Poisson Models

Dr Josh Hodge

J.Hodge@imperial.ac.uk

Intended Learning Outcomes

Students will be able to:

- Fit and interpret a generalized linear model of family Poisson
- Validate a Poisson model
- Explain when an alternative model should be fitted
- Fit and interpret a quasi-Poisson and negative binomial models

Fitting a Poisson GLM

- Number of eggs laid and female body size of vapourer moth:

1. Distribution of response = Poisson

2. Predictor function:

$$\text{Number of Eggs} = \beta_0 + \beta_1 X\ logbodymass$$

3. Link between the predictor and the mean of the distribution: log-linear:

$$\mathbf{log}(\text{Number of Eggs}) = \beta_0 + \beta_1 X\ logbodymass$$

Moth Eggs Example

```
Call:
glm(formula = RedEggs ~ logBodyMass, family = "poisson", data = motheggs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0076  -1.7330  -0.3676   1.5132   2.2396

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.3994     0.2977  -24.86   <2e-16 ***
logBodyMass   6.6991     0.1625   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3151.44  on 38  degrees of freedom
Residual deviance:  180.11  on 37  degrees of freedom
AIC: 365

Number of Fisher Scoring iterations: 5
```

$$\log(\textit{Number of Eggs}) = -7.40 + 6.70X\ \textit{logbodymass}$$

Interpreting Coefficients

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.3994     0.2977  -24.86   <2e-16 ***
logBodyMass   6.6991     0.1625   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\log(\text{Number of Eggs}) = -7.40 + 6.70X\ logbodymass$$
$$\text{Number of Eggs} = e^{-7.40+6.70X\ logbodymass}$$

- For a log increase in body mass increases moth eggs by a natural log factor of 6.70 **OR**
- For a log increase in body mass increases moth eggs by a factor of $e^{6.70}$ or 812.41-fold. ← **EFFECT SIZE**

Fold Change

- For fold changes if e^{β_1} is greater than 1 = positive effect, if less than 1 negative effect

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.3994     0.2977  -24.86   <2e-16 ***
logBodyMass   6.6991     0.1625   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

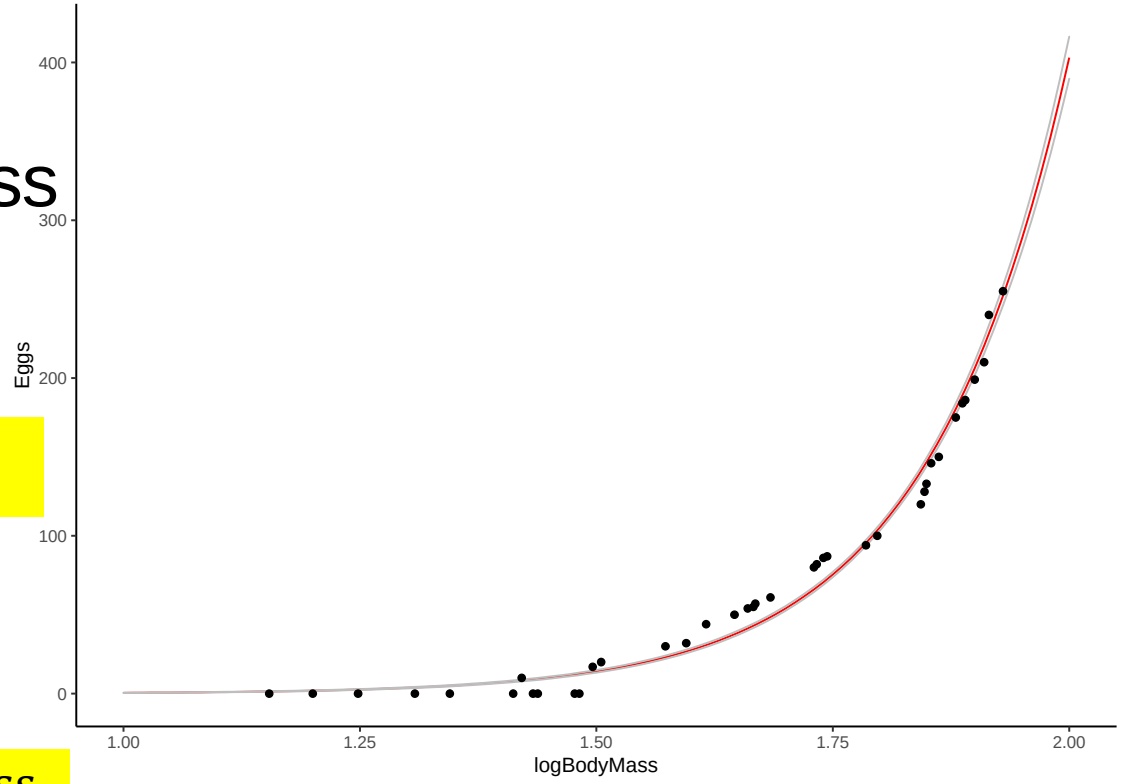
- $e^{6.70} = 812.41$ fold (positive effect)
- $e^{-6.70} = 0.001$ fold (negative effect)

Why Fold Change?

- For a log increase in body mass increases moth eggs by a factor of $e^{6.70}$ or 812.41-fold **???**
- The fitted slope is not the same across the log body mass distribution – it is multiplicative.

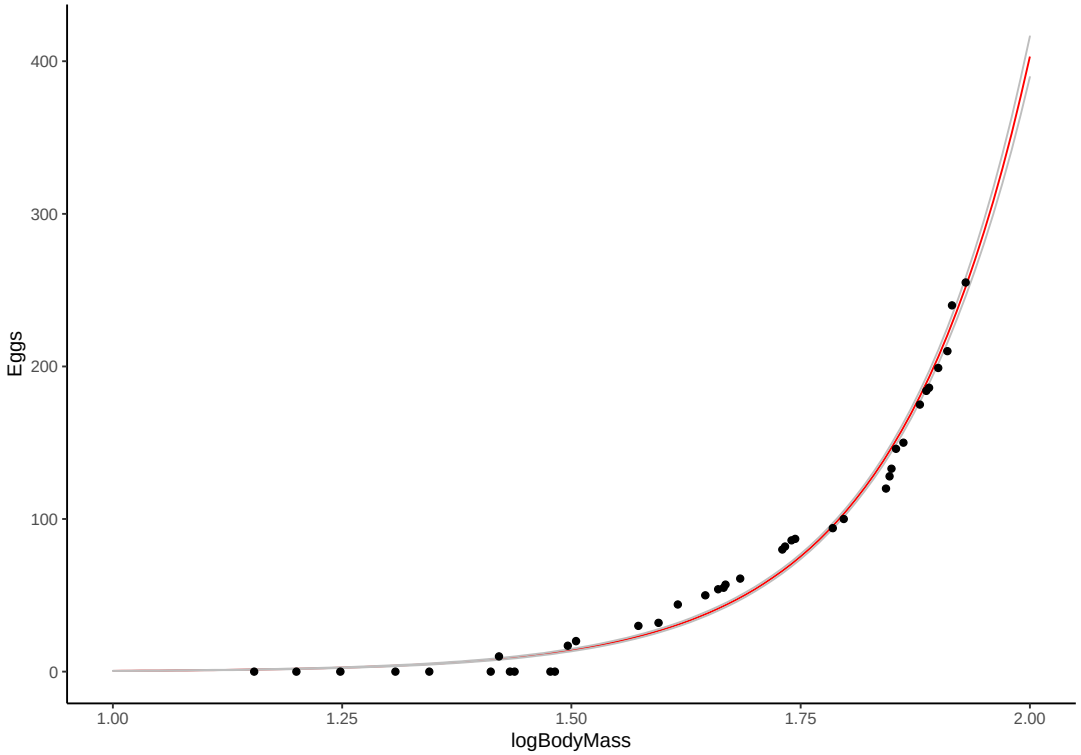
$$\text{Number of Eggs} = e^{-7.40+6.70X\ logbodymass}$$

$$\text{Number of Eggs} = e^{-7.40} * e^{6.70X\ logbodymass}$$



Moth Eggs Example

```
Call:
glm(formula = RedEggs ~ logBodyMass, family = "poisson", data = motheggs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0076  -1.7330  -0.3676   1.5132   2.2396

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.3994     0.2977  -24.86   <2e-16 ***
logBodyMass   6.6991     0.1625   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3151.44  on 38  degrees of freedom
Residual deviance:  180.11  on 37  degrees of freedom
AIC: 365

Number of Fisher Scoring iterations: 5
```



Null & Residual Deviance

```
    Null deviance: 3151.44  on 38  degrees of freedom
Residual deviance:  180.11  on 37  degrees of freedom
```

- Null summarises how well the response variable is predicted by a null model
- Residual summarises how well the response variable is predicted by current model
- Both used to estimate goodness-of-fit for model
- Pseudo-R^2 : $1 - \left({}^{Residual\ Deviance}/_{Null\ Deviance}\right)$
  - 1-(180.11/3151.44) = 0.94

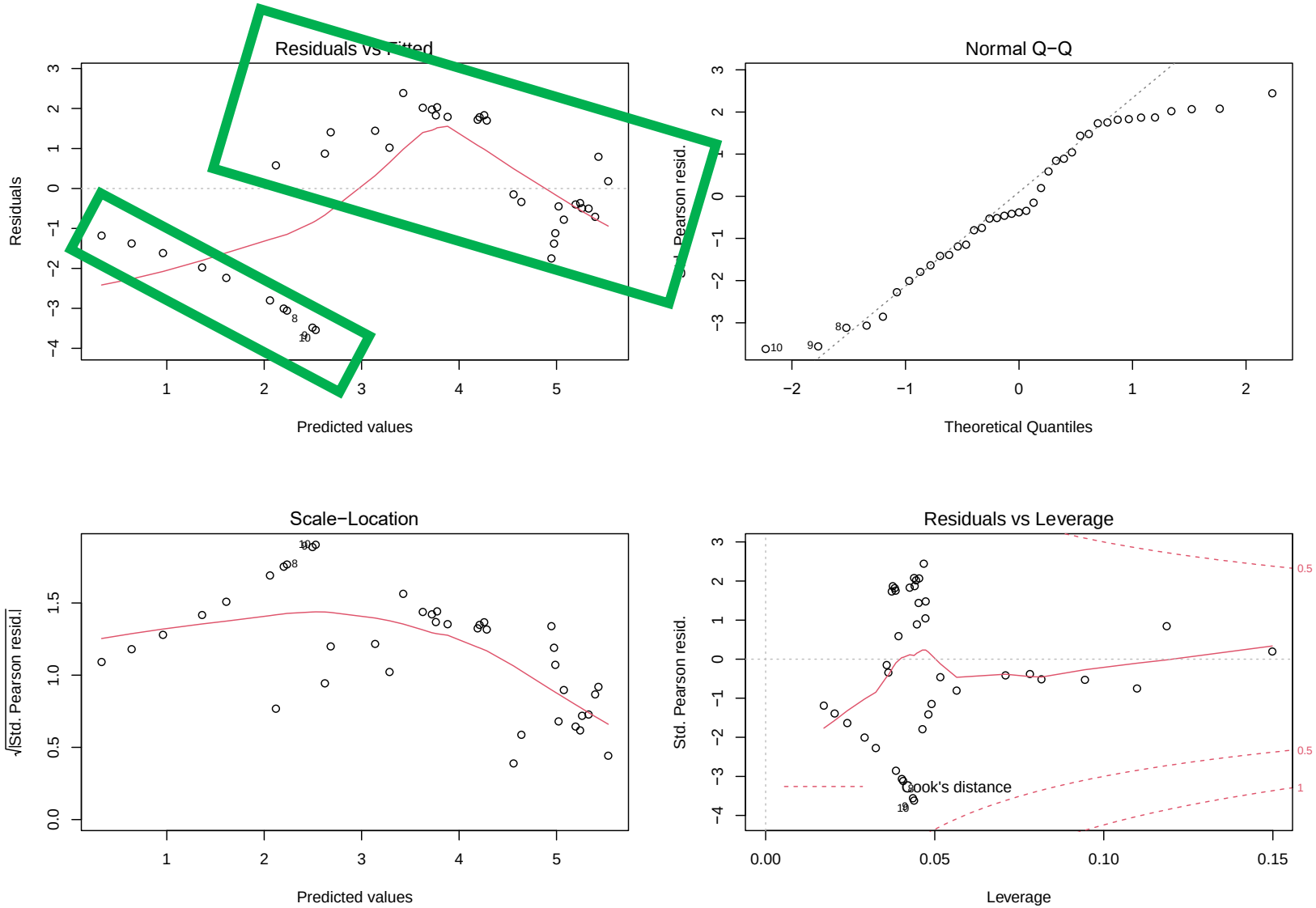
Goodness-of-fit

- Estimated using goodness-of-fit **chi-squared test** synomynous with the F-test for linear models

- Tests H_0 that **fitted model and variables** is not different from the **null model**

```
> anova(M1, test = "Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: RedEggs

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            38    3151.44
logBodyMass  1   2971.3        37     180.11 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Validation - Diagnostics



Dispersion

```
(Dispersion parameter for poisson family assumed to be 1)
```

- Dispersion refers to the theoretical or expected amount of variability based on Poisson distribution assumption.
- Overdispersion is very common in Poisson models and can be due to a whole host of factors:
  - Too simplistic (missing explanatory variables and/or interaction terms)
  - Explanatory variables measured on different scales
  - A covariate has a non-linear effect
  - One or more outliers
  - Zero inflation
  - Inherent dependency in the data – i.e. pseudoreplication

Dispersion Parameter

- Tells us how much larger or smaller is our conditional variance to our conditional mean
- Dispersion parameter should equal 1
  - >1 overdispersion
  - <1 underdispersion

```
    Null deviance: 3151.44  on 38  degrees of freedom
Residual deviance:  180.11  on 37  degrees of freedom
```

180.11/37=4.87

- What level of dispersion is acceptable? **Still debated**

- So, what next?

Accounting for Overdispersion

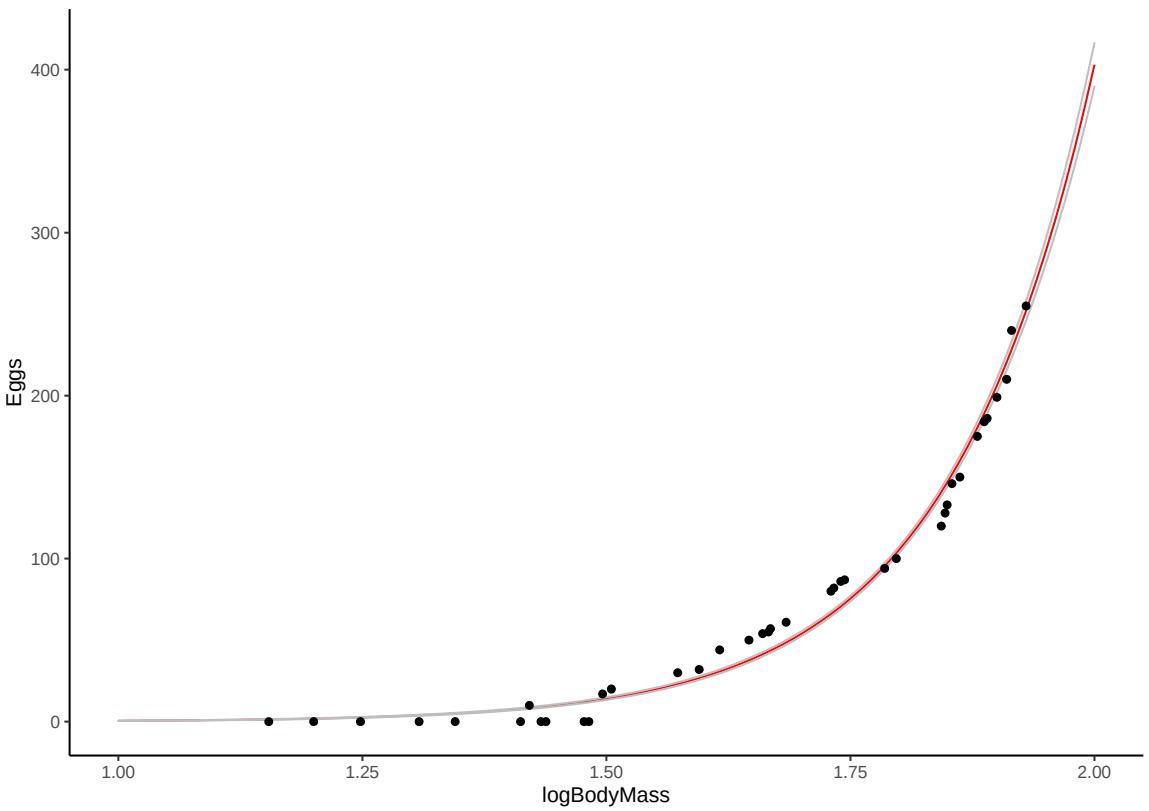
- Quasi-likelihood methods
  - Assumes overdispersion only impacts the standard errors
  - Standard errors are adjusted by scaling

```
Call:
glm(formula = RedEggs ~ logBodyMass, family = "poisson", data =

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0076  -1.7330  -0.3676   1.5132   2.2396

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.3994     0.2977  -24.86   <2e-16 ***
logBodyMass   6.6991     0.1625   41.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3151.44  on 38  degrees of freedom
Residual deviance:  180.11  on 37  degrees of freedom
AIC: 365

Number of Fisher Scoring iterations: 5
```

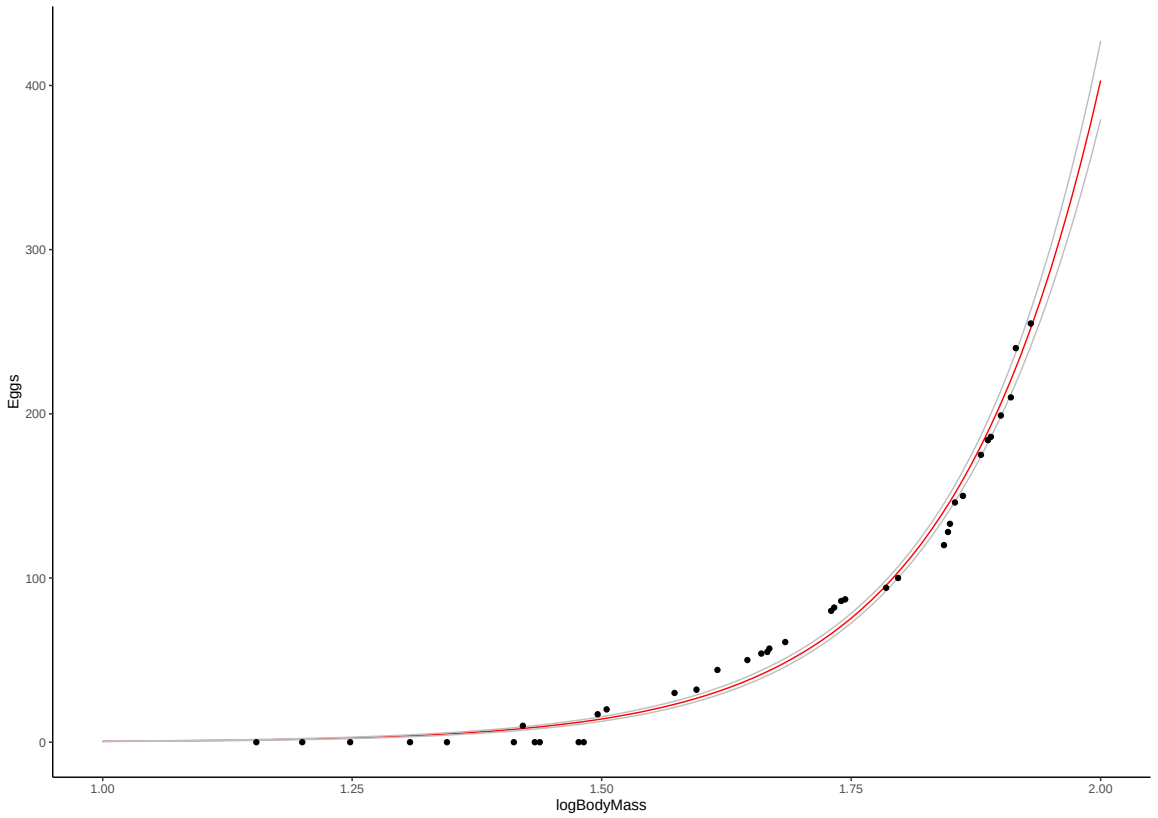
```
Call:
glm(formula = RedEggs ~ logBodyMass, family = "quasipoisson"
    data = motheggs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0076  -1.7330  -0.3676   1.5132   2.2396

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.3994     0.5310  -13.94 2.68e-16 ***
logBodyMass   6.6991     0.2898   23.12  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 3.181806)

    Null deviance: 3151.44  on 38  degrees of freedom
Residual deviance:  180.11  on 37  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Accounting for Dispersion

Poisson



Quasi-Poisson



Accounting for Overdispersion – Negative Binomial Models

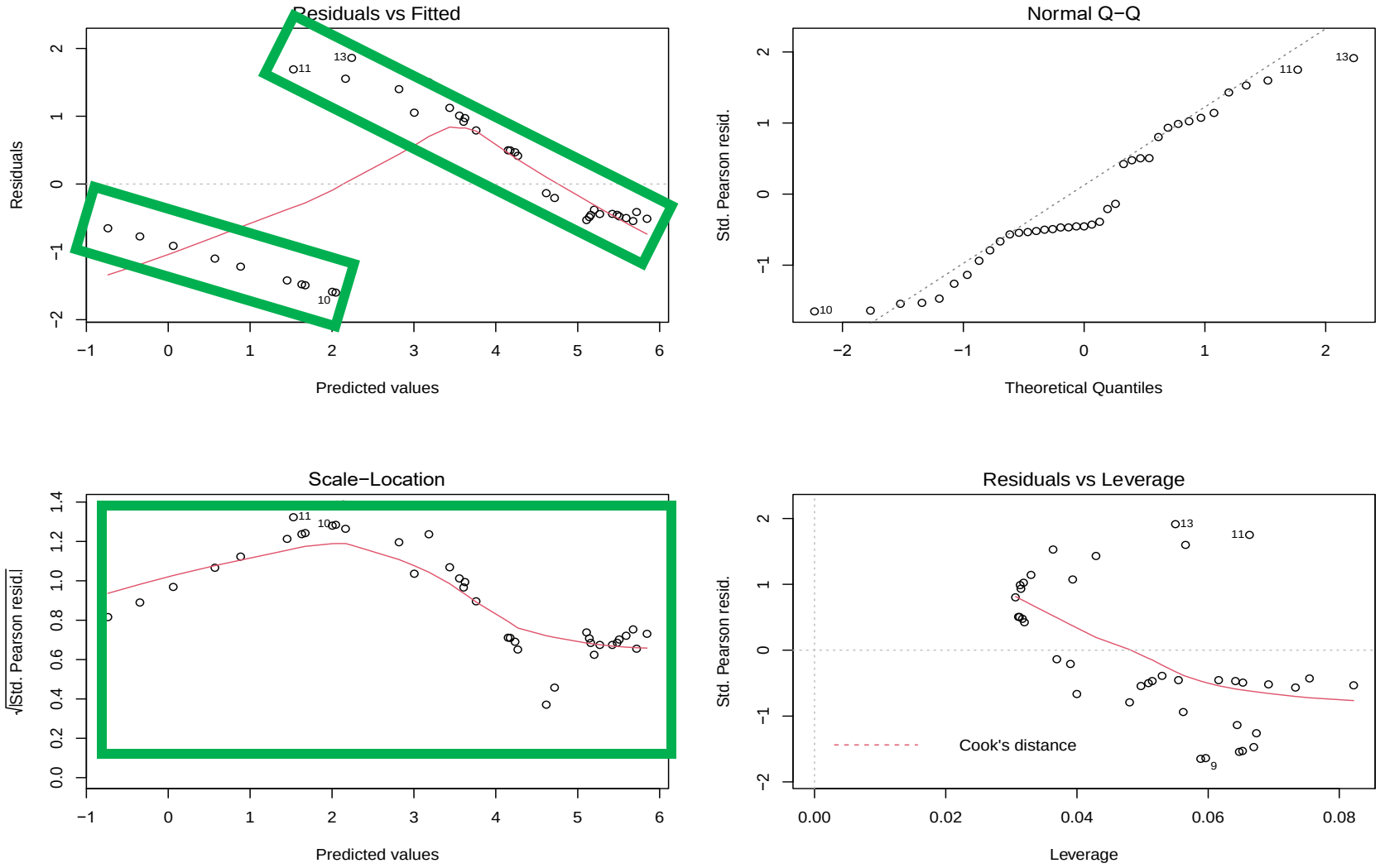
- Negative binomial approach
- Essentially, an overdispersed Poisson distribution converges with a negative binomial distribution.
- A negative binomial will give you different estimates and standard errors.
- Model Equation:

$$\text{Number of Eggs} = e^{-10.53+8.48X\ logbodymass}$$

- Pseudo-R^2 = 0.79

```
Call:
glm.nb(formula = RedEggs ~ logBodyMass, data = motheggs, init.theta = 3.833483026,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9100  -0.7783  -0.4785   0.7541   1.4682

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.5262     0.9025  -11.66   <2e-16 ***
logBodyMass   8.4830     0.5247   16.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.8335) family taken to be 1)

    Null deviance: 302.801  on 38  degrees of freedom
Residual deviance:  63.208  on 37  degrees of freedom
AIC: 341.68

Number of Fisher Scoring iterations: 1


              Theta:  3.83
          Std. Err.:  1.65

 2 x log-likelihood:  -335.679
```

Negative Binomial Models



What next?

- Might want to explore other options:
  - Too simplistic (missing explanatory variables and/or interaction terms)
  - Explanatory variables measured on different scales
  - A covariate has a non-linear effect
  - One or more outliers
  - Zero inflation
  - Inherent dependency in the data – i.e. pseudoreplication
- If you're interested look up:
  - Offsets
  - Zero inflated and truncated models
  - Generalised linear mixed models

1. "A Beginner's Guide to GLM and GLMM with R" – Zuur, Hilbe & N leno (2015 - Book)
2. "A brief introduction to mixed effects modelling and multi-model inference in ecology" Harrison *et al* (2018- Paper)

Summary

- Poisson models with the log-linear link function are able to handle Poisson data that are positive integers
- The estimates are on this log-linear scale and require exponentiating, which expresses the slope coefficient as a factor due to the multiplicative nature
- Poisson models can be validated by examining the diagnostic plots and the dispersion parameter
- Overdispersion can be caused by a whole host of factors
- A Quasi-likelihood approach or a negative binomial model can be fitted to account for overdispersion