Poisson and Binomial Models

Dr Josh Hodge

J.Hodge@imperial.ac.uk

Intended Learning Outcomes

Students will be able to:

- Differentiate between a linear model and generalised linear models of binomial and poisson data
- Interpret the R outputs of a generalized linear model
- Modify GLM approach according to the dispersion parameter

LMs vs GLMS

- Response variable data type
  - Unconstrained vs Constrained

- Model fitting approach
  - Ordinary Least Squares vs Maximum Likelihood

- Assumptions
  - Means and variance

Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Response Variable

Intercept

Slope of Explanatory Variable

Error Term

Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

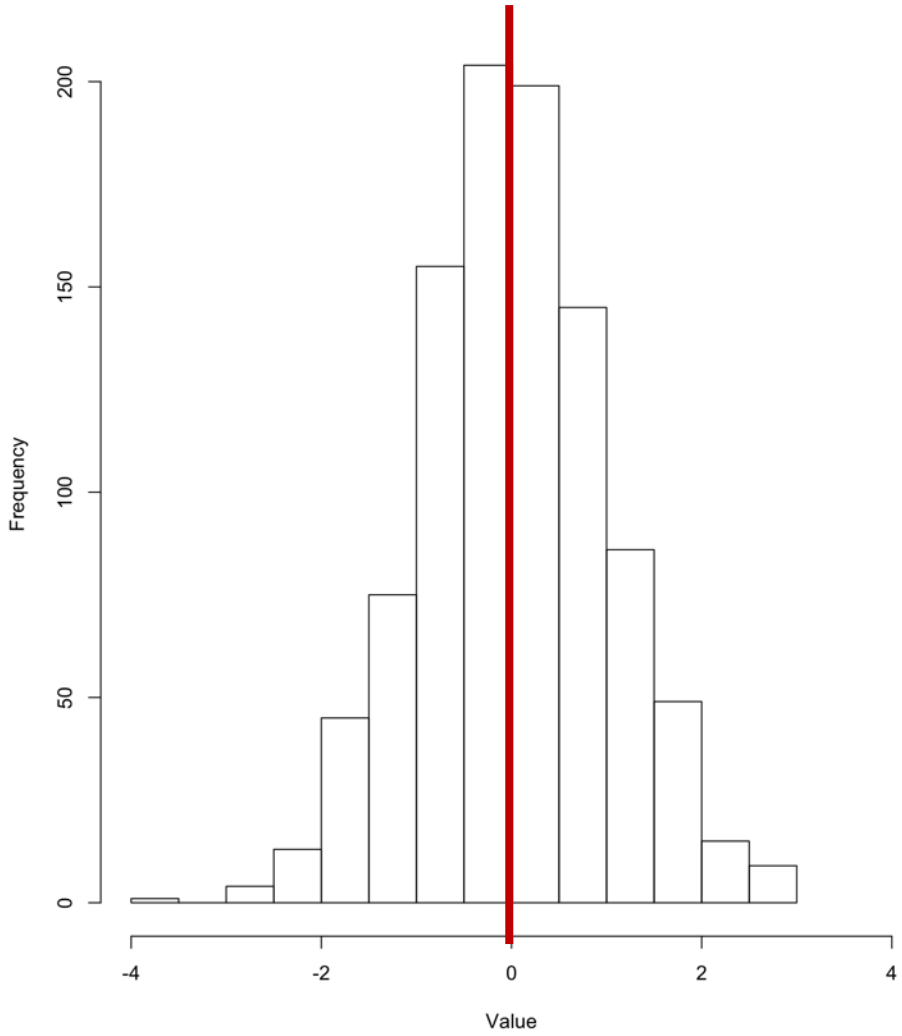
Response Variable

Intercept

Slope of Explanatory Variable

Error Term

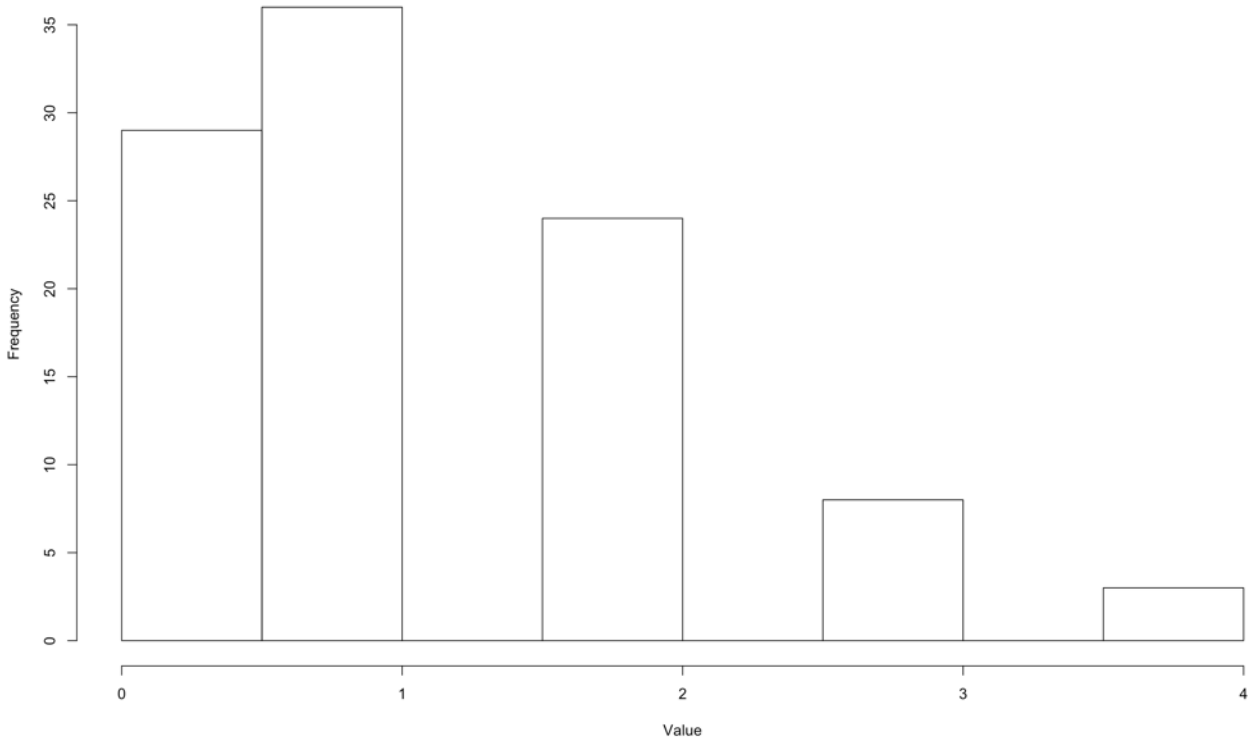
Data Distributions- Normal



- -Infinity to Infinity
- Mean represents the centre of the data

Data Distributions- Poisson

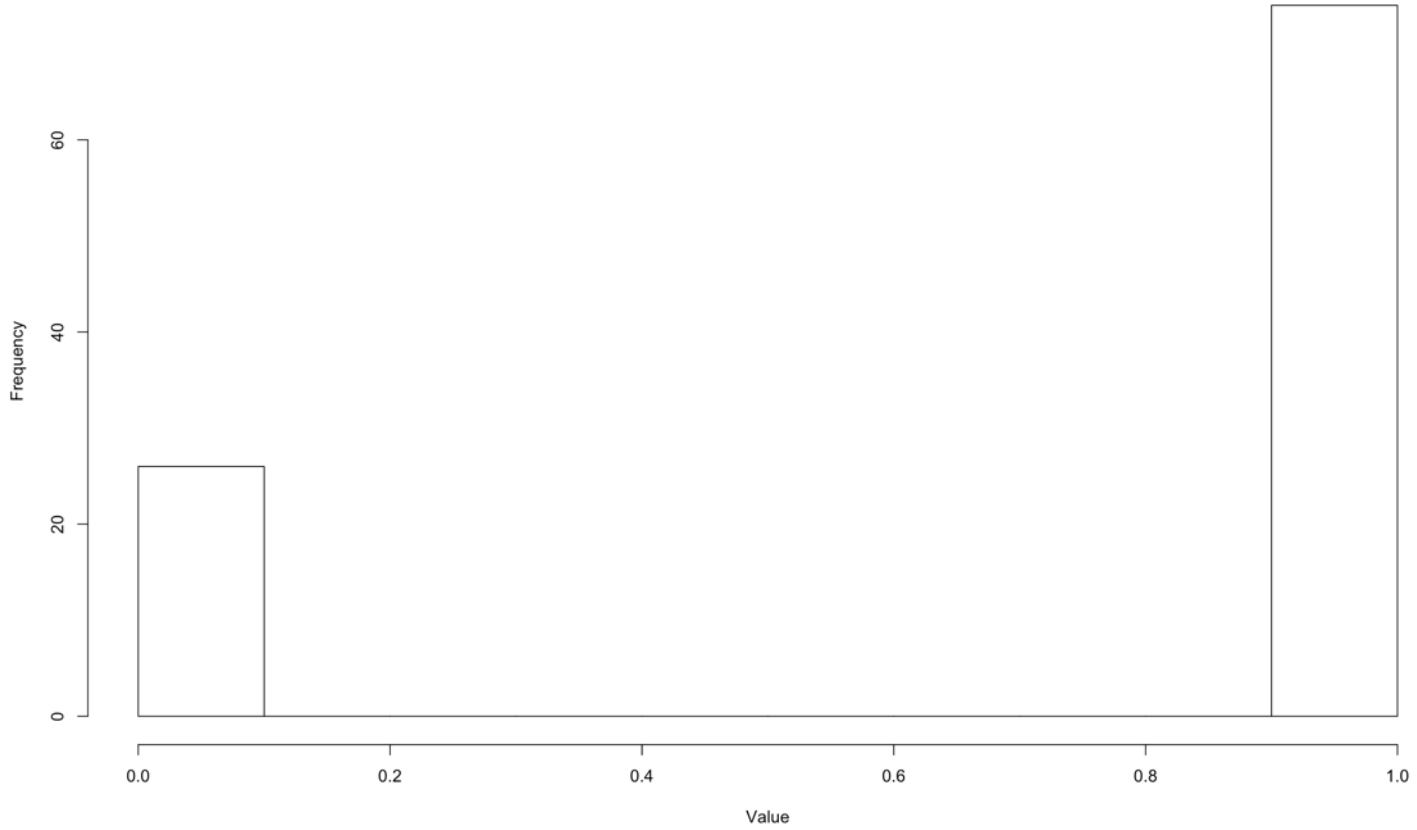
- Count Data
- **Constrained** to absolute whole numbers
- Typically right skewed
- Examples:
  - Number of Species
  - Number of Enzymes
  - Heartbeat
  - Number of Offspring



Data Distributions- Binomial

- **Constrained** between 0 and 1

- Examples:
  - Proportions
  - Presence/Absence Data



LMs vs GLMS

- Response variable data type
  - Unconstrained vs Constrained

- Model fitting approach
  - Ordinary Least Squares vs Maximum Likelihood Estimation

- Assumptions
  - Means and variance

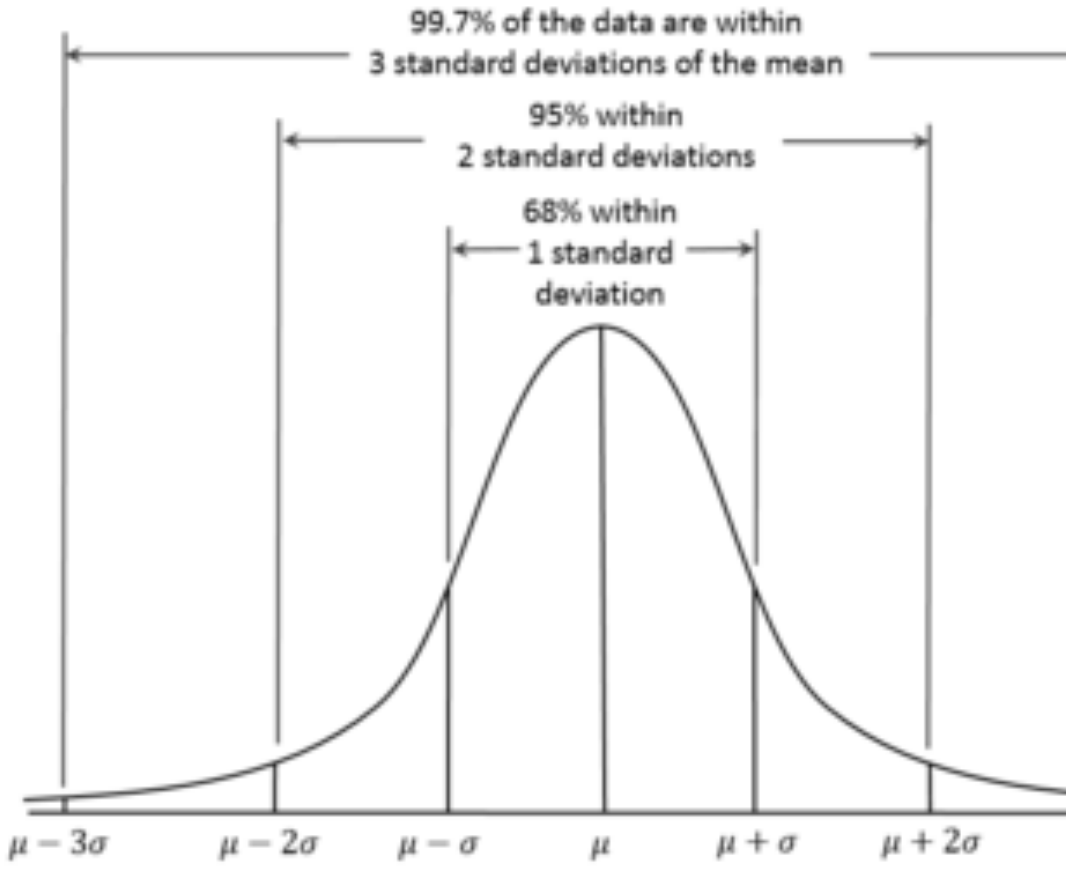
OLS vs MLE

- Model fitting approach
  - Ordinary Least Squares vs Maximum Likelihood Estimation

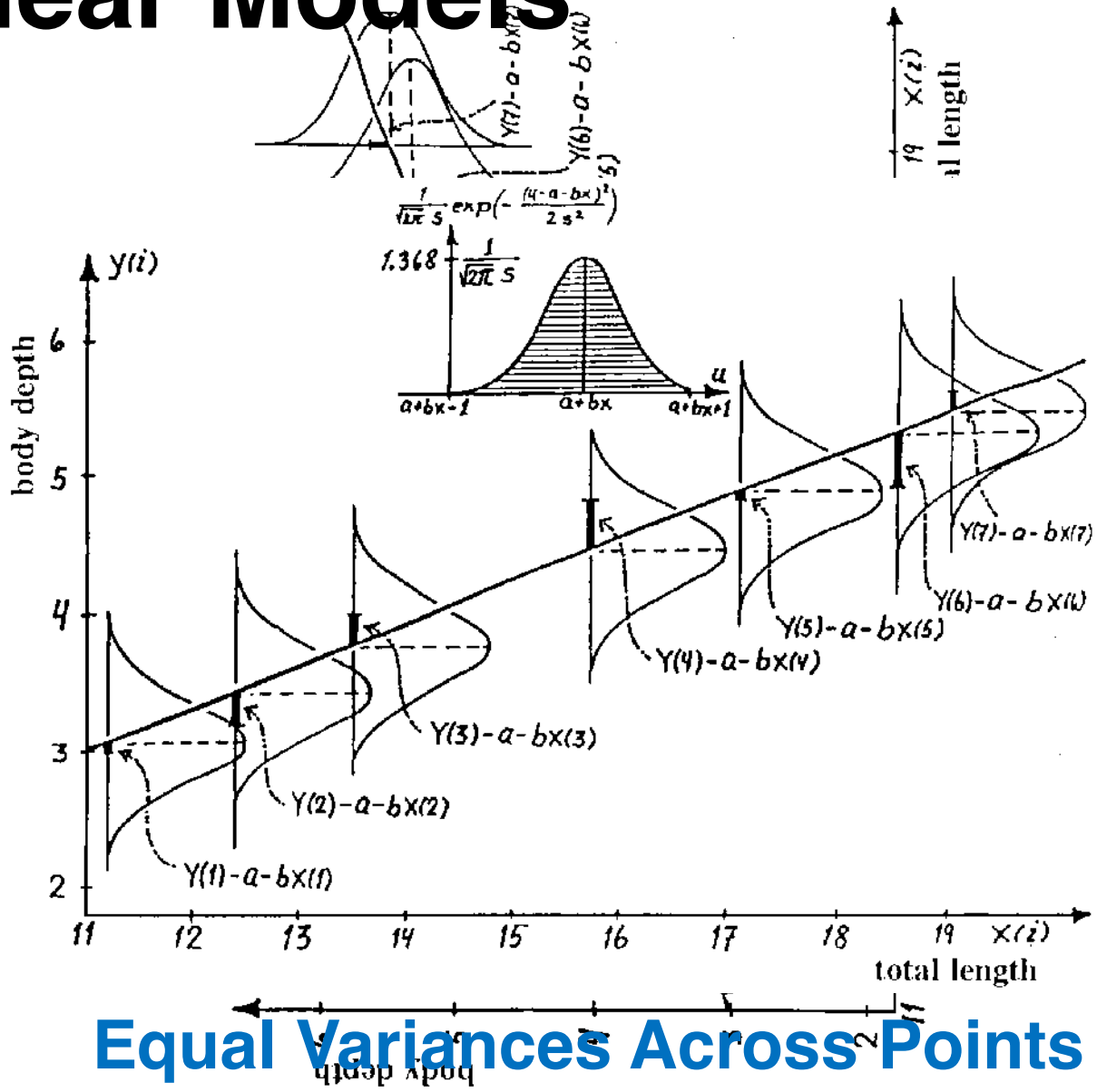
Fits a line that minimizes the residual sum of squares

Fits a line that maximises the log-likelihood

Assumptions of Linear Models



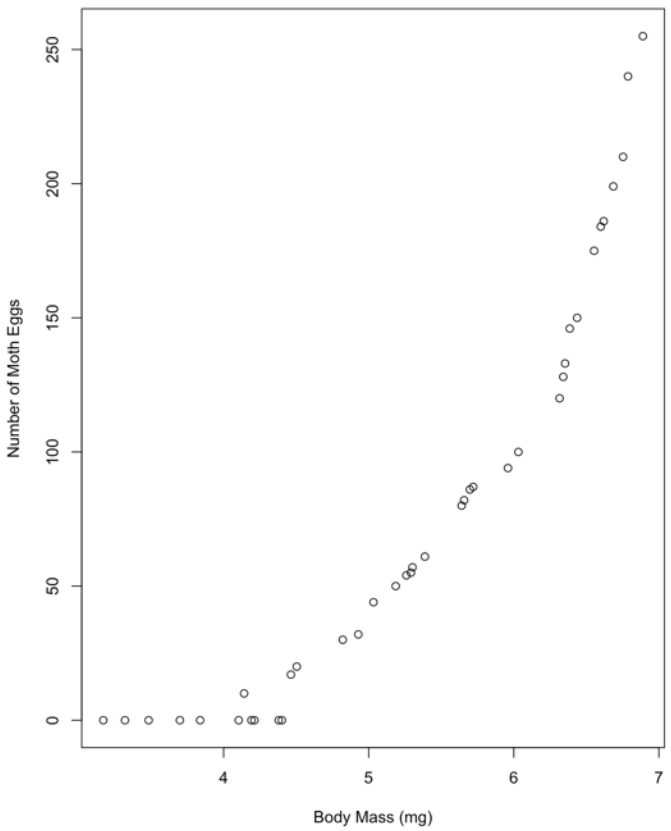
Normally Distributed



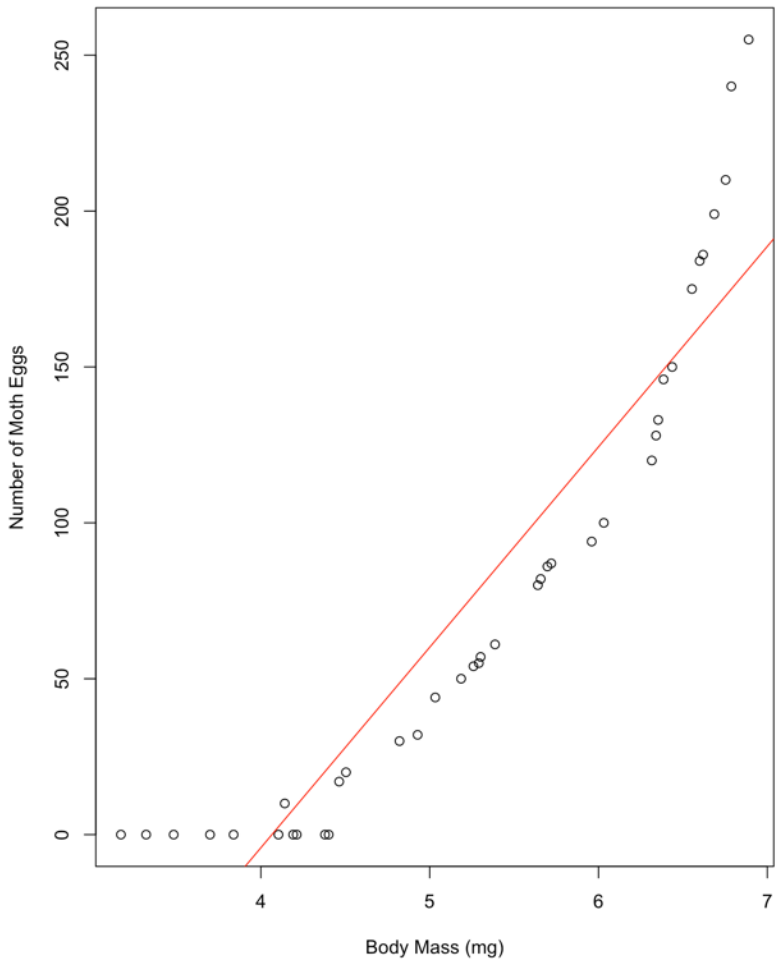
Equal Variances Across Points

Example

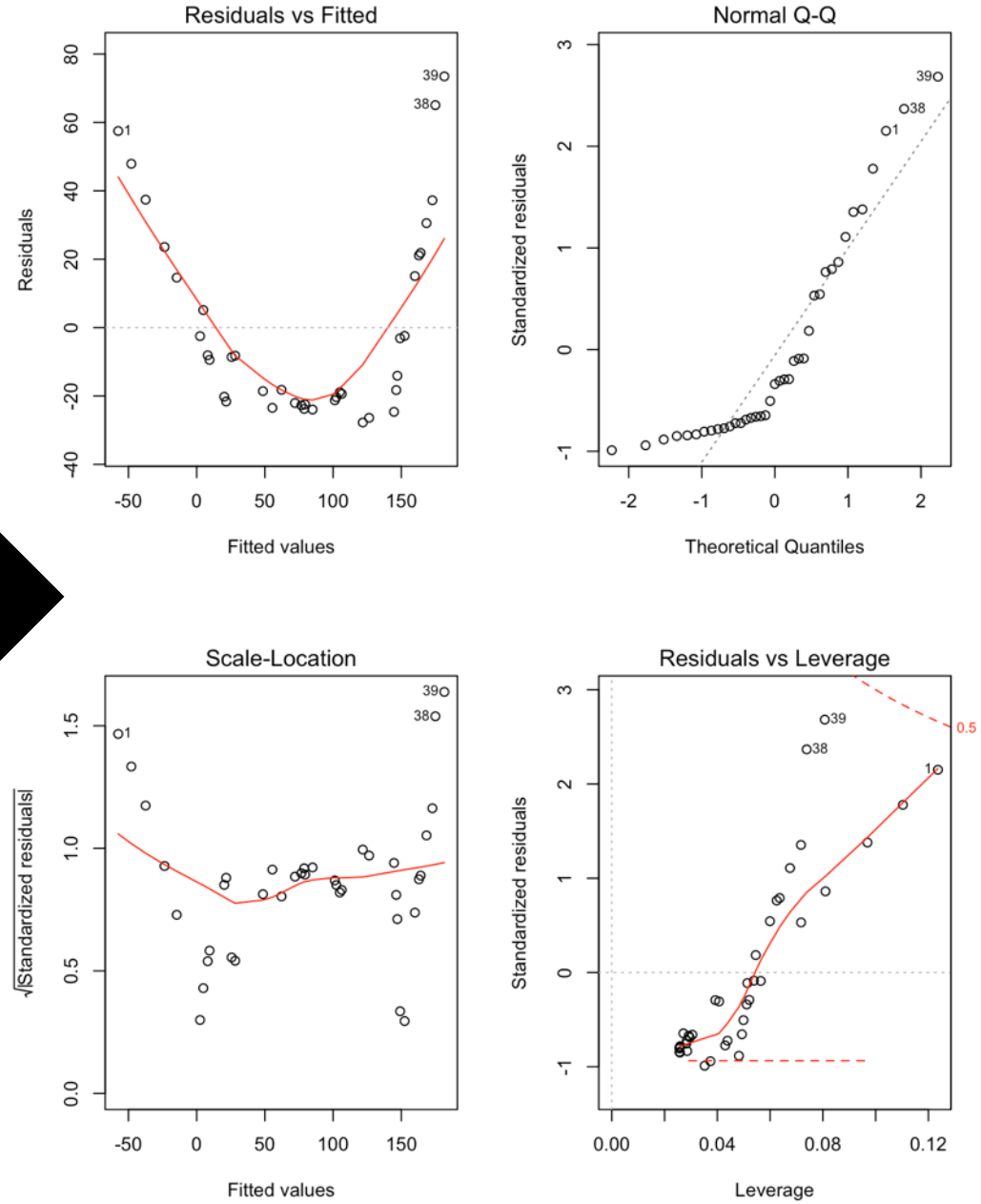
- Number of eggs laid and female body size of vapourer moth



Example



Diagnostics

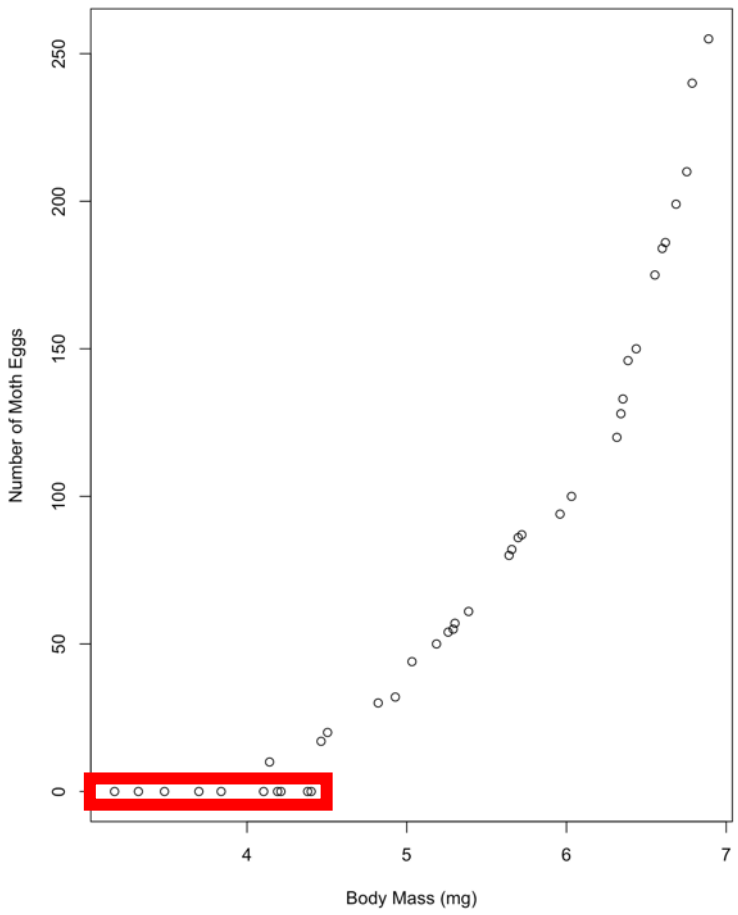


Mentimeter

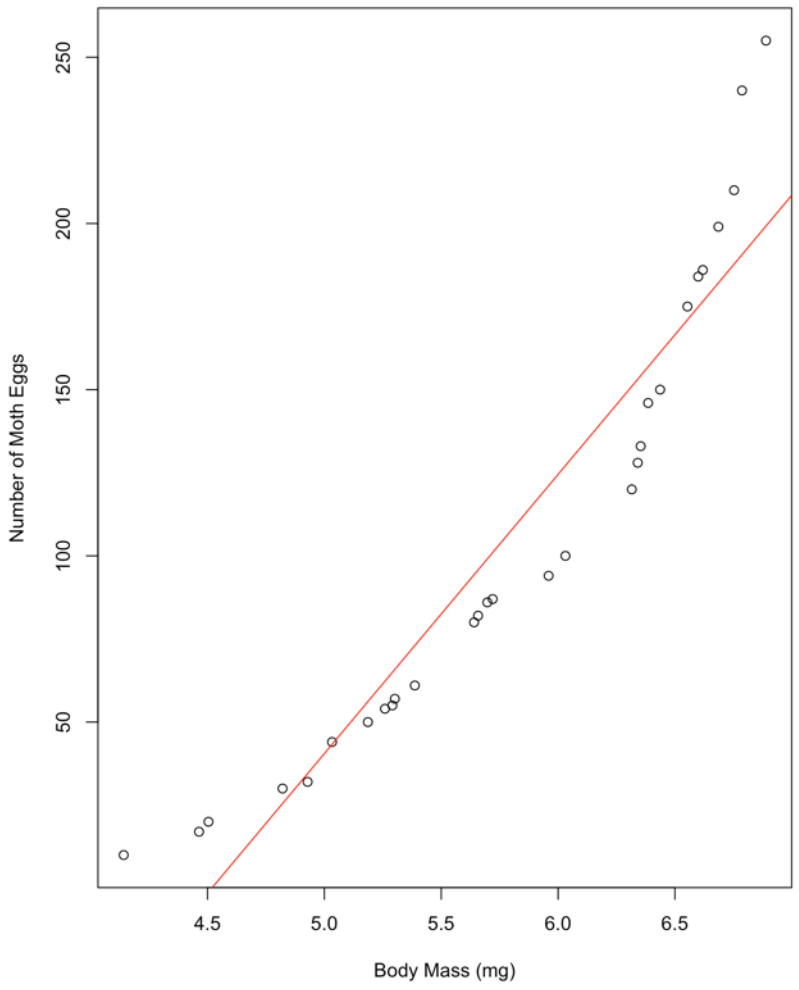
Mentimeter

Solutions?

- Log transformations of count data



Lose ~26% of Data



Generalised Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Rather than transforming y_i a generalised linear model transforms the linear predictor.

$$\hat{Y} = h(\underline{\beta_0 + \beta_1 x_i + \varepsilon_i})$$

Predicted Response

Linear predictor

Generalised Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Rather than transforming y_i a generalised linear model transforms the linear predictor/model.

$$\hat{Y} = \boldsymbol{h}(\underline{\beta_0 + \beta_1 x_i + \varepsilon_i})$$

Link function

Predicted Response

Linear predictor

Link Functions

Normal	Poisson (Counts)	Binomial
Leaf Mass	Number of Species	0,1,1,0
Bill Width	Number of Enzymes	Female, Male, Female, Male
Height	Number of Offspring	Absent, Present, Present, Absent
Weight	0, 1, 2, 3, 7	Heads, Heads, Heads, Tails
Identity	**Log-linear (natural)**	**Logit**

Link Functions - Normal

$$\hat{Y} = \underline{\beta_0 + \beta_1 x_i + \varepsilon_i}$$

Predicted Response

Linear predictor

Approx. follows

Constant variance

$$y_i \sim N(\bar{Y}_i, \sigma^2)$$

Observed Data

Mean of the Response

Normal Distribution

Link Functions

Normal

$$y_i \sim N(\bar{Y}_i, \sigma^2)$$

Poisson (Counts)

Scale Parameter

$$y_i \sim Poisson(\bar{Y}_i, \phi)$$

Mean==Variance

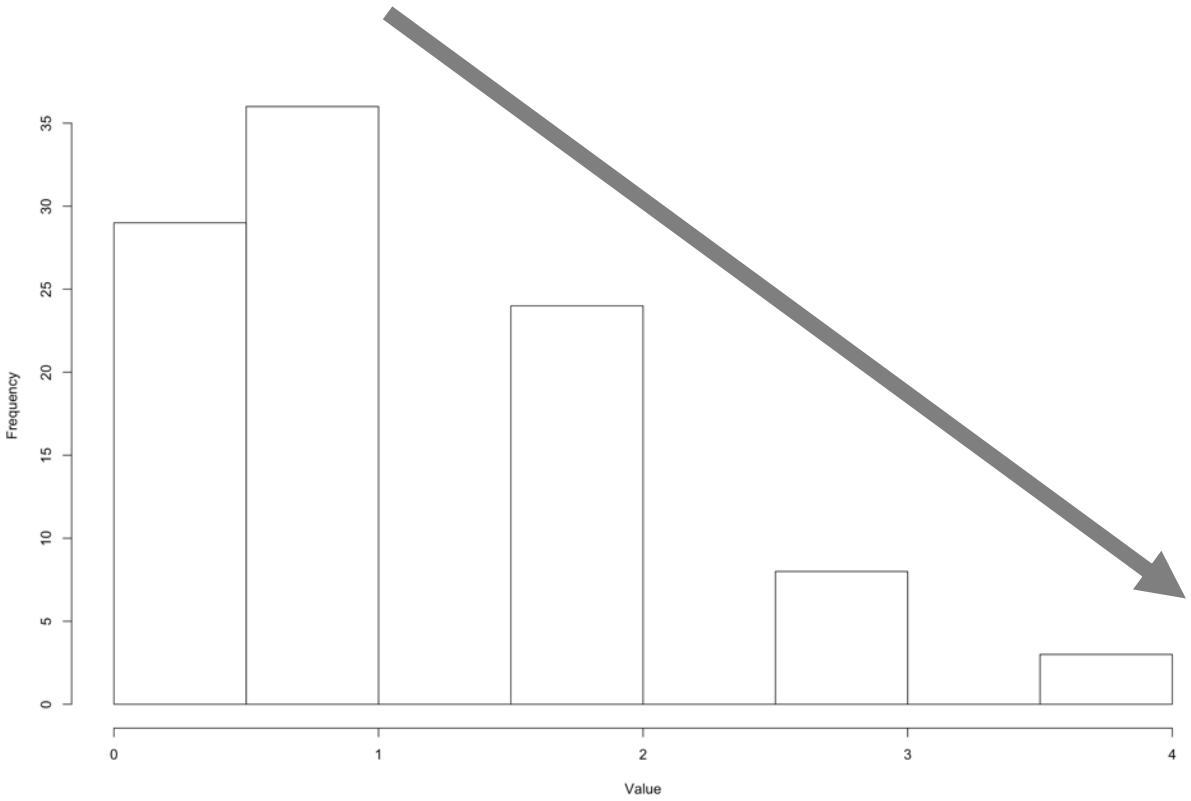
Binomial

Probability of success

$$y_i \sim Bernoulli(p)$$

Poisson Models

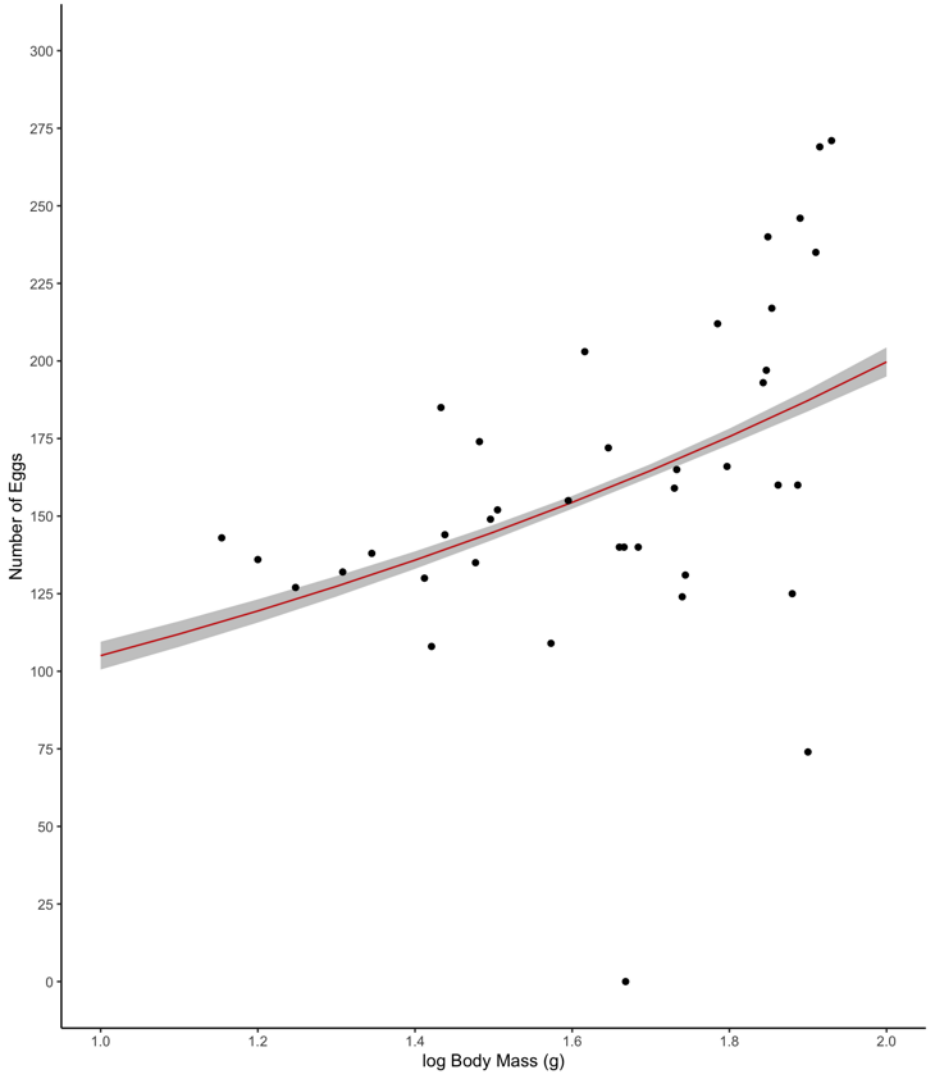
Recap of Poisson Data



- Cannot be less than zero
- Right-skewed distribution
- Mean == Variance
  - Increasing mean increases variance

Moth Eggs Example

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-17.9631   -1.7114    0.3496    2.4567    5.4615

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.01194    0.10205   39.31   <2e-16 ***
logBodyMass  0.64242    0.06051   10.62   <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: 775.07  on 38  degrees of freedom
Residual deviance: 658.93  on 37  degrees of freedom
AIC: 925.38

Number of Fisher Scoring iterations: 4
```



Interpreting Coefficients

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.01194    0.10205   39.31   <2e-16 ***
logBodyMass  0.64242    0.06051   10.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\hat{Y} \sim (\exp(4.01 + 0.64 * x)$$

- For a milligram increase in body mass increases moth eggs by a natural log factor of 0.64.

- $e^{0.64} = 1.90$ fold or by 90% ← **EFFECT SIZE**

Interpreting Coefficients

- For a milligram increase in body mass increases moth eggs by a natural log factor of 0.64.

- $e^{0.64} = 1.90$ fold or by 90% **????**

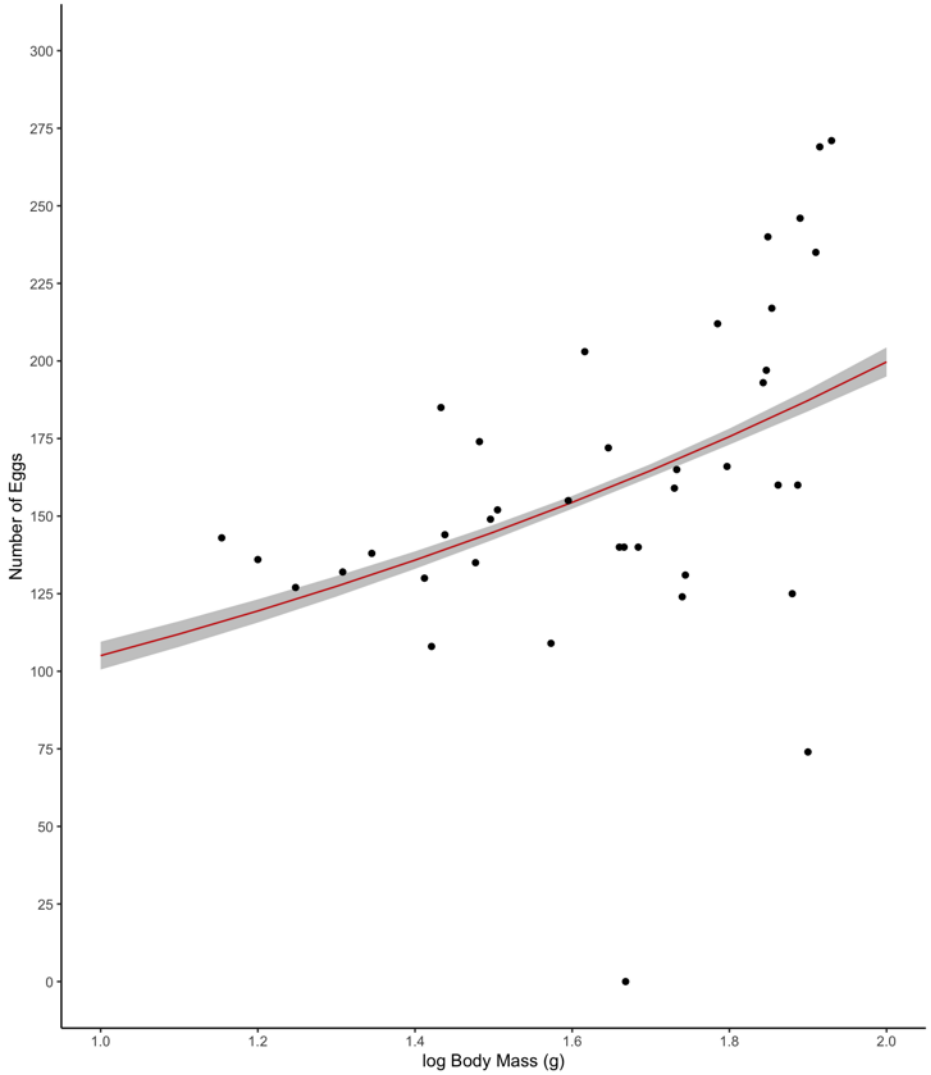
- How many eggs would a moth lay weighing log 1.5g?

- How many eggs would a moth lay weighing log 2.5g?

- What is the percentage difference between the log 1.5g and 2.5g?

Moth Eggs Example

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-17.9631   -1.7114    0.3496    2.4567    5.4615

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.01194    0.10205   39.31   <2e-16 ***
logBodyMass  0.64242    0.06051   10.62   <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: 775.07  on 38  degrees of freedom
Residual deviance: 658.93  on 37  degrees of freedom
AIC: 925.38

Number of Fisher Scoring iterations: 4
```



Null & Residual Deviance

```
    Null deviance: 775.07  on 38  degrees of freedom
Residual deviance: 658.93  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 : $1 - \left({}^{Residual\ Deviance}/_{Null\ Deviance}\right)$
  - 0.15

Goodness-of-fit

- Estimated using goodness-of-fit **chi-squared test**

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

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

Model: poisson, link: log

Response: Eggs

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            38     775.07
logBodyMass  1   116.14        37     658.93 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dispersion Parameter

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

- Poisson GLMs assumes the variance at a point in the model is equal to the prediction – the mean.

- If this is violated the model will suffer from dispersion

- Dispersion parameter should equal 1
  - >1 overdispersion
  - <1 underdispersion

```
> poisson$deviance/poisson$df.residual
[1] 17.80895
```

- Can be accounted for using quasi-Poisson family of GLMs

Dispersion Parameter

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

- Poisson GLMs assumes the variance at a point in the model is equal to the prediction – the mean.

- If this is violated the model will suffer from dispersion

- Dispersion parameter should equal 1
  - **>1 overdispersion**
  - <1 underdispersion

```
> poisson$deviance/poisson$df.residual
[1] 17.80895
```

- Can be accounted for using quasi-Poisson family of GLMs

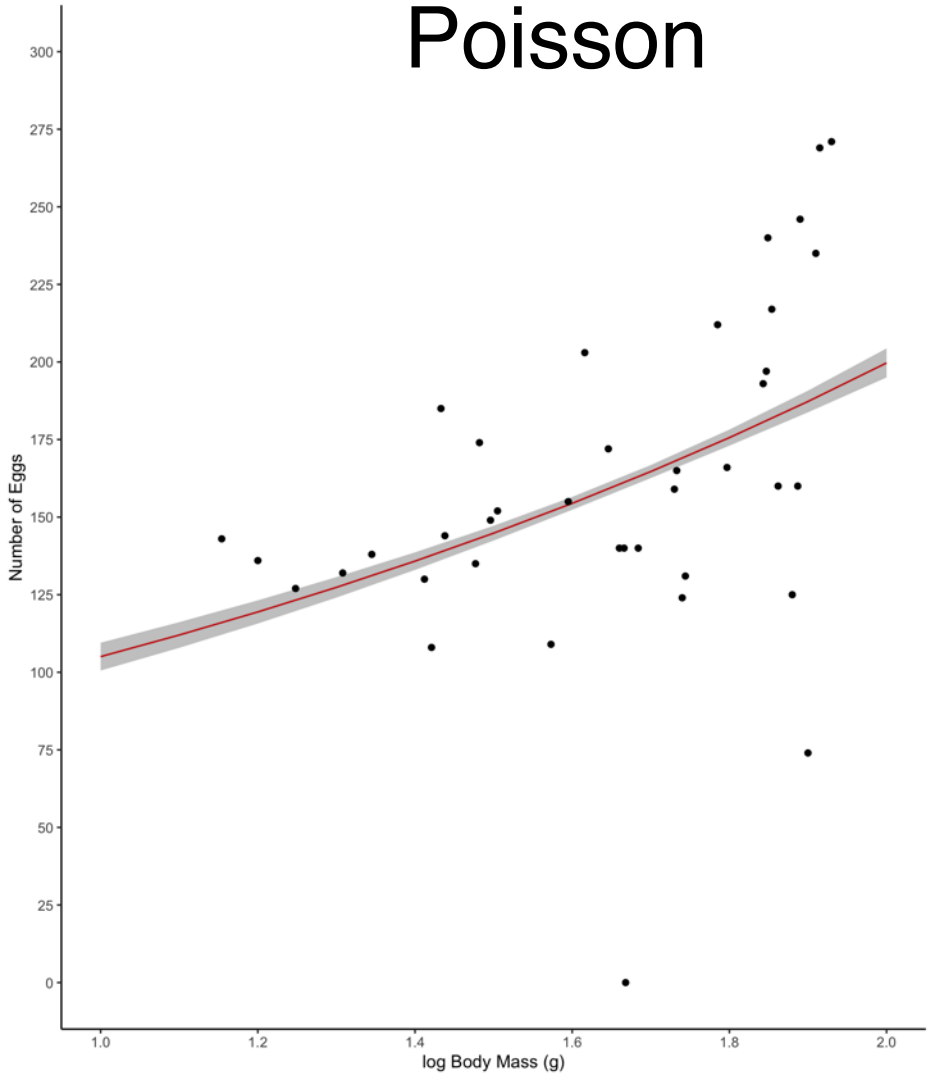
Accounting for Dispersion

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-17.9631   -1.7114    0.3496    2.4567    5.4615

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.01194    0.10205   39.31   <2e-16 ***
logBodyMass  0.64242    0.06051   10.62   <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: 775.07  on 38  degrees of freedom
Residual deviance: 658.93  on 37  degrees of freedom
AIC: 925.38

Number of Fisher Scoring iterations: 4
```

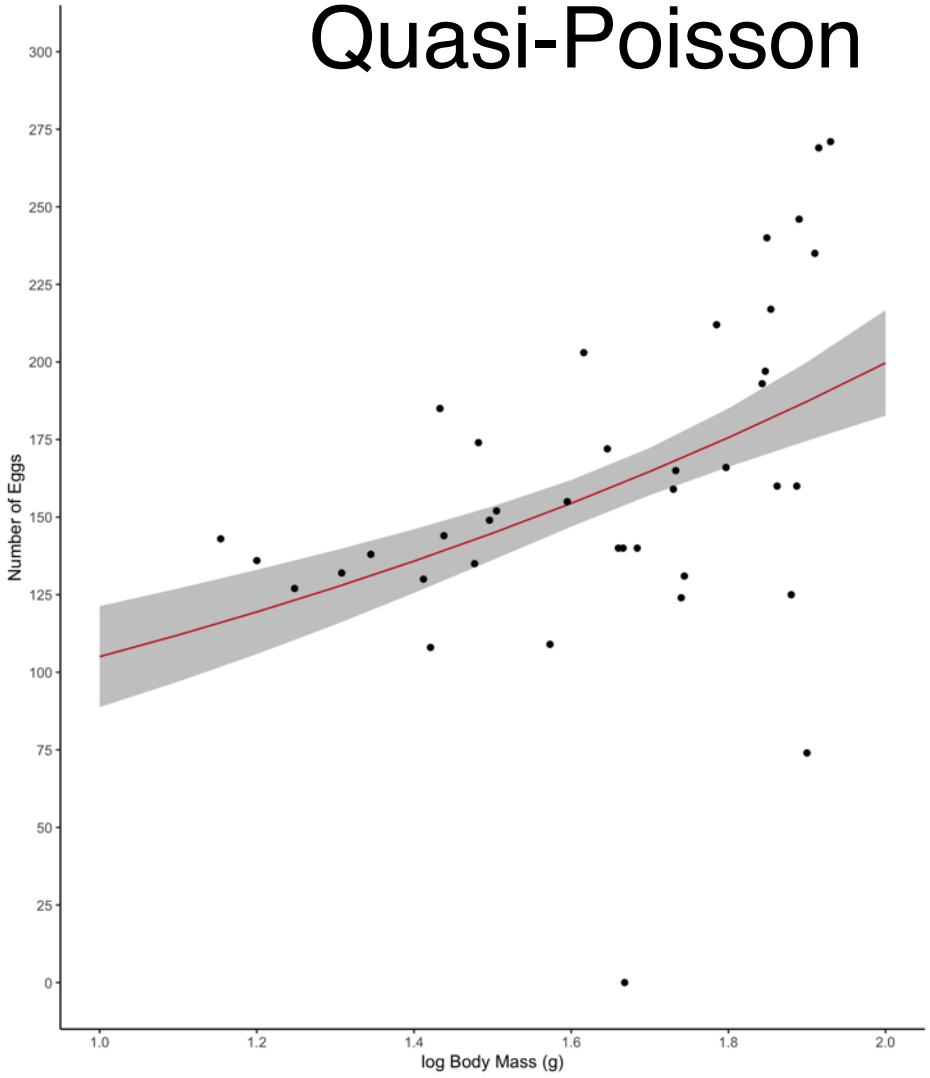
```
Call:
glm(formula = Eggs ~ logBodyMass, family = "quasipoisson", data = motheggs)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-17.9631   -1.7114    0.3496    2.4567    5.4615

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0119     0.3702  10.837  4.9e-13 ***
logBodyMass   0.6424     0.2195   2.927  0.00583 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 13.15988)

    Null deviance: 775.07  on 38  degrees of freedom
Residual deviance: 658.93  on 37  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Accounting for Dispersion

Poisson



Quasi-Poisson



HO 3

- Species richness on the Galapagos Islands
- Focus on the <u>statistical concepts and themes</u>

- Extra:
  - Amphibian roadkills in Portugal
  - Species richness in grassland plot

Binomial Models

Recap of Binomial Data

- Can be:
  - Binary: 0,1 encoding absence/presence, survived/died
  - Binomial: probability value → 3 out of 10 survived → 0.3

$$p = \frac{k}{n} = \frac{Number\ of\ successes}{number\ of\ trials}$$

Binomial Models

$$y_i \sim Bernoulli(p)$$

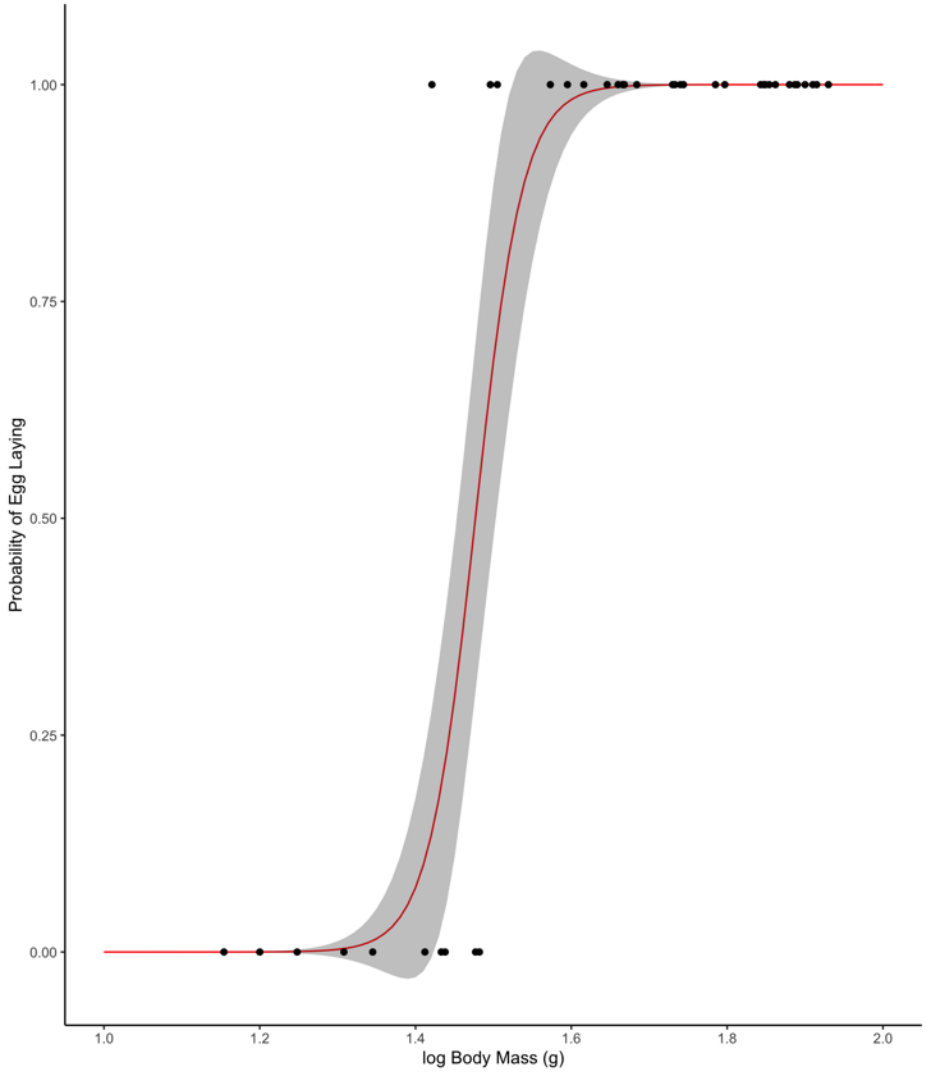
The Odds Ratio

$$Bernoulli(p) = \ln\left(\boxed{\frac{k}{n-k}}\right) \qquad \frac{number\ of\ successes}{number\ of\ failures}$$

$$Bernoulli(p) = \ln\left(\boxed{\frac{p_s}{1-p_f}}\right) \qquad \frac{Probability\ of\ success}{Probability\ of\ failure}$$

Moth Eggs Example- Binomial

```
Call:
glm(formula = BinaryEggs ~ logBodyMass, family = "binomial",
    data = motheggs)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.25014  -0.00311   0.00314   0.05421   1.99088

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -48.49      26.06  -1.861   0.0628 .
logBodyMass    32.83      17.70   1.855   0.0635 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.4029  on 38  degrees of freedom
Residual deviance:  9.7883  on 37  degrees of freedom
AIC: 13.788

Number of Fisher Scoring iterations: 9
```



Interpreting Coefficients

- The most important thing to remember with coefficients is that they are still in **the log odds ratios.**

- 3 ways to interpret the logistic model coefficients

1) Probability change

2) Divide by 4 Rule

3) Probability change (derived from the mean)

1) Probability Change

- **Intercept:** Not biologically meaningful without standardisation

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -48.49      26.06  -1.861   0.0628 .
logBodyMass    32.83      17.70   1.855   0.0635 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Slope:**

Log Odds Ratio **Probability**

$$\log\left(\frac{p_s}{1-p_s}\right) \longrightarrow p_s$$

$$\frac{e^{\beta_1}}{1+e^{\beta_1}} = \frac{e^{32.83}}{1+e^{32.82}} = 1 \text{ or } 100\%$$

1) Probability Change

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -48.49      26.06  -1.861   0.0628 .
logBodyMass    32.83      17.70   1.855   0.0635 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For every 1 log increase in body mass, the probability of a vapourer moth laying an egg increases by 100%.

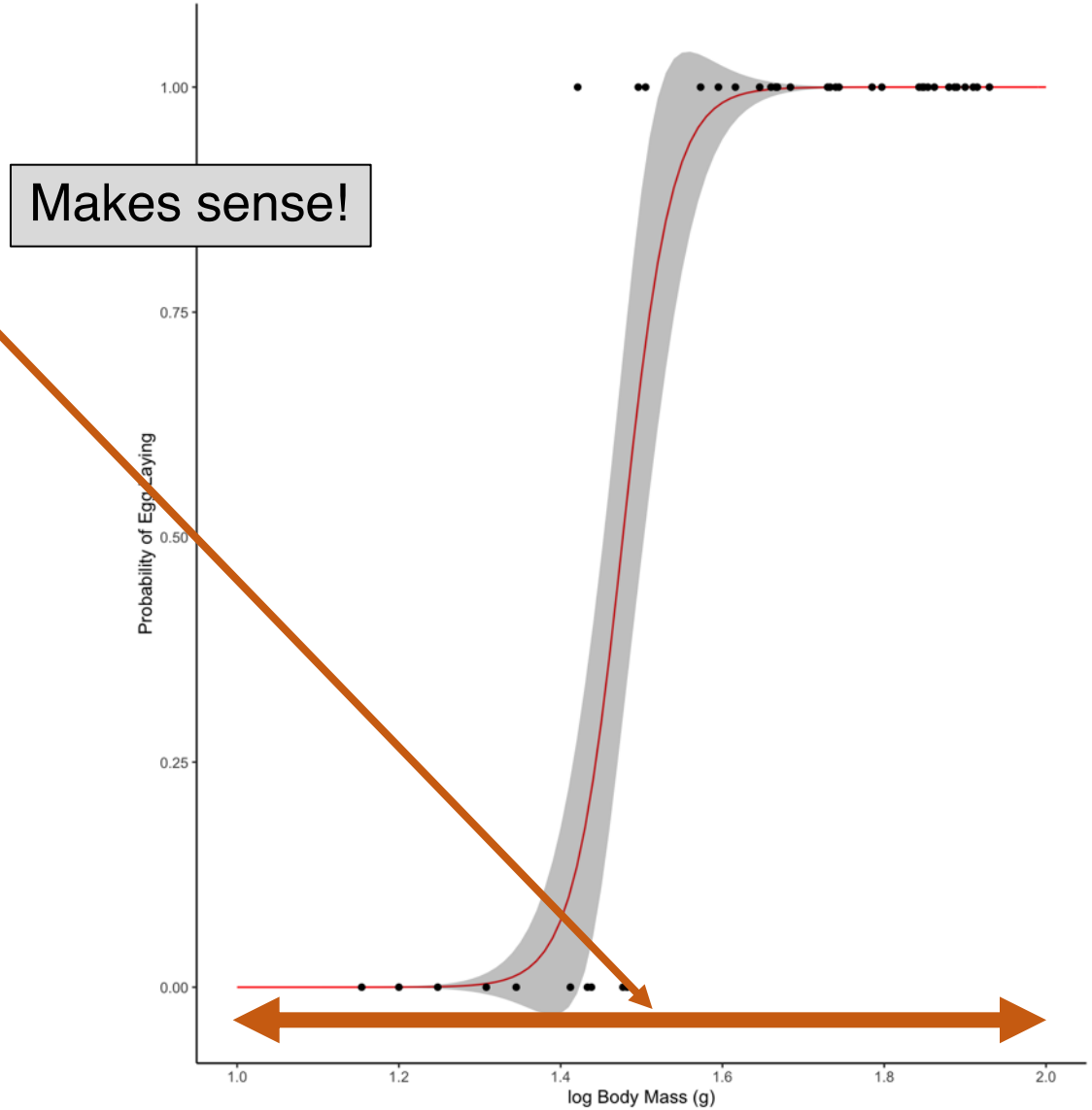
$$\frac{e^{\beta_1}}{1+e^{\beta_1}} = \frac{e^{32.83}}{1+e^{32.82}} = 1 \text{ or } 100\%$$

1) Probability Change

For every 1 log increase in body mass, the probability of a vapourer moth laying an egg increases by 100%.

Maybe think about units?

For every 0.1 log increase in body mass, the probability of a vapourer moth laying an egg increases by 10%.

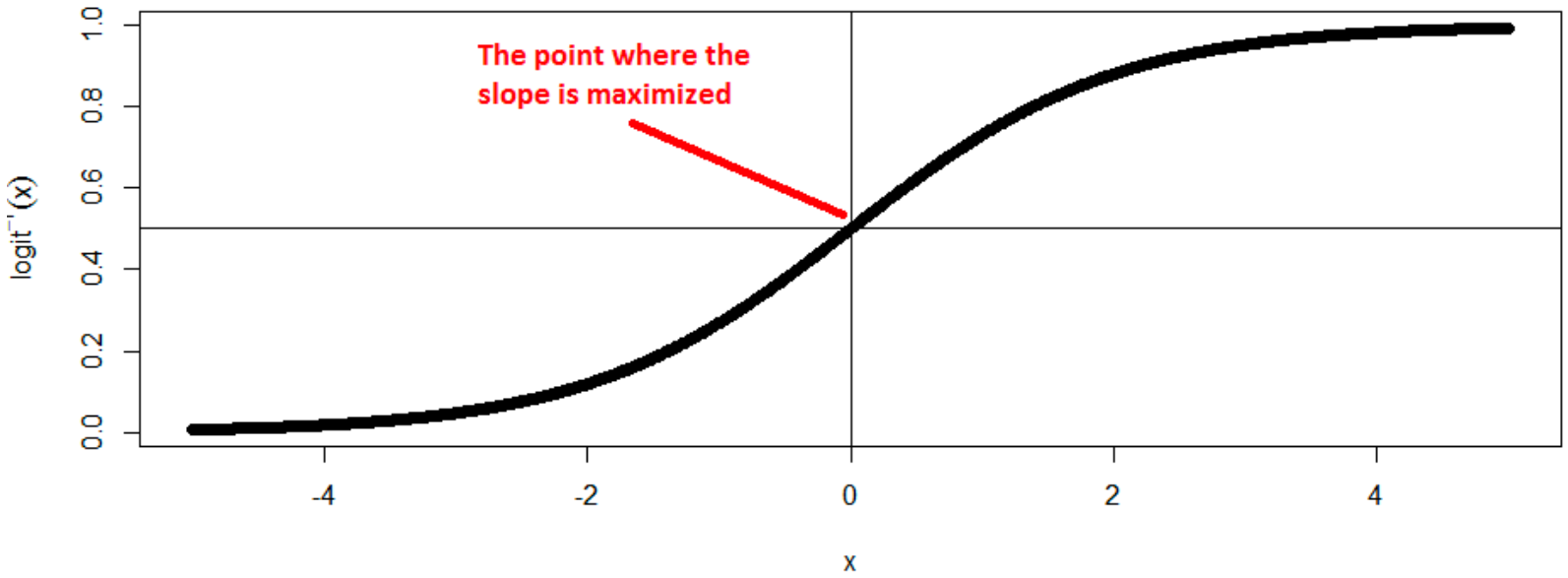


2) “Divide by 4” Rule

- Assumes the curve is steepest at its center and the slope of the curve is a derivative of the logistic function

- Derivative of the logistic curve and differentiated with respect to response variable.

$$\frac{\beta e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})^2}$$



2) "Divide by 4" Rule

- Now if we take $x = 0$, when the explanatory variable has been standardised.

- We get:

$$\frac{\beta e^0}{(1+e^0)^2} = \frac{\beta}{(1+1)^2} = \frac{\beta}{4}$$

- Interpretation gives you maximum increase or decrease in probability with one unit increase.

- **NOTE:** the divide by four rule is only applicable if $\beta \leq 1$

3) Probability Change from mean of x

- This method differs in that it makes a predictions of $\hat{y}$ from x_i and x_j- both derived from the mean:

- Mean logBodyMass $= 1.64$, so $x_i = 1$ and $x_j = 2$.

- So to predict $\hat{y}$:

$$\hat{y} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \qquad \hat{y} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

3) Probability Change from mean of x

- To predict $\hat{y}$ when $x = 1$:

$$\hat{y} = \frac{e^{-48.49+32.83*1}}{1 + e^{-48.49+32.83*1}}$$

$$\hat{y} = 0$$

- To predict $\hat{y}$ when $x = 2$:

$$\hat{y} = \frac{e^{-48.49+32.83*2}}{1 + e^{-48.49+32.83*2}}$$

$$\hat{y} = 1$$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -48.49      26.06  -1.861   0.0628 .
logBodyMass    32.83      17.70   1.855   0.0635 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Therefore:

For every 1 log increase in body mass, the probability of a vapourer moth laying an egg increases by 100%.

Goodness of Fit

```
    Null deviance: 44.4029  on 38  degrees of freedom
Residual deviance:  9.7883  on 37  degrees of freedom
AIC: 13.788
```

- Goodness-of-fit:
  - pseudo-R^2: $1 - ({}^{9.78}/_{44.40}) = 0.78$
  - Chi-squared

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

Model: binomial, link: logit

Response: BinaryEggs

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           38     44.403
logBodyMass  1   34.615        37      9.788 4.019e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dispersion Parameter

```
    Null deviance: 44.4029  on 38  degrees of freedom
Residual deviance:  9.7883  on 37  degrees of freedom
AIC: 13.788
```

- $\left({}^{9.78}/_{37}\right) = 0.26$ → **underdispersed**

HO 4

- Endemicity on the Galapagos Islands
- Focus on the <u>statistical concepts and themes</u>

- Extra:
  - Predicting threat in Galliformes