TMA4315 Generalized linear models H2018

Module 4: Count and continuous positive response data (Poisson and gamma regression)

Mette Langaas, Department of Mathematical Sciences, NTNU – with contributions from Ingeborg Hem

27.09.2018 and 04.10.2018 [PL], 28.09.2018 and 05.10.2018 [IL]

Contents

Overview **3**
- Learning material . . . 3
- Topics . . . 3

Examples of count data **3**
- Sales of newspapers . . . 4
- Female crabs with satellites . . . 4

Modelling counts with the Poisson distribution **5**
- The Poisson process . . . 5
- The Poisson distribution . . . 6
- Properties of the Poisson distribution . . . 7
- Exponential family . . . 7

Regression with count data **7**
- Aim . . . 7
- What is we instead want to use the multiple linear regression model? . . . 8
- The linear Poisson model . . . 8
- The log-linear Poisson model . . . 8
- Interpreting parameters in the log-linear Poisson model . . . 9
- Example: interpreting parameters for the female crabs with satellites . . . 9

Parameter estimation with maximum likelihood **10**
- Likelihood $L(\beta)$ . . . 10
- Loglikelihood $l(\beta)$ . . . 10
- Score function $s(\beta)$ . . . 11
- The expected Fisher information matrix $F(\beta)$ . . . 11
- Observed Fisher information matrix $H(\beta)$ . . . 12

Parameter estimation - in practice **13**
- Fisher scoring . . . 13
- Requirements for convergence . . . 13

Statistical inference **13**
- Asymptotic properties of ML estimates . . . 13
- Confidence intervals . . . 14
- Example: Female crabs with satellites . . . 14
- Hypothesis testing . . . 15

Interactive session - first week **19**
- Problem 1: Exam 2005 (Problem 1d-f - slightly modified) - Female crabs and satellites . . . 19
- Problem 2: Exam 2017 (Problem 1) - Poisson regression . . . 21
- Problem 3: Exam December 2017 from UiO, Problem 1. . . . 23

Poisson regression for count data **23**

Residuals **23**
Deviance residuals . . . 24
Pearson residuals . . . 24
Plotting residuals . . . 24

Model assessment and model choice **28**
Deviance test . . . 28
Pearson test . . . 29
Example: goodness of fit with female crabs . . . 29
AIC . . . 30
Analysis of deviance . . . 30

Overdispersion **30**

Rate models and offset **31**
Example: British doctors and rate models . . . 32

Modelling continuous positive response data **35**
Examples of continuous positive responses . . . 35
Models for continuous positive responses . . . 35
Time to blood coagulation . . . 35

Lognormal distribution **36**

Gamma regression **37**
The gamma distribution . . . 37
Gamma GLM model . . . 39

Gamma regression: likelihood and derivations thereof **39**

Scaled and unscaled deviance **40**

Comparing models **41**
Comparing models based on deviance . . . 41
Comparing models based on AIC . . . 41

Interactive session - second week **43**
Problem 1: Exam 2007 (Problem 1, a bit modified) - Smoking and lung cancer . . . 43
Problem 2: TMA4315 Exam 2012, Problem 3: Precipitation in Trondheim, amount . . . 46
Problem 3: Taken from UiO, STK3100, 2015, problem 2 . . . 47

Work on your own: Exam questions **49**
December 2013 (Essay exam) . . . 49

R packages **49**

Further reading **49**

(Latest changes: 06.10: solutions added. 01.10: small changes for second week. 27.09: added one Problem for ILw1, moved stuff to w2, added a few dimensions to score test.)

Overview

Learning material

- Textbook: Fahrmeir et al (2013): Chapter 5.2, 5.3.
- Classnotes 27.09.2018
- Classnotes 04.10.2018

Topics

First week

- examples of count data
- the Poisson distribution
- regression with count data
- Poisson regression with log-link
- parameter estimation (ML): log-likelihood, score vector, information matrix to give iterative calculations
- asymptotic MLE properties
- confidence intervals and hypothesis tests (Wald, score and LRT)

Jump to interactive (week 1)

Second week

- Count data with Poisson regression (continued)
- deviance, model fit and model choice
- overdispersion
- rate models and offset
- Modelling continuous response data: lognormal and gamma
- the gamma distribution
- the gamma GLM model
- gamma likelihood and derivations thereof
- dispersion parameter: scaled and unscaled deviance

Jump to week 2 and interactive (week 2).

FIRST WEEK

Examples of count data

- the number of automobile thefts pr city worldwide
- the number of UFO sightings around the world
- the number of visits at web pages
- the number of male crabs (satellites) residing nearby a female crab
- the number of goals by the home team and the number of goals for the away team in soccer

- the number of newspapers sold at newsagents

Sales of newspapers

This is a short description of a project run at the Norwegian Computing Centre a few years ago.

The aim of the project was to provide a statistical model to predict the number of newspapers to be sold at each of 11 thousand outlets all over Norway for a given day ("tomorrow" - maybe scaled to match front page issues). And then based on the prediction to decide on how many newspapers to be delivered to each outlet in order to optimize the overall profit (lost sales if outlet are sold out, return costs if unsold papers).

Response data: number of newspapers (delivered) sold at each outlet. Covariate data: type of outlet, but mainly calendar information= weekday, month, season, public holidays, winter/autumn/easter/xmas, ...

Female crabs with satellites

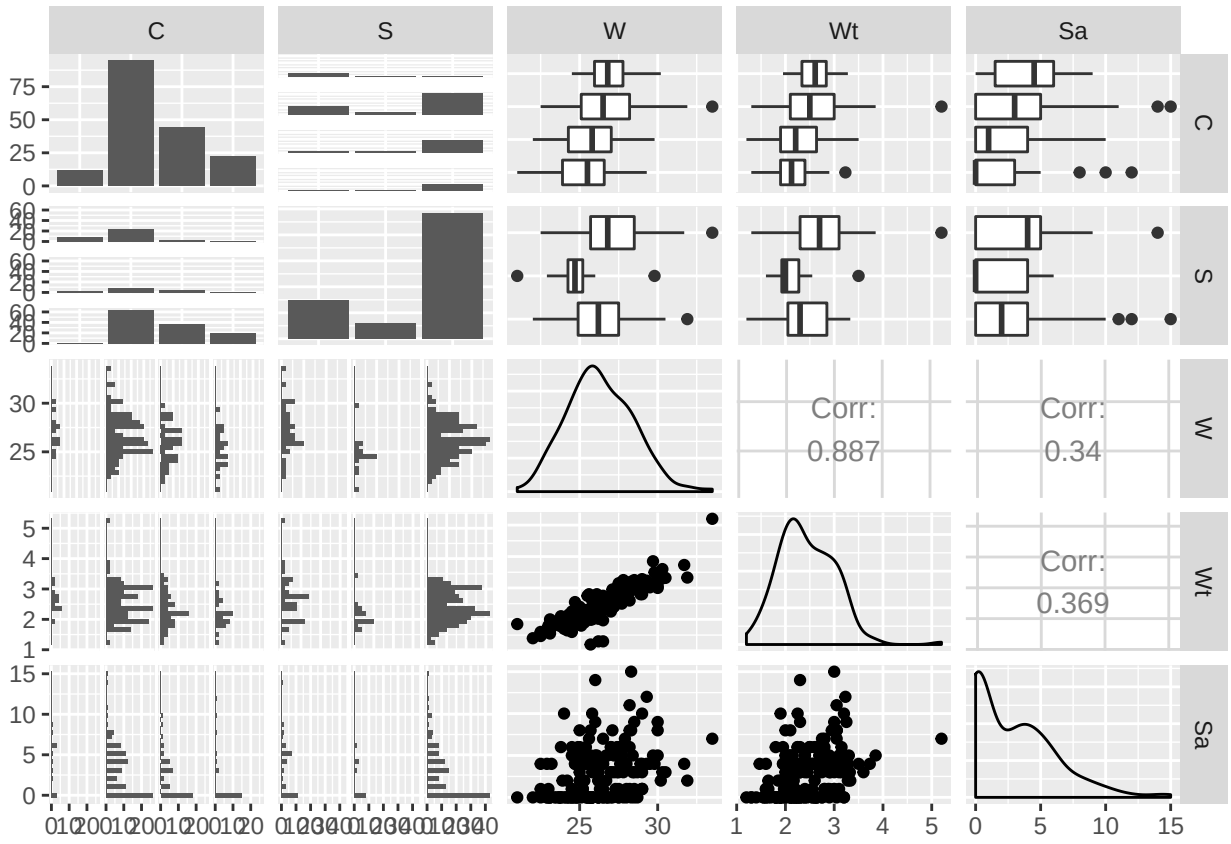
The example is taken from Agresti (1996): "An Introduction to Categorical Data Analysis", and the data is from a study of nesting horseshoe crabs (J. Brockmann, Ethology 1996)

First, the study objects were female crabs (horseshoe crabs). Each female crab had a male crab attached to her in her nest. The objective of the study was to investigate factors that affect whether the female crab had any other males, called satellites, residing near her. The following covariates were collected for 173 female crabs:

- `C`: the color of the female crab (1=light medium, 2=medium, 3=dark medium, 4=dark)
- `S`: spine condition (1=both good, 2=one worn or broken, 3=both worn or broken)
- `W`: width of carapace (cm)
- `Wt`: weight (kg)

The response was the number of satellites, `Sa`= male crabs residing nearby.

```
library(ggplot2)
library(GGally)
crab = read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/crab.txt")
colnames(crab) = c("Obs", "C", "S", "W", "Wt", "Sa")
crab = crab[, -1]  #remove column with Obs
crab$C = as.factor(crab$C)
crab$S = as.factor(crab$S)
ggpairs(crab)
```



Q: Discuss what you see. Any potential covariates to influence `Sa`? Which distribution can `Sa` have?

```
## **A:** Seems that `C`, `W` and `Wt` might be correlated with `Sa`. Also `W` and `Wt` are very correla
```

Modelling counts with the Poisson distribution

The Poisson process

We observe events that may occur within a time interval or a region.

1. The number of events occuring within a time interval or a region, is independent of the number of events that occurs in any other disjoint (non-overlapping) time interval or region.
2. The probability that a single event occurs within a small time interval or region, is proportional to the length of the interval or the size of the region.
3. The probability that more than one event may occur within a small time interval or region is negligable.

When all of these three properties are funfilled we have a *Poisson process*. This leads to three distributions

- The number of events in a Poisson process follows a Poisson distribution.
- Time between two events in a Poisson process follows an exponential distribution.
- Time between many events in a Poisson process follows a gamma distribution.

We will first study the Poisson distribution - and link it to a regression setting.

The Poisson distribution

We study a Poisson process within a time interval or a region of specified size. Then, the number of events, Y, will follow a *Poisson distribution* with parameter λ

$$f(y) = \frac{\lambda^y}{y!}e^{-\lambda} \text{ for } y = 0, 1, 2, \ldots$$

Here the parameter λ is the proportionality factor in the requirement 2 (above) for the Poisson process. Another popular parameterization is μ, or given some interval λt, but we will stick with λ. In R we calculate the Poisson point probabilities using `dpois`.

If you want to see how this distribution function is derived form the binomial distribution you may watch this video: Poisson process and distribution

---

Cumulative distribution (cdf): The cumulative distribution is calculated by summing $F(y) = \sum_{t \leq y} f(t)$, and we might calculate the Poisson cdf in R with `ppois`.

Expected value and variance: Let Y follow a Poisson distribution with parameter λ. Then

$$\mathrm{E}(Y) = \lambda \text{ and } \mathrm{Var}(Y) = \lambda$$

Proof:

$$E(y) = \sum_{y=0}^{\infty} y\frac{\lambda^y}{y!}e^{-\lambda} = \sum_{y=1}^{\infty} y\frac{\lambda^y}{y!}e^{-\lambda} = \sum_{y=1}^{\infty} \frac{\lambda\lambda^{y-1}}{(y-1)!}e^{-\lambda} = \lambda\sum_{z=0}^{\infty} \frac{\lambda^z}{z!}e^{-\lambda} = \lambda$$

In the first transition we use that the term $y = 0$ gives no contribution to the sum. Then we cancel out y in the numerator with the first term of $y!$ in the denominator. Then we let $z = y - 1$, and finally we use that the sum of the Poisson distribution for all possible outcomes equal 1.

To calculate the variance we first use that

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2$$

and then that

$$Y^2 = Y(Y-1) + Y$$

so that

$$\mathrm{Var}(Y) = \mathrm{E}(Y(Y-1)) + \mathrm{E}(Y) - (\mathrm{E}(Y))^2$$

The reason for this is that we may use the same type of trick as for $\mathrm{E}(Y)$ since the sum is over all values of the distribution function.

$$\mathrm{E}(Y(Y-1)) = \sum_{y=0}^{\infty} y(y-1)\frac{\lambda^y}{y!}e^{-\lambda} = \sum_{y=2}^{\infty} y(y-1)\frac{\lambda^y}{y!}e^{-\lambda}$$

$$= \sum_{y=2}^{\infty} \frac{\lambda^{y-2}\lambda^2}{(y-2)!}e^{-\lambda} = \lambda^2\sum_{y=0}^{\infty} \frac{\lambda^y}{y!}e^{-\lambda} = \lambda^2$$

Putting it all togehter

$$\mathrm{Var}(Y) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

which was to be shown.

Properties of the Poisson distribution

- A sum of n independent Poisson distributed random variables, Y_i with means λ_i are Poisson distributed with mean $\sum_{i=1}^{n} \lambda_i$.
- When the mean increases the Poisson distribution becomes more and more symmetric and for large λ the Poisson distribution can be approximated by a normal distribution.

Exponential family

In Module 1 we introduced distributions of the Y_i, that could be written in the form of a *univariate exponential family*

$$f(y_i \mid \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} \cdot w_i + c(y_i, \phi, w_i)\right)$$

where we said that

- θ_i is called the canonical parameter and is a parameter of interest
- ϕ is called a nuisance parameter (and is not of interest to us=therefore a nuisance (plage))
- w_i is a weight function, in most cases $w_i = 1$
- b and c are known functions.

It can be shown that $\mathrm{E}(Y_i) = b'(\theta_i)$ and $\mathrm{Var}(Y_i) = b''(\theta_i) \cdot \frac{\phi_i}{w_i}$, see derivation from Module 1.

In Module 1 we found that the Poisson distribution $Y_i \sim \text{Poisson}(\lambda_i)$ is an exponential family derivation from Module 1,

and that

- $\theta_i = \ln(\lambda_i)$ is the canonical parameter
- $\phi = 1$, no nuisance
- $w_i = 1$
- $b(\theta_i) = \exp(\theta)$
- $\mu_i = \mathrm{E}(Y_i) = \lambda_i$

For a GLM with linear predictor η_i - to have a canonical link we need

$$\theta_i = \eta_i$$

Since $\eta_i = g(\mu_i) = g(\lambda_i)$ this means to us that we need

$$g(\mu_i) = g(\lambda_i) = \theta_i$$

saying that with the Poisson the canonical link is $\ln(\lambda_i)$.

Q: Why may we want to choose a canonical link?

Regression with count data

Aim

1. Construct a model to help understand the relationship between a count variable and one or many possible explanatory variables. The response measurements are counts.

2. Use the model for understanding what can explain count, and for prediction of counts.

What is we instead want to use the multiple linear regression model?

When modelling the counts some times the normal approximation might be used, especially when the counts are high. It is also possible to use a transformation of the response to get a constant variance. The transformation $\sqrt{Y_i}$ give an approximate constant variance, but then it is not clear if there then is a linear relationship between $\mathrm{E}(\sqrt{Y_i})$ and the covariates.

In general, with count data, we instead use a Possion GLM regression.

The linear Poisson model

It is possible to construct a *linear Poisson model* where we have the direct relationship

$$\lambda_i = \eta_i$$

between the mean of the Poisson distribution and the linear predictor.

This means that the covariates have an additive effect on the rate λ_i.

However, since the rate λ_i can not be negative, then to use this model restrictions on the parameter space of the β is needed.

We will not use the linear Poisson model, but instead the *log-linear Poisson model*.

The log-linear Poisson model

Assumptions:

1. $Y_i \sim \text{Poisson}(\lambda_i)$, with $\mathrm{E}(Y_i) = \lambda_i$, and $\mathrm{Var}(Y_i) = \lambda_i$.
2. Linear predictor: $\eta_i = \mathbf{x}_i^T \beta$.
3. Log link
$$\eta_i = \ln(\lambda_i) = g(\lambda_i)$$
and (inverse thereof) response function
$$\lambda_i = \exp(\eta_i)$$

Assumptions 1 and 3 above can be written as

$$Y_i \sim \text{Poisson}(\exp(\eta_i)),\ i = 1, \ldots, n$$

Interpreting parameters in the log-linear Poisson model

In the log-linear model the mean, $\mathrm{E}(Y_i) = \lambda_i$ satisfy an exponential relationship to covariates

$$\lambda_i = \exp(\eta_i) = \exp(\mathbf{x}_i^T \beta) = \exp(\beta_0) \cdot \exp(\beta_1)^{x_{i1}} \cdots \exp(\beta_k)^{x_{ik}}.$$

Let us look in detail at β_1 with covariate x_{i1} for observation i.

1. If x_{i1} increases by one unit to $x_{i1}+1$ then the mean $\mathrm{E}(Y_i)$ will in our model change by a factor $\exp(\beta_1)$.
2. If β_1=0 then $\exp(\beta_1) = 1$, so that a change in x_{i1} does not change $\mathrm{E}(Y_i)$.
3. If $\beta_1 < 0$ then $\exp(\beta_1) < 1$ so if x_{i1} increase then $\mathrm{E}(Y_i)$ decrease.
4. If $\beta_1 > 0$ then $\exp(\beta_1) > 1$ so if x_{i1} increase then $\mathrm{E}(Y_i)$ increase.

Thus, the covariates have a multiplicative effect on the rate λ_i.

Example: interpreting parameters for the female crabs with satellites

We fit a log-linear model to `Sa`, assuming the number of satellites follows a Poisson distribution with log-link.

Q:

1. First the model is fitted with intercept only. What do we assume then? Interpret the fit.
2. Then width `W` is added as a covariate in the log-linear model. What happens if the width increase by one unit (cm)?
3. What is the predicted number of satellites for the average width?

```
model1 = glm(Sa ~ 1, family = poisson(link = log), data = crab)
cat("Intercept only\n")
print(model1$coefficients)
cat("Intercept only, exp\n")
exp(model1$coefficients)
print(mean(crab$Sa))
model2 = glm(Sa ~ W, family = poisson(link = log), data = crab)
cat("Intercept + W\n")
print(model2$coefficients)
cat("Intercept+W, exp\n")
exp(model2$coefficients)
cat("summary of width\n")
summary(crab$W)
cat("what is this?\n")
print(exp(model2$coefficients[1] + model2$coefficients[2] * mean(crab$W)))
```

```
## Intercept only
## (Intercept)
##    1.071267
## Intercept only, exp
## (Intercept)
##    2.919075
## [1] 2.919075
## Intercept + W
## (Intercept)           W
##  -3.3047572   0.1640451
## Intercept+W, exp
```

```
## (Intercept)           W
##  0.03670812  1.17826744
## summary of width
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    21.0    24.9    26.1    26.3    27.7    33.5
## what is this?
## (Intercept)
##    2.744061
```

Parameter estimation with maximum likelihood

Our parameter of interest is the vector β of regression coefficients, and we have no nuisance parameters. We would like to estimate β from maximizing the likelihood - the presentation here is essentially the same as for Module 3: Binary regression - with "Poisson and log" instead of "Bernoulli and logit". And, also here we will not have a closed form solution for $\hat{\beta}$ (except for a few special cases).

Likelihood $L(\beta)$

We assume that pairs of covariates and response are measured independently of each other: $(\mathbf{x}_i, Y_i)$, and Y_i follows the distribution specified above, and $\mathbf{x}_i$ is fixed.

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} f(y_i; \beta) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$$

Note: still a slight misuse of notation - where is β?

A: $\eta_i = \ln(\lambda_i) = \mathbf{x}_i^T \beta$, so replace λ_i with $\exp(\mathbf{x}_i^T \beta)$.

Loglikelihood $l(\beta)$

The log-likelihood is just the natural log of the likelihood, and we work with the log-likelihood because this makes the mathematics simpler - since we work with exponential families. The main aim with the likelihood is to maximize it to find the maximum likelihood estimate, and since the log is a monotone function the maximum of the log-likelihood will be in the same place as the maximum of the likelihood.

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \ln L_i(\beta) = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} [y_i \ln(\lambda_i) - \lambda_i - \ln(y!)]$$

Observe that the log-likelihood is a sum of invidual contributions for each observation pair i. We often omit the last term since it is not a function of model parameters, only data.

If we want a function of $\eta_i = \ln(\lambda_i)$ or β:

$$l(\beta) = \sum_{i=1}^{n} [y_i \eta_i - \exp(\eta_i) + C_i] = \sum_{i=1}^{n} y_i \mathbf{x}_i^T \beta - \sum_{i=1}^{n} \exp(\mathbf{x}_i^T \beta) + C$$

Score function $s(\beta)$

The score function is a $p \times 1$ vector, $s(\beta)$, with the partial derivatives of the log-likelihood with respect to the p elements of the β vector. Remember, the score function is linear in the individual contributions:

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta)$$

We work with $l_i(\beta) = l_i(\pi_i(\eta_i(\beta)))$ and use the chain rule to find $s_i(\beta)$.

$$\begin{aligned} s_i(\beta) = \frac{\partial l_i(\beta)}{\partial \beta} &= \frac{\partial l_i(\beta)}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta} = \frac{\partial [y_i \eta_i - \exp(\eta_i) + C_i]}{\partial \eta_i} \cdot \frac{\partial [\mathbf{x}_i^T \beta]}{\partial \beta} \\ &= [y_i - \exp(\eta_i)] \cdot \mathbf{x}_i = (y_i - \lambda_i)\mathbf{x}_i \end{aligned}$$

See Module 3 for rules for partial derivatives of scalar wrt vector.

The score function is given as:

$$s(\beta) = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n} (y_i - \lambda_i)\mathbf{x}_i$$

Again, observe that $\mathrm{E}(s_i(\beta)) = \mathrm{E}((Y_i - \lambda_i)\mathbf{x}_i) = 0$ because $\mathrm{E}(Y_i) = \lambda_i$, and thus also $\mathrm{E}(s(\beta)) = 0$.

To find the maximum likelihood estimate $\hat{\beta}$ we solve the set of p non-linear equations:

$$s(\hat{\beta}) = 0$$

And, as before we do that using the Newton-Raphson or Fisher Scoring iterative methods, so we need the derivative of the score vector (our Fisher information).

The expected Fisher information matrix $F(\beta)$

We saw in Module 3 that the expected Fisher information matrix, $F(\beta)$ is equal the covariance matrix of the score function.

$$\begin{aligned} F(\beta) = \mathrm{Cov}(s(\beta)) &= \sum_{i=1}^{n} \mathrm{Cov}(s_i(\beta)) && (1) \\ &= \sum_{i=1}^{n} E\left[\Big(s_i(\beta) - E(s_i(\beta)) \Big) \Big(s_i(\beta) - E(s_i(\beta)) \Big)^T \right] && (2) \\ &= \sum_{i=1}^{n} E(s_i(\beta) s_i(\beta)^T) = \sum_{i=1}^{n} F_i(\beta) && (3) \end{aligned}$$

where it is used that the responses Y_i and Y_j are independent, and that $E(s_i(\beta)) = 0 \ \forall i$.

Remember that $s_i(\beta) = (Y_i - \lambda_i)\mathbf{x}_i$, then:

$$F_i(\beta) = E(s_i(\beta) s_i(\beta)^T) = E((Y_i - \lambda_i)\mathbf{x}_i (Y_i - \lambda_i)\mathbf{x}_i^T) = \mathbf{x}_i \mathbf{x}_i^T E((Y_i - \lambda_i)^2) = \mathbf{x}_i \mathbf{x}_i^T \lambda_i$$

where $E((Y_i - \lambda_i)^2) = \text{Var}(Y_i) = \lambda$ is the variance of Y_i. Thus

$$F(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \lambda_i.$$

Observed Fisher information matrix $H(\beta)$

We do not really need the observed version of the Fisher information matrix, and since we use canonical link $H(\beta) = F(\beta)$ - so we already have it.

But, for completeness, we add the direct derivation of $H(\beta)$.

$$\begin{aligned} H(\beta) &= -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\frac{\partial s(\beta)}{\partial \beta^T} = \frac{\partial}{\partial \beta^T} \left[\sum_{i=1}^{n} (\lambda_i - y_i) \mathbf{x}_i \right] \\ &= \frac{\partial}{\partial \beta^T} \left[\sum_{i=1}^{n} (\exp(\eta_i) - y_i) \mathbf{x}_i \right] \end{aligned}$$

because $s(\beta) = \sum_{i=1}^{n} (y_i - \lambda_i)\mathbf{x}_i$ and hence $-s(\beta) = \sum_{i=1}^{n} (\lambda_i - y_i)\mathbf{x}_i$. Note that $\lambda_i = \exp(\eta_i)$.

$$H(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta^T} [\mathbf{x}_i \lambda_i - \mathbf{x}_i y_i] = \sum_{i=1}^{n} \frac{\partial}{\partial \beta^T} \mathbf{x}_i \lambda_i = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial \lambda_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta^T}$$

Use that

$$\frac{\partial \eta_i}{\partial \beta^T} = \frac{\partial \mathbf{x}_i^T \beta}{\partial \beta^T} = \left(\frac{\partial \mathbf{x}_i^T \beta}{\partial \beta} \right)^T = \mathbf{x}_i^T$$

and

$$\frac{\partial \lambda_i}{\partial \eta_i} = \frac{\partial \exp(\eta_i)}{\partial \eta_i} = \exp(\eta_i) = \lambda_i$$

And thus

$$H(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \lambda_i.$$

Note that the observed and the expected Fisher information matrix are equal (see below - canonical link - that this is a general finding).

Parameter estimation - in practice

To find the ML estimate $\hat{\beta}$ we need to solve

$$s(\hat{\beta}) = 0$$

We have that the score function for the log-linear model is:

$$s(\beta) = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \lambda_i) = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \exp(\mathbf{x}_i^T\beta)).$$

Observe that this is a non-linear function in β, and has no closed form solution (except for a few special cases).

Fisher scoring

To solve this we use the Fisher scoring algorithm, were we at interation $t+1$ have

$$\beta^{(t+1)} = \beta^{(t)} + F(\beta^{(t)})^{-1}s(\beta^{(t)})$$

Remark: what do we need to do to use the Newton-Raphson method instead? Well, replace F with H, but for canonical link (which is the log-link for the Poisson) $F = H$.

Requirements for convergence

For the Fisher scoring algorithm the expected Fisher information matrix F needs to be invertible, and analogous to what we saw in Module 3 this is possible if $\lambda_i > 0$ for all i and that the design matrix has full rank (p). Since we have the log-link we have that $\lambda_i = \exp(\mathbf{x}_i^T\beta)$ which is always positive, so all good. Note, with the linear link $\lambda_i = \eta_i$ this might be a challenge, and restrictions on β must be set.

Again, it is possible that the algorithm does not converge. This may happen for "unfavorable" data configurations (especially for small samples). Accoring to our text book, Fahrmeir et al (2013), page 284, the conditions for uniqueness and existence of ML estimators are very complex, and the authors suggest that the GLM user instead checks for convergence in practice by performing the iterations - also for the Poisson log-linear model.

Statistical inference

Asymptotic properties of ML estimates

We repeat what we found for Module 3: Under some (weak) regularity conditions:

Let $\hat{\beta}$ be the maximum likelihood (ML) estimate in the GLM model. As the total sample size increases, $n \to \infty$:

1. $\hat{\beta}$ exists
2. $\hat{\beta}$ is consistent (convergence in probability, yielding asymptotically unbiased estimator, variances goes towards 0)
3. $\hat{\beta} \approx N_p(\beta, F^{-1}(\hat{\beta}))$

Observe that this means that asymptotically $\text{Cov}(\hat{\beta}) = F^{-1}(\hat{\beta})$: the inverse of the expected Fisher information matrix evaluated at the ML estimate.

In our case we have

$$F(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \lambda_i = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

where $\mathbf{W} = \text{diag}(\lambda_i)$. This means $\text{Cov}(\hat{\beta}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ (remember that $\hat{\beta}$ comes in with $\hat{\lambda}_i$ in $\mathbf{W}$).

Let $\mathbf{A}(\beta) = F^{-1}(\beta)$, and $a_{jj}(\beta)$ is diagonal element number j.

For one element of the parameter vector:

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{a_{jj}(\hat{\beta})}}$$

is standard normal, which can be used to make confidence intervals - and test hypotheses.

Confidence intervals

In addition to providing a parameter estimate for each element of our parameter vector β we should also report a $(1-\alpha)100\%$ confidence interval (CI) for each element.

Let $z_{\alpha/2}$ be such that $P(Z_j > z_{\alpha/2}) = \alpha/2$. We then use

$$P(-z_{\alpha/2} \leq Z_j \leq z_{\alpha/2}) = 1 - \alpha$$

insert Z_j and solve for β_j to get

$$P(\hat{\beta}_j - z_{\alpha/2}\sqrt{a_{jj}(\hat{\beta})} \leq \beta_j \leq \hat{\beta}_j - z_{\alpha/2}\sqrt{a_{jj}(\hat{\beta})}) = 1 - \alpha$$

A $(1-\alpha)\%$ CI for β_j is when we insert numerical values for the upper and lower limits.

Example: Female crabs with satellites

Q:

1. Explain what is done in the R-print-out.
2. What if we instead want a CI for $\beta_0 + \beta_1 x_{i1}$?
3. What if we instead want a CI for $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1})$?

```
model2 = glm(Sa ~ W, family = poisson(link = log), data = crab)
summary(model2)
cat("lower\n")
lower = model2$coefficients - qnorm(0.975) * sqrt(diag(vcov(model2)))
lower
cat("upper\n")
upper = model2$coefficients + qnorm(0.975) * sqrt(diag(vcov(model2)))
upper
confint(model2)
```

```
##
## Call:
## glm(formula = Sa ~ W, family = poisson(link = log), data = crab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8526  -1.9884  -0.4933   1.0970   4.9221
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
## W            0.16405    0.01997   8.216  < 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: 632.79  on 172  degrees of freedom
## Residual deviance: 567.88  on 171  degrees of freedom
## AIC: 927.18
##
## Number of Fisher Scoring iterations: 6
##
## lower
## (Intercept)           W
##  -4.3675312   0.1249137
## upper
## (Intercept)           W
##  -2.2419833   0.2031764
##                   2.5 %     97.5 %
## (Intercept) -4.3662326 -2.2406858
## W            0.1247244  0.2029871
```

Hypothesis testing

There are three methods that are mainly used for testing hypotheses in GLMs - these are called Wald test, likelihood ratio test and score test. In Module 3 we looked at the Wald and likelihood ratio test - and what we found there still applies for the Poisson GLM. We only repeat the previous findings, and then add an example with the crab data.

We will look at null hypotheses and alternative hypotheses that can be written

$$H_0 : \mathbf{C}\beta = \mathbf{d} \text{ vs. } \mathbf{C}\beta \neq \mathbf{d}$$

by specifying $\mathbf{C}$ to be a $r \times p$ matrix and $\mathbf{d}$ to be a column vector of length r,

and/or where we define

- A: the larger model and
- B: the smaller model (under H_0), and the smaller model is nested within the larger model (that is, B is a submodel of A).

The Wald test

The Wald test statistic is given as:

$$w = (\mathbf{C}\hat{\beta} - \mathbf{d})^{\mathrm{T}}[\mathbf{C}F^{-1}(\hat{\beta})\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{C}\hat{\beta} - \mathbf{d})$$

and measures the distance between the estimate $\mathbf{C}\hat{\beta}$ and the value under then null hypothesis $\mathbf{d}$, weighted by the asymptotic covariance matrix of $\mathbf{C}\hat{\beta}$, and under the null follows a χ^2 distribution with r degrees of freedom (where r is the number of hypotheses tested).

P-values are calculated in the upper tail of the χ^2-distribution.

The likelihood ratio test

Notation:

- A: the larger model and
- B: the smaller model (under H_0), and the smaller model is nested within the larger model (that is, B is a submodel of A).

The likelihood ratio statistic is defined as

$$-2\ln\lambda = -2(\ln L(\hat{\beta}_B) - \ln L(\hat{\beta}_A))$$

which under the null is asymptotically χ^2-distributed with degrees of freedom equal the difference in the number of parameters in the large and the small model.

Again, p-values are calculated in the upper tail of the χ^2-distribution.

The score test

We continue to use the notation that A: the larger model (under H_1) and B: the smaller model (under H_0), and still assume that the smaller model is nested within the larger model (that is, B is a submodel of A).

The *score statistics* is based on the *score function*, and measures the distance to the score function at the maximum likelihood for model A (which is 0) and scales with the covariance to form the test statistic.

- Under the null hypothesis investigated let $\tilde{\boldsymbol{\beta}}$ be the ML estimate (that is, model B, the smaller model) - that means that this is a restricted ML estimate, and
- under H_1 we have the larger model (A) with maximum likelihood $\hat{\boldsymbol{\beta}}$.

The score statistics is:

$$U = (s(\tilde{\boldsymbol{\beta}}) - \mathbf{0})^T\mathbf{F}^{-1}(\tilde{\boldsymbol{\beta}})(s(\tilde{\boldsymbol{\beta}}) - \mathbf{0})$$

Here $s(\tilde{\boldsymbol{\beta}})$ represents a subvector of the score function where only the elements that is in H_1 and not in H_0 are present (so dimension is the difference in number of parameters between A and B models), but the score function is evaluated based on parameter estimates under H_0 (i.e. the value of $\hat{\lambda}$ in the score and expected Fisher information is based on $\tilde{\boldsymbol{\beta}}$.

To calculate $\mathbf{F}^{-1}(\tilde{\boldsymbol{\beta}})$ this is a submatrix of the full inverted matrix (not invert just the submatrix). The dimension of this matrix is "difference in number of parameters between A and B models" squared.

When the null hypothesis is true U as an asymptotic χ^2-distribution with r degrees of freedom (difference in number of estimated parameters between the large and small model).

Remark: In R the function `glm.scoretest` in the package `statmod` calculates the score test for a GLM when the difference between H_0 and H_1 is one parameter. The output from the `glm.scoretest` is the square root of U and must be squared and related to a χ^2_1 distribution (see example below).

The score test is very useful for special situations when the smaller model is to be tested towards many larger models, because only the smaller model has to be fitted.

The score test is perhaps the the most complex and least studied of the three tests, and in this course the main focus will be on the Wald and LRT tests. But, it is important for you to have heard of the score test, because in special situation it may be the preferred test.

Example: Female crabs with satellites - the different tests.

We fit a model with two covariates, one is categorical and we use effect coding. We want to test if we need to add these covariates.

Q: Comment on what you see in the print-out. Which hypotheses are we testing? Compare the test statistics.

```
library(statmod)
model3 = glm(Sa ~ W + C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum"))
summary(model3)

# possible to use type III in Anova
library(car)
Anova(model3, type = "III", test.statistic = "Wald")  #Q:same as summary?
Anova(model3, type = "III", test.statistic = "LR")  #same as comparing deviances

# compare to type I with anova - just to remember the difference between
# type I and III anova(model3,test='Chisq') anova(model3,test='LRT')
```

```
##
## Call:
## glm(formula = Sa ~ W + C, family = poisson(link = log), data = crab,
##     contrasts = list(C = "contr.sum"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0415  -1.9581  -0.5575   0.9830   4.7523
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.92089    0.56010  -5.215 1.84e-07 ***
## W            0.14934    0.02084   7.166 7.73e-13 ***
## C1           0.27085    0.11784   2.298   0.0215 *
## C2           0.07117    0.07296   0.975   0.3294
## C3          -0.16551    0.09316  -1.777   0.0756 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 559.34  on 168  degrees of freedom
## AIC: 924.64
##
## Number of Fisher Scoring iterations: 6
##
## Analysis of Deviance Table (Type III tests)
##
## Response: Sa
##             Df  Chisq Pr(>Chisq)
## (Intercept)  1 27.196  1.839e-07 ***
## W            1 51.350  7.726e-13 ***
## C            3  8.518    0.03644 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Deviance Table (Type III tests)
##
## Response: Sa
##   LR Chisq Df Pr(>Chisq)
## W   49.794  1  1.707e-12 ***
## C    8.534  3    0.03618 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Optional: We also include the results for the score test, but to test an hypothesis for a factor (more than one column in our design matrix) the `glm.scoretest` can not be used and then some manual programming is done below. There exists other packages to perform score tests for GLMs, but is not considered here.

```
# for score test we need ML under H0, and we need to fit two different H0
# models here when H0 is model with intercept and W:
model2a = glm(Sa ~ W, family = poisson(link = log), data = crab)
# when H0 is model with intercept and C:
model2b = glm(Sa ~ C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum"))

# First H0 is intercept and C and H1 is to add W (only one parameter)
testobs3b = glm.scoretest(model2b, x2 = crab$W)^2
testobs3b
1 - pchisq(testobs3b, 1)

# alternatively, implement from scratch
X = model.matrix(model3)
colnames(X)
estlambda = model2b$fitted.values  # the means under H0 (equals the variances)
W = diag(estlambda)  # variance on diagonal, estimated under H0
XH0 = X[, -2]  # model under H0
XH1 = X[, 2]  # part of model under H1 but not under H0
scorefunH1 = t(XH1) %*% (crab$Sa - estlambda)
expFH1full = t(X) %*% W %*% X
fullinv = solve(expFH1full)
subfullinv = fullinv[2, 2]
scoreteststat = t(scorefunH1) %*% subfullinv %*% scorefunH1
scoreteststat
```

```r
1 - pchisq(scoreteststat, 1)
# difference within numerical accuracy

# test if C should be included in a model with W
mm = model.matrix(Sa ~ C, data = crab, , contrasts = list(C = "contr.sum"))
testobs3a = glm.scoretest(model2a, x2 = mm[, -1])^2  #only test one level at a time, not all together -
testobs3a
1 - pchisq(testobs3a, 1)
# need to use other implementation (like above)
X = model.matrix(model3)
colnames(X)
estlambda = model2a$fitted.values  # the means under H0 (and variances since Poisson)
W = diag(estlambda)  # variance on diagonal, estimated under H0
XH0 = X[, 1:2]  # model under H0
XH1 = X[, 3:5]  # part of model under H1 but not under H0
scorefunH1 = t(XH1) %*% (crab$Sa - estlambda)
expFH1full = t(X) %*% W %*% X
fullinv = solve(expFH1full)
subfullinv = fullinv[3:5, 3:5]
scoreteststat = t(scorefunH1) %*% subfullinv %*% scorefunH1
scoreteststat
1 - pchisq(scoreteststat, 3)
```

```
## [1] 51.51423
## [1] 7.107648e-13
## [1] "(Intercept)" "W"           "C1"          "C2"          "C3"
##          [,1]
## [1,] 51.51418
##              [,1]
## [1,] 7.107648e-13
##        C1        C2        C3
## 4.2333184 2.6565569 0.5677206
##         C1         C2         C3
## 0.03963787 0.10312375 0.45116612
## [1] "(Intercept)" "W"           "C1"          "C2"          "C3"
##          [,1]
## [1,] 8.578756
##            [,1]
## [1,] 0.03544893
```

Interactive session - first week

Problem 1: Exam 2005 (Problem 1d-f - slightly modified) - Female crabs and satellites

(Only a subset of 20 crabs in the original data was used in the exam problems, but we will use the full data set of 173 crabs - so results will not be the same. Permitted aids for the exam was "all printed and handwritten material and all calculators" - NB that is not the case for 2018!)

We assume that the number of satellites for each female crab follows a Poisson distribution and want to perform a Poisson regression using a log-link to find if there is a connection between the expected number of satellites `Sa` and the width `W` and colour `C` of the carapace of the female crab.

- `C`: the color of the female crab (1=light medium, 2=medium, 3=dark medium, 4=dark)
- `W`: width of carapace (cm)

The following model was fitted.

```
crab <- read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/crab.txt")
colnames(crab) <- c("Obs", "C", "S", "W", "Wt", "Sa")
crab <- crab[,-1] #remove column with Obs
crab$C <- as.factor(crab$C)
modelEXAM <- glm(Sa ~ W + C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum
summary(modelEXAM)
library(car)
Anova(modelEXAM, type = "III", test.statistic = "Wald")
```

```
##
## Call:
## glm(formula = Sa ~ W + C, family = poisson(link = log), data = crab,
##     contrasts = list(C = "contr.sum"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0415  -1.9581  -0.5575   0.9830   4.7523
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.92089    0.56010  -5.215 1.84e-07 ***
## W            0.14934    0.02084   7.166 7.73e-13 ***
## C1           0.27085    0.11784   2.298   0.0215 *
## C2           0.07117    0.07296   0.975   0.3294
## C3          -0.16551    0.09316  -1.777   0.0756 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 559.34  on 168  degrees of freedom
## AIC: 924.64
##
## Number of Fisher Scoring iterations: 6
##
## Analysis of Deviance Table (Type III tests)
##
## Response: Sa
##             Df  Chisq Pr(>Chisq)
## (Intercept)  1 27.196  1.839e-07 ***
## W            1 51.350  7.726e-13 ***
## C            3  8.518    0.03644 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Write down the estimated model and evaluate the significance of the estimated coefficients.

How would you proceed to evaluate the significance of the coefficient for the colour C4. Explain.

Perform a test to evaluate the fit of the model. (In Problem 3d we derive the mathematical formula for the test statistic this test is based on.)

Use a 5% significance level for your evaluations. What is your conclusion?

b) Make a sketch *by hand* illustrating the connection between the expected number of satellites and the width of the carapace for each of the four colours of the crab.

The estimated multiplicative change in the expected number of satellites when the width of the carapace is changed by 1 cm is, according to the model, independent of the colour. Explain why.

Also find a 95% confidence interval for this change.

c) Let $\hat{\eta}(x_W)$ be the estimated linear predictor when the width of the carapace is x_W and the crab is "light medium". What is the distribution of $\hat{\eta}(x_W)$? Which value of x_W would give estimated mean number of satellites equal to 5?

Optional: Using R – for this value of x_W, construct a 95% confidence interval for the mean number of satellites.

Optional: Use `ggplot` to create (and improve) the sketch from b).

Problem 2: Exam 2017 (Problem 1) - Poisson regression

Consider a random variable Y. In our course we have considered the univariate exponential family having distribution (probability density function for continuous variables and probability mass function for discrete variables)

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w)\right)$$

where θ is called the *natural parameter* (or parameter of interest) and ϕ the *dispersion parameter*.

The Poisson distribution is a discrete distribution with probability mass function

$$f(y) = \frac{\lambda^y}{y!} \exp(-\lambda), \text{ for } y = 0, 1, \ldots$$

where $\lambda > 0$.

a) (10 points)

- Show that the Poisson distribution is a univariate exponential family, and specify what the elements of the exponential family (θ, ϕ, $b(\theta)$, w, $c(y, \phi, w)$) are.
- What is the connections between $\mathrm{E}(Y)$ and elements of the exponential family?
- What is the connections between $\mathrm{Var}(Y)$ and elements of the exponential family?
- Use these connections to derive the mean and variance for the Poisson distribution.
- If the Poisson distribution is used as the distribution for the response in a generalized linear model, what is the *canonical link* function?

b) (15 points)

We consider a Poisson regression with log link $\eta_i = g(\mu_i) = \ln(\mu_i)$, and linear predictor equal to $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$. Further, let p be the number of regression parameters in $\boldsymbol{\beta}$ (intercept included). The response–covariate pairs $(Y_i, \mathbf{x}_i)$ are independent for $i = 1, \ldots, n$.

- Does this set-up satisfy the *requirements* of a GLM model? Explain.
- Write down the log-likelihood.
- From the log-likelihood, *derive* the formula for the score function $\mathbf{s}(\boldsymbol{\beta})$ and the expected Fisher information matrix, $\mathbf{F}(\boldsymbol{\beta})$.
- What are the dimensions of $\mathbf{s}(\boldsymbol{\beta})$ and $\mathbf{F}(\boldsymbol{\beta})$?
- How can $\mathbf{s}(\boldsymbol{\beta})$ and $\mathbf{F}(\boldsymbol{\beta})$ be used to arrive at a maximum likelihood estimate for $\boldsymbol{\beta}$?

c) (10 points)

We now look at a data set giving the number of species of tortoise on the various Galapagos Islands (Data taken from the book "Practical Regression and Anova using R" by Julian J. Faraway.).

The data set contains measurements on 30 islands, and we study the following variables:

- `Species`: The number of species of tortoise found on the island.
- `Area`: The area of the island (km^2).
- `Elevation`: The highest elevation of the island (m).
- `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).

We have fitted a Poisson regression with log link to `Species` as response, and the other five variables are used as continuous covariates. Print-out from the fitted model is given in below.

```
library(faraway)
fit <- glm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala,
    family = poisson(link = log))
summary(fit)
```

```
##
## Call:
## glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, family = poisson(link = log), 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 ***
## Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
## Elevation    3.541e-03  8.741e-05  40.507  < 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
```

Let $\boldsymbol{\beta}$ be a 6×1 column vector with the regression coefficients (intercept included), and let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator for $\boldsymbol{\beta}$.

- Write down the asymptotic distribution for $\hat{\boldsymbol{\beta}}$, and specify how the covariance matrix for $\hat{\boldsymbol{\beta}}$ is estimated.

We will focus on the effect of `Elevation`, and denote the corresponding regression coefficient β_2.

- Write down the maximum likelihood estimate for β_2 in the print-out above.

- How can you explain this value to a biologist interested in understanding the effect of `Elevation` on the number of species of tortoise found on the islands?
- What is numerical value for the estimated standard deviation of $\hat{\beta}_2$ given in above?
- Construct an approximate 95% confidence interval for β_2.

Problem 3: Exam December 2017 from UiO, Problem 1.

(written out - with small changes to fit the notation we use - from https://www.uio.no/studier/emner/matnat/math/STK3100/h17/stk3100-4100_2017_2eng.pdf)

Assume that the random variable Y is Poisson distributed with probability mass function (pmf)

$$\mathrm{P}(Y = y|\lambda) = \frac{\lambda^y}{y!}\exp(-\lambda),\ y = 0, 1, 2, \ldots.$$

a) Show that the distribution of Y is an exponential family, that is, show that the pmf can be written on the form

$$\exp\left\{\frac{y\theta - b(\theta)}{\phi}w + c(y, \phi, w)\right\},$$

and determine θ, ϕ, w, $b(\theta)$ and $c(y, \phi, w)$.

We then assume that $Y_1, Y_2, \ldots, Y_n$ are independent with the pmf from a), and let $\mu_i = \mathrm{E}(Y_i)$, $i = 1, \ldots, n$.

b) Explain what we mean by a generalized linear model (GLM) for $Y_1, Y_2, \ldots, Y_n$ with link function $g(\cdot)$, and determine the canonical link function.

c) Derive an expression for the log-likelihood function $l(\mu; \mathbf{y})$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the observed value of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and $\mu = (\mu_1, \ldots, \mu_n)^T$.

d) Explain what we mean by a saturated model and determine the maximum of $l(\mu; \mathbf{y})$ for the saturated model.

e) Explain what we mean by the deviance $D(\mathbf{y}; \hat{\mu})$ of a Poisson GLM, find an expression for the deviance, and discuss how it may be used.

SECOND WEEK

Poisson regression for count data

What did we do last week?

Examples — GLM model — loglikelihood, score function and Fisher information matrix — asympototic results for $\hat{\boldsymbol{\beta}}$ and Wald, score and LRT.

Residuals

Two types of residuals are popular: *deviance* and *Pearson*.

Deviance residuals

Recall that the deviance was defined by relating a *candidate* model to a *saturated model* and calculating the likelihood ratio statistic with these two models.

Saturated model: If we were to provide a perfect fit to our data then we would estimate the mean λ_i by the observed count for observation i. That is, $\tilde{\lambda}_i = y_i$ Then $\tilde{\lambda}$ is an n dimensional column vector with the elements $\tilde{\lambda}_i$.

Candidate model: The model that we are investigated can be thought of as a *candidate* model. Then we maximize the likelihood and get $\hat{\beta}$ which through our linear predictor and link function we turn into $\hat{\lambda}_i$, also called $\hat{y}_i$. Then $\hat{\lambda}$ is an n-dimensional column vector with the elements $\hat{\lambda}_i$, also called $\hat{y}$.

$$D = -2(\ln L(\text{candidate model}) - \ln L(\text{saturated model})) = -2(l(\hat{\lambda}) - l(\tilde{\lambda})) = -2\sum_{i=1}^{n}(l_i(\hat{\lambda}_i) - l_i(\tilde{\lambda}_i))$$

Inserting the Poisson likelihood and λ_i estimates this gives:

$$D = 2\sum_{i=1}^{n}[y_i \ln(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)]$$

Where $\hat{y}_i = \exp(\mathbf{x}_i^T\hat{\beta})$. Verify this by yourself.

The deviance residuals are given by a signed version of each element in the sum for the deviance, that is

$$d_i = \text{sign}(y_i - \hat{y}_i) \cdot \left\{2[y_i \ln(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)]\right\}^{1/2}$$

where the term $\text{sign}(y_i - \hat{y}_i)$ makes negative residuals possible - and we get the same sign as the *Pearson residuals*

Pearson residuals

The Pearson residuals are given as

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}$$

where o_i is the observed count for observation i and e_i is the estimated expected count for observation i. We have that $o_i = y_i$ and $e_i = \hat{y}_i = \hat{\lambda}_i = \exp(\mathbf{x}_i^T\hat{\beta})$.

Remark: A standardized version scales the Pearson residuals with $\sqrt{1 - h_{ii}}$ similar to the standardized residuals for the normal model. Here h_{ii} is the diagonal element number i in the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

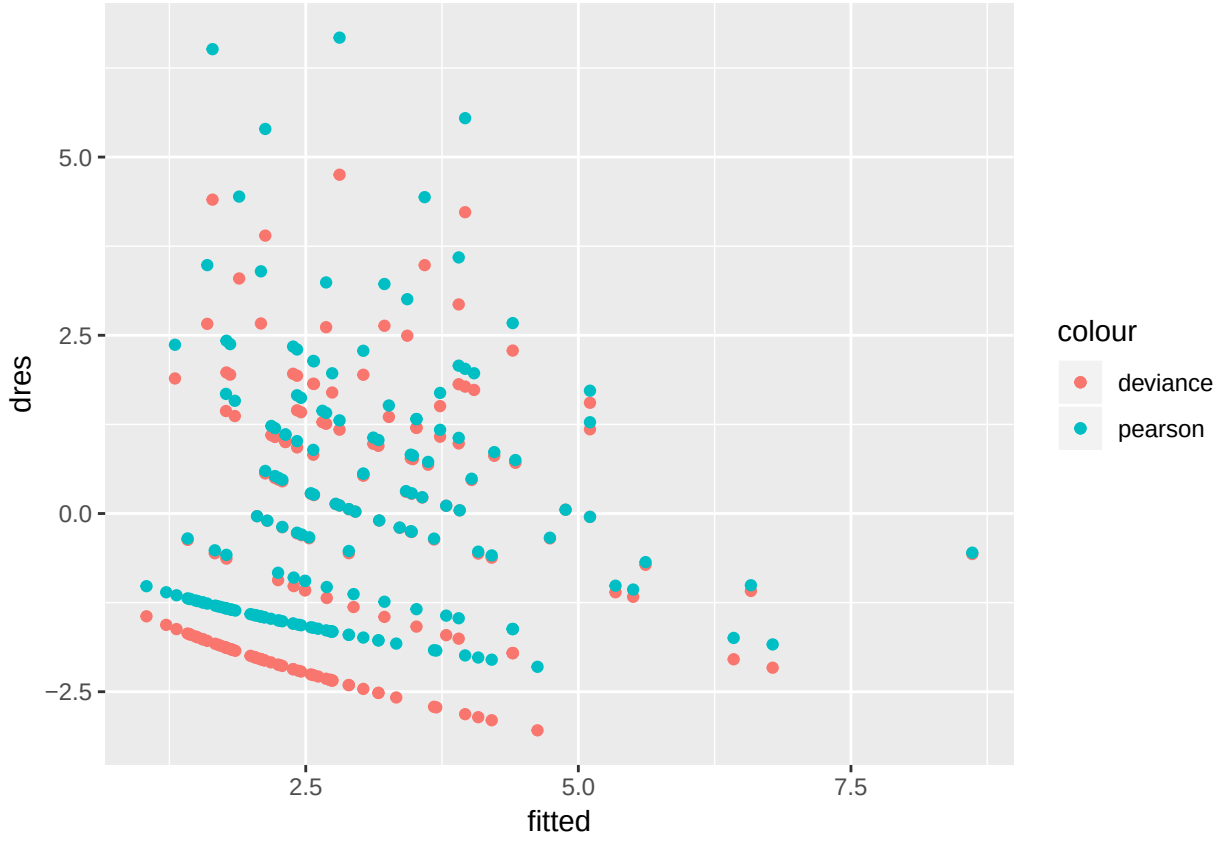
Plotting residuals

Deviance and Pearson residuals can be used for checking the fit of the model, by plotting the residuals against fitted values and covariates. Normality of residuals are also assumed, and can be checked using qq-plots as for the MLR in Module 2.

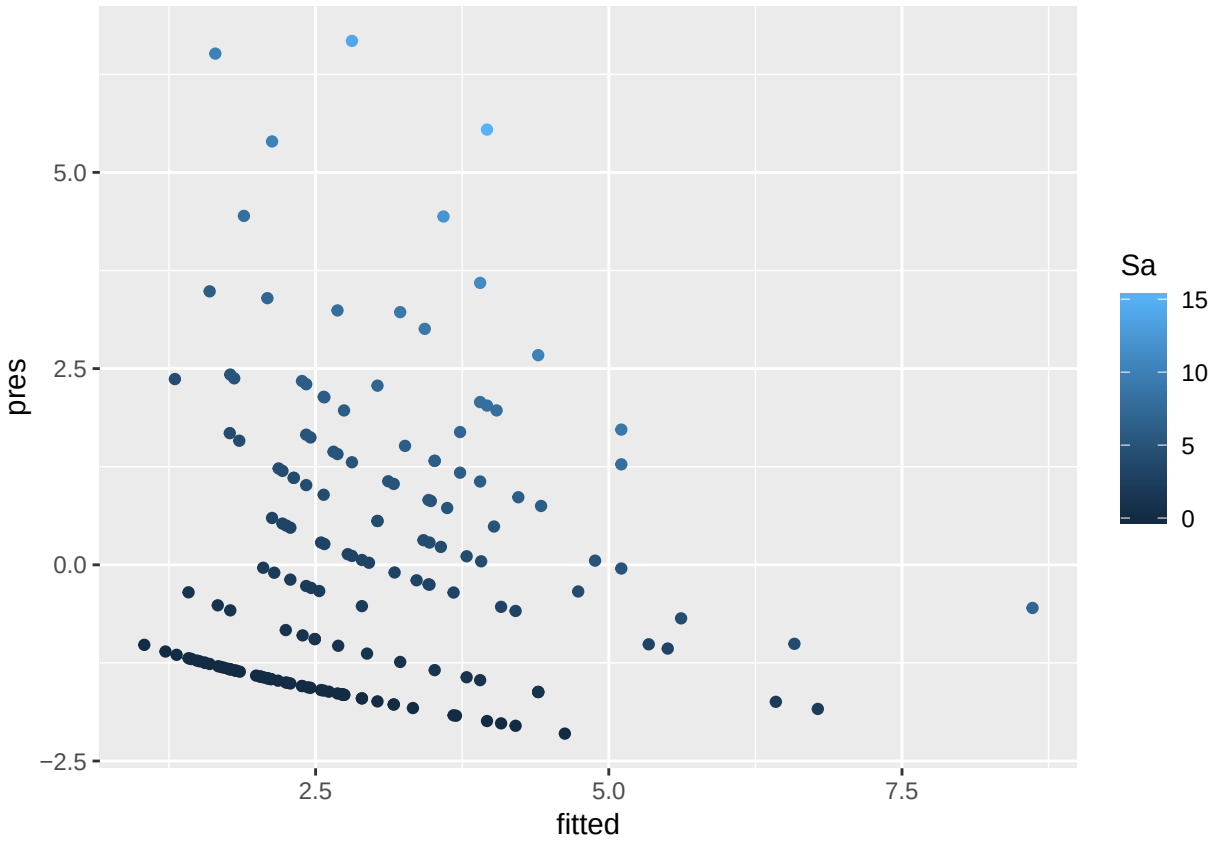
Below - notice the trend in the residuals, this is due to the discrete nature of the response. The plot with different shades of blue shows that the structures are for equal values of y.

```
model3 = glm(Sa ~ W + C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum",
    S = "contr.sum"))
df = data.frame(Sa = crab$Sa, fitted = model3$fitted.values, dres = residuals(model3,
    type = "deviance"), pres = residuals(model3, type = "pearson"))
```

```
library(ggplot2)

gg1 = ggplot(df) + geom_point(aes(x = fitted, y = dres, color = "deviance")) +
    geom_point(aes(x = fitted, y = pres, color = "pearson"))
gg1
```



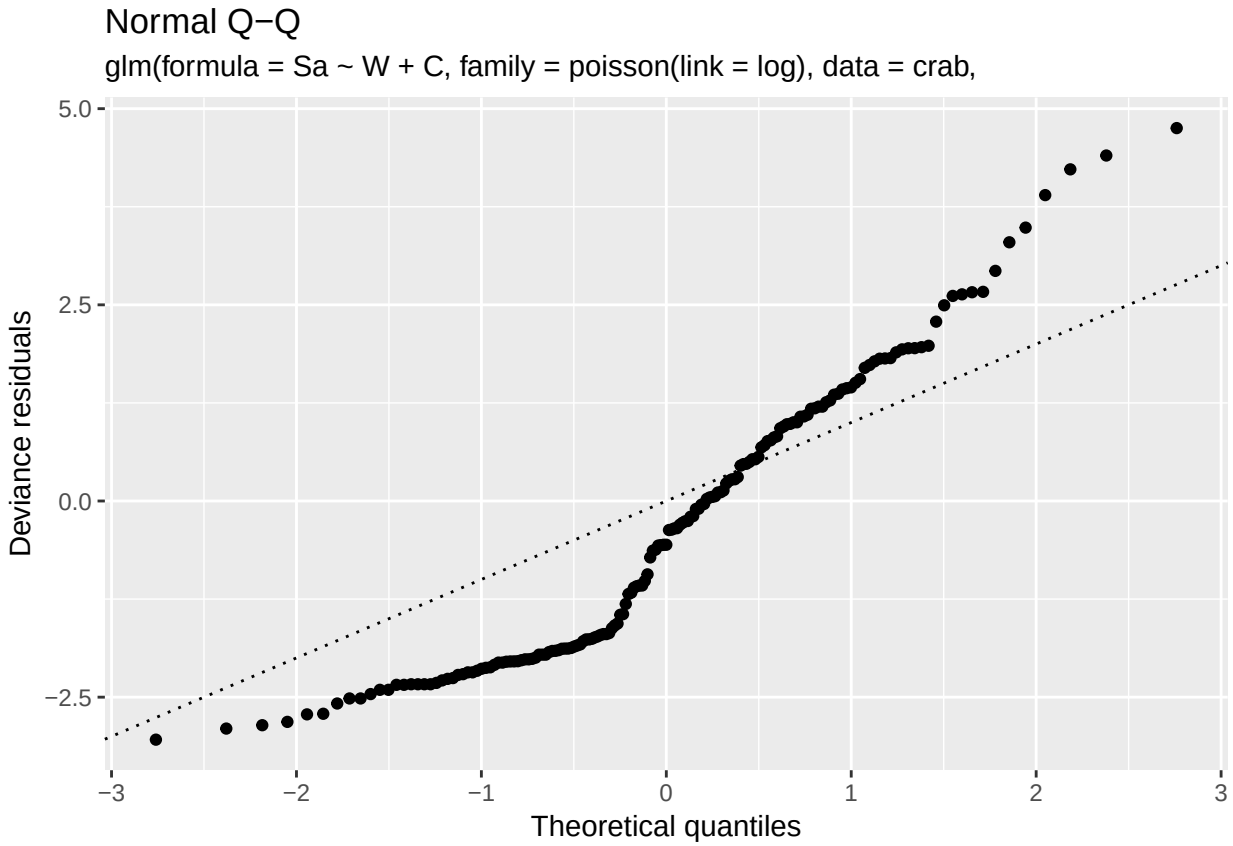
```
gg2 = ggplot(df) + geom_point(aes(x = fitted, y = pres, color = Sa))
gg2
```



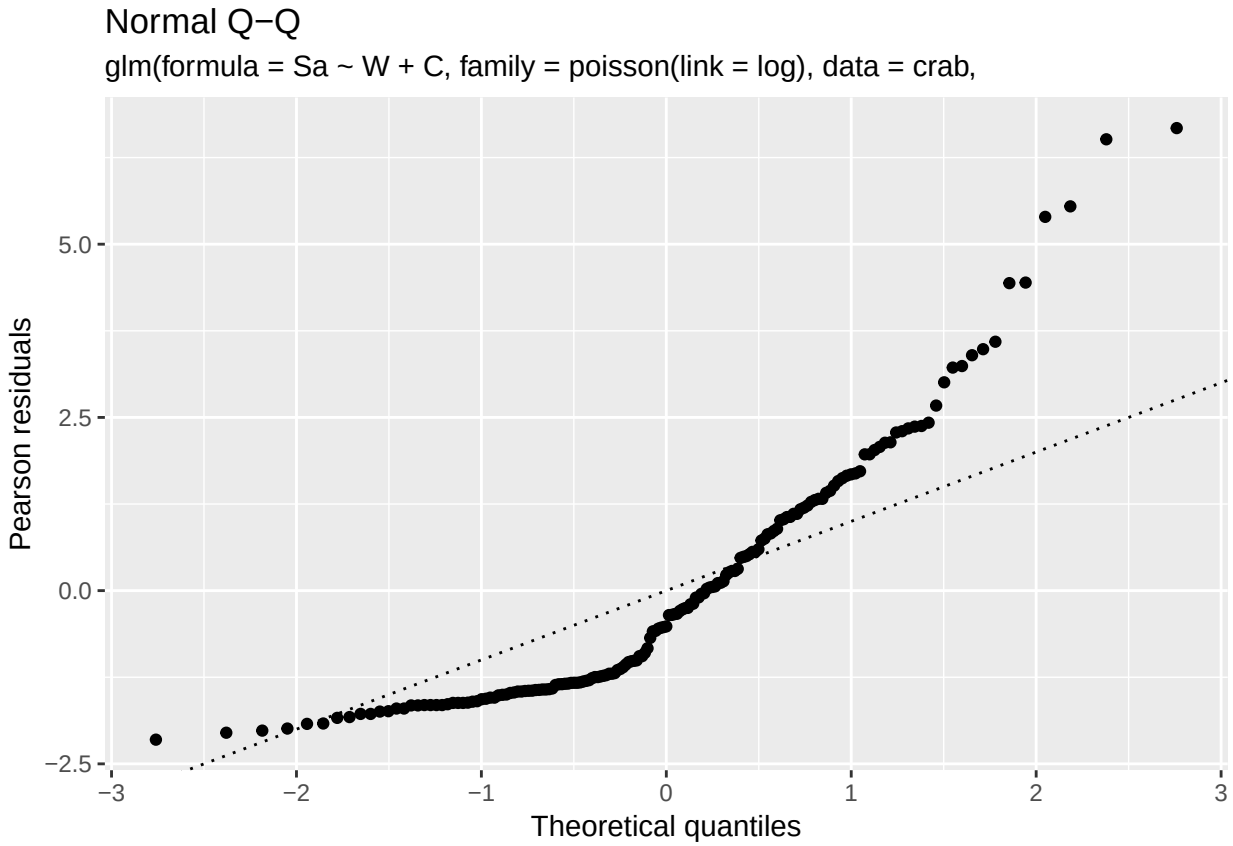

```
# qqnorm(residuals(model3,type='deviance'))
# qqline(residuals(model3,type='deviance'))
# qqnorm(residuals(model3,type='pearson'))
# qqline(residuals(model3,type='pearson'))
```

Pearson residuals

```
dff = data.frame(devres = residuals(model3, type = "deviance"), pearsonres = residuals(model3,
    type = "pearson"))
ggplot(dff, aes(sample = devres)) + stat_qq(pch = 19) + geom_abline(intercept = 0,
    slope = 1, linetype = "dotted") + labs(x = "Theoretical quantiles", y = "Deviance residuals",
    title = "Normal Q-Q", subtitle = deparse(model3$call))
```




```
ggplot(dff, aes(sample = pearsonres)) + stat_qq(pch = 19) + geom_abline(intercept = 0,
    slope = 1, linetype = "dotted") + labs(x = "Theoretical quantiles", y = "Pearson residuals",
    title = "Normal Q-Q", subtitle = deparse(model3$call))
```



Model assessment and model choice

The fit of the model can be assessed based on goodness of fit statistics (and related tests) and by residual plots. Model choice can be made from analysis of deviance, or by comparing the AIC for different models.

Deviance test

We may use the deviance test presented in Module 3 to test if the model under study is preferred compared to the saturated model.

We may write the deviance test as a sum of the squared deviance residuals.

$$D = 2\sum_{i=1}^{n}[y_i \ln(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)]$$

Remark: if $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ then deviance residuals will be equal to

$$D = 2\sum_{i=1}^{n} y_i \ln(\frac{y_i}{\hat{y}_i})$$

Q: is this the case for the log-linear model?

A: yes, but only if an intercept is included in the model?

The deviance statistic might be approximately χ^2_{n-p}, at least when the counts are high.

Note: see below for remark.

Remember that the likelihood ratio test can be performed using the difference between two deviances.

Pearson test

The Pearson χ^2-goodness of fit statistic is given as the sum of the squared Pearson residuals

$$X_P^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(y_j - \hat{y}_i)^2}{\hat{y}_i}$$

where $\hat{y}_i = \hat{\lambda}_i = \exp(\mathbf{x}_i^T \hat{\beta})$. The Pearson χ^2 statistic is asymptotically equivalent to the deviance statistic and thus is asymptotically χ^2_{n-p}.

Remark: See connection with Pearson residuals.

Remark: the asymptotic distribution of both statistics (deviance and Pearson) are questionable when there are many low counts. Agresti (1996, page 990) suggest analysing grouped data, for example by grouping by width in the crab example.

Remark: The Pearson statistic is also used for testing independence in contingency tables - we will do that in Compulsory Exercise 2.

Example: goodness of fit with female crabs

Q: Comment on the print-out. Is this a good fit? What might a bad fit be due to?

```
model3 = glm(Sa ~ W + C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum"))
summary(model3)
1 - pchisq(model3$deviance, model3$df.residual)
Xp = sum(residuals(model3, type = "pearson")^2)
Xp
1 - pchisq(Xp, model3$df.residual)
```

```
##
## Call:
## glm(formula = Sa ~ W + C, family = poisson(link = log), data = crab,
##     contrasts = list(C = "contr.sum"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0415  -1.9581  -0.5575   0.9830   4.7523
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.92089    0.56010  -5.215 1.84e-07 ***
## W            0.14934    0.02084   7.166 7.73e-13 ***
## C1           0.27085    0.11784   2.298   0.0215 *
## C2           0.07117    0.07296   0.975   0.3294
```

```
## C3          -0.16551    0.09316  -1.777   0.0756 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 559.34  on 168  degrees of freedom
## AIC: 924.64
##
## Number of Fisher Scoring iterations: 6
##
## [1] 0
## [1] 543.249
## [1] 0

## **A:** Reject the null that the model is good. This may be due to wrong model or missing covariate,
```

AIC

Identical to Module 3 - we may use the Akaike informations criterion. Let p be the number of regression parameters in our model.

$$\text{AIC} = -2 \cdot l(\hat{\beta}) + 2p$$

A scaled version of AIC, standardizing for sample size, is sometimes preferred. And, we may also use the BIC, where $2p$ is replaced by $\log(n) \cdot p$.

Analysis of deviance

Identical to Module 3 we may also sequentially compare models, and use analysis of deviance for this.

Overdispersion

Count data might show greater variability in the response counts than we would expect if the response followed a Poisson distribution. This is called *overdispersion*.

Example: newspaper sales with tourist bus.

Our model states that the variance $\text{Var}(Y_i) = \lambda_i$. If we change the model to $\text{Var}(Y_i) = \phi\lambda_i$ we may allow for an increased variance due to heterogeniety among subjects.

Agresti (1996, page 92) explains: "For our crab data set, what if width, weight, colour and spine affect the number of satellites for a female crab, and we only fitted a model with width as covariate. Then the crabs with a certain width are a mixture of crabs of various weights, colours and spine condition - that is, a mixture of several Poisson populations, each with its own mean for the response. This heterogeniety may give an overall response distribution where the variance is greater than the standard Poisson variance."

The overdispersion parameter can be estimated as the average Pearson statistic or average deviance

$$\hat{\phi}_D = \frac{1}{n-p} D$$

where D is the deviance. Note that similarity to $\hat{\sigma^2} = 1/(n-p) \cdot \text{SSE}$ in the MLR. The $\text{Cov}(\hat{\beta})$ can then be changed to $\hat{\phi} F^{-1}(\hat{\beta})$, so we multiply the standard error by the square root of $\hat{\phi}_D$.

Remark: We are now moving from likelihood to quasi-likelihood theory, where only $\text{E}(Y_j)$ and $\text{Var}(Y_j)$ - and not the distribution of Y_j - are used in the estimation.

```
model.disp = glm(Sa ~ W, family = quasipoisson(link = log), data = crab)
summary.glm(model.disp)
summary.glm(model.disp)$dispersion
```

```
##
## Call:
## glm(formula = Sa ~ W, family = quasipoisson(link = log), data = crab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8526  -1.9884  -0.4933   1.0970   4.9221
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.30476    0.96729  -3.417 0.000793 ***
## W            0.16405    0.03562   4.606 7.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.182205)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 567.88  on 171  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
##
## [1] 3.182205
```

Rate models and offset

In the Poisson process we might analyse an event that occurs within a time interval or region in space, and therefore it is often of interest to model the *rate* at which events occur.

Examples:

- crime rates in cities
- death rate for smokers vs. non-smokers
- rate of auto thefts in cities

Agresti (1996, page 86): "what if we want to model the number of auto thefts for a year in a sample of cities. We would make a rate for each city by dividing the number of auto thefts by the population size of the city. The model could then describe how this rate depends on unemployment rate, median income, percentage of residents having completed high school.""

Now we don't want a model for Y_i but for Y_i/t_i, where

- Let t_i denote the index (population size in the example) associated with observation i.
- We still assume that Y_i follows a Poisson distribution, but we now include the index in the modelling and focus on Y_i/t_i.
- The expected value of Y_i/t_i would then be $\mathrm{E}(Y_i)/t_i = \lambda_i/t_i$.

A log-linear model would be

$$\log(\lambda_i/t_i) = \mathbf{x}_i^T\beta$$

We may equivalently write the model as

$$\log(\lambda_i) - \log(t_i) = \mathbf{x}_i^T\beta$$

This adjustment term is called an *offset* and is a known quantity.

The expected number of outcomes will then satisfy

$$\mathrm{E}(Y_i) = \lambda_i = t_i \exp(\mathbf{x}_i^T\beta).$$

Example: British doctors and rate models

Count data - the number of times and event occurs - is common. In one famous study British doctors were in 1951 sent a questionnaire about whether they smoked tobacco - and later information about their death were collected.

Research questions that were asked were: 1) Is the death rate higher for smokers than for non-smokers? 2) If so, by how much? 3) And, how is this related to age?

```
library(boot)
`?`(breslow)
# n=person-year, ns=smoker-years, age=midpoint 10 year age group, y=number
# of deaths due to cad, smoke=smoking status
head(breslow, n = 10)
```

```
##    age smoke     n   y    ns
## 1   40     0 18790   2     0
## 2   50     0 10673  12     0
## 3   60     0  5710  28     0
## 4   70     0  2585  28     0
## 5   80     0  1462  31     0
## 6   40     1 52407  32 52407
## 7   50     1 43248 104 43248
## 8   60     1 28612 206 28612
## 9   70     1 12663 186 12663
## 10  80     1  5317 102  5317
```

To investigate this we will look at different ways of relating the expected number of deaths and the number of doctors at risk in the observation period for each smoke and age group. The aim is to model the rate of cardiovascular mortality.

```
# first age and smoke (but not interaction thereof)
fit1 <- glm(y ~ factor(age) + factor(smoke), offset = log(n), family = poisson,
    data = breslow)
summary(fit1)
```

```r
# do we need interaction?
fit2 <- update(fit1, . ~ . + factor(smoke) * factor(age))
summary(fit2)
anova(fit1, fit2, test = "Chisq")
# yes, significant interaction between age and smoking - how does this
# compare to a deviance test for fit1?

# reporting on final model - give rate in each possible group
cbind(fit2$fitted.values, breslow$y)  #perfect fit since number of coeffs equal number of groups
exp(predict(fit2, type = "link"))
predict(fit2, type = "response")

# I want to see the estimated value of lambda for a population size of 1 and
# of 1000
length(fit2$coefficients)
# year 40 nonsmokers should only be the intercept
exp(fit2$coefficients[1])  # expected number of deaths pr individual in a population of 40-year olds wh
# pr 1000
exp(fit2$coefficients[1]) * 1000
# 80 year olds who smoke
exp(sum(fit2$coefficients[c(1, 5, 6, 10)]))
# pr 1000
1000 * exp(sum(fit2$coefficients[c(1, 5, 6, 10)]))
```

```
##
## Call:
## glm(formula = y ~ factor(age) + factor(smoke), family = poisson,
##     data = breslow, offset = log(n))
##
## Deviance Residuals:
##       1        2        3        4        5        6        7
## -2.17978  -1.30800  -0.13791   0.22882   1.91902   0.90160   0.51038
##       8        9       10
##  0.05135  -0.08732  -0.91237
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.9193     0.1918 -41.298  < 2e-16 ***
## factor(age)50    1.4840     0.1951   7.606 2.82e-14 ***
## factor(age)60    2.6275     0.1837  14.301  < 2e-16 ***
## factor(age)70    3.3505     0.1848  18.131  < 2e-16 ***
## factor(age)80    3.7001     0.1922  19.249  < 2e-16 ***
## factor(smoke)1   0.3545     0.1074   3.302  0.00096 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 935.067  on 9  degrees of freedom
## Residual deviance:  12.132  on 4  degrees of freedom
## AIC: 79.2
##
## Number of Fisher Scoring iterations: 4
##
```

```
##
## Call:
## glm(formula = y ~ factor(age) + factor(smoke) + factor(age):factor(smoke),
##     family = poisson, data = breslow, offset = log(n))
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -9.1479     0.7071 -12.937  < 2e-16 ***
## factor(age)50                   2.3574     0.7638   3.087  0.00203 **
## factor(age)60                   3.8302     0.7319   5.233 1.67e-07 ***
## factor(age)70                   4.6227     0.7319   6.316 2.69e-10 ***
## factor(age)80                   5.2944     0.7296   7.257 3.96e-13 ***
## factor(smoke)1                  1.7469     0.7289   2.397  0.01654 *
## factor(age)50:factor(smoke)1   -0.9866     0.7901  -1.249  0.21174
## factor(age)60:factor(smoke)1   -1.3628     0.7562  -1.802  0.07151 .
## factor(age)70:factor(smoke)1   -1.4423     0.7565  -1.906  0.05659 .
## factor(age)80:factor(smoke)1   -1.8470     0.7572  -2.439  0.01471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 9.3507e+02  on 9  degrees of freedom
## Residual deviance: 2.9754e-14  on 0  degrees of freedom
## AIC: 75.068
##
## Number of Fisher Scoring iterations: 3
##
## Analysis of Deviance Table
##
## Model 1: y ~ factor(age) + factor(smoke)
## Model 2: y ~ factor(age) + factor(smoke) + factor(age):factor(smoke)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         4     12.132
## 2         0      0.000  4   12.132  0.01639 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##    [,1] [,2]
## 1     2    2
## 2    12   12
## 3    28   28
## 4    28   28
## 5    31   31
## 6    32   32
## 7   104  104
## 8   206  206
## 9   186  186
## 10  102  102
##   1   2   3   4   5   6   7   8   9  10
##   2  12  28  28  31  32 104 206 186 102
##   1   2   3   4   5   6   7   8   9  10
```

```
##   2  12  28  28  31  32 104 206 186 102
## [1] 10
##  (Intercept)
## 0.0001064396
## (Intercept)
##   0.1064396
## [1] 0.01918375
## [1] 19.18375
```

Modelling continuous positive response data

Examples of continuous positive responses

- Insurance: Claim sizes
- Medicine: Time to blood coagulation (main example)
- Biology: Time in various development stages for fruit fly
- Meteorology: Amount of precipitation (interactive session - exam question 2012)

Models for continuous positive responses

- Lognormal distribution on response
- Gamma distribution on response
- Inverse Gaussian distribution on response (we will not consider this here)

Time to blood coagulation

This data is described in McCullagh and Nelder (1989, page 300). The data represents clotting time of blood (in seconds) `y` for normal plasma diluted to nine different percentage concentrations `u` with socalled prothrombin-free plasma. To induce the clotting a chemical called thromboplasting was used, and in the experiment two different lots of the chemical were used - denoted `lot`. Our aim is to investigate the relationship between the clotting time and the dilution percentage, and look at differences between the lots.

```
clot = read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/clot.txt", header = T)
clot$lot = as.factor(clot$lot)
summary(clot)
```

```
##        u            time       lot
##  Min.   :  5   Min.   : 12.0   1:9
##  1st Qu.: 15   1st Qu.: 18.0   2:9
##  Median : 30   Median : 23.0
##  Mean   : 40   Mean   : 32.5
##  3rd Qu.: 60   3rd Qu.: 35.0
##  Max.   :100   Max.   :118.0
```

Lognormal distribution

Let Y_i be the response on the original scale, where $Y_i > 0$.

Transform the response to a logaritmic scale: $Y_i^* = \ln(Y_i)$. Then, assume that transformed responses follow a normal distribution (or follows approximately) and use ordinary MLR. This means we have a GLM with normal response and identity link (on logarithmic scale of reponse).

1. $Y_i^* \sim N(\mu_i^*, \sigma^{*2})$
2. $\eta_i = \mathbf{x}_i^T \beta$
3. $\mu_i^* = \eta_i$ (identity link)

There are two ways of looking at this,

1. either this is just a transformation to achieve approximate normality, or
2. we assume that the original data follows a lognormal distribution.

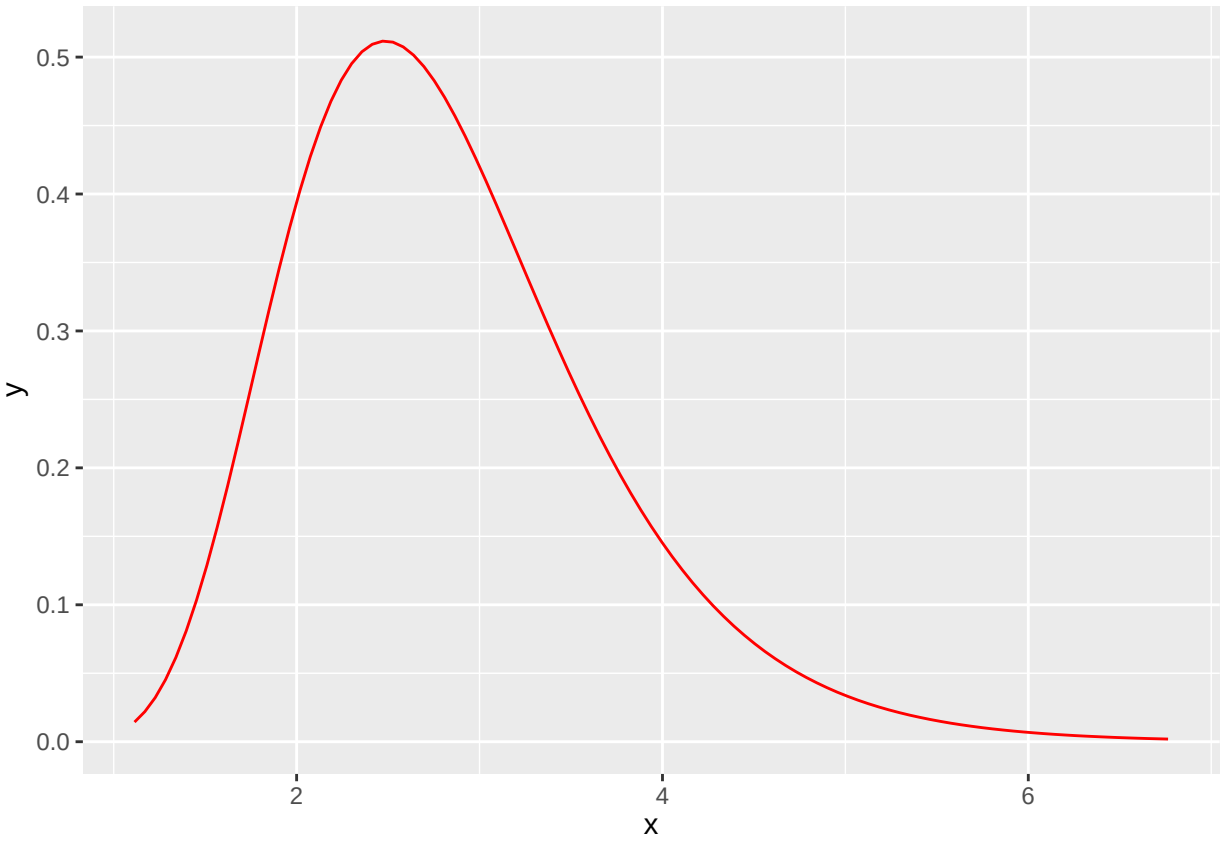
In genomics one usually assume the former, and reports back results on the exponential scale - just say that the mean of original data is $\exp(\mu_i^*)$.

However, if on instead assume that the original data really comes from a lognormal distribution, then it can be shown that

$$\mathrm{E}(Y_i) = \exp(\mu_i^*) \cdot \exp(\sigma^{*2}/2)$$
$$\mathrm{Var}(Y_i) = \exp(\sigma^{*2} - 1) \cdot \mu_i^2$$

i.e. standard deviation proportional to expectation. That is in general a put-off for using the lognormal model.

```
orgmu = 1
orgsd = 0.3
library(ggplot2)
xrange = range(rlnorm(1000, orgmu, orgsd))
ggplot(data.frame(x = xrange), aes(xrange)) + xlab(expression(x)) + stat_function(fun = dlnorm,
    args = list(meanlog = orgmu, sdlog = orgsd), geom = "line", colour = "red")
```



Gamma regression

The gamma distribution

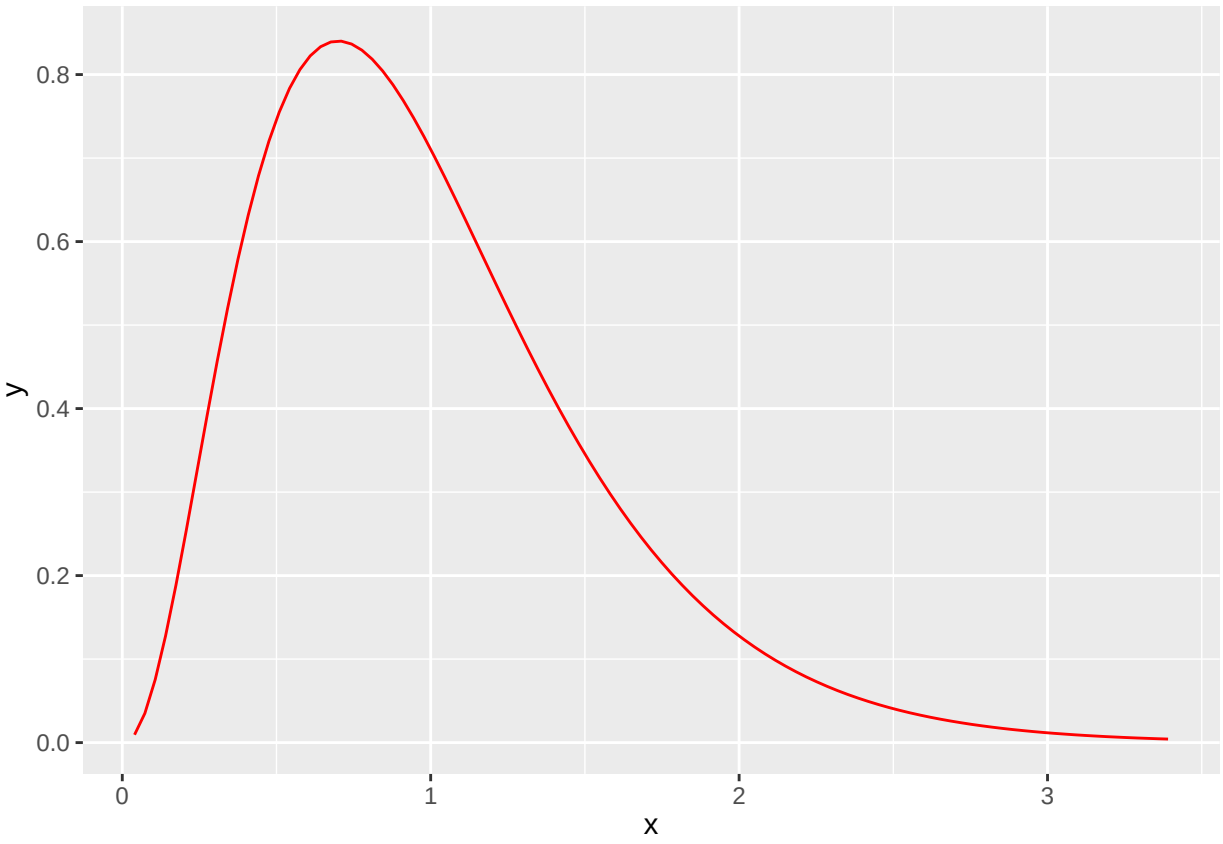
We have seen that a gamma distributed variable may be the result of the time between events in a Poisson process. The well known χ^2_δ-distribution is a special case of the gamma distribution ($\frac{\nu}{\mu_i} = 2$, $\nu = \frac{\delta}{2}$).

There are many parameterization for the gamma distribution, but we will stick with the one used in our textbook (page 643):

$Y_i \sim Ga(\mu_i, \nu)$ with density

$$f(y_i) = \frac{1}{\Gamma(\nu)}(\frac{\nu}{\mu_i})^\nu y_i^{\nu-1} \exp(-\frac{\nu}{\mu_i} y_i) \text{ for } y_i > 0$$

```
mu = 1
nu = 0.3
library(ggplot2)
xrange = range(rgamma(1000, shape = mu/nu, scale = nu))
ggplot(data.frame(x = xrange), aes(xrange)) + xlab(expression(x)) + stat_function(fun = dgamma,
    args = list(shape = mu/nu, scale = nu), geom = "line", colour = "red")
```



We found in Module 1, see exponential family that the gamma distribution is an exponential family, with

- $\theta_i = -\frac{1}{\mu_i}$ is the canonical parameter
- $\phi = \frac{1}{\nu}$,
- $w_i = 1$
- $b(\theta_i) = -\ln(-\theta_i)$
- $\mathrm{E}(Y_i) = b'(\theta_i) = -\frac{1}{\theta_i} = \mu_i$
- $\mathrm{Var}(Y_i) = b''(\theta_i)\frac{\psi}{w_i} = \frac{\mu_i^2}{\nu}$

For a GLM model we have canonical link if

$$\theta_i = \eta_i$$

Since $\eta_i = g(\mu_i)$ this means to us that we need

$$\theta_i = g(\mu_i) = -\frac{1}{\mu_i}$$

saying that with the canonical link is $-\frac{1}{\mu_i}$.

However, the most commonly used link is $g(\mu_i) = \ln(\mu_i)$, and the identity link is also used.

Q: Discuss the implications on η_i when using the canonical link. What about using log-link?

Remark: often the inverse and not the negative inverse is used, and since

$$g(\mu_i) = -\frac{1}{\mu_i} = \mathbf{x}_i^T\beta$$

then

$$\frac{1}{\mu_i} = -\mathbf{x}_i^T\beta = \mathbf{x}_i^T\beta^*$$

where $\beta^* = -\beta$.

Gamma GLM model

1. $Y_i \sim Ga(\mu_i, \nu)$
2. $\eta_i = \mathbf{x}_i^T\beta$
3. Popular link functions:

- $\eta_i = \mu_i$ (identity link)
- $\eta_i = \frac{1}{\mu_i}$ (inverse link)
- $\eta_i = \ln(\mu_i)$ (log-link)

Remark: In our model the parameter μ_i varies with i but ν is the same for all observations.

Example: Time to blood coagulation

A simple model to start with is as follows (dosages often analysed on log scale):

```
fit1 = glm(time ~ lot + log(u), data = clot, family = Gamma(link = log))
summary(fit1)
```

```
##
## Call:
## glm(formula = time ~ lot + log(u), family = Gamma(link = log),
##     data = clot)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.17470  -0.11596  -0.04281   0.06919   0.27749
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.44660    0.13453   40.48  < 2e-16 ***
## lot2        -0.47034    0.07095   -6.63 8.02e-06 ***
## log(u)      -0.58476    0.03772  -15.50 1.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.02265072)
##
##     Null deviance: 7.7087  on 17  degrees of freedom
## Residual deviance: 0.3211  on 15  degrees of freedom
## AIC: 104.28
##
## Number of Fisher Scoring iterations: 5
```

Q: describe what you see in the print-out.

Gamma regression: likelihood and derivations thereof

Likelihood:

$$L(\beta) = \prod_{i=1}^{n} \exp(-\frac{\nu y_i}{\mu_i} - \nu\ln\mu_i + \nu\ln\nu + (\nu-1)\ln y_i - \ln(\Gamma(\nu)))$$

Log-likelihood:

$$l(\beta) = \sum_{i=1}^{n} [-\frac{\nu y_i}{\mu_i} - \nu\ln\mu_i + \nu\ln\nu + (\nu-1)\ln y_i - \ln(\Gamma(\nu))]$$

Observe that we now- for the first time - have a nuisance parameter ν here.

To produce numerical estimates for the parameter of interest β we may proceed to the score function, and solve using Newton Raphson or Fisher scoring. If we do not have the canonical link the observed and expected Fisher information matrix may not be equal.

What about $\phi = 1/\nu$? Also estimated using maximum likelihood.

Further analyses: as before we use asymptotic distribution of parameter estimates, and of Wald, LRT and score test.

Scaled and unscaled deviance

We have defined the deviance as

$$D = -2(\ln L(\text{candidate model}) - \ln L(\text{saturated model}))$$

This is often called the *scaled deviance*.

The *unscaled deviance* is then defined as ϕD, but is sadly sometimes also called the deviance - for example by `R`.

1. For the normal model the

- scaled deviance is $D = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$, while
- unscaled deviance is $\phi D = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$

2. For the binomial and Poisson model $\phi = 1$ so the scaled and unscaled deviance are equal.
3. What about the Gamma model?

Some calculations - see IL week 2, problem 2: 1b.

$$D = \frac{-2\sum_{i=1}^{n}[\ln(\frac{y_i}{\hat{\mu}_i}) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}]}{\phi}$$

and unscaled as $\phi D = -2\sum_{i=1}^{n}[\ln(\frac{y_i}{\hat{\mu}_i}) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}]$.

Compare to print-out from R: the deviance in R is the *unscaled deviance*.

```
deviance(fit1)
nu1 = 1/summary(fit1)$dispersion
nu1
D = -2 * nu1 * sum(log(fit1$y/fit1$fitted.values) - ((fit1$y - fit1$fitted.values)/fit1$fitted.values))
D
deviance(fit1) * nu1
```

```
## [1] 0.3210963
## [1] 44.14871
## [1] 14.17599
## [1] 14.17599
```

Comparing models

Comparing models based on deviance

```
fit2 = glm(time ~ lot + log(u) + lot:log(u), data = clot, family = Gamma(link = log))
anova(fit1, fit2)
```

```
## Analysis of Deviance Table
##
## Model 1: time ~ lot + log(u)
## Model 2: time ~ lot + log(u) + lot:log(u)
##   Resid. Df Resid. Dev Df  Deviance
## 1        15    0.32110
## 2        14    0.31576  1 0.0053352
```

The deviance table does not include ϕ, so the unscaled deviance is reported. If significance testing is done, the estimated ϕ from the largest model is used, and p-values are based on the scaled deviance.

```
anova(fit1, fit2, test = "Chisq")
1 - pchisq((deviance(fit1) - deviance(fit2))/summary(fit2)$dispersion, fit1$df.residual -
    fit2$df.residual)
anova(fit1, fit2, test = "F")
1 - pf((deviance(fit1) - deviance(fit2))/summary(fit2)$dispersion, fit1$df.residual -
    fit2$df.residual, fit2$df.residual)
```

```
## Analysis of Deviance Table
##
## Model 1: time ~ lot + log(u)
## Model 2: time ~ lot + log(u) + lot:log(u)
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1        15    0.32110
## 2        14    0.31576  1 0.0053352   0.6355
## [1] 0.6355477
## Analysis of Deviance Table
##
## Model 1: time ~ lot + log(u)
## Model 2: time ~ lot + log(u) + lot:log(u)
##   Resid. Df Resid. Dev Df  Deviance      F Pr(>F)
## 1        15    0.32110
## 2        14    0.31576  1 0.0053352 0.2246 0.6429
## [1] 0.642854
```

Comparing models based on AIC

```
AIC(fit1, fit2)
```

```
##      df      AIC
## fit1  4 104.2763
## fit2  5 105.9738
```

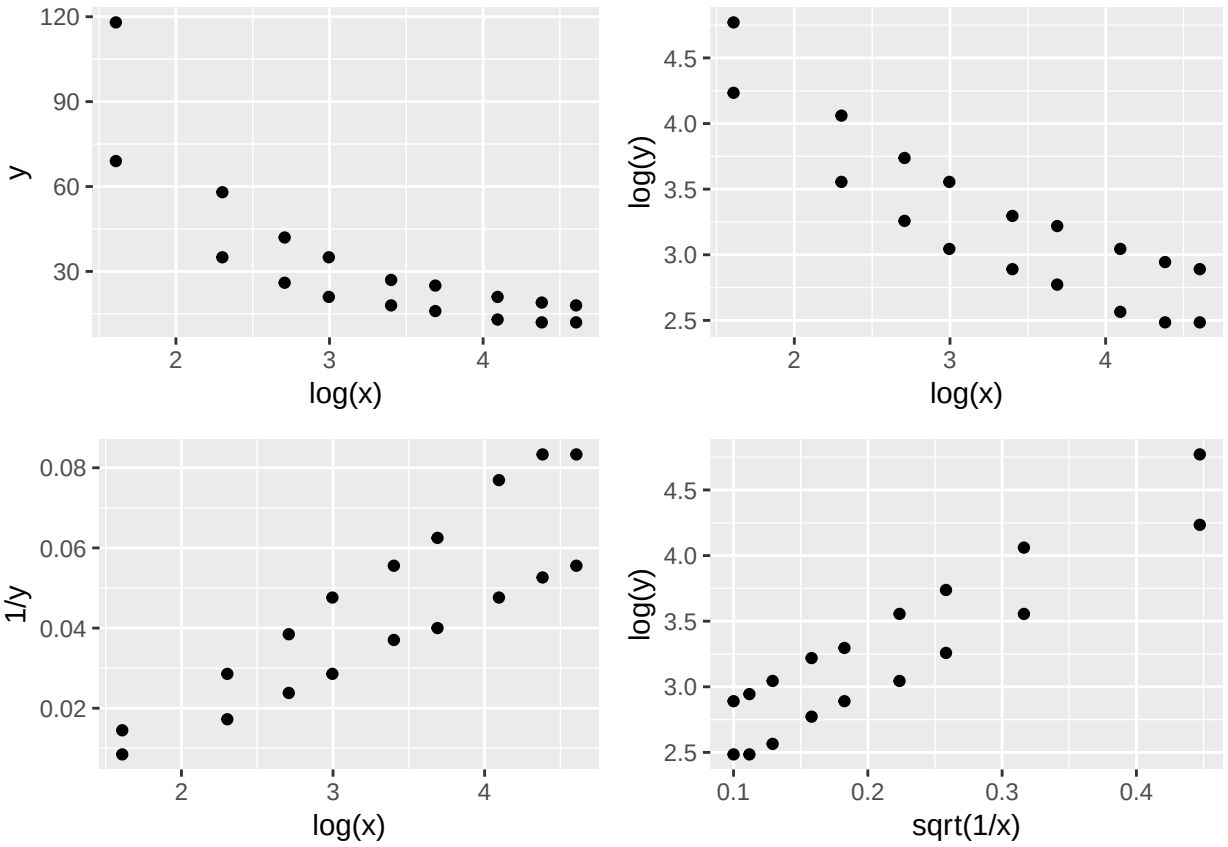
Q: would you prefer `fit1` or `fit2`?

AIC can also be used when we compare models with different link functions (models that are not nested).

The literature suggests to plot y_i vs. each covariate to get a hint about which link function or transformation to use.

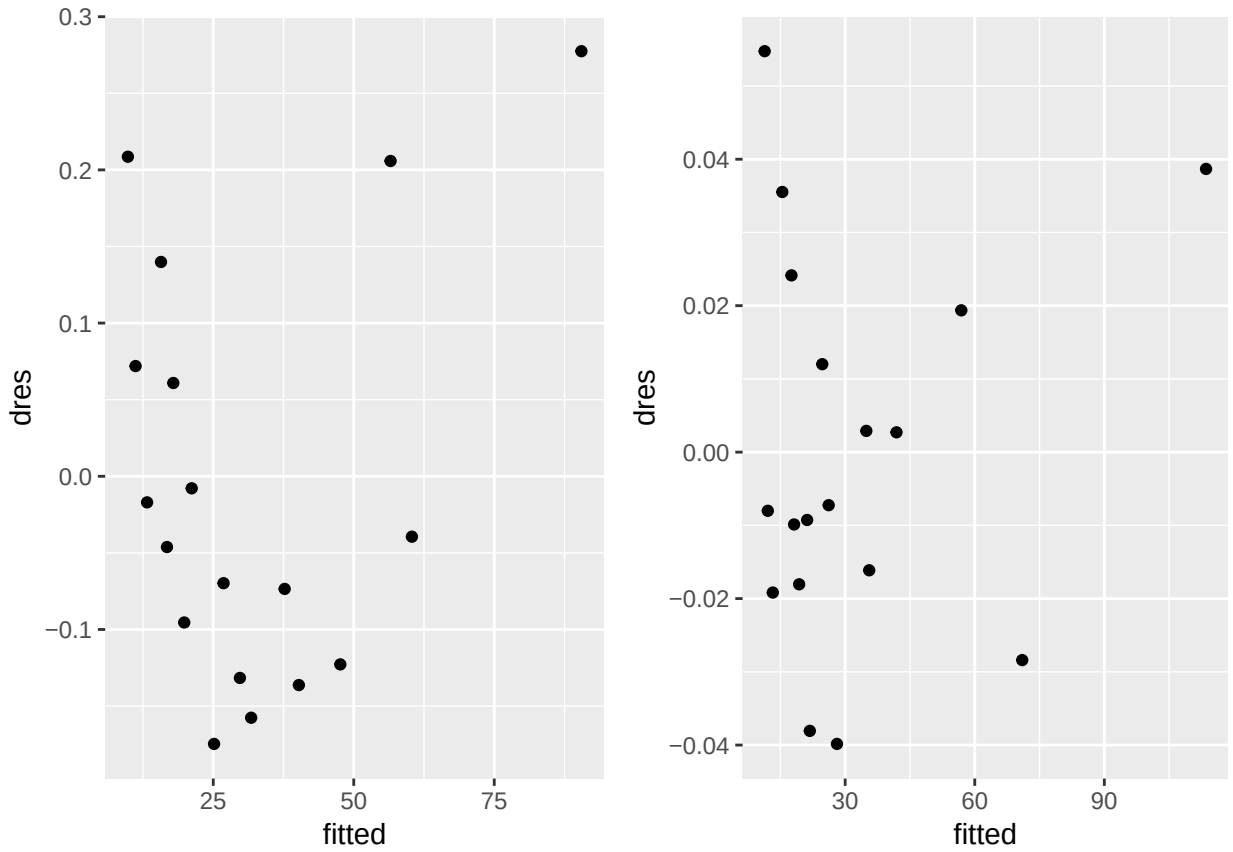
- Identity: Plot of y_i vs x_i should be close to linear
- ln : Plot of $\ln(y_i)$ vs x_i should be close to linear
- Inverse (reciprocal): Plot of $1/y_i$ vs x_i should be close to linear

```
library(ggplot2)
library(ggpubr)
y = clot$time
x = clot$u

df = data.frame(y = y, x = x)
gg1 = ggplot(df) + geom_point(aes(x = log(x), y = y))
gg2 = ggplot(df) + geom_point(aes(x = log(x), y = log(y)))
gg3 = ggplot(df) + geom_point(aes(x = log(x), y = 1/y))
gg4 = ggplot(df) + geom_point(aes(x = sqrt(1/x), y = log(y)))
ggarrange(gg1, gg2, gg3, gg4)
```




```
fit4 = glm(time ~ lot + sqrt(1/u), data = clot, family = Gamma(link = log))
AIC(fit1, fit4)

df4 = data.frame(fitted = fit4$fitted.values, dres = residuals(fit4, type = "deviance"))
gg4 = ggplot(df4) + geom_point(aes(x = fitted, y = dres)) + scale_color_discrete("")
```

```
df1 = data.frame(fitted = fit1$fitted.values, dres = residuals(fit1, type = "deviance"))
gg1 = ggplot(df1) + geom_point(aes(x = fitted, y = dres)) + scale_color_discrete("")
ggarrange(gg1, gg4)
```




```
##      df       AIC
## fit1  4 104.27633
## fit4  4  45.01688
```

Interactive session - second week

Problem 1: Exam 2007 (Problem 1, a bit modified) - Smoking and lung cancer

(Permitted aids for the exam was "Tabeller og formler i statistikk", Matematisk formelsamling (Rottmann), one A5 sheet with your own handwritten notes, and a simple calculator.)

The dataset given in `smoking.txt` consists of four variables:

- `deaths`: number of lung cancer deaths over a period of six years [remark: incorrectly 1 year in exam question]
- `population`: the number of people [remark: incorrectly in 100 000 people in exam question]
- `age`: in five-year age groups (`40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+`)
- `ageLevel`: age group as numbers from 1 to 9 (1 corresponds to 40-44, 2 to 45-59, and so on)
- `smoking status`: doesn't smoke (`no`), smokes cigars or pipe only (`cigarPipeOnly`), smokes cigarettes and cigar or pipe (`cigarettePlus`), and smokes cigarettes only (`cigaretteOnly`)

You can look at the dataset here: https://www.math.ntnu.no/emner/TMA4315/2018h/smoking.txt, and can be found here as well: http://data.princeton.edu/wws509/datasets/#smoking

Remark: the data set is probably taken from https://www.jstor.org/stable/41983444?seq=1#page_scan_tab_contents and there it is said the the persons under study are males who have contributed in wars before 1956 and who answered a questionary. The authors point out that the dataset is not representative for the whole population.

We are interested in studying if the mortality rate due to lung cancer (the number of deaths due to lung cancer per individual during six year) controlled for age group and smoking status. Assume that the number of deaths for each set of covariate values, Y_i, can be considered Poisson distributed, $Y_i \sim \text{Poisson}(\lambda_i)$. We fit the following model:

```
# load data:
smoking <- read.table(file = "https://www.math.ntnu.no/emner/TMA4315/2018h/smoking.txt")
head(smoking)
nrow(smoking)
model1 <- glm(dead ~ age + smoke, family = "poisson", data = smoking, offset = log(pop))
# note that the size of the population for each combination of the
# covaraites is the offset here
summary(model1)
```

```
##   dead  pop   age ageLevel smoke
## 1   18  656 40-44        1    no
## 2   22  359 45-59        2    no
## 3   19  249 50-54        3    no
## 4   55  632 55-59        4    no
## 5  117 1067 60-64        5    no
## 6  170  897 65-69        6    no
## [1] 36
##
## Call:
## glm(formula = dead ~ age + smoke, family = "poisson", data = smoking,
##     offset = log(pop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.06055  -0.54773   0.06431   0.29963   1.48348
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.63222    0.06783 -53.552  < 2e-16 ***
## age45-59             0.55388    0.07999   6.924 4.38e-12 ***
## age50-54             0.98039    0.07682  12.762  < 2e-16 ***
## age55-59             1.37946    0.06526  21.138  < 2e-16 ***
## age60-64             1.65423    0.06257  26.439  < 2e-16 ***
## age65-69             1.99817    0.06279  31.824  < 2e-16 ***
## age70-74             2.27141    0.06435  35.296  < 2e-16 ***
## age75-79             2.55858    0.06778  37.746  < 2e-16 ***
## age80+               2.84692    0.07242  39.310  < 2e-16 ***
## smokecigarretteOnly  0.36915    0.03791   9.737  < 2e-16 ***
## smokecigarrettePlus  0.17015    0.03643   4.671 3.00e-06 ***
## smokeno             -0.04781    0.04699  -1.017    0.309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4055.984  on 35  degrees of freedom
## Residual deviance:   21.487  on 24  degrees of freedom
## AIC: 285.51
##
## Number of Fisher Scoring iterations: 4
```

a) What is an offset?
For 53-year old non-smokers, what is the estimated number of deaths due to lung cancer (per person over 6 years)?
Why is the number of degrees of freedom for the deviance of this model 24? Does the model give a good fit?

b) Let $\lambda(a, s)$ denote the expected number of lung cancer deaths per person in age group a with smoking status s. For two different smoking statuses s_1 and s_2, define

$$r(a, s_1, s_2) = \frac{\lambda(a, s_1)}{\lambda(a, s_2)}.$$

Explain why $r(a, s_1, s_2)$ does **not** vary as a function of a in `model1`.
For $s_1 =$ `cigarPipeOnly` and $s_2 =$ `cigaretteOnly`, find an estimate value for $r(a, s_1, s_2)$ and an approximate 90 % confidence interval. Is there a significant difference in the expected number of lung cancer deaths for individuals that smoke cigarettes `cigaretteOnly` versus those that smoke cigar/pipe `cigarPipeOnly`?

c) We will now consider two alternative models, `model2` and `model3`:

```
model2 <- glm(dead ~ smoke, family = "poisson", data = smoking, offset = log(pop))
model3 <- glm(dead ~ ageLevel + smoke, family = "poisson", data = smoking, offset = log(pop))

summary(model2)
summary(model3)
```

```
##
## Call:
## glm(formula = dead ~ smoke, family = "poisson", data = smoking,
##     offset = log(pop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -24.546   -5.892   -2.310    8.343   15.612
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.47319    0.03099 -47.532  < 2e-16 ***
## smokecigarretteOnly -0.31219    0.03576  -8.729  < 2e-16 ***
## smokecigarrettePlus -0.43013    0.03468 -12.402  < 2e-16 ***
## smokeno              -0.36678    0.04669  -7.855 3.98e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4056.0  on 35  degrees of freedom
## Residual deviance: 3910.7  on 32  degrees of freedom
## AIC: 4158.7
##
```

```
## Number of Fisher Scoring iterations: 5
##
##
## Call:
## glm(formula = dead ~ ageLevel + smoke, family = "poisson", data = smoking,
##     offset = log(pop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0918  -1.1673  -0.2755   0.7803   2.6364
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.705950   0.050717 -73.071  < 2e-16 ***
## ageLevel             0.333006   0.005591  59.559  < 2e-16 ***
## smokecigarretteOnly  0.405019   0.037463  10.811  < 2e-16 ***
## smokecigarrettePlus  0.203426   0.035996   5.651 1.59e-08 ***
## smokeno             -0.032927   0.046894  -0.702    0.483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4055.984  on 35  degrees of freedom
## Residual deviance:   75.734  on 31  degrees of freedom
## AIC: 325.76
##
## Number of Fisher Scoring iterations: 4
```

Why does `model2` and `model3` have 32 and 31 degrees of freedom, respectively? If we want to compare the three models `model1`, `model2` and `model3`, which model would you choose as the best? Justify your answer by formulating relevant hypotheses and perform hypothesis tests, based on the print-outs above. Hint: remember that to compare two models then one model needs to be nested within the other model.

New: For `model3`: Plot the regression line for the expected number of lung cancer deaths per person as a function of `age` for each of the four different smoking levels in the same plot. First use all coefficients, and then only the ones that are significant. Use `ggplot`! Discuss what you see.

Problem 2: TMA4315 Exam 2012, Problem 3: Precipitation in Trondheim, amount

Remark: the text is slightly modified from the original exam since we parameterized the gamma as in our textbook.

Remark: Problems 1 (binomial) and 2 (multinomial) at the 2012 exam also asked about percipitation.

We want to model the amount of daily precipitation given that it *is* precipitation, and denote this quantity Y. It is common to model Y as a gamma distributed random variable, $Y \sim Gamma(\nu, \mu)$, with density

$$f_Y(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right)$$

In this problem we consider N observations, each gamma distributed with $Y_i \sim Gamma(\nu, \mu_i)$ (remark: common ν). Here ν is considered to be a known nuisance parameter, and the μ_is are unknown.

a) Show that the gamma distribution function is member of the exponential family when μ_i is the parameter of interest.
Use this to find expressions for the expected value and the variance of Y_i, in terms of (ν, μ_i), and interpret ν.
New: What is the canonical link for the Gamma distribution?

Hint: if you want to focus on discussing - you may look at the solutions from Module 1 together.

b) Explain what a saturated model is.
Set up the log-likelihood function expressed by μ_i, and use it to find the maximum likelihood estimators for μ_i-s of the saturated model.
Find the deviance (based on all N observations).

Hint: Do you see directly that $\hat{\mu}_i = y_i$? If not, you may look at the likelihood for one observation and solve that the derivative equals 0.

c) We now want to construct a model for amount of precipitation (given that there are occurrence) with precipitation forecast as explanatory variable.
Let Y_i be amount of precipitation for day i, and let x_i be the precipitation forecast valid for day i. Set up a GLM for this, and argue for your choice of link function and linear predictor.

Hint: explain why this fits into our GLM-gamma framework and set up our 3 equation model (random, systematic, link).

Problem 3: Taken from UiO, STK3100, 2015, problem 2

(For reference: here is the original exam).

This is a problem on the logistic regression.

Do not look at the dataset before the end of the exercise! You should solve the exercise without using `R`, just as if you were at an exam.

In this problem you shall consider data of survivals from a study of treatment for breast cancer. The response is the number that survived for three years. The covariates were the four factors

- `app`: appearance of tumor, two levels, 1 = malignant, 2 = benign
- `infl`: inflammatory reaction, two levels, 1 = minimal, 2 = moderate or severe
- `age`: age of patients, three levels, 1 = under 50, 2 = 50 to 69, 3 = 70 or older
- `country`: hospital of treatment, three levels, 1 = Japan, 2 = US, 3 = UK

The dataset we have used differs slightly from the one they used at UiO (the number of survivals and non-survivals differ).

The number of survivors is modelled as a binomially distributed variable using a canonical logit link. Dummy variable coding is used for all factors, with the level coded as "1" as the reference category.

a) The output from fitting the model where only appearance and country are used as covariates, i.e., a model with predictor on the form

$$\eta = \beta_0 + \beta_1 \texttt{ fapp} + \beta_2 \texttt{ fcountry2} + \beta_3 \texttt{ fcountry3}$$

is displayed below. What is the interpretation of the estimate of the coefficient of appearance, `fapp` (`f` means factor)? Explain also how this coefficient can be expressed in terms of an odds ratio.

Hint: first explain what is the predicted odds of survivial for country j for a benign tumor, then the same for a malignant tumor- and then make an odds ratio.

New: The main difference between our dataset and UiO's dataset is the number of degrees of freedom for the null model; we have 34, they have 35. How many observations do we have in the dataset? And how many does UiO have? Hint: All the covariates in the dataset are categorical, two with 3 levels, and two with 2 levels. How many possible combinations of observations does that make?

```
Call:
glm(formula = cbind(surv, nsurv) ~ fapp + fcountry, family = binomial,
    data = brc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8142  -0.7279   0.2147   0.7576   1.8715

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0803     0.1656   6.522 6.93e-11 ***
fapp2         0.5157     0.1662   3.103 0.001913 **
fcountry2    -0.6589     0.1998  -3.298 0.000972 ***
fcountry3    -0.4945     0.2071  -2.388 0.016946 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 57.588  on 34  degrees of freedom
Residual deviance: 36.625  on 31  degrees of freedom
```

b) The output below is an analysis of deviance for comparing various model spesifications. Fill out the positions indicated `i-iv`.
Note: `y ~ x1*x2` gives both the linear components and the interaction, i.e., it is the same as `y ~ x1 + x2 + x1:x2` (`x1:x2` gives interaction only).

Remark: remember that `anova` gives the sequential comparison of deviances - that is - comparing each model to the previous model.

```
Analysis of Deviance Table

Model 1: cbind(surv, nsurv) ~ fapp + fage + fcountry
Model 2: cbind(surv, nsurv) ~ fapp + fage + finfl + fcountry
Model 3: cbind(surv, nsurv) ~ fapp + finfl + fage * fcountry
Model 4: cbind(surv, nsurv) ~ fapp * finfl + fage * fcountry
Model 5: cbind(surv, nsurv) ~ fapp * finfl + fapp * fage + fage * fcountry
Model 6: cbind(surv, nsurv) ~ fapp * finfl * fage * fcountry
  Resid. Df Resid. Dev Df Deviance
1        29     33.102
2         i     33.095  1   0.0065
3        24     25.674 ii   7.4210
4        23     25.504  1      iii
5        21     22.021  2   3.4823
6         0      0.000 iv  22.0214
```

In the remaining parts of this problem we return to the model in part a) and consider the hypothesis

$$H_0 : \beta_2 + \beta_3 = -1 \text{ versus } H_1 : \beta_2 + \beta_3 \neq -1$$

Note: what are you testing now?

c) The estimated covariance matrix between the estimators of the coefficients β_2 and β_3 above is $\begin{pmatrix} 0.040 & 0.021 \\ 0.021 & 0.043 \end{pmatrix}$. Use a Wald test to test the null hypothesis above.

Note: remember - this was a written exam so you do this by hand!

d) Explain how the null hypothesis can be tested with a likelihood ratio test by fitting two suitable models. No numerical calculations are necessary, but it must be specified how the predictors should be defined. Remark: rather technical - and hint: offset term?

New: Do this test in `R`. Here is the data set!

Work on your own: Exam questions

December 2013 (Essay exam)

We will consider the following Poisson regression

$$Y_i \sim \text{Poisson}(\exp(\eta_i)),\ i = 1, \ldots, n$$

where the linear predictor is $\eta_i = \mathbf{x}_i^T \beta$. Here $\mathbf{x}_i$ is a vector of the p covariates for the ith observation Y_i and β is unknown p dimensional column vector of unknown regression coefficients.

Write an introduction to Poisson regression and its practical usage, for a student with a good background in statistics, but no knowledge about Generalized Linear Models (GLM). Topics you may want to consider, are

- When to use it? Underlying assumptions.
- Parameter estimation, limiting results for the MLE, Fisher information and observed Fisher information, confidence intervals and hypothesis testing.
- Output analysis, residual plots and interpretation of results.
- Deviance and its usage.
- What do we do when a covariate is a factor, and should the results be interpreted?
- The use of Poisson regression in the analysis of contingency tables.

R packages

```
install.packages(c("tidyverse", "ggplot2", "statmod", "corrplot", "ggplot2",
    "GGally", "boot"))
```

Further reading

- A. Agresti (1996): "An Introduction to Categorical Data Analysis".
- A. Agresti (2015): "Foundations of Linear and Generalized Linear Models." Wiley.
- A. J. Dobson and A. G. Barnett (2008): "An Introduction to Generalized Linear Models", Third edition.
- J. Faraway (2015): "Extending the Linear Model with R", Second Edition. http://www.maths.bath.ac.uk/~jjf23/ELM/
- P. McCullagh and J. A. Nelder (1989): "Generalized Linear Models". Second edition.