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)

Outline

▶ Normal distribution
▶ Linear models ...

The binomial GLM

Data: r the number of successes in N trials
Parameters: probability p (now: π_i)
Goal: estimate π_i for each observation

Binomial GLM use

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

In ecology

▶ Predicting species' distributions
▶ Number of germinated plant seeds
▶ Prevalence of disease in a population
▶ Probability of observing a behavior
▶ Proportion of orchids 🙃

Is binomial regression really a GLM?

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

All GLMs can be formulated as:

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

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

Bernoulli

▶ $r \in \{0, 1\}$: Bernoulli (presence-absence)
▶ $N = 1$

$$\mathcal{L}(y_i; \Theta) = \exp\left\{\log(1) + \frac{r\log(\frac{\pi_i}{1-\pi_i}) + \log(1-\pi_i)}{1}\right\} \quad (4)$$

All GLMs can be formulated as:

$$\mathcal{L}(y_i; \Theta) = \exp\left\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right\} \quad (5)$$

▶ for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$
▶ for Bernoulli distribution: $\eta_i = \log(\frac{\pi_i}{1-\pi_i})$, $a(\phi) = 1$,
$b(\eta_i) = -\log(1-\pi_i)$, $c(y_i, \phi) = \log(1) = 0$
▶ canonical link

The binomial GLM

Link functions:

▶ Logit: $\log(\frac{\pi_i}{1-\pi_i})$ and inverse $\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ - *the canonical link*

▶ Probit: $\Phi^{-1}(\pi_i)$ and inverse $\Phi(\eta_i)$

▶ Complementary log-log: $\log(-\log(1-\pi_i)$ 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
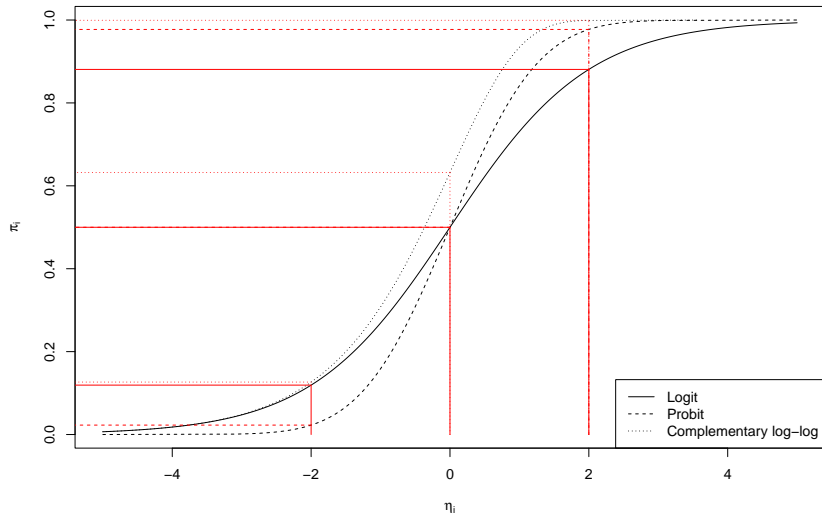
Each is a little different, but all translate from $(0, 1)$ to $(-\infty, \infty)$
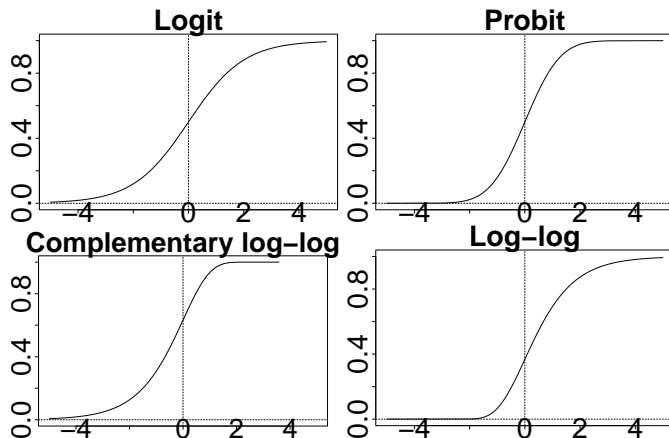
Binomial link functions (2)
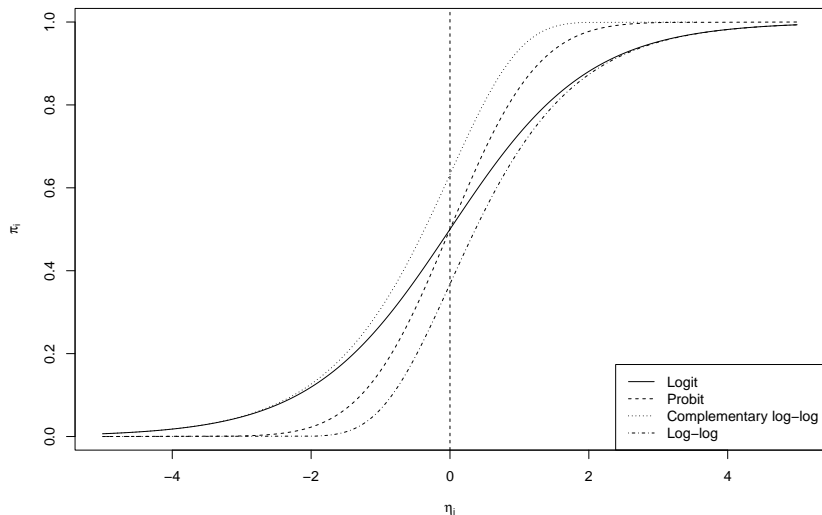
Binomial link functions (2)
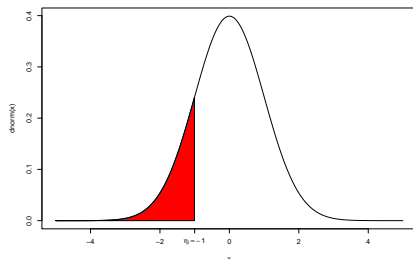
Binomial link functions: logit

$$\text{pr}(y_i = 1) = \pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \mathbf{x}_i \boldsymbol{\beta}$$

(6)

▶ η is the log odds
▶ Odds ratio: $\frac{\pi_i}{1 - \pi_i} = \exp(\eta_i)$
  ▶ 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
▶ $\text{pr}(y_i = 0) = 1 - \pi$

Binomial link functions: probit

$$\pi_i = \Phi(\eta_i)$$
$$\eta_i = \mathbf{x}_i\boldsymbol{\beta} \qquad\qquad (7)$$
$$\eta_i^\star = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0,1)$$

$$y_i = \begin{cases} 1, & \text{if } \eta_i^\star > 0 \\ 0, & \text{otherwise} \end{cases} \qquad (8)$$

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

Binomial: Logit LVM

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

$$\eta_i = \mathbf{x}_i \boldsymbol{\beta} \tag{9}$$

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

$$y_i = \begin{cases} 1, & \text{if } \eta_i^\star > 0 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Binomial link functions: complementary log-log

$$\log(\lambda_i) = \log(-\log(1 - \pi_i))$$
$$\lambda_i = -\log(1 - \pi_i) = \exp(\eta_i) \tag{11}$$

▶ 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
22	Shade	D > 2	H >= 5	Early	12	1

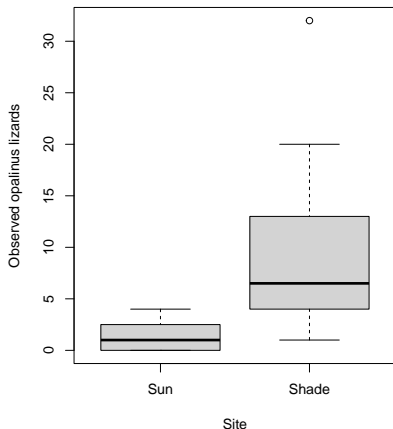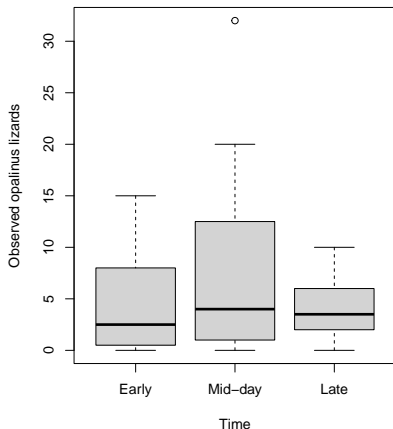## Lizards: visually inspect the data
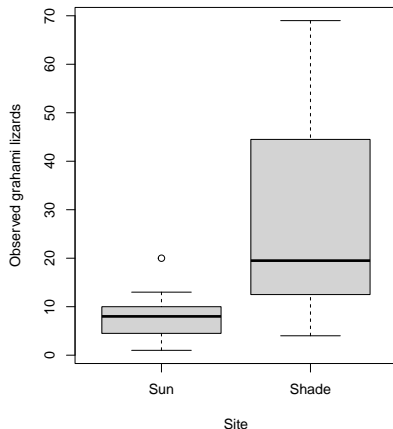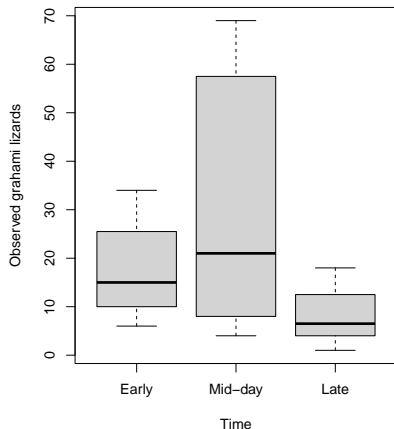
What can we tell about grahami lizards in Sunny or Shady sites?

Lizards: visually inspect the data
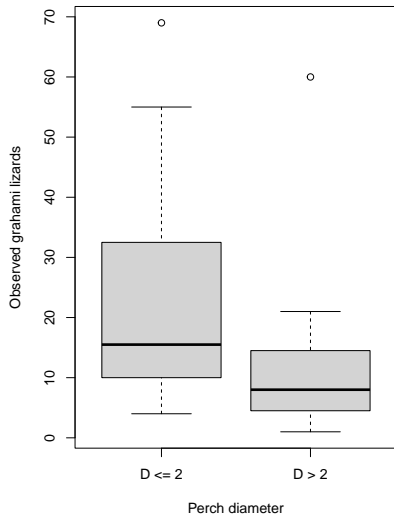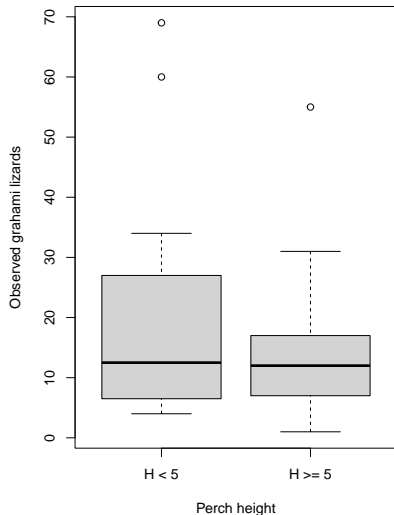
Lizards: visually inspect the data

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

Lizards: fit the model

```
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

```
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
```
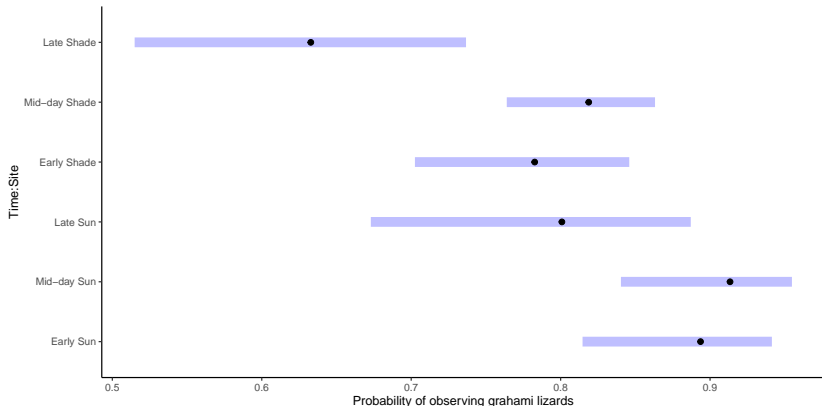
Lizards: parameter estimates with different link functions

```
##                   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
▶ $\text{var}(y_i) = N_i \pi_i (1 - \pi_i)$

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

```
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 by Physalia

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

```
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
```

Choosing the wrong link

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