Binomial regression

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- Binomial regression
- Model comparison
- P-values
- R^2
- Q & A

Questions about yesterday?

Recap

Generalised linear models are a unifying class

- They have many things in common
- Two key components: the distribution and the link function
- They link the sytematic and the random components

$$g\{\mathsf{E}(y_i|x_i)\} = \alpha + x_i\beta \tag{1}$$

Link function

- $g(\cdot)$ the link function, can be e.g.,:
    - identity (linear regression)
    - log
    - logit
    - probit
    - inverse
    - many more

Makes sure that the mean is on the right scale (e.g., positive only, or between 0 and 1)

The binomial GLM

Response data: r the number of successes in N trials
Predictor variables: x_i albeit continuous and/or categorical
Parameters: probability of success p_i in trial i
Goal: estimate p_i for each observation

Binomial GLM use

- When a linear regression is not appropriate :)
- For binary data or (bounded) counts of successes/failures

Common examples

- Occurrence of species: orchids 🙃
- Number of germinated plant seeds
- Prevalence of disease in a population: cancer rates
- Drug trials (effect or no effect)
- Probability of observing a behavior

Binomial GLM data example

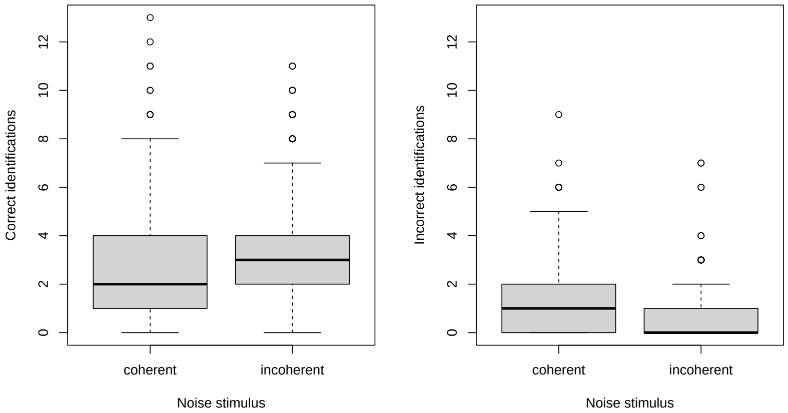
Dataset from the MASS R-package

- OME: Otitis Media with Effusion
- Children were stimulated with sound signals
- Signals included either coherent or incoherent noise
- The number of correct signal identifications

ID	Age	OME	Loud	Noise	Correct	Trials
1	30	low	35	coherent	1	4
1	30	low	35	incoherent	4	5
1	30	low	40	coherent	0	3
1	30	low	40	incoherent	1	1
1	30	low	45	coherent	2	4
1	30	low	45	incoherent	2	2

Which is the response and which are the covariates?

OME: visually inspect the data

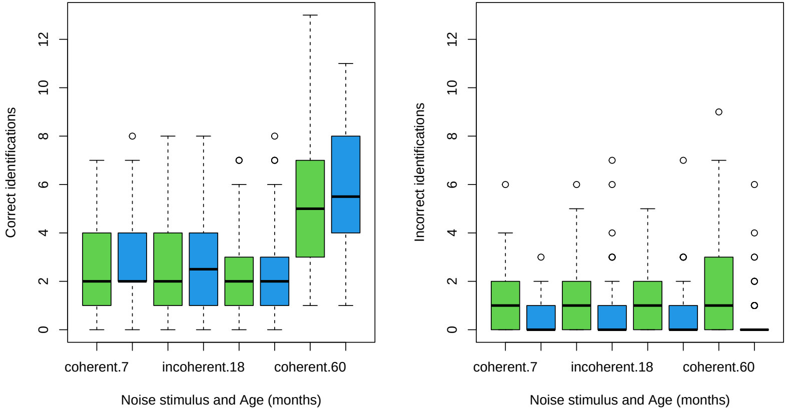


OME: fit the model

```
model <- glm(cbind(Correct, Trials-Correct)~Noise,
             data = OME, family="binomial")
```

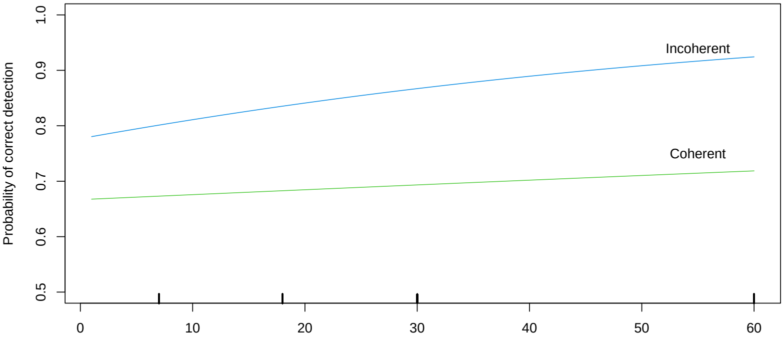
- cbind: combines two vectors as columns
- Incorrect identifications = Total - Correct identifications

OME: visually inspect the data



OME: predicting with age

```
model <- glm(cbind(Correct, Trials-Correct)~Noise*Age,
             data = OME, family="binomial")
```



Binomial GLM data example

ID	Age	OME	Loud	Noise	Correct	Trials
1	30	low	35	coherent	1	4
1	30	low	35	incoherent	4	5
1	30	low	40	coherent	0	3
1	30	low	40	incoherent	1	1
1	30	low	45	coherent	2	4
1	30	low	45	incoherent	2	2

What would this data look like if there was only one child in each row?

Is binomial regression in the EF?

$$\mathcal{L}(y_i;\Theta) = \exp\left\{\log\binom{N}{r} + \frac{\frac{r}{N}\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N}\right\} \quad (2)$$

Is binomial regression in the EF?

$$\mathcal{L}(y_i;\Theta) = \textsf{exp}\left\{\log\binom{N}{r} + \frac{\frac{r}{N}\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N}\right\} \quad (2)$$

All GLMs can be formulated as:

$$\mathcal{L}(y_i;\Theta) = \textsf{exp}\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \quad (3)$$

Is binomial regression in the EF?

$$\mathcal{L}(y_i;\Theta) = \exp\left\{\log\binom{N}{r} + \frac{\frac{r}{N}\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N}\right\} \quad (2)$$

All GLMs can be formulated as:

$$\mathcal{L}(y_i;\Theta) = \exp\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \quad (3)$$

- for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$
- for binomial distribution: $\eta_i = \log(\frac{p_i}{1-p_i})$, $a(\phi) = 1/N$, $b(\eta_i) = -\log(1-p_i)$, $c(y_i,\phi) = \log\binom{N}{r}$
- canonical link

Link functions

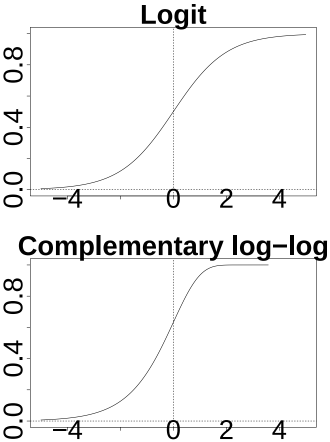
In a binomial GLM we can change the link function

- Logit: $\log(\frac{p_i}{1-p_i})$ and inverse $\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ - *the canonical link*
- Probit: $\Phi^{-1}(p_i)$ and inverse $\Phi(\eta_i)$
- Complementary log-log: $\log(-\log(1-p_i)$ and inverse $1-\exp(-\exp(\eta_i))$
- ~~Log-log~~

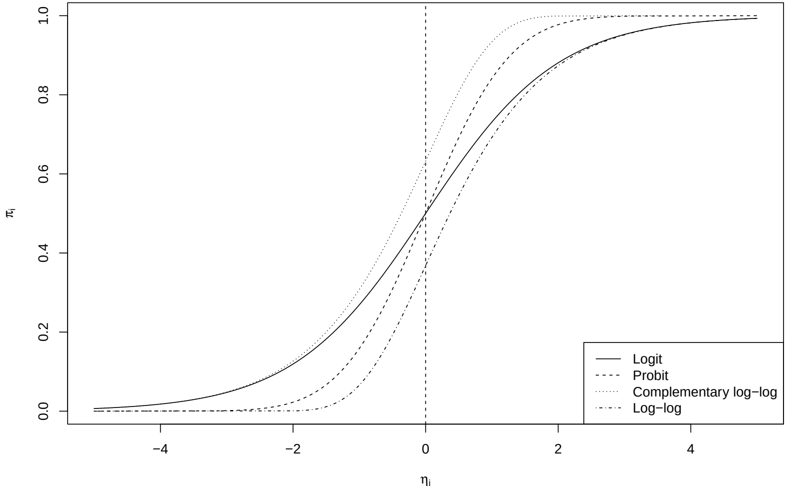
- Logit is canonical and easier to interpret
- Probit is sometimes easier mathematically than Logit
- Complementary log-log for counts

Binomial link functions

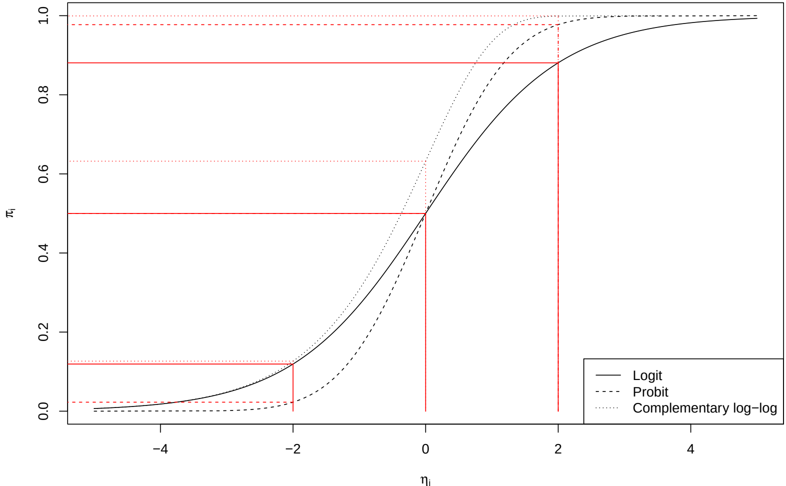
Each is a little different, but all translate from $(0, 1)$ to $(-\infty, \infty)$



Binomial link functions (2)



Binomial link functions (2)



Link functions: logit

$$\mathsf{pr}(y_i = 1) = p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

Link functions: logit

$$\begin{aligned}\mathsf{pr}(y_i = 1) = p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \\ \log(\frac{p_i}{1 - p_i}) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}\end{aligned} \qquad (4)$$

Link functions: logit

$$\begin{aligned}\text{pr}(y_i = 1) = p_i &= \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \\ \log(\frac{p_i}{1-p_i}) &= \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}\end{aligned} \tag{4}$$

- η is the log odds
- Odds ratio: $\frac{p_i}{1-p_i} = \exp(\eta_i)$
  - E.g., horse races: 10:1 (for every 1 dollar spent, you get 10 if won)
  - I.e., you think that horse 1 is 10 times more likely to win than horse 2
- $\text{pr}(y_i = 0) = 1 - p_i$

Link functions: probit

In probit regression, we use a latent variable $\eta_i^\star$ for thresholding.

Link functions: probit

In probit regression, we use a latent variable $\eta_i^\star$ for thresholding.

$$y_i = \begin{cases} 1, & \text{if } \eta_i^\star > 0 \\ 0, & \text{otherwise} \end{cases}$$

Link functions: probit

In probit regression, we use a latent variable $\eta_i^\star$ for thresholding.

$$y_i = \begin{cases} 1, & \text{if } \eta_i^\star > 0 \\ 0, & \text{otherwise} \end{cases}$$

and we model this latent variable:

$$\eta_i^\star = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0,1)$$

Link functions: probit

In probit regression, we use a latent variable $\eta_i^\star$ for thresholding.

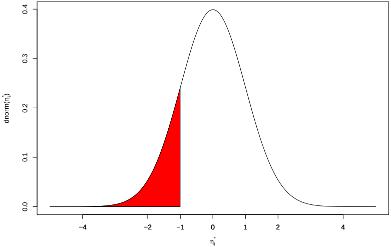
$$y_i = \begin{cases} 1, & \text{if } \eta_i^\star > 0 \\ 0, & \text{otherwise} \end{cases}$$

and we model this latent variable:

$$\eta_i^\star = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0,1)$$

Which is the same as:

$$p_i = \Phi(\eta_i)$$
$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$



Link functions: probit

In probit regression, we use a latent variable $\eta_i^\star$ for thresholding.

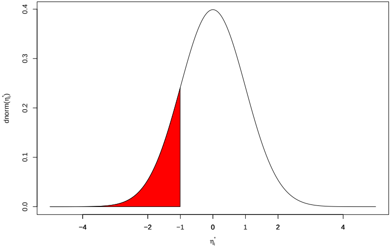
$$y_i = \begin{cases} 1, & \text{if } \eta_i^\star > 0 \\ 0, & \text{otherwise} \end{cases}$$

and we model this latent variable:

$$\eta_i^\star = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0,1)$$

Which is the same as:

$$p_i = \Phi(\eta_i)$$
$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$



if $\eta_i^\star$ is positive, we have 1 and 0 if it is negative

Logit latent variable model

In logistic regression, we can also use a latent variable $\eta_i^\star$ for thresholding.

Logit latent variable model

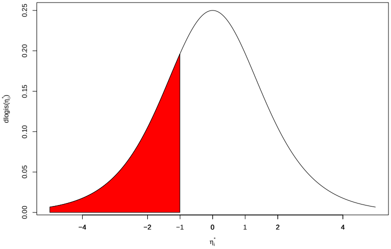
In logistic regression, we can also use a latent variable $\eta_i^\star$ for thresholding.

Probit regression: $\epsilon \sim \mathcal{N}(0, 1)$
Logistic regression: $\epsilon \sim \mathcal{L}ogistic(0, 1)$

Logit latent variable model

In logistic regression, we can also use a latent variable $\eta_i^\star$ for thresholding.

Probit regression: $\epsilon \sim \mathcal{N}(0,1)$
Logistic regression: $\epsilon \sim \mathcal{Logistic}(0,1)$



$$\begin{aligned}
p_i &= \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \\
\eta_i &= \mathbf{x}_i^\top \boldsymbol{\beta} \\
\eta_i^\star &= \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \sim \mathcal{Logistic}(0,1)
\end{aligned} \tag{5}$$

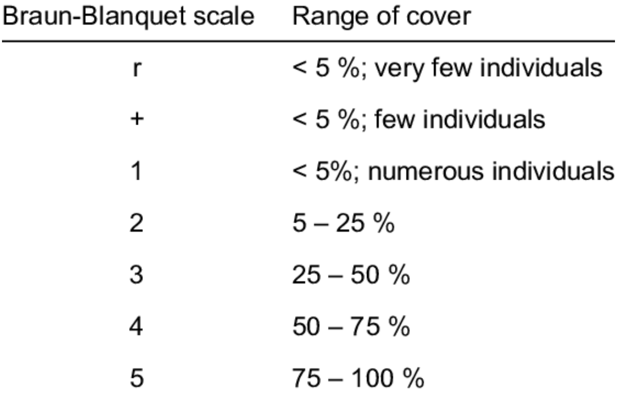
Binomial link functions: complementary log-log

$$\begin{aligned} \log(\lambda_i) &= \log(-\log(1-p_i)) \\ \lambda_i &= -\log(1-p_i) = \exp(\eta_i) \end{aligned} \tag{6}$$

- Has a link to count data
  - Just look at the Poisson likelihood! **Thursday**
- When binary data are really counts
- Probability to get a count larger than 0
- Can also be interpreted as LVM (reverse extreme value distribution)

Cumulative link

Suppose you have ordered data:



Braun-Blanquet scale	Range of cover
r	< 5 %; very few individuals
+	< 5 %; few individuals
1	< 5%; numerous individuals
2	5 – 25 %
3	25 – 50 %
4	50 – 75 %
5	75 – 100 %

Cumulative link functions facilitate this kind of response by introducing **order**. **More on Thursday**.

Example: Lizard habitat preference

- Data originally by Schoener (1970)
- Counts of two species of lizard in Jamaica

Figure 2: wikipedia.org: grahami

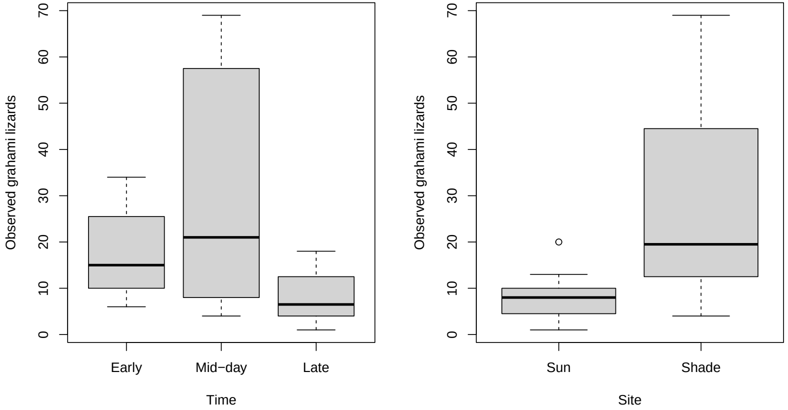
Figure 3: wikipedia.org: opalinus

Lizards: the data

Dataset also covered in McCullagh and Nelder (1989)

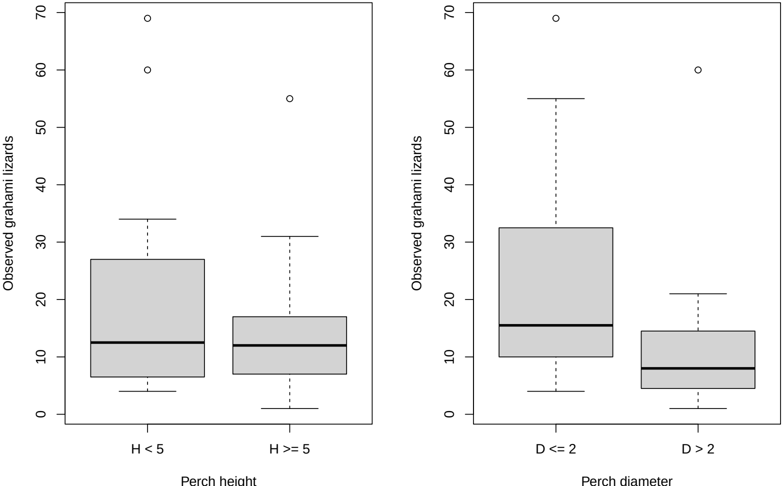
	Site	Diameter	Height	Time	grahami	opalinus
1	Sun	D <= 2	H < 5	Early	20	2
2	Sun	D <= 2	H < 5	Mid-day	8	1
3	Sun	D <= 2	H < 5	Late	4	4
4	Sun	D <= 2	H >= 5	Early	13	0
5	Sun	D <= 2	H >= 5	Mid-day	8	0
6	Sun	D <= 2	H >= 5	Late	12	0
7	Sun	D > 2	H < 5	Early	8	3
8	Sun	D > 2	H < 5	Mid-day	4	1
9	Sun	D > 2	H < 5	Late	5	3
10	Sun	D > 2	H >= 5	Early	6	0
12	Sun	D > 2	H >= 5	Late	1	1
13	Shade	D <= 2	H < 5	Early	34	11
14	Shade	D <= 2	H < 5	Mid-day	69	20
15	Shade	D <= 2	H < 5	Late	18	10
16	Shade	D <= 2	H >= 5	Early	31	5
17	Shade	D <= 2	H >= 5	Mid-day	55	4
18	Shade	D <= 2	H >= 5	Late	13	3
19	Shade	D > 2	H < 5	Early	17	15
20	Shade	D > 2	H < 5	Mid-day	60	32
21	Shade	D > 2	H < 5	Late	8	8

Lizards: visually inspect the data

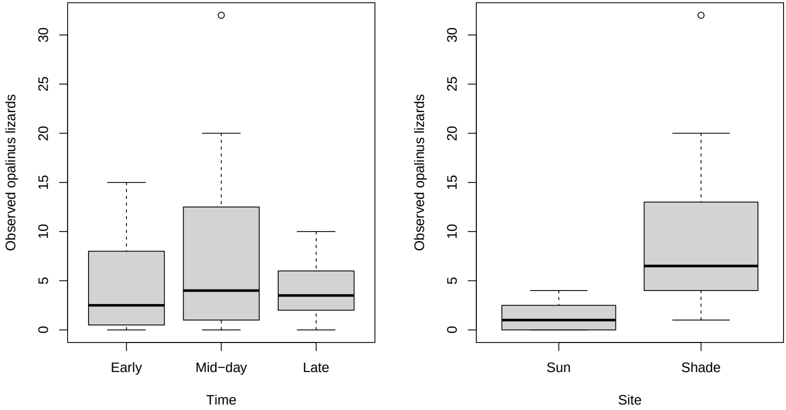


What can we tell about grahami lizards in Sunny or Shady sites?

Lizards: visually inspect the data



Lizards: visually inspect the data



What can we tell about opalinus lizards in Sunny or Shady sites?

Lizards: fit the model

```r
model <- glm(cbind(grahami, opalinus)~Time+Diameter+Height+Site,
             data = lizards, family="binomial")
```

- cbind: combines two vectors as columns
- N_i is the row sum (total lizards per site)
- Canonical link is used by default

Lizards: interpreting parameters

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.90       0.34    5.70  1.2e-08
## TimeMid-day        0.23       0.25    0.91  3.6e-01
## TimeLate          -0.74       0.30   -2.50  1.4e-02
## DiameterD > 2     -0.76       0.21   -3.60  3.1e-04
## HeightH >= 5       1.10       0.26    4.40  1.1e-05
## SiteShade         -0.85       0.32   -2.60  8.6e-03
```

- (Intercept) = Early, small diameter low perches
- Change in odds for observing grahami lizards mid-day: `exp(0.23) = 1.26`
- Down for late-day: `exp(-0.74) = 0.48`
- What else?

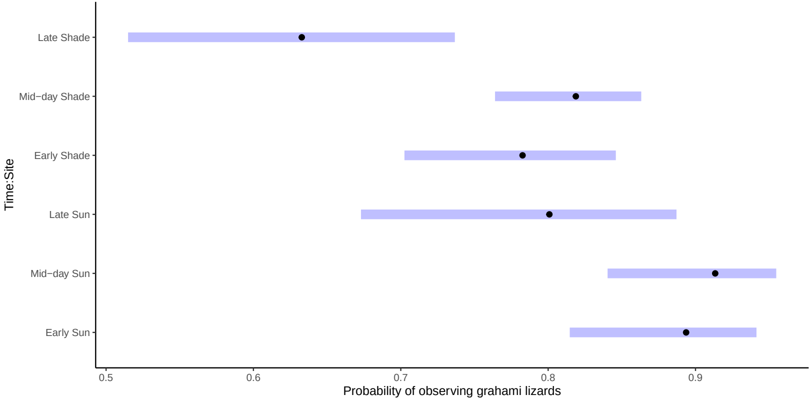
Lizards: inspecting the groups

```r
emmeans::emmeans(model, ~Time+Site, type = "response")
```

```
##  Time    Site   prob     SE  df asymp.LCL asymp.UCL
##  Early   Sun   0.894 0.0314 Inf     0.815     0.941
##  Mid-day Sun   0.913 0.0280 Inf     0.840     0.955
##  Late    Sun   0.801 0.0545 Inf     0.673     0.887
##  Early   Shade 0.783 0.0366 Inf     0.702     0.846
##  Mid-day Shade 0.819 0.0253 Inf     0.764     0.863
##  Late    Shade 0.633 0.0574 Inf     0.515     0.737
##
## Results are averaged over the levels of: Diameter, Height
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
```

Lizards: inspecting the groups visually

```
plot(emmeans::emmeans(model, ~Time+Site, type = "response"))+ggplot2::theme_classic()+ggplot2::xlab("Probabili
```



Lizards: parameter estimates with different link functions

```
##                 logit probit cloglog
## (Intercept)      1.90   1.20    0.78
## TimeMid-day      0.23   0.13    0.12
## TimeLate        -0.74  -0.44   -0.42
## DiameterD > 2 -0.76  -0.45   -0.41
## HeightH >= 5     1.10   0.65    0.58
## SiteShade       -0.85  -0.51   -0.49
```

Lizards: group estimates with different link functions

```
##                  logit probit cloglog
## Early Sun         0.89   0.90    0.91
## Mid-day Sun       0.91   0.92    0.93
## Late Sun          0.80   0.80    0.79
## Early Shade       0.78   0.78    0.77
## Mid-day Shade     0.82   0.82    0.81
## Late Shade        0.63   0.63    0.62
```

Overdispersion

In a binomial glm we (amongst others) assume:

- Independent Bernoulli trials
- var$(y_i) = N_i p_i (1 - p_i)$

If these are violated, we can have overdispersion. And due to omitted covariates.

Methods for diagnosing:

- Residual diagnostics
- Overdispersion factor
- Overdispersion test

Ignoring overdispersion is likely to cause inflated type I error.

This will come back for Poisson responses.
It is not possible to detect overdispersion in binary responses.

Lizards: checking overdispersion

```
deviance(model)/df.residual(model)
```

```
## [1] 0.7103637
```

```
sum(residuals(model,"pearson")^2)/df.residual(model)
```

```
## [1] 0.6431694
```

- Deviance is (approx.) χ^2-distributed
- Can do a test for overdispersion

Lizards: checking overdispersion

```r
performance::check_overdispersion(model)
```

```
## # Overdispersion test
##
##  dispersion ratio = 0.994
##           p-value =  0.92
```

```
## No overdispersion detected.
```

Accouting for overdispersion

- Beta-binomial GLM
- Quasibinomial GLM
- Mixed-effects model See GLMM workshop

Perfect separation

```
Warning message: glm.fit: fitted probabilities
numerically 0 or 1 occurred
```

Complete separation occurs whenever a linear function of x_i can generate perfect predictions of y_i

In essence: MLE is on the boundary of valid parameter space. We perfectly classify the response.

Solution:

- Sometimes it happens
- Change model (simplify)
- Collapse categories in covariates
- Penalised regression
- Other

Perfect separation: example

```r
x <- seq(-3, 3, by=0.1)
y <- x > 0
model <- glm(y ~ x, family=binomial)
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Perfect separation (2)

```
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -17.89    4642.58  -0.004    0.997
## x             357.72   65657.28   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.4548e+01  on 60  degrees of freedom
## Residual deviance: 6.8270e-08  on 59  degrees of freedom
## AIC: 4
##
```

Perfect separation: example (3)

```
cbind(y, signif(predict(model, type = "response"),2))
```

```
##    y
## 1  0 2.2e-16
## 2  0 2.2e-16
## 3  0 2.2e-16
## 4  0 2.2e-16
## 5  0 2.2e-16
## 6  0 2.2e-16
## 7  0 2.2e-16
## 8  0 2.2e-16
## 9  0 2.2e-16
## 10 0 2.2e-16
## 11 0 2.2e-16
## 12 0 2.2e-16
## 13 0 2.2e-16
## 14 0 2.2e-16
```

Summary

- Binomial GLMs for binary of count (succeses/failures) responses
- Four potential link functions: logit, probit, cloglog, loglog
- Fixed dispersion: can exhibit overdispersion issues
- Perfect separation