounded appropriately
- Use a **binomial family** to specify the appropriate error variance
- If the data are counts of two outcomes, then bind columns to create a two-vector response
- If the data are a binary outcome (e.g., survive or not), then leave as is

Linear predictors and link functions: Binomial data

The logit link function linearises the binomial probability function

$$\ln\left(\frac{p}{q}\right) = \beta_0 + \beta_1 x$$



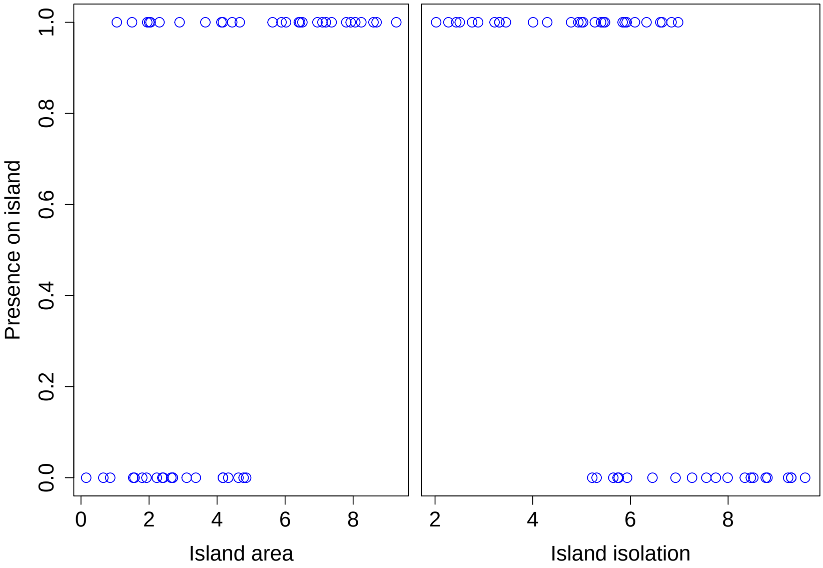
Linear predictors and link functions: Binomial data

Data on the incidence of breeding birds on islands (present or absent) as a function of area and isolation (distance from mainland)

```
## Parsed with column specification:
## cols(
##   INCIDENCE = col_double(),
##   AREA = col_double(),
##   ISOLATION = col_double()
## )
```

```
## # A tibble: 50 x 3
##    INCIDENCE  AREA ISOLATION
##        <dbl> <dbl>     <dbl>
##  1         1  7.93      3.32
##  2         0  1.92      7.55
##  3         1  2.04      5.88
##  4         0  4.78      5.93
##  5         0  1.54      5.31
##  6         1  7.37      4.93
##  7         1  8.60      2.88
##  8         0  2.42      8.77
##  9         1  6.40      6.09
## 10         1  7.20      6.98
## # ... with 40 more rows
```

Linear predictors and link functions: Binomial data

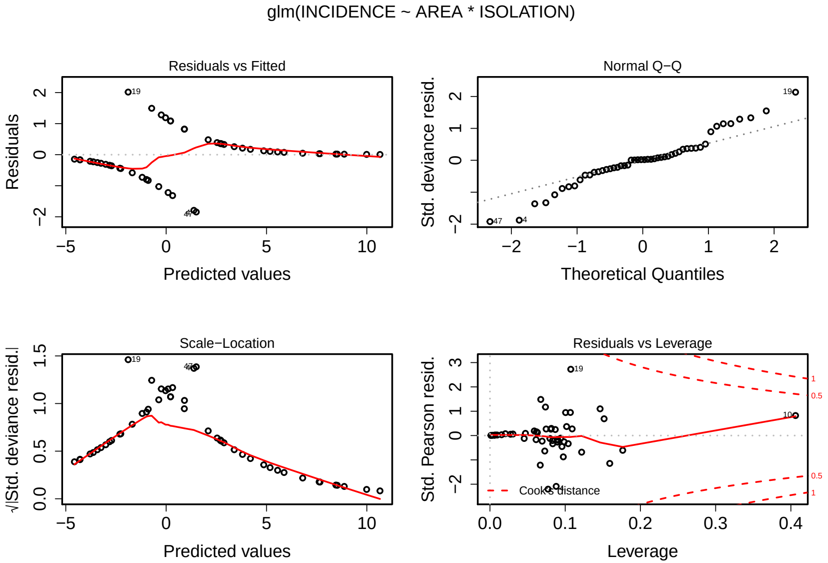


Linear predictors and link functions: Binomial data

```
model1 <- glm(INCIDENCE ~ AREA * ISOLATION, data = ISLAND, family = binomial);
model2 <- glm(INCIDENCE ~ AREA + ISOLATION, data = ISLAND, family = binomial);
anova(model1, model2, test = "Chi");
```

```
## Analysis of Deviance Table
##
## Model 1: INCIDENCE ~ AREA * ISOLATION
## Model 2: INCIDENCE ~ AREA + ISOLATION
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        46     28.252
## 2        47     28.402 -1 -0.15043   0.6981
```

Linear predictors and link functions: Binomial data

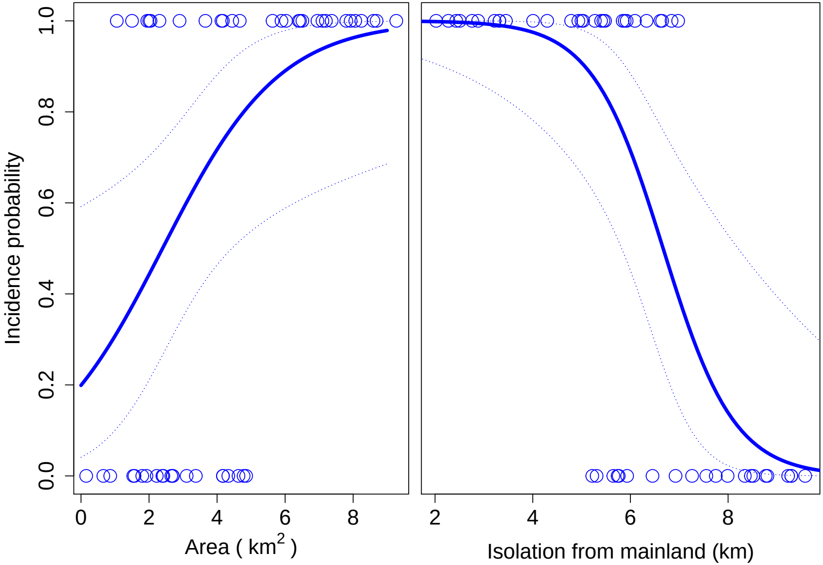


Linear predictors and link functions: Binomial data

Output from `summary(model2)`:

```
##
## Call:
## glm(formula = INCIDENCE ~ AREA + ISOLATION, family = binomial,
##     data = ISLAND)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8189  -0.3089   0.0490   0.3635   2.1192
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.6417     2.9218   2.273  0.02302 *
## AREA          0.5807     0.2478   2.344  0.01909 *
## ISOLATION    -1.3719     0.4769  -2.877  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.029  on 49  degrees of freedom
## Residual deviance: 28.402  on 47  degrees of freedom
## AIC: 34.402
##
## Number of Fisher Scoring iterations: 6
```

Linear predictors and link functions: Binomial data



Further reading suggestions

- Logan, M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley & Sons. **(Chapter 17)**
- Crawley, MJ. 2012. *The R book*. John Wiley & Sons. **(Chapters 13, 14, 16)**
- Generalised Linear Mixed Models: http://glmm.wikidot.com
- Rencher, AC, & Schaalje, GB. 2008. *Linear models in statistics*. John Wiley & Sons, 446-448.
- Nelder, JA, & Wedderburn, RW. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135:370-384. [PDF]