

Binary and Binomial Models (Logistic 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 to binary and binomial data
- Validate these models
- Calculate the pseudo- R^2 and dispersion parameter

Recap

- Any GLM consists of three steps:
 1. Choosing a distribution for the response variable = Binary or Binomial

Binary and Binomial Data

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

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



Recap

- Any GLM consists of three steps:
 1. Choosing a distribution for the response variable = Binary or Binomial
 2. Specifying the linear function of covariates and/or fixed factors

$$h(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Recap

- Any GLM consists of three steps:
 1. Choosing a distribution for the response variable = Binary or Binomial
 2. Specifying the linear function of covariates and/or fixed factors
 3. Choosing a link between the predictor function and the mean of the distribution

Logit link function

Logit Link Function

The Odds Ratio

$$\ln \left(\frac{k}{n - k} \right)$$

$\frac{\text{number of successes}}{\text{number of failures}}$

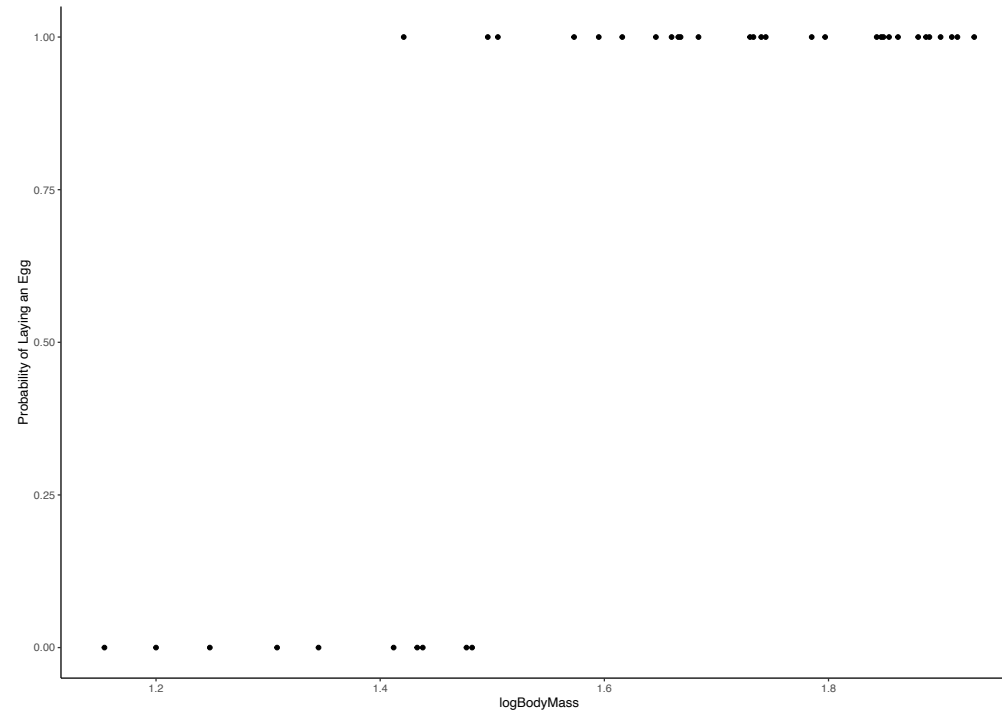
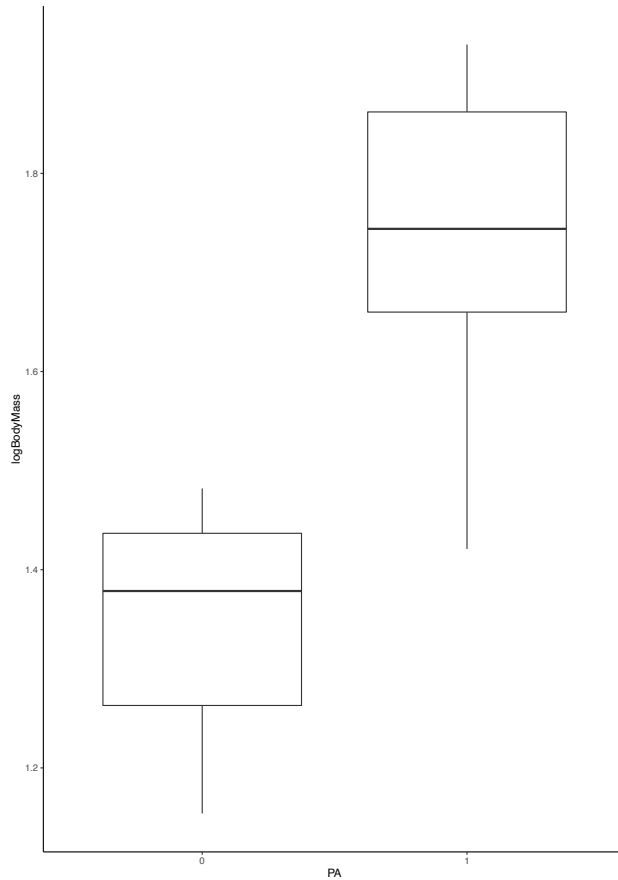
$$\ln \left(\frac{p}{1 - p} \right)$$

$\frac{\text{Probability of successes}}{\text{Probability of failures}}$

Binary Data – Moth Eggs

- Does the probability of laying an egg increase with log body mass?

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \log \text{body mass}$$



Moth Eggs Example- Binary

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

Deviance Residuals:

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

Coefficients:

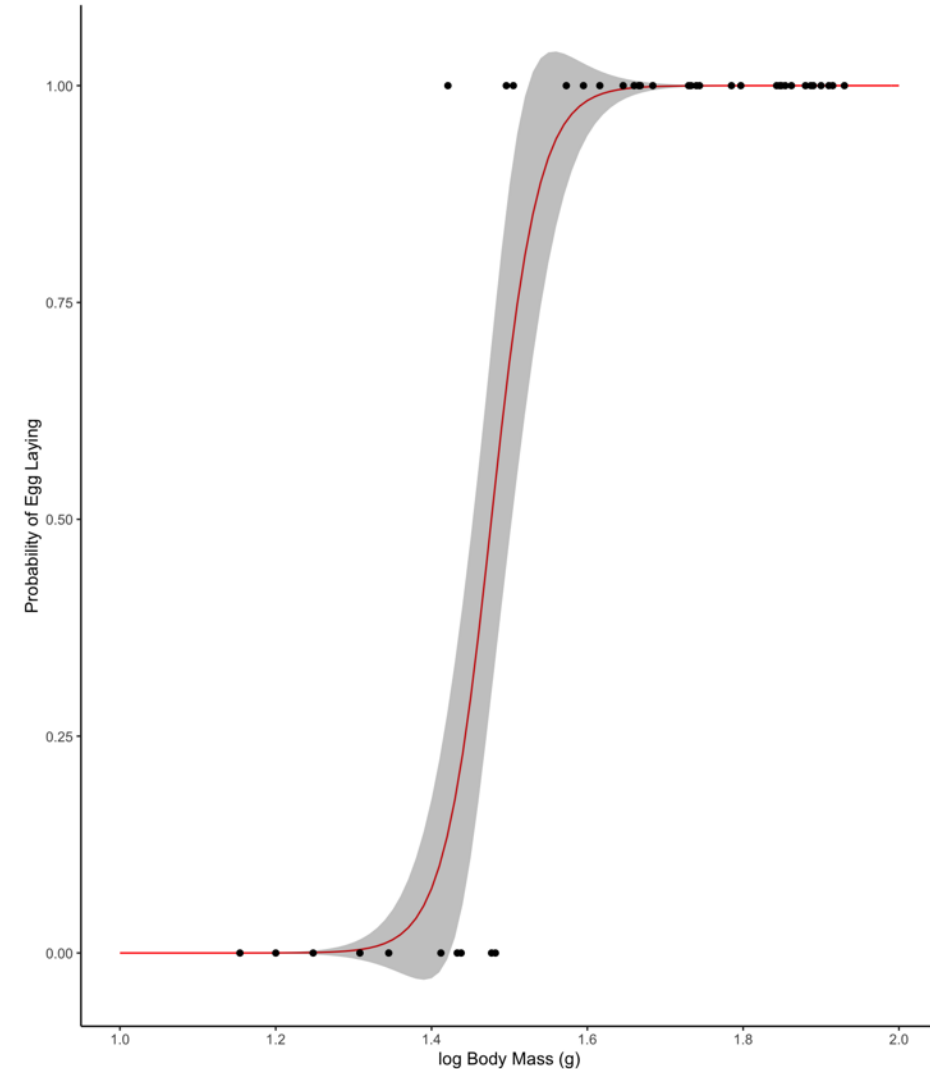
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-48.49	26.06	-1.861	0.0628 .
logBodyMass	32.83	17.70	1.855	0.0635 .

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**.

$$\ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X \log \text{ body mass}$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X \log \text{ body mass}}$$

$$p = \frac{e^{\beta_0 + \beta_1 X \log \text{ body mass}}}{1 + e^{\beta_0 + \beta_1 X \log \text{ body mass}}}$$

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 .

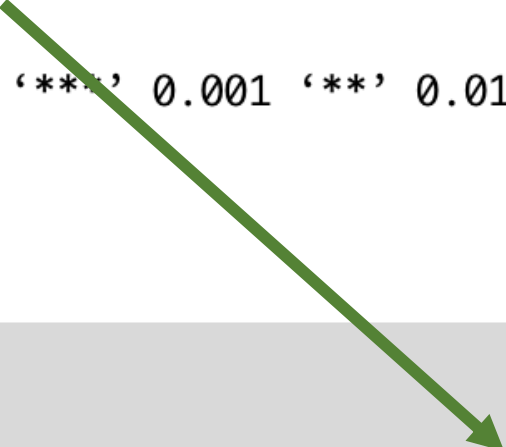
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- **Slope:**

Log Odds Ratio

Probability

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


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

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

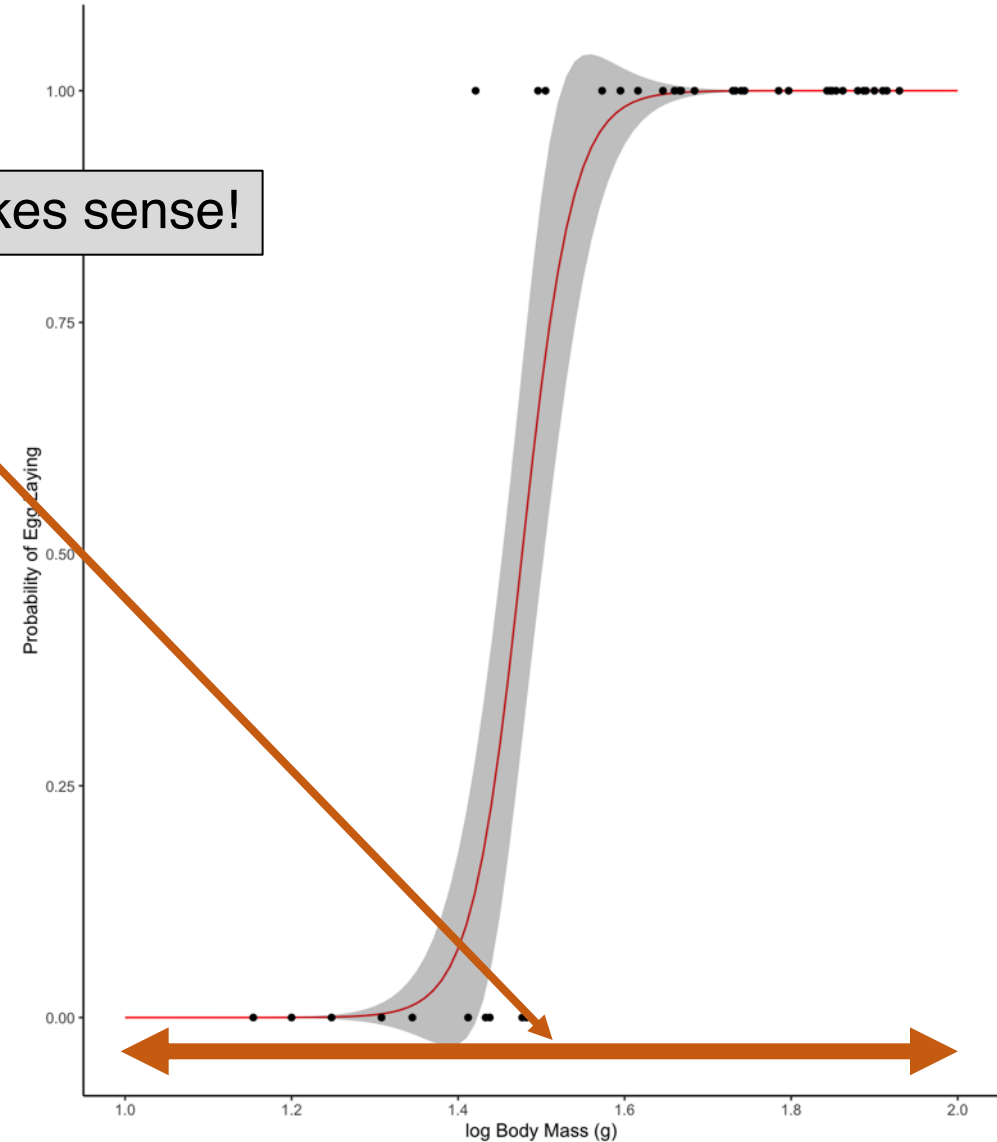
For every 1 log increase in body mass, the probability of a vapourer moth laying an egg increases by 100%.

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

Probability Change

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

Makes sense!

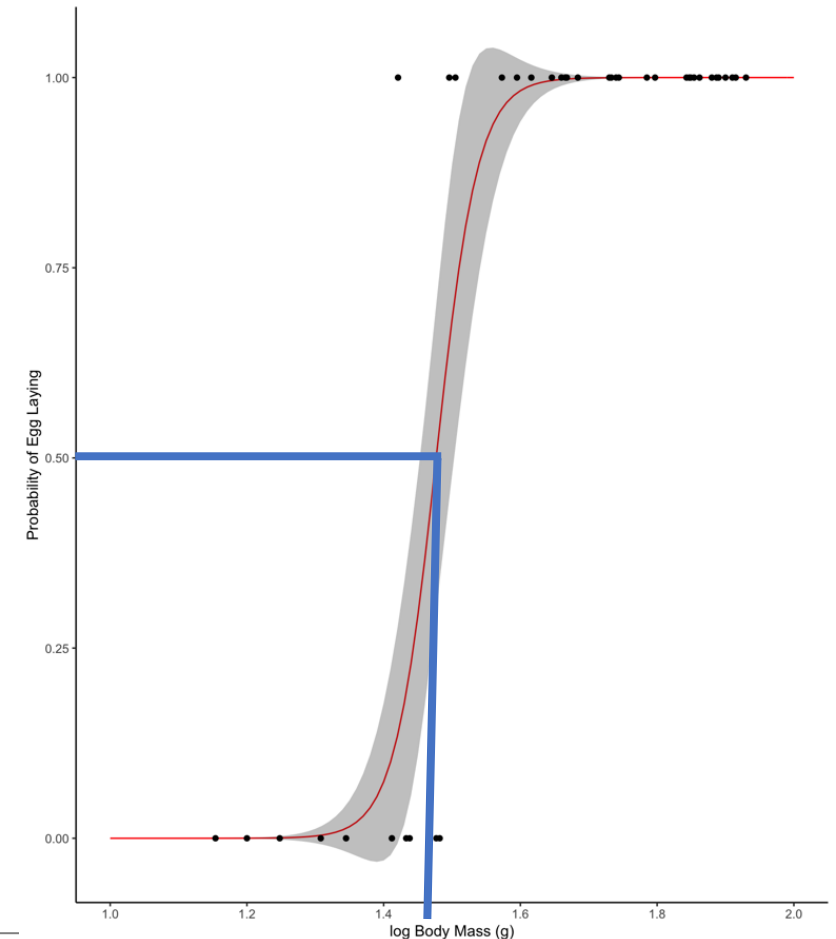


Finding the flip

- When thinking about probabilities we are interested in the point where the probability flips, i.e. you are more likely to do something.
- This occurs when the probability is greater than 0.5
- Solve for log body mass

$$0.5 = \frac{e^{-48.49+32.83 \times \log \text{ body mass}}}{1 + e^{-48.49+32.83 \times \log \text{ body mass}}}$$

- Value of $\frac{\beta_0}{\beta_1} = \frac{48.49}{32.83} = 1.48$



Finding the flip – other solutions

- Where the values of β_0 and β_1 are absolute.

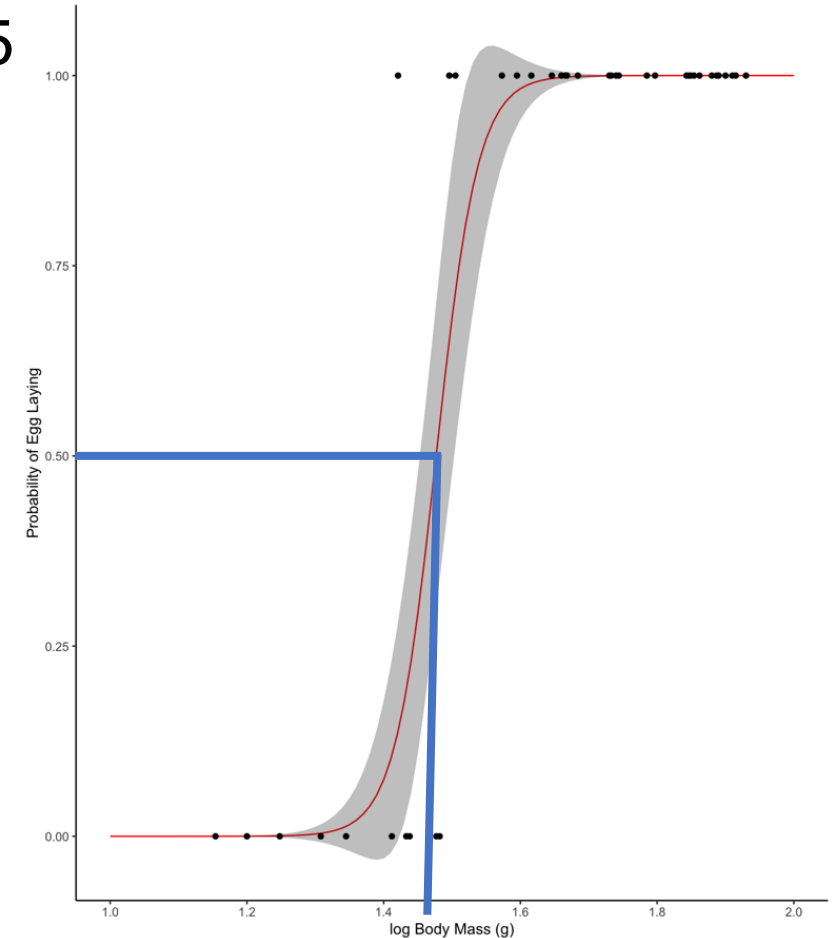
		Slope	
		Negative	Positive
Intercept	Negative	$-\frac{\beta_0}{\beta_1}$	$\frac{\beta_0}{\beta_1}$
	Positive	$\frac{\beta_0}{\beta_1}$	$-\frac{\beta_0}{\beta_1}$

Finding the flip

- When thinking about probabilities we are interested in the point where the probability flips, i.e. you are more likely to do something.
- This occurs when the probability is greater than 0.5
- Solve for log body mass

$$0.5 = \frac{e^{-48.49+32.83 \times \log \text{ body mass}}}{1 + e^{-48.49+32.83 \times \log \text{ body mass}}}$$

- Value of $\frac{\beta_0}{\beta_1} = \frac{48.49}{32.83} = 1.48$
- We can infer then that vapourer moths weighing over 1.48 logs of body mass are more likely to lay an egg.



Pseudo-R² and Goodness-of-fit

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

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

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

■ Pseudo-R²:

- $1 - (9.89 / 44.40) = 0.78$

Model Validation

- Validation is difficult as the response variable consists only of 0's and 1's and so the diagnostic plots aren't reliable.
- There is still debate regarding the use of quasi-likelihood approaches for binary data – quasi-binomial but these are regarded as controversial – they are supported by the `glm` function in R however.

Moth Eggs Example- Binomial

```
Call:
glm(formula = cbind(Hatched, RedEggs - Hatched) ~ logBodyMass,
     family = "binomial", data = motheggs)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.21468	-0.46512	-0.07996	0.00000	2.76868

Coefficients:

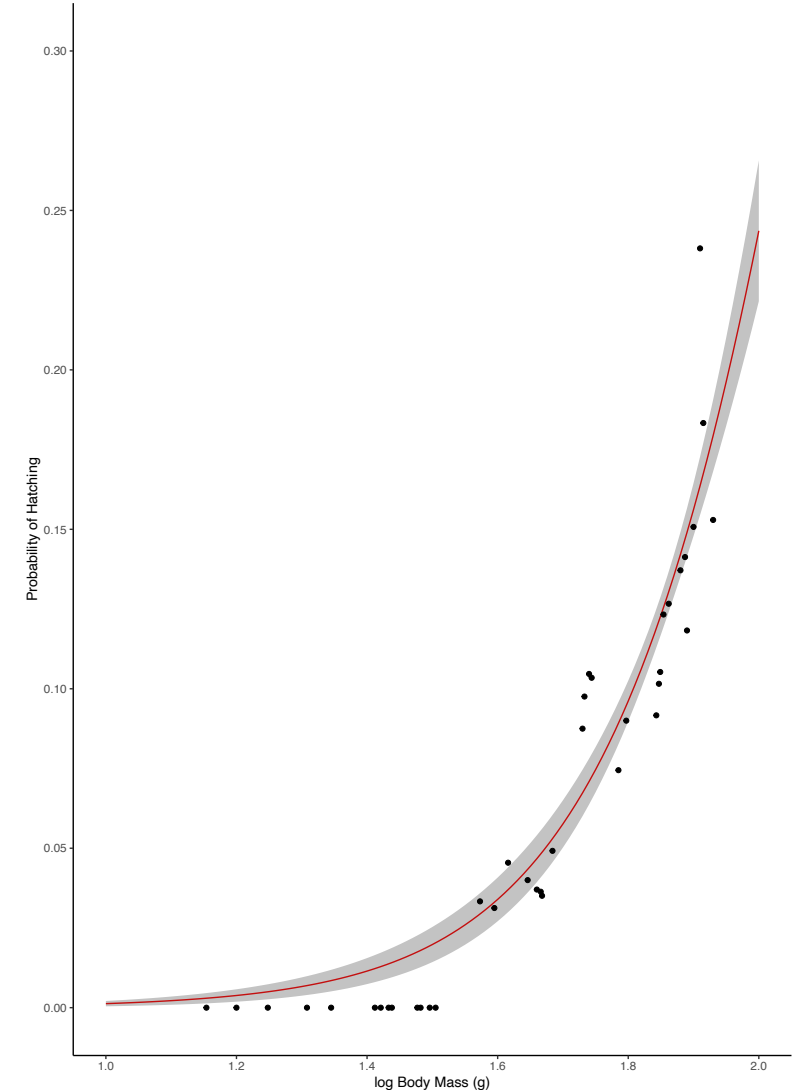
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-12.2095	1.4332	-8.519	<2e-16 ***
logBodyMass	5.5383	0.7691	7.201	6e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 85.659 on 28 degrees of freedom
Residual deviance: 19.372 on 27 degrees of freedom
AIC: 124.31

Number of Fisher Scoring iterations: 4

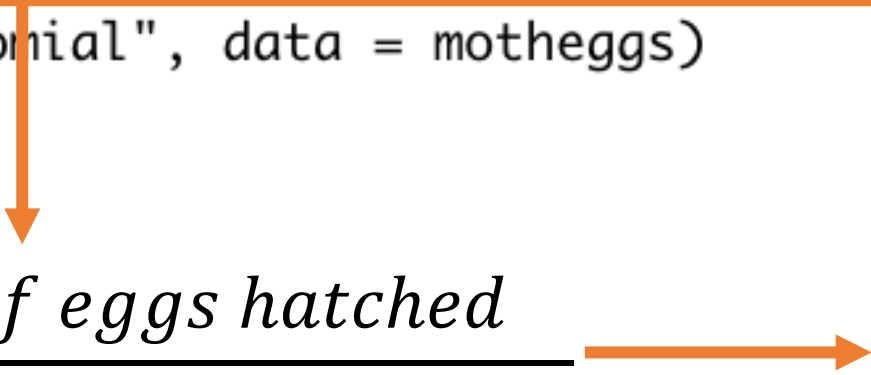


Model Formula

- Does maternal body weight affect the probability of eggs hatching?

Call:

```
glm(formula = cbind(Hatched, RedEggs - Hatched) ~ logBodyMass,  
     family = "binomial", data = motheegs)
```


$$\frac{\text{number of eggs hatched}}{\text{number of eggs not hatched}} \longrightarrow \ln \left(\frac{k}{n - k} \right)$$

Interpreting Coefficients

Coefficients:

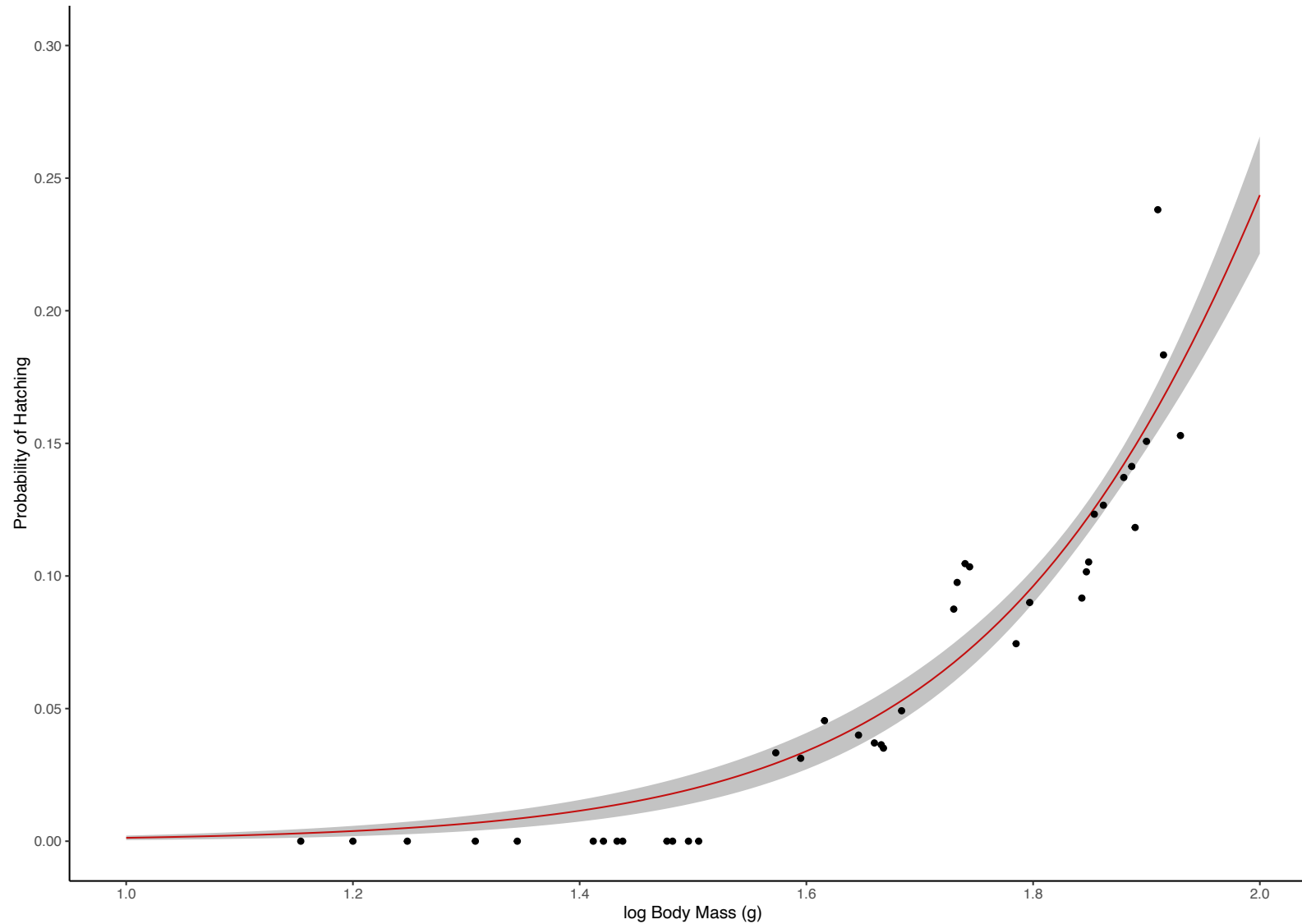
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-12.2095	1.4332	-8.519	<2e-16	***
logBodyMass	5.5383	0.7691	7.201	6e-13	***

- For a milligram increase in body mass increases the log odds of a moth egg hatching by 5.54

$$\text{Probability of egg hatching} = \frac{e^{-12.21+5.54 \times \log \text{body mass}}}{1 + e^{-12.21+5.54 \times \log \text{body mass}}}$$

- Because the relationship between logit and probability is not linear we can't just back transform the slope coefficient to get a probability change instead it is more common to plot to model and infer from that.

Interpreting Coefficients



Pseudo R² and Goodness-of-Fit

```
Call:
glm(formula = cbind(Hatched, RedEggs - Hatched) ~ logBodyMass,
     family = "binomial", data = motheegs)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.21468	-0.46512	-0.07996	0.00000	2.76868

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-12.2095	1.4332	-8.519	<2e-16 ***
logBodyMass	5.5383	0.7691	7.201	6e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 85.659 on 28 degrees of freedom
Residual deviance: 19.372 on 27 degrees of freedom
AIC: 124.31

Number of Fisher Scoring iterations: 4

Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Hatched, RedEggs - Hatched)

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			28	85.659	
logBodyMass	1	66.288	27	19.372	3.896e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

■ Goodness-of-fit:

- pseudo-R²: $1 - (19.37 / 85.66) = 0.77$

Model Validation

```
Null deviance: 85.659  on 28  degrees of freedom  
Residual deviance: 19.372  on 27  degrees of freedom  
AIC: 124.31
```

- $(19.372/27) = 0.71 \rightarrow$ **underdispersed**
- Underdispersion methods are less developed
- If overdispersed can fit a quasi-binomial approach and investigate the following 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

Summary

- Logistic models can handle binary outcomes (0,1) and binomials via the log odds ratio
- The logit function abstracts the interpretation of the coefficients and instead plotting is preferable
- Logistic models fitted to binary data are hard to validate and quasi-likelihood approaches are generally avoided, however, these can be implemented with binomial data