evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a Státním rozpočtem ČR
InoBio – CZ.1.07/2.2.00/28.0018

Logistic regression

14. 11. 2013

Ing. Daniel Volařík, Ph.D.

european social fund in the czech republic

EUROPEAN UNION

MINISTRY OF EDUCATION, YOUTH AND SPORTS

OP Education for Competitiveness

INVESTMENTS IN EDUCATION DEVELOPMENT

Recup

- Linear model
  - Constant variance
  - Normality
  - independence
- Generalised linear models
  - Linear predictor
  - Link function
  - Distribution of errors (Gamma, Poisson, binomial)

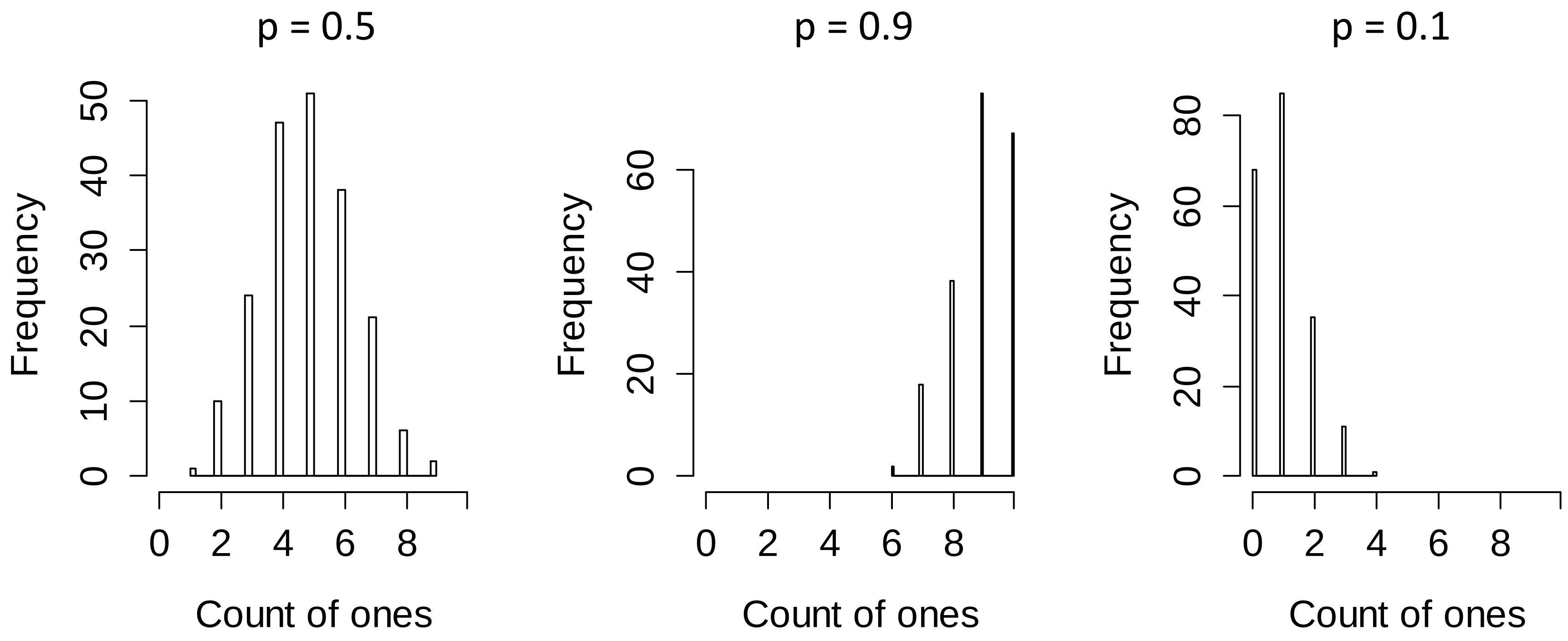
Logistic regression

- GLM with binomial error distribution
- Presence-absence data
- Proportional data

Binomial distribution

- A sequence of independent Bernoulli trials (like tossing a coin).
- A two-parameter distribution: the number of trials, *N*, and the probability of success, *p,* in any given trial.
- Mean is given by $N \times p$
- Variance by $N \times p \times (1 – p)$
- Assumption: probability of success does not change from trial to trial.

Binomial error distribution



- All examples are from samples with N = 10
- Each sampling has 200 runs
- In R you can explore this using rbinom() function

Presence-absence data

- binary (0 or 1) responses such as presence versus absence of an organism, alive versus dead, male versus female and so on.
- Binary data can only take two possible values - zero or one - and can therefore not be normally distributed.
- Special version of the binomial distribution known as the Bernoulli distribution where the number of trials is one (n = 1).

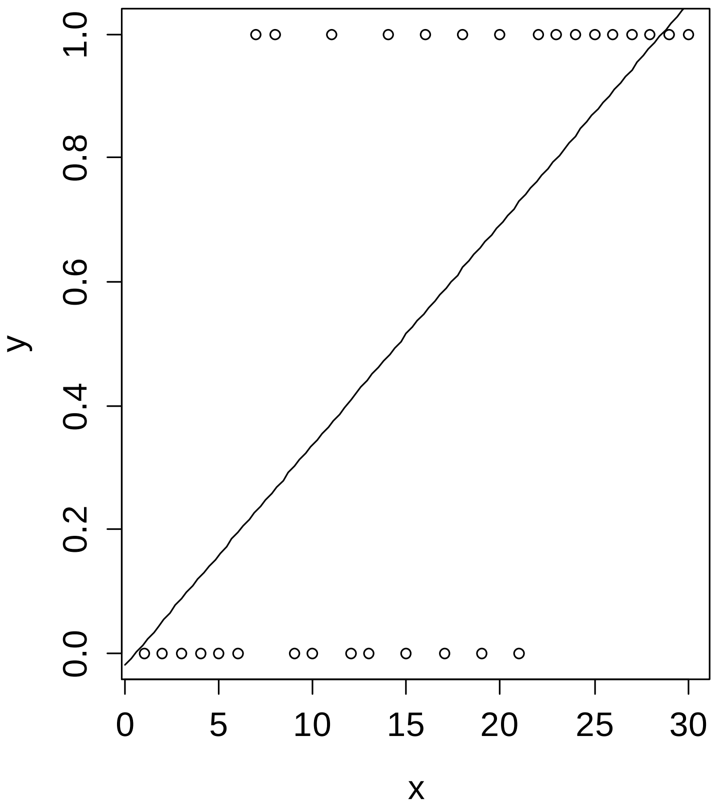
Proportional data

- For example we sample N animals for presence-absence of some disease, or we sample N trees weather they are dead or alive
- We have proportion of animals with disease or proportion of dead trees
- binomial distribution

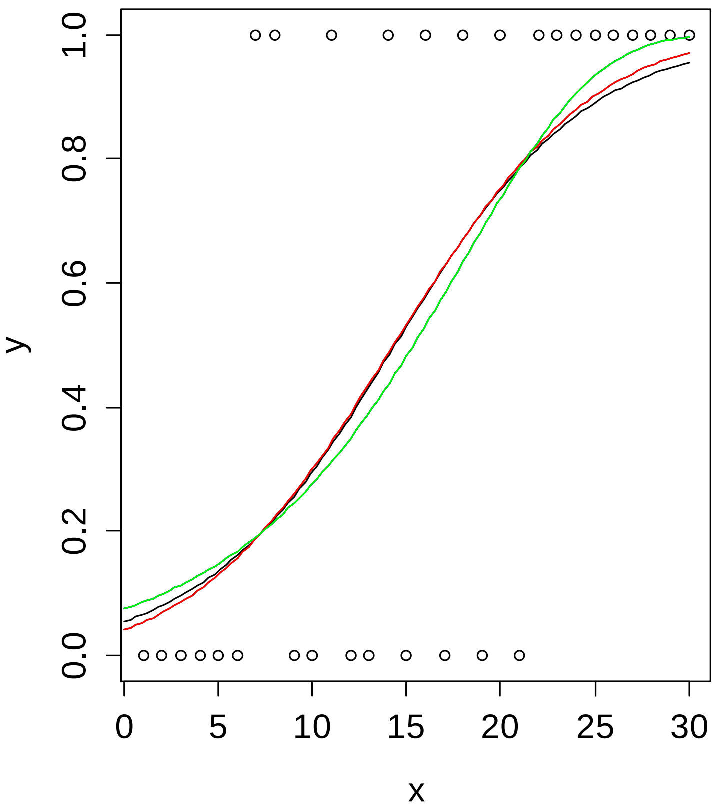
Link functions

- logit: $\mathrm{g}(p) = \log\left(\frac{p}{1-p}\right)$
- probit: $\mathrm{g}(p) = \phi^{-1}(p)$, where ϕ is the cumulative density function of the standard normal distribution
- cloglog – complementary log-log:
$$\mathrm{g}(p) = \log(-\log(1-p))$$
- Sigmoidal shape curves bounded between 0 and 1
- Differente links have slightly differente shape
- Cloglog could be good choise when there is considerable more zeros than ones or vice versa.

Link functions – example



Linear regression



GLM

Logit – black curve

Probit – red curve

Cloglog – green curve

Inverse link – antilogit

- $\log\left(\frac{p}{1-p}\right)$ = A + B*x
- If you want to make predictions on probability scale, you have to use inverse link – in case to logit it is antilogit
- $\left(\mathrm{logit}^{-1} = \frac{1}{1+ e^{-(linear\ predictor)}}\right)$

Overdispersion

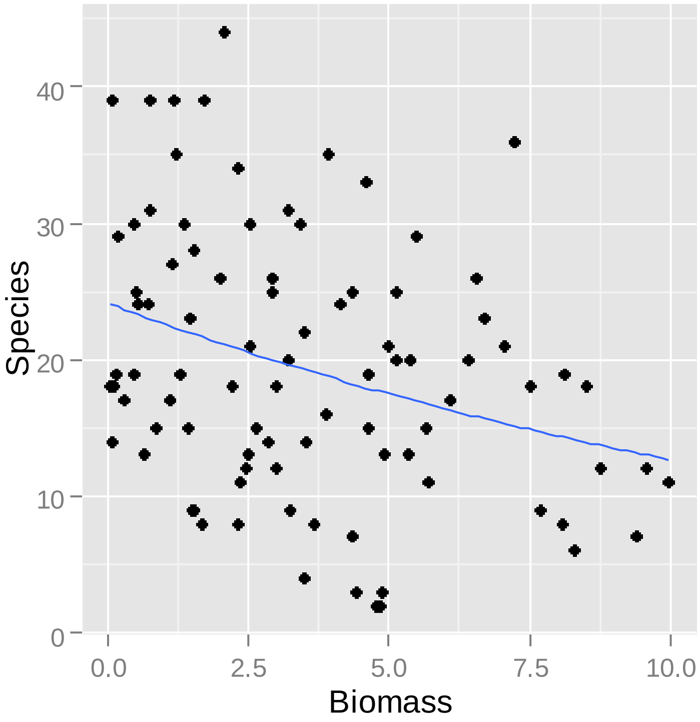
- In binomial distribution, mean is given by $N \times p$, variance by $N \times p \times (1 – p)$,
- for binary data mean is given by p and variance by p × (1-p)
- Overdispersion – when the variance is higher than expected
- (underdispersion – variance is lower than expected)
- Causes of overdispersion:
  - Important covariates or interactions are missing, outliers, wrong link function, non-linear effect entered as linear effect, …
  - When we can not find any of previos causes, it is real overdispersion – the variance is really larger than expected

Overdispersion - solution

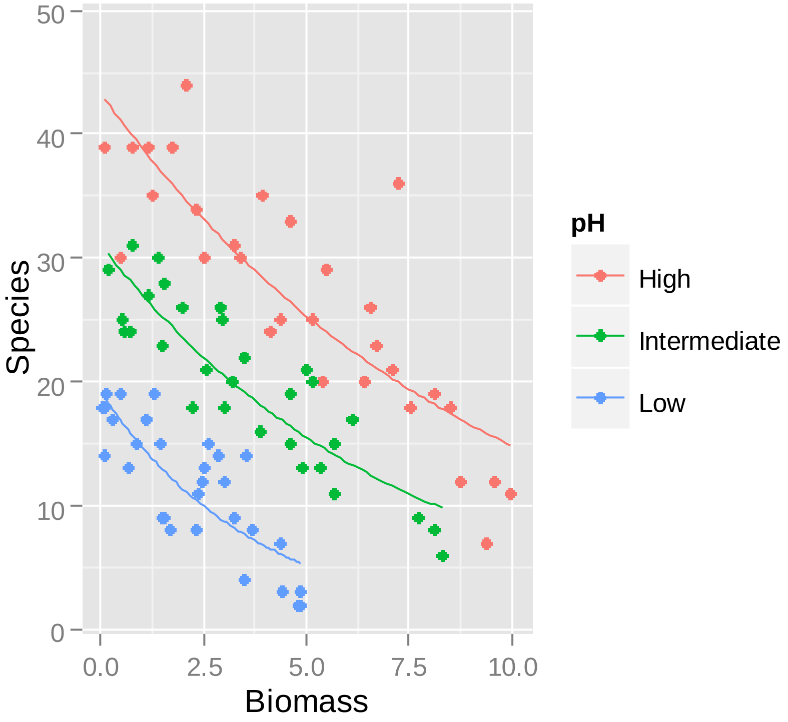
- We can detect overdispersion by comparing residual deviance with residual degrees of freedom – they should be approximately equal
- Quasi-binomial distribution
- Dispersion parameter ϕ

Overdispersion

- Nice example from yesterday



Residual deviance: 407.67
on 88 degrees of freedom



Residual deviance: 83.20
on 84 degrees of freedom

Logistic regression in R

- Function glm() with parameters:
  - Formula (for linear predictor – the same as in lm()
  - Family – binomial
  - Link function – one of logit, probit, cloglog
- Example:

```
glm(response ~ x + z, data = data,
 family = binomial(link = "logit"))
```

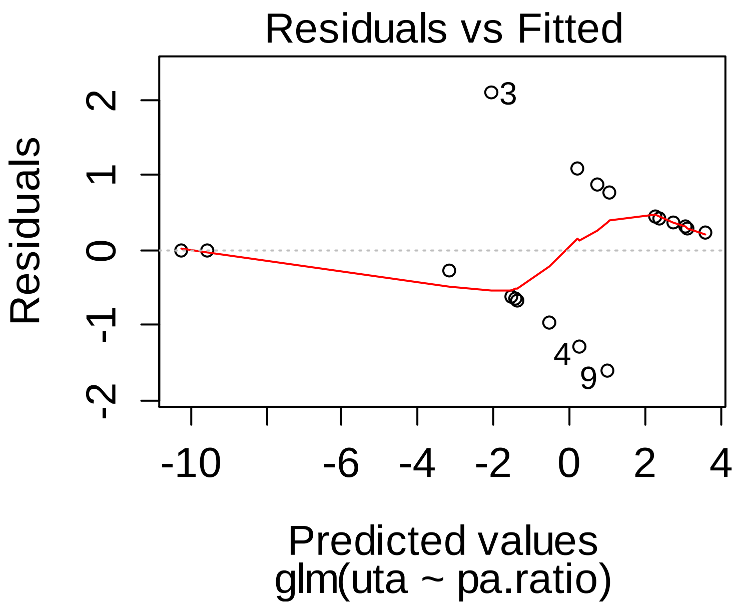
- For presence-absence data, the response is vector of zeros and ones
- For proportional data, the response is a list of two vector – one with positive cases and one with negative

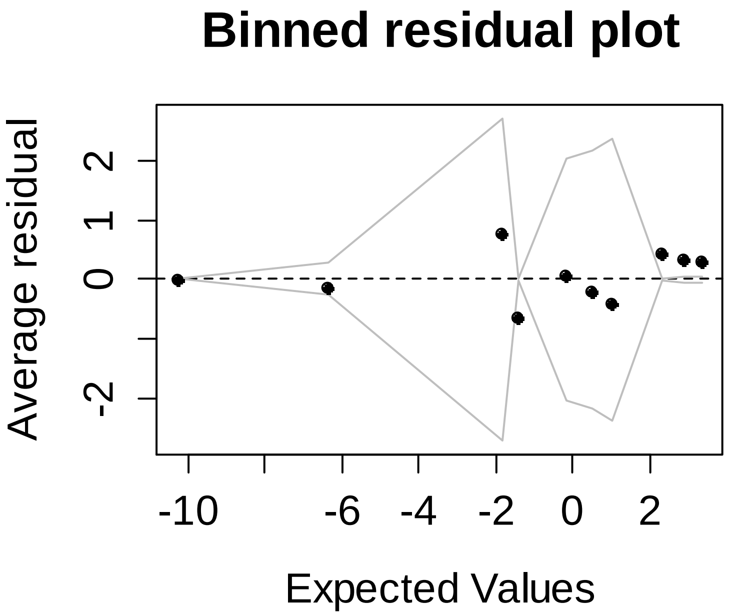
```
Response = cbind(positive_cases, negative_cases)
```

Model checking

- Check for equality of residual deviance and residual degrees of freedom for proportional data.
- binned plot
  - plots of raw residuals from logistic regression are usually not useful – because data are discrete and so are residuals
  - Instead we can plot binned (grouped) residuals vs fitted values
  - There is a degree of arbitraries in the size of bins
  - In R function binnedplot from arm package

Model checking





- Lines in the binned residual plot are +/- 2 standard-error bounds within which 95% of the binned residuals are expected to lie if the model is ‘true’

Exercices

- Presence-absence of lizards Uta in relation to perimeter-to-area ratio
- Proportional data – germination of Orobanche seed stimulated by extract of cucumber and beam