# Linear and Generalized Linear Models

Week 7, Lecture 3

Saskia le Cessie

Leiden University Medical Centre

# Yesterday

- There was an error on the slide on the AIC → repaired

```
glm(formula = agvhd ~ agedon + as.factor(diag) + sexdon + sexrec +    sexdon * sexrec, family = binomial, data = bonemarrow)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.70008    0.55629  -1.258  0.20821
agedon             0.03672    0.01515   2.423  0.01538 *
as.factor(diag)2 -1.43816    0.53595  -2.683  0.00729 **
as.factor(diag)3 -2.20872    0.67311  -3.281  0.00103 **
sexdon             0.61243    0.51447   1.190  0.23389
sexrec             0.90014    0.50119   1.796  0.07249 .
sexdon:sexrec     -1.41596    0.72550  -1.952  0.05098 .

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.94  on 165  degrees of freedom
Residual deviance: 184.75  on 159  degrees of freedom
AIC: 198.75
```

# Today

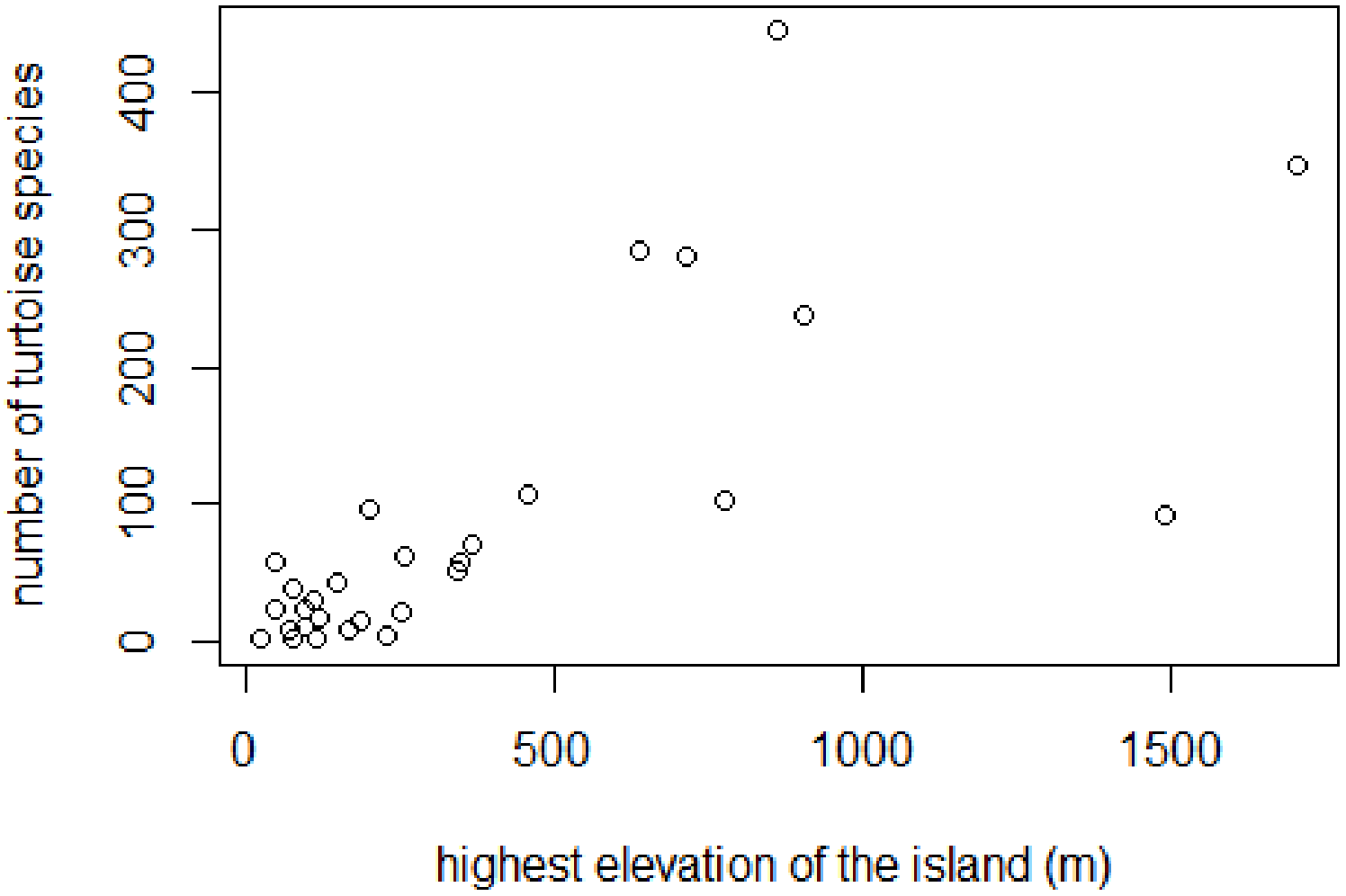
- Poisson regression
- Generalized linear models
- General remarks

# Models for counts

Counts of species of tortoises at the Galapagos Islands.

- 30 cases (different Islands)
- Outcome: number of species of tortoises at the island
- 5 geographic variables
  - Elevation:  The highest elevation of the island (m)
  - Area: Area of the island ($km^2$ )
  - Nearest: The distance from the nearest island (km)
  - Scruz: The distance from Santa Cruz island (km)
  - Adjacent: The area of the adjacent island ($km^2$ )

# Plot


number of turtoise species
0
100
200
300
400
0
500
1000
1500
highest elevation of the island (m)

# What do we observe?

- Standard deviation clearly increases with increasing elevation.
- Standard deviation seems to depend on mean(Y)

- Transformation? Use log(Y) or sqrt(Y) ?

- Here, use a model for counts, based on the Poisson distribution.

# Poisson distribution

- $Y_i \sim Pois(\mu_i)$
- Probability density function is $P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$
- $E(Y) = \mu$ and $Var(Y) = \mu$.
- Expected count is usually modeled on the log-scale
- $\log(\mu_i) = \eta_i = x_i' \beta$

# Poisson distribution

Arises naturally in several ways:

- Approximation of binomial distribution if $n$ is large and $p$ is small
    - Examples: modeling incidence of rare forms of diseases in large populations
- If probability of occurrence of events in time interval is proportional to the length of time interval, and events occur independently. Then number of events in any specified time interval has Poisson distribution
    - Examples: the number of incoming telephone calls per day, number of earthquakes per month).
- If time between events is independent and identically exponentially distributed. The number of events in given time period then has Poisson distribution.

```
> model1 <- glm(Species~Elevation, family=poisson, data=gala)
> summary(model1)
Call:
glm(formula = Species ~ Elevation, family = poisson, data = gala)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-15.6209   -6.3990   -2.5511    0.6435   20.9499

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.665e+00  3.157e-02  116.08   <2e-16 ***
Elevation   1.436e-03  3.184e-05   45.11   <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: 3510.7  on 29  degrees of freedom
Residual deviance: 1826.2  on 28  degrees of freedom
AIC: 1991

Number of Fisher Scoring iterations: 5
```

# Interpretation coefficients

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.665e+00  3.157e-02  116.08   <2e-16 ***
Elevation   1.436e-03  3.184e-05   45.11   <2e-16 ***
---
```

Model: $\log(\mu) = \beta_0 + \beta_1 x_1$ ; $\mu = e^{\beta_0 + \beta_1 x_1}$

Additive

- One meter increase in elevation corresponds to a 1.436e-03 increase in expected log(count) of tortoise
- One meter increase in elevation corresponds to a exp(1.436e-03)=1.0014 times higher expected number of tortoise species

Multiplicative

# Add more x-variables to the model

```
glm(formula = Species ~ Elevation + Area + Nearest + Scruz +  Adjacent, family = poisson, data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 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: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

All covariates are significantly associated with the expected number of turtoises

## Many issues are very similar to logistic and linear regression

- Parameter estimation via Maximum likelihood
- Deviance D = -2 log likelihood fitted model – -2 log likelihood saturated model
- Saturated model: a model with as many parameters as observations so that the data are fitted exactly. $\hat{\mu}_i^{sat} = y_i$
- Test overall fit of the model by comparing D with $\chi^2_{n-p}$ distribution.
- Testing individual predictors and making confidence intervals using likelihood methods or Wald test (or score tests)

# Goodness of fit, our example

```
Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68
```

n=30 islands
p= 6 number of estimated parameters
n-p= 24

P-value of goodness of fit test:

1-pchisq(q=716.85, df=24) = 0.00000000

Model does not fit ….

- Check linearity assumptions
- Interactions needed?
- Poisson distribution too restrictive?

# Overdispersion in Poisson regression

- Poisson distribution: very strict relationship between variance and mean: mean$(Y)$ =var$(Y)$
- Often var $(Y) > \mu$
- Easy, simplistic, way out is, to introduce extra scale parameter $\phi$.
  - $Var(Y) = \phi E(Y)$.
  - If $\phi > 1$, then overdispersion, if $\phi < 1$ then underdispersion.
  - Standard errors of estimates will change

```
glm(formula = Species ~ Elevation + Area + Nearest + Scruz +
    Adjacent, family = quasipoisson(), data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.1548079  0.2915901  10.819 1.03e-10 ***
Elevation    0.0035406  0.0004925   7.189 1.98e-07 ***
Area        -0.0005799  0.0001480  -3.918 0.000649 ***
Nearest      0.0088256  0.0102622   0.860 0.398292
Scruz       -0.0057094  0.0035251  -1.620 0.118380
Adjacent    -0.0006630  0.0001653  -4.012 0.000511 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 31.74921)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
```

Standard errors have increased
Effect of `Nearest` no longer statistically significant

Var(Y) = 31.7 $\mu$

# Overdispersion in logistic regression


Can occur if there are replicated binary responses with same covariate values.

- Example: Winter wheat (certain grain)
- Pots with $n_i$ = 5 plants
- $Y_{ij}$: outcome of the *j*th plant in pot *i*: survived the winter (yes/no)
  - $Y_{ij}$~ binomial(1, $\pi_i$) .
- $M_i$: number of plants that survived in pot $i$
- $M_i$ ~ binomial($n_i$, $\pi_i$) .

Binomial distribution assumes E($M_i$)= $n_i\pi_i$, var($M_i$)= $n_i\pi_i(1-\pi_i)$.

→ overdispersion: var($M_i$)= $\phi n_i\pi_i(1-\pi_i)$.

# Poisson regression for rates

Examples of rates

- Number of burglaries reported in different cities depend on number of households in cities
- Interest in rate: number of burglaries per 10,000 households

- Number of surgery-complications in hospital depend on total number of surgeries performed
- Interest in rate: number of burglaries per 1000 surgeries

- Number of deaths depend on duration of follow up of patients
- Interest in Rate: number of deaths per person-year follow up

# Rate models

- Poisson distribution for counts: $\mathrm{Y} \sim Pois(\mu)$
- Model for rate = $\mu$/total
- $\log \frac{\mu}{total} = x'\beta.$
- $\log(\mu) = x'\beta + \log(total).$
- $\log(total)$ is a regression variable with regression coefficient identical to 1. This is called an offset. Other regression coefficient(s) will be estimated as before.
- → more in the exercises

# Generalized linear models

# Different regression models for different types of responses

| Type of outcome | Regression model |
|---|---|
| Numerical | Linear or non-linear regression |
| Binary (1/0, success/failure) | Logistic regression |
| Counts | Poisson regression |
| Rates (number of new cases per time period) | Poisson regression |

## Similarities between these models

Three elements

1. The effect of the different variables is summarized in a linear predictor:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = x'\beta$$

2. There is a link function g, which links the mean $\mu = E(Y)$ to the linear predictor: $\mathbf{g}(\mu) = \eta$.

3. Random part of model: distribution of response variable $Y$

Generalized linear models

# Generalized linear models

1. Linear predictor $\eta = x'\beta$
2. Link function $\mathbf{g}(\mu) = \eta$.

3. Distribution of Y is from the exponential family

Density function can be written as $f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$

# Exponential family, some examples

| | Normal | Poisson | Binomial |
|---|---|---|---|
| Notation | $N(\mu, \sigma^2)$ | $P(\mu)$ | $Bin(1, \pi)$ |
| Range | $(-\infty, \infty)$ | $(0, \infty)$ | $(0,1)$ |
| Dispersion parameter $\phi$ | $\phi = \sigma^2$ | 1 | 1 |
| Variance V($\mu$) | 1 | $\mu$ | $\mu$(1-$\mu$) |
| Canonical link function | | | |
| .. | | | |
| .. | | | |

# GLMs and canonical link function

- Canonical link function is a link function that follows naturally from exponentially family notation

Canonical links:

- Normal distribution: Identity: $\eta = \mu$
- Poisson distribution: Log: $\eta = \log(\mu)$
- Binomial distribution Logit $\eta = \log(\mu/(1-\mu))$

- Gamma distribution: $\eta = \mu^{-1}$
- Inverse Gausian: $\eta = \mu^{-2}$

Canonical link is mathematically and computationally efficient. Often natural choice, but not required.

# GLM: a general framework for fitting and handling regression models

- Parameters and standard errors are estimated with maximum likelihood
- Numerical iterative optimalisation methods are used to find maximum
- Testing parameters with Wald or Likelihood ratio tests (or score tests)
- Confidence intervals for regression coefficients with Wald or profile likelihood methods
- Model diagnostics like residuals, influential points, points with high leverage

# Final comments

# Why do we need regression models?

- To predict
- To estimate an effect of one variable, adjusted for other variables
  - Handling confounding
- To increase precision
  - Residual variance can be reduced by adjusting for other predictors of the outcome
- As a general tool

# Why do we need regression models?

- To predict
  - Regression models are basis of several machine learning methods. More in Course Statistical Learning
- To estimate an effect of one variable, adjusted for other variables
  - Confounding More in course Causal Inference
- To increase precision
  - Residual variance can be reduced by adjusting for other predictors of the outcome
  - More in course Design and analysis of Biological Experiments
- As a general tool More in course Causal Inference

# What did we not discuss?

- We assumed independent observations (except when discussing overdispersion)
  - More on correlated observations (multi level models, repeated measurements) in course Essentials of Mixed and Longitudinal Modelling

- We did not discuss censored observations (time-to-event data, values below a detection limit)
  - More in survival analysis

# Exam

- I will make some practice exercises on logistic regression and GLM for the exam
- Will be quite similar to questions asked in practicals and in lecture

# Evaluations

- You will receive online evaluation form for this course and for all courses
- Please fill them in, it is needed to improve courses