TMA4315 Generalized linear models H2018

Module 3: BINARY REGRESSION

Mette Langaas, Department of Mathematical Sciences, NTNU - with contributions from Øyvind Bakke, Thea Bjørnland and Ingeborg Hem

13.09 and 20.09 [PL], 14.09 and 21.09 [IL]

(Latest changes: 09.11: clarified one sentence on the devianc. 23.09: score test moved to M4. 20.09: typos, and added solutions to Qs in class. 18.09: typos and added sententence ILw2Problem3c. 16.09: edited and added material for week 2, 13.09 moved material not lectured to after the ILw1, and added one sentence to Problem 5 ILw1.)

Overview

Learning material

- Textbook: Fahrmeir et al (2013): Chapter 2.3, 5.1, B4.1-3
- Classnotes 13.09.2018
- Classnotes 20.09.2018

Topics

First week

- aim of binary regression
- how to model a binary response
- three ingredients of a GLM model
- the logit model: logistic regression
- interpreting the logit model - with odds
- grouped vs. individual data
- parameter estimation with maximum likelihood
  - likelihood, log-likelihood,
  - score function

Second week

- Parameter estimation
  - score function- and mean and covariance thereof,
  - observed and expected information matrix
- comparison with the normal distribution - score function and Fisher information
- exponential family and canonical link
- iterative calculation of ML estimator (Newton-Raphson and Fisher scoring) - and in R with `optim`
- asymptotic properties of ML estimators - how to use in inference?
- statistical inference
  - confidence intervals
  - hypothesis testing: Wald, and likelihood ratio
- deviance: definition, analysis of deviance, deviance residuals
- model fit and model choice
- overdispersion and estimating overdispersion parameter
- sampling strategy: cohort, but also case-control data good for logit model

SECOND WEEK

Remember the beetle and infant respitory disease examples? First, we look back at the model requirements for the binary regression - and the loglikelihood and score function.

Likelihood and derivations thereof - continued

Individual data (not grouped):

Loglikelihood:

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

Score function:

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

Properties of the score function

Since the score function depends on $Y_i = y_i$ we may regard the score function as a random vector. We will now calculate the mean and covariance matrix for the score function.

$E(s(\beta))$

The expected value of the score function is

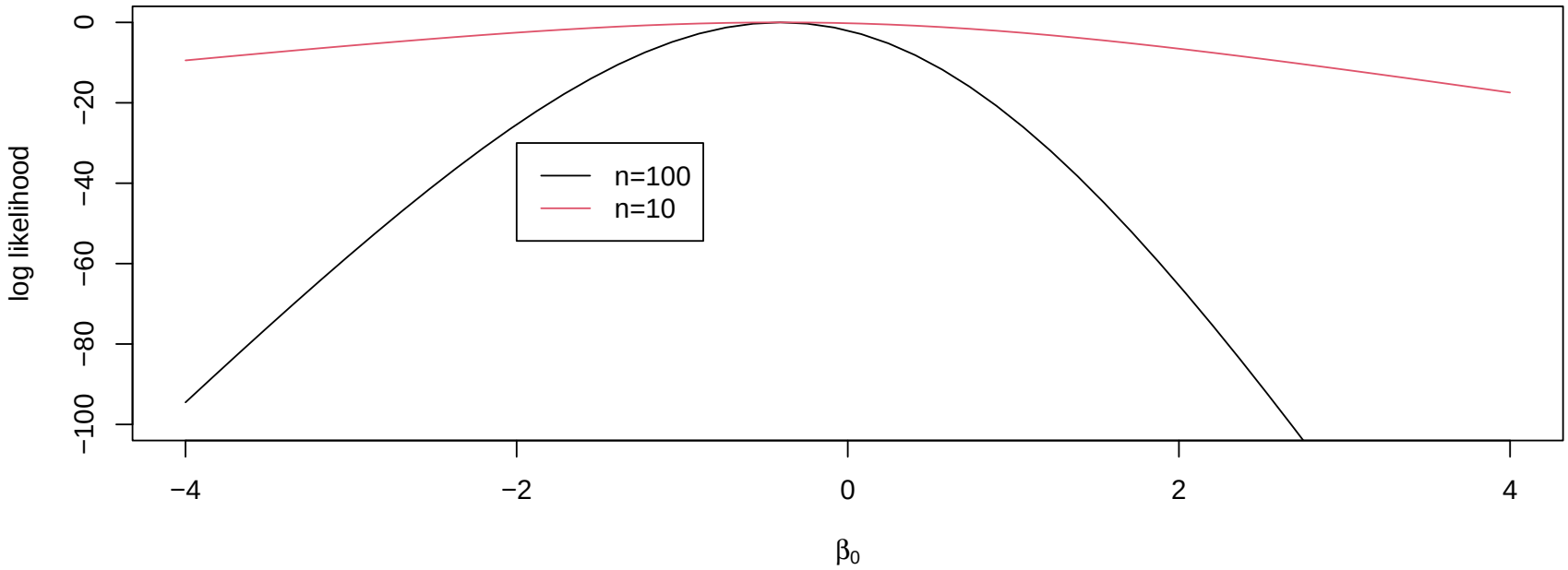
$$\begin{aligned} E(s(\beta)) &= E(\sum_{i=1}^{n}(Y_i - \pi_i)\mathbf{x}_i) \\ &= \sum_{i=1}^{n} E((Y_i - \pi_i)\mathbf{x}_i) \\ &= \sum_{i=1}^{n}(E(Y_i) - \pi_i)\mathbf{x}_i = 0 \end{aligned}$$

as $E(Y_i) = \pi_i$.

We also see that $E(s_i(\beta)) = E((Y_i - \pi_i)\mathbf{x}_i) = 0, \forall i$.

Fisher Information and Variances of Estimates

The "amount of information" that the data carry about the parameters, β, can be summarised by the curvature in the likelihood surface



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

The covariance of $s(\beta)$ is called the expected Fisher information matrix, $F(\beta)$ and is given by

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

assuming that the responses Y_i and Y_j are independent

$F_i(\beta)$

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

$$\begin{aligned} F_i(\beta) &= E(s_i(\beta)s_i(\beta)^T) = E((Y_i - \pi_i)\mathbf{x}_i(Y_i - \pi_i)\mathbf{x}_i^T) \\ &= \mathbf{x}_i\mathbf{x}_i^T E((Y_i - \pi_i)^2) \\ &= \mathbf{x}_i\mathbf{x}_i^T \pi_i(1 - \pi_i) \end{aligned}$$

where $E((Y_i - \pi_i)^2) = \text{Var}(Y_i) = \pi_i(1 - \pi_i)$ is the variance of Y_i. Thus

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

A useful relationship: Under mild regularity conditions (so we can change the order of $\int$ and $\frac{\partial}{\partial \beta}$):

$$\text{Cov}(s(\beta)) = F(\beta) = E\left(-\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right) = E(-\text{Hessian matrix of } l)$$

which relates the expected to the observed Fisher information matrix.

Observed Fisher information matrix $H(\beta)$

What is the observed Fisher information matrix? i.e. don't take expectations...

$$H(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = -\frac{\partial s(\beta)}{\partial\beta^T} = \frac{\partial}{\partial\beta^T}\left[\sum_{i=1}^{n}(\pi_i - y_i)\mathbf{x}_i\right]$$

because $s(\beta) = \sum_{i=1}^{n}(y_i - \pi_i)\mathbf{x}_i$, so $-s(\beta) = \sum_{i=1}^{n}(\pi_i - y_i)\mathbf{x}_i$.
Note that $\pi_i = \pi_i(\beta)$.

Calculating $H(\beta)$

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

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

$$\begin{aligned}
\frac{\partial \pi_i}{\partial \eta_i} &= \frac{\partial \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)} \right)}{\partial \eta_i} \\
&= \frac{(1+\exp(\eta_i))\exp(\eta_i) - \exp(\eta_i)\exp(\eta_i)}{(1+\exp(\eta_i))^2} \\
&= \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \frac{1+\exp(\eta_i)-\exp(\eta_i)}{1+\exp(\eta_i)} \\
&= \pi_i(1-\pi_i).
\end{aligned}$$

And thus

$$
\begin{aligned}
H(\beta) &= \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta^T} \\
&= \sum_{i=1}^{n} \mathbf{x}_i \pi_i (1 - \pi_i) \mathbf{x}_i^T \\
&= \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \pi_i (1 - \pi_i).
\end{aligned}
$$

So the observed and the expected Fisher information matrix are equal.

This is not the case for the probit or complementary log-log models.

Overview of the results for individual and grouped data

- Individual data: $i = 1, \ldots, n$, and pairs $(\mathbf{x}_i, y_i)$.
- Grouped data: $j = 1, \ldots, G$ with n_j observations for group j, and $Y_j = \sum Y_i$ for all i member of group j. In total $\sum_{j=1}^{G} n_j$ observations. For each pair $(\mathbf{x}_j, y_j)$, where $\mathbf{x}_j$ the covariate pattern for group j.

NB: we keep that $\eta_i = \ln(\frac{\pi_i}{1-\pi_i})$ - not changed for grouped data (but now $\mu_j = n_j \pi_j$).

Log-likelihood:

Individual:

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

Grouped:

$$l(\beta) = \sum_{j=1}^{G} [y_j \ln \pi_j - y_j \ln(1 - \pi_j) + n_j \ln(1 - \pi_j) + \ln \binom{n_j}{y_j}]$$

The last part is usually not include in calculations.

Score function:

Individual:

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

Grouped:

$$s(\beta) = \sum_{j=1}^{G} \mathbf{x}_j (y_j - n_j \pi_j)$$

Expected Fisher information matrix:

Individual:

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

Grouped:

$$F(\beta) = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T n_j \pi_j (1 - \pi_j)$$

The observed Fisher information matrix equals the expected Fisher information matrix - because the logit model is the *canonical link* for the binomial distribution.

Fitting the Models

There is no analytic solution, so we have to resort to numerics.
Luckily, these models behave well enough

Iterative gradient-based methods

Use a first order multivariate Taylor approximation for $s(\hat{\beta})$, around some chosen reference value $\beta^{(0)}$:

$$s(\hat{\beta}) \approx s(\beta^{(0)}) + \frac{\partial s(\beta)}{\partial \beta}\Big|_{\beta=\beta^{(0)}} (\hat{\beta} - \beta^{(0)})$$

Let $H(\beta^{(0)}) = -\frac{\partial s(\beta)}{\partial \beta}\Big|_{\beta=\beta^{(0)}}$. Setting $s(\hat{\beta}) = 0$ solving for $\hat{\beta}$ gives

$$\hat{\beta} = \beta^{(0)} + H(\beta^{(0)})^{-1} s(\beta^{(0)})$$

where $H(\beta^{(0)})^{-1}$ is the matrix inverse of $H(\beta^{(0)})$.

Enter nNewton and Raphson

If we start with some value $\beta^{(0)}$ and then find a new value $\beta^{(1)}$ by applying this equation, and then continue applying the equation until convergence we have the *Newton-Raphson* method:

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

Replacing the observed Fisher information matrix $\mathbf{H}$ with the expected Fisher information matrix $\mathbf{F}$ yields the *Fisher-scoring* method.

For the logit model these two methods are the same since the observed and expected Fisher information matrix is the same for canonical link functions (like the logit is for binary regression).

This algorithm is run until the relative difference in Euclidean distance between two iterations "(new-old)/old" is smaller than some chosen constant.

Requirements for convergence

For the Newton-Raphson algorithm we see that the observed Fisher information matrix H needs to be invertible for all β, alternatively for the Fisher scoring algorithm the expected Fisher information matrix F needs to be invertible.

Proof of convergence

In our logit model

$$F(\beta) = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T n_j \pi_j (1 - \pi_j)$$

Let $\mathbf{X}$ be the design matrix, where the rows are $\mathbf{x}_j^T$. Then $\mathbf{X}^T\mathbf{X} = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T$.
If we require that the design matrix has full rank (G) then also $\mathbf{X}^T\mathbf{X}$ will have full rank (it will also be positive definite) and and in addition $\pi_j(1 - \pi_j) > 0$ for all $\pi_j \in (0, 1)$, so then $F(\beta)$ will be positive definite and all is good.

Why is $F(\beta)$ positive definite if we require that the design matrix has full rank?

Formally, let $\mathbf{X}$ be a $n \times p$ matrix and Λ a $n \times n$ diagonal matrix where all the diagonal elements are positive (like our $\pi_j(1-\pi_j)$, yes, put them on the diagonal). Let $\mathbf{X}$ have independent columnes (full rank) $\Leftrightarrow \mathbf{X}^T\Lambda\mathbf{X}$ is positive definite.

Proof: $\Rightarrow$: Let $\mathbf{v}$ be a p dimensional column vector. Assume $0 = \mathbf{v}^T\mathbf{X}^T\Lambda\mathbf{X}\mathbf{v} = (\Lambda^{1/2}\mathbf{X}\mathbf{v})^T(\Lambda^{1/2}\mathbf{X}\mathbf{v}) = \sum_{i=1}^n w_i^2$ where $\mathbf{W} = \Lambda^{1/2}\mathbf{X}\mathbf{v}$. Then, w must be 0, that is $\Lambda^{1/2}\mathbf{X}\mathbf{v} = 0$ since multiplication with $\Lambda^{1/2}$ is to multiply each element in $\mathbf{X}\mathbf{v}$ with a number different from 0. That is, we must have $\mathbf{v} = \mathbf{0}$ since $\mathbf{X}$ has independent columns.

$\Leftarrow$: Assume that $\mathbf{X}\mathbf{v} = \mathbf{0}$. Then $\mathbf{v}^T\mathbf{X}^T\Lambda\mathbf{X}\mathbf{v} = \mathbf{0}$ so $\mathbf{v} = \mathbf{0}$ since $\mathbf{X}^T\Lambda\mathbf{X}$ is positive definite. This is, $\mathbf{X}$ has independent columns.

End of proof

We need a full rank

Therefore, for GLMs we will also - as for the multiple linear regression model in Module 2 - assume that the design matrix has full rank!

We will see in Module 5 that this is the requirement needed for GLMs in general.

Convergence

Convergence is still not guarnateed, especially for small samples. According 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. In practice, the logit model most often causes problems, when (for grouped data) $y_i = 0$ or $y_i = n_i$, because $\hat{\pi}_i = 0/1$, so $\hat{\eta}_i = \pm\infty$. Computers do not like infinity

stopped here

Asymptotic properties of ML estimates

Results

We need some weak regularity conditions, including

- β falls in the interior of the parameter space and
- p is fixed that n increases

(Agresti (2015) page 125):

The Results

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. $\hat{\beta} \approx N_p(\beta, F^{-1}(\hat{\beta}))$

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

The *proof* (for the univariate case) is given in the course TMA4295 Statistical Inference course, Casella and Berger (2002):"Statistical inference", page 472.

Here we will sketch the proof. The strategy:

- make a first order Taylor expansion of the score function around the true parameter,
- use the fact that the maximum likelihood estimate is defined as the zero of the score function.

The sketch

Use θ as the parameter of interest
(there is the connection to μ, then to η and finally to β)
We start with the multivariate version of the first order Taylor expansion of the score around the true parameter value θ:

$$s(\theta) \approx s(\hat{\theta}) + s'(\theta)(\hat{\theta} - \theta)$$

As $s'(\theta) = \mathbf{H}(\theta)$, and $s(\hat{\theta}) = 0$,

$$s(\hat{\theta}) \approx s(\theta) - \mathbf{H}(\theta)(\hat{\theta} - \theta) = 0$$

$$s(\theta) \approx \mathbf{H}(\theta)(\hat{\theta} - \theta)$$

And premultiplying with $\mathbf{H}^{-1}(\theta)$ gives

$$(\hat{\theta} - \theta) \approx \mathbf{H}^{-1}(\theta)s(\theta)$$

Then, to use the central limit theorem we need some smart manipulations with n, so we start by multiplying with $\sqrt{n}$ and split that into n and $\frac{1}{\sqrt{n}}$.

$$\sqrt{n}(\hat{\theta} - \theta) \approx \sqrt{n}\mathbf{H}^{-1}(\theta)s(\theta) = (\frac{1}{n}\mathbf{H}(\theta))^{-1}\frac{1}{\sqrt{n}}s(\theta)$$

From the central limit theorem:

1) $\frac{1}{n}\mathbf{H}(\theta)$ goes to the expected value which is $\mathbf{F}(\theta)$ (in probability),
2) the part $\frac{1}{\sqrt{n}}s(\theta)$ asymptotically goes to a random variable W that follows a multivariate normal with $\mathbf{W} \sim N(\mathbf{0}, \frac{1}{n}\mathbf{F}(\theta))$:
   - mean $\mathrm{E}(\frac{1}{\sqrt{n}}s(\theta)) = \mathbf{0}$ and the
   - covariance matrix is $\mathrm{Cov}(\frac{1}{\sqrt{n}}s(\theta)) = \frac{1}{n}\mathbf{F}(\theta)$

$$\sqrt{n}(\hat{\theta} - \theta) \approx \mathbf{F^{-1}}(\theta)\mathbf{W}$$

On the right side here we have a multivariate normal distributed random variable $\mathbf{F^{-1}}(\theta)\mathbf{W}$ with mean $\mathbf{0}$ and covariance matrix

$$\text{Cov}(\mathbf{F}^{-1}(\theta)\mathbf{W}) = \mathbf{F}^{-1}(\theta)\frac{1}{n}\mathbf{F}(\theta)\mathbf{F}^{-1}(\theta) = \frac{1}{n}\mathbf{F}^{-1}(\theta)$$

This leads to the wanted result:

$$\hat{\theta} \approx N(\theta, \mathbf{F}^{-1}(\theta))$$

Due to the Slutsky theorem (from TMA4295 Statistical inference) this also holds when $\mathbf{F}^{-1}(\theta))$ is replaced by $\mathbf{F}^{-1}(\hat{\theta}))$.

Parameter estimation

Parameter estimation can be based on grouped data - so now we use $Y_j \sim \text{bin}(n_j, \pi_j)$ from 1 above, but keep 2 and 3 unchanged. The number of groups is G and the total number of observations is $\sum_{j=1}^{G} n_j$.

- Likelihood=joint distribution, exponential family.

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

where we have that $\theta = \ln(\frac{\pi}{1-\pi})$ for the binomial distribution, which means that our logit model is gives the canonical link (remember, good properties!).

- Log-likelihood

$$l(\beta) = \sum_{j=1}^{G} [y_j \ln \pi_j - y_j \ln(1 - \pi_j) + n_j \ln(1 - \pi_j) + \ln \binom{n_j}{y_j}]$$

- Score function=vector of partial derivatives of log-likelihood. Find ML by solving $s(\hat{\beta})) = 0$ - but no closed form solutions.

$$s(\beta) = \sum_{j=1}^{G} \mathbf{x}_j (y_j - n_j \pi_j)$$

- Expected Fisher information matrix

$$F(\beta) = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T n_j \pi_j (1 - \pi_j)$$

- $\hat{\beta}$ found iteratively using Newton-Raphson or Fisher scoring

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

- $\hat{\beta} \approx N_p(\beta, F^{-1}(\hat{\beta}))$

Further statistical inference

Our further statistical inference (confidence intervals and hypotheses tests) are based on the asymptotic distribution of the parameter estimates

$$\widehat{\beta} \approx N_p(\beta, F^{-1}(\widehat{\beta}))$$

where $F^{-1}(\widehat{\beta}))$ is the inverse of the expected Fisher information matrix inserted $\widehat{\beta}$.

For the logit model we found that

$$F(\beta) = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T n_j \pi_j (1 - \pi_j)$$

So we would need to do $\pi_j = \frac{\exp(\eta - j)}{1 + \exp(\eta_j)}$ and $\eta_j = \mathbf{x}_j^T \beta$ as "usual", and then replace β with $\widehat{\beta}$.

The asymptotic distribution still holds when we replace β with $\widehat{\beta}$ in $\mathbf{F}$.

If we make a diagonal matrix $\mathbf{W}$ with $n_j\pi_j(1-\pi_j)$ on the diagonal, then we may write the matrix $F(\beta)$ in matrix notation. As before $\mathbf{X}$ is the $G \times p$ design matrix.

$$F(\beta) = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T n_j \pi_j (1-\pi_j) = \mathbf{X}^T \mathbf{W} \mathbf{X}.$$

which means that $\text{Cov}(\hat{\beta}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ for the binomial model (remember that $\hat{\beta}$ comes in with $\hat{\pi}_j$ in $\mathbf{W}$).

Q: How is this compared to the normal case?

A: $F(\beta) = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}$, and the inverse $\text{Cov}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.

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

For one element of the parameter vector:

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{a}_{kk}(\beta)}}$$

is asymptotically standard normal. We will use this now!

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.

We focus on element k of β, called β_k. It is known that asympotically

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{a_{kk}(\hat{\beta})}} \sim N(0,1)$$

We use that to form confidence intervals.

Let $z_{\alpha/2}$ be such that $P(Z_k > z_{\alpha/2}) = \alpha/2$.

We then use

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

insert Z_k and solve for β_k to get

$$P(\hat{\beta}_k - z_{\alpha/2}\sqrt{a_{kk}(\hat{\beta})} \leq \beta_k \leq \hat{\beta}_k - z_{\alpha/2}\sqrt{a_{kk}(\hat{\beta})}) = 1 - \alpha$$

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

Q: We write $a_{kk}(\hat{\beta})$. Why not $a_{kk}(\hat{\beta}_{kk})$?

Example with the beetle data

Again, we study our beetle data - in the grouped version.
Here we calculate the upper and lower limits of the confidence interval using the formula.

```
fitgrouped=glm(cbind(y, n-y) ~ ldose, family = "binomial",
coeff=fitgrouped$coefficients
sds=sqrt(diag(summary(fitgrouped)$cov.scaled))
alpha=0.05
lower=coeff-qnorm(1-alpha/2)*sds
upper=coeff+qnorm(1-alpha/2)*sds
cbind(lower,upper)
```

```
##                    lower     upper
## (Intercept) -70.87144 -50.56347
## ldose         28.56265  39.97800
```

Q: Explain what is done in the R-print-out.

Hypothesis testing

There are three methods that are mainly used for testing hypotheses in GLMs: - Wald test, - likelihood ratio test and - score test.

We will look at the first two.

First, look at linear hypotheses: We study a binary regression model with $p = k + 1$ covariates, and refer to this as model A (the larger model). As for the multiple linear model we then want to investigate the null and alternative hypotheses of the following type(s):

$$
\begin{aligned}
H_0 : \beta_j &= 0 \text{ vs. } H_1 : \beta_j \neq 0 \\
H_0 : \beta_1 &= \beta_2 = \beta_3 = 0 \text{ vs. } H_1 : \text{ at least one of these } \neq 0 \\
H_0 : \beta_1 &= \beta_2 = \cdots = \beta_k = 0 \text{ vs. } H_1 : \text{ at least one of these } \neq 0
\end{aligned}
$$

We call the restricted model (when the null hypotesis is true) model B, or the smaller model.

These null hypotheses and alternative hypotheses can all be rewritten as a linear hypothesis

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

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

The Wald test

The Wald test statistic is given as:

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

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 $\mathbf{C}\hat{\beta}$.
Remember: $\text{Cov}(\mathbf{C}\hat{\beta}) = \mathbf{C}F^{-1}(\hat{\beta})\mathbf{C}^{\mathsf{T}}$.
Asymptotically, under the null hypothesis $w \sim \chi^2_r$ distribution with (where r is the number of hypotheses tested).
P-values are calculated in the upper tail of the χ^2-distribution.
Observe: to perform the test you only need to fit the larger model (and not the smaller).

For the special case that we only test one regression parameter, for example β_k:

$$H_0 : \beta_k = 0 \text{ vs. } H_1 : \beta_k \neq 0.$$

Now $\mathbf{C}\hat{\beta} = \beta_k$ and $\mathbf{C}[F(\hat{\beta})]^{-1}\mathbf{C}^\mathsf{T} = \mathbf{C}\mathbf{A}(\beta)\mathbf{C}^\mathsf{T} = a_{kk}(\beta)$, and the Wald test becomes

$$(\hat{\beta}_k - \beta_k)[a_{kk}(\hat{\beta})]^{-1}(\hat{\beta}_k - \beta_k) = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{a_{kk}(\hat{\beta})}}\right)^2 = Z_k^2$$

so, asymptotically the square of the standard normal, which we know follows a χ^2-distribution with 1 degree of freedom.

Q: Explain what you find in the columns named `z value` and `Pr(>|z|)` below, and which hypothesis tests these are related to. Are the hypothesis tests performed using the Wald test?

```
library(investr)
fitgrouped=glm(cbind(y,n-y) ~ ldose, family="binomial",
               data = investr::beetle)
knitr::kable(summary(fitgrouped)$coefficients, digits=2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-60.72	5.18	-11.72	0
ldose	34.27	2.91	11.77	0

The likelihood ratio test

The likelihood ratio test (LRT), which compares the likelihood of two models.

- First we maximize the likelihood for model A (the larger model) to get $L(\hat{\beta}_A)$ and $\hat{\beta}_A$.
- Then we maximize the likelihood for model B (the smaller model) to get $L(\hat{\beta}_B)$ and $\hat{\beta}_B$.

The likelihood of the larger model (A) will always be larger or equal to the likelihood of the smaller mode (B). (Why?)

The likelihood ratio statistic is defined as

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

(so, -2 times small minus large).

Under weak regularity conditions the test statistic is approximately χ^2-distributed with degrees of freedom equal the difference in the number of parameters in the large and the small model. This is general - and not related to the GLM! More in TMA4295 Statistical Inference!

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

Observe: to perform the test you need to fit both the small and the large model.

Notice: asymptotically the Wald and likelihood ratio test statistics have the same distribution, but the value of the test statistics might be different. How different?

For the beetle data we compare model A=model with `ldose` as covariate with model B=model with only intercept. We use the loglikelihood-function that we made for the lecture session for week 2.

```
library(investr)
fitgrouped=glm(cbind(y, n-y) ~ ldose, family = "binomial",
fitnull=glm(cbind(y, n-y) ~ 1, family = "binomial", data =

loglik <- function(par, args){
  y <- args$y; x <- args$x; n <- args$n
  res <- sum(y*x%*%par - n*log(1 + exp(x%*%par)))
  return(res)
}
```

```r
# call this with parameters estimated under model A=larger
beetleargs = list(y = investr::beetle$y,
                  x = cbind(rep(1, nrow(investr::beetle)), invest
                  n = investr:: beetle$n)

llA=loglik(matrix(fitgrouped$coefficients,ncol=1),args=beet

# then the smaller model, then we set the coeff for ldose
llB=loglik(matrix(c(fitnull$coefficients,0),ncol=1),args=be
lrt=-2*(llB-llA)
lrt
```

```
## [1] 272.9702
```

```r
pchisq(lrt,df=1, lower.tail = FALSE)
```

```
## [1] 2.556089e-61
```

```r
anova(fitnull, fitgrouped, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(y, n - y) ~ 1
## Model 2: cbind(y, n - y) ~ ldose
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         7    284.202
## 2         6     11.232  1   272.97 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Q and A: Here the small model is the model with only intercept and the large is the one with dose as covariate. This means that the null hypothesis is that "the small model is preferred" and our p-value is very small, so we reject the null hyptheses and stick with the model with dose as covariate. Observe that the LRT can be performed using anova.

Deviance

The *deviance* is used to assess model fit and also for model choice, and is based on the likelihood ratio test statistic.

Saturated model: One parmeter per group: estimate π_j by the observed frequency for the group: $\tilde{\pi}_j = \frac{y_j}{n_j}$. Then $\tilde{\pi}$ is a G-dimensional column vector with the elements $\tilde{\pi}_j$.

This "imaginary model" is called the *saturated* model.

Candidate model: The model that we are investigated can be thought of as a *candidate* model. Then we maximize the likelihood to get $\hat{\beta}$ & thus $\hat{\pi}_j$. Then $\hat{\pi}$ is a G-dimensional column vector with the elements $\hat{\pi}_j$.

The *deviance* is defined as the likelihood ratio statistic, where we put the saturated model in place of the larger model A and our candidate model in place of the smaller model B:

$$\begin{aligned} D &= -2(\ln L(\text{candidate model}) - \ln L(\text{saturated model})) \\ &= -2(l(\hat{\pi}) - l(\tilde{\pi})) = -2\sum_{j=1}^{G}(l_j(\hat{\pi}_j) - l_j(\tilde{\pi}_j)) \end{aligned}$$

For our logit model this can be written as (after some maths):

$$D = 2\sum_{j=1}^{G}[y_j \ln(\frac{y_j}{n_j\hat{\pi}_j}) + (n_j - y_j)\ln(\frac{n_j - y_j}{n_j - n_j\hat{\pi}_j})]$$

Verify this by yourself.

If our model is good, it should not be too far from the saturated model, and we measure this distance by the deviance.

If we want to investigate the null hypothesis that "our model fits the data well" to the negation, it is useful to know that asymptotically D is distributed as χ^2 with $G - p$ degrees of freedom (same reason as for the likelihood ratio test statistic).

This result depends on n_j being large, hard to say how large (at least 5 is a rule of thumb).

The deviance is in `summary.glm` outputted as "Residual deviance", which we read off as 11.2322311. Let's check for our beetle example by computing the formula for D directly:

```
D=deviance(fitgrouped)
D
```

```
## [1] 11.23223
```

```
G=dim(investr::beetle)[1]
G
```

```
## [1] 8
```

```
p=2
1-pchisq(D,G-p)
```

```
## [1] 0.08145881
```

So, do we have a good fit?

The null hypothesis is that the candiate model is equally good as the saturated model. We do not reject this hypothesis at level 0.05. This means that we are satisfied with the candidate model.

In the summary from glm also the socalled *NULL deviance* is given. This is the deviance when the candicate model is the model with only intercept term present. This deviance asymptotically distributed as χ^2 with $G - 1$ degrees of freedom.

```
summary(fitgrouped)
```

```
##
## Call:
## glm(formula = cbind(y, n - y) ~ ldose, family = "binomial", data = i
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## ldose         34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

Q: where is the deviance(s) here and how do we use these?

Analysis of deviance

In MLR we have seen that we may produce a sequential analysis of variance (Type I) by adding more and more terms to the model and comparing the scaled decrease in SSE by the scaled SSE of a full model.

For the binary regression we may adapt a similar strategy, but with using the scaled change in deviance instead of the SSE.

We use the infant respiratory disease data as an example

```r
library(faraway)
fit=glm(cbind(disease, nondisease)~sex*food,family=binomial(link=logit))
summary(fit)
```

```
##
## Call:
## glm(formula = cbind(disease, nondisease) ~ sex * food, family = bino
##     data = babyfood)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.59899    0.12495 -12.797  < 2e-16 ***
## sexGirl            -0.34692    0.19855  -1.747 0.080591 .
## foodBreast         -0.65342    0.19780  -3.303 0.000955 ***
## foodSuppl          -0.30860    0.27578  -1.119 0.263145
## sexGirl:foodBreast -0.03742    0.31225  -0.120 0.904603
## sexGirl:foodSuppl   0.31757    0.41397   0.767 0.443012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.6375e+01  on 5  degrees of freedom
## Residual deviance: 4.2144e-13  on 0  degrees of freedom
## AIC: 43.518
```

Deviance residuals

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

$$d_k = \text{sign}(y_k - n_k\hat{\pi}_k) \cdot \left\{ 2[y_k \ln(\frac{y_k}{n_k\hat{\pi}_k}) + (n_k - y_k)\ln(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k})] \right\}^{1/2}$$

where the term $\text{sign}(y_k - n_k\hat{\pi}_k)$ makes negative residuals possible.

Model assessment and 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 for grouped data

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

Pearson test and residuals

An alternative to the deviance test is the Pearson test. We will look in more detail at this test in a Module 4. The Pearson test statistic can be written as a function of the Pearson residuals, which for the binomial regression is given as:

$$r_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

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

The Pearson χ^2-goodness of fit statistic is given as

$$X_P^2 = \sum_{j=1}^{G} r_j^2 = \sum_{j=1}^{G} \frac{(y_j - n_j\hat{\pi}_j)^2}{n_j\hat{\pi}_j(1-\hat{\pi}_j)}$$

The Pearson χ^2 statistic is asymptotically equivalent to the deviance statistic and thus is asymptotically χ^2_{G-p}.

The Pearson χ^2 statistic is not a good choice if any of the groups have a low expected number of observations, i.e. $n_j\hat{\pi}_j$ is small (below 1).

Model assessment with continuous covariates

If data have continuous covariates it is possible to form groups based making intervals for continuous covariates. Alternatively grouping on predicted probabilites can be done.

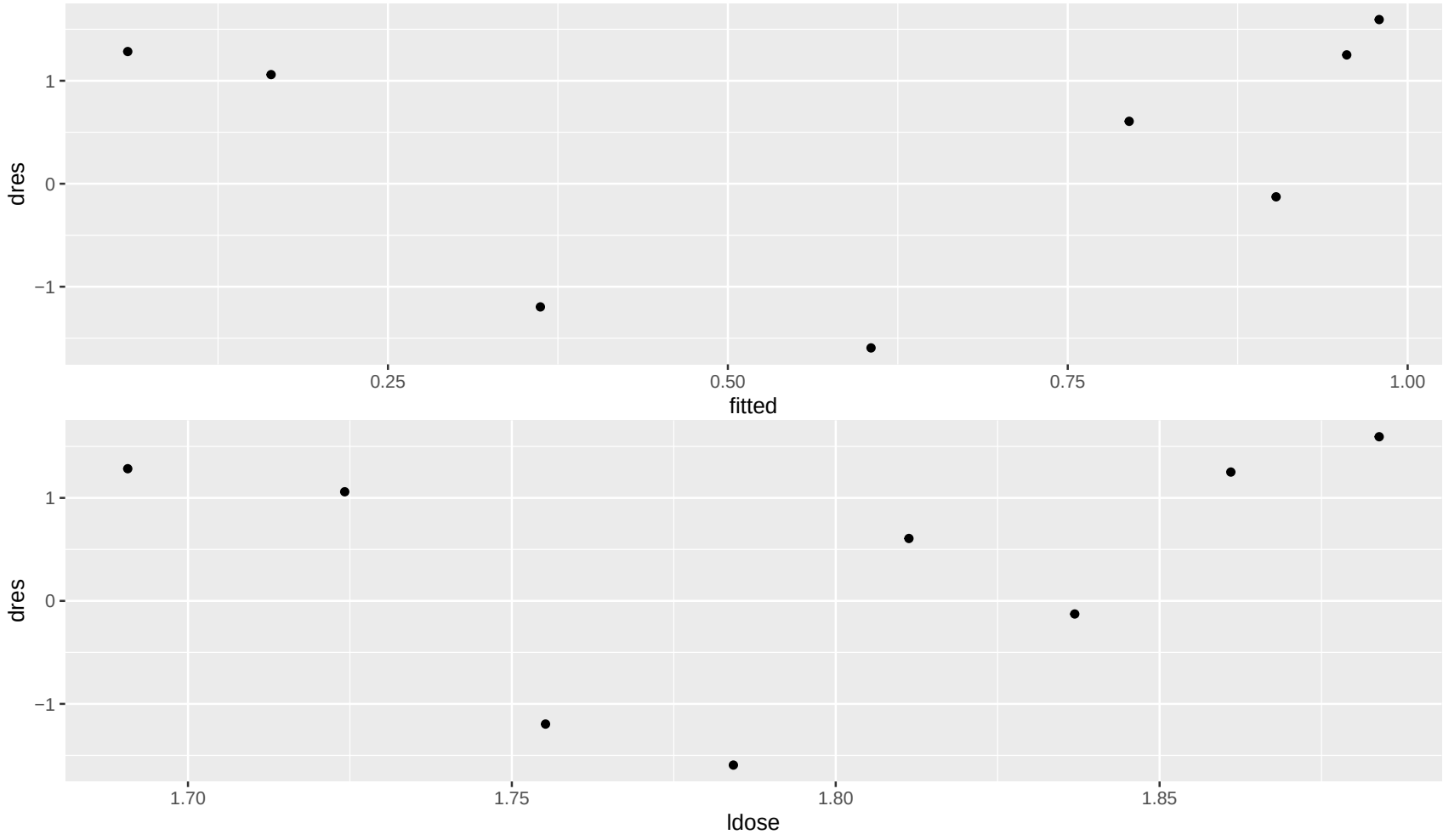
For continuous data the Hosmer Lemeshow test can be used - not on our reading list.

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.

If n_j is small for the covariate patterns the residual plots may be relatively uninformative.

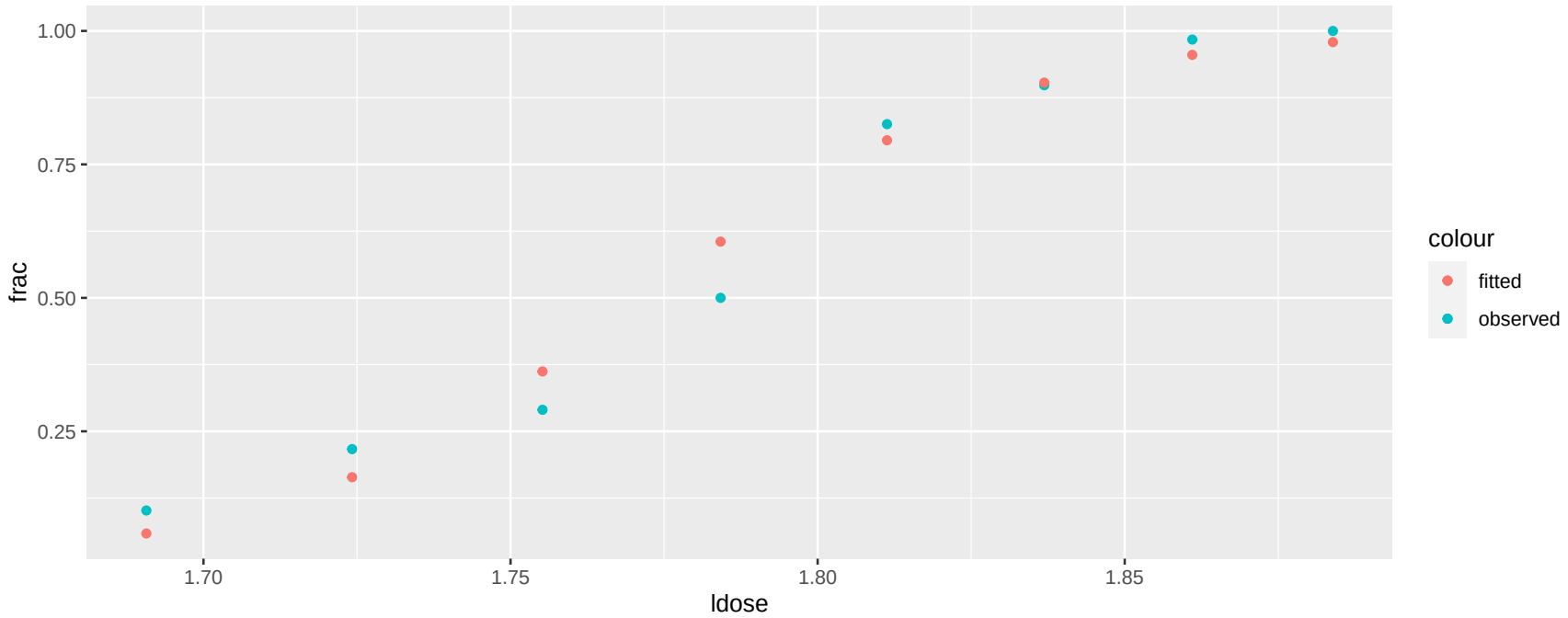
Residual plots for the logistics regression - and for the GLM in general - is highly debated, and we will not put much emphasis on residual plots for this module.


dres
1
0
−1
0.25
0.50
0.75
1.00
fitted
dres
1
0
−1
1.70
1.75
1.80
1.85
ldose

Other plots

A useful plot is to show observed and fitted proportions (grouped data) plotted against the linear predictor or covariates.

```
library(ggplot2)
df=data.frame("fitted"=fitgrouped$fitted.values,"dres"=resi
ggplot(df,aes(x=ldose))+geom_point(aes(y=frac,colour="obse
```



AIC

It is known to us from multiple linear regression that if a model is chosen based on a goodness of fit statistic (like the SSE or R^2 in multiple linear regression) will in general result in us choosing a to big model (to many parameters fit). The Akaike informations criterion - that we studied for multiple linear regression - can also be used for binary regression: Let p be the number of regression parameters in our model.

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

A scaled version of AIC, standardizing for sample size, is sometimes preferred.

To use AIC for model selection you use the model with the *smallest* AIC.

We may also use the BIC, where $2p$ is replaced by $\log(G) \cdot p$ or $\log(n) \cdot p$.

```r
library(faraway)
fit1=glm(cbind(disease, nondisease)~1,family=binomial(link=logit),data=
fit2=glm(cbind(disease, nondisease)~sex,family=binomial(link=logit),dat
fit3=glm(cbind(disease, nondisease)~food,family=binomial(link=logit),da
fit4=glm(cbind(disease, nondisease)~food+sex,family=binomial(link=logit
fit5=glm(cbind(disease, nondisease)~food*sex,family=binomial(link=logit
AIC(fit1,fit2,fit3,fit4,fit5)
```

```
##      df      AIC
## fit1  1 59.89324
## fit2  2 56.41710
## fit3  3 43.21693
## fit4  4 40.23987
## fit5  6 43.51795
```

Q: Which of these 5 models would you prefer?

Overdispersion and estimating overdispersion parameter

When we have grouped data: $Y_j \sim \text{Bin}(n_j, \pi_j)$ and $\text{Var}(Y_j) = n_j\pi_j(1-\pi_j)$.

It is possible to estimate the variance (within a group) by $n_j\bar{y_j}(1-\bar{y_j})$ where $\bar{y_j} = y_j/n_j$ (this is an estimate of π_j for group j). We call this the *empirical variance*.

In a logistic regresson we estimate $\hat{\pi}_j = h(\mathbf{x}_j^T\hat{\beta})$ ($h(\cdot)$ is the inverse link function) which is

$$\hat{\pi_j} = \frac{\exp(x_j^T\hat{\beta})}{1+\exp(x_j^T\hat{\beta})}$$

for a logistic regression. This would give the *estimated binomial variance* for Y_j as $n_j\hat{\pi}_j(1-\hat{\pi}_j)$.

Some times the empirical variance is much larger than the estimated binomial variance of the model. This is called *overdispersion* and may occur when the individual responses within the groups are correlated, or when the model could be improved upon (missing/unobserved covariates?).

Positively correlated binary variables will give a variance of the sum that is larger than for uncorrelated variables, e.g.

$$\text{Var}(\sum_{k=1}^{K} Y_k) = \sum_{k=1}^{K} \text{Var}(Y_k) + 2 \sum_{k<l} \text{Cov}(Y_k, Y_l).$$

This can be handeled by including an *overdispersion parameter*, named ϕ, in the variance formula:

$$\text{Var}(Y_j) = \phi n_j \pi_j (1 - \pi_j)$$

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

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

where D is the deviance. Note that similarity to $\widehat{\sigma^2} = 1/(n-p) \cdot \text{SSE}$ in the MLR. The Cov$(\widehat{\beta})$ can then be changed to $\hat{\phi} F^{-1}(\widehat{\beta})$.

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

In Modules 7 and 8 we will look at using multilevel models to handle correlated observations.

```r
library(investr)
estpi=investr::beetle$y/investr::beetle$n
empvars=investr::beetle$n*estpi*(1-estpi)
fit=glm(cbind(y, n-y) ~ ldose, family = "binomial", data =
modelestvar=investr::beetle$n*fit$fitted.values*(1-fit$fitt
cbind(empvars,modelestvar)
```

```
##      empvars modelestvar
## 1   5.389831    3.254850
## 2 10.183333    8.227364
## 3 12.774194   14.321308
## 4 14.000000   13.378891
## 5   9.079365   10.261038
## 6   5.389831    5.156652
## 7   0.983871    2.653383
## 8   0.000000    1.230704
```

```r
est.dispersion=fit$deviance/fit$df.residual
est.dispersion
```

```
## [1] 1.872039
```

References for further reading

- 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.

If we have time

Look back at MLR - what is $s(\beta)$ and $F(\beta)$ then?

1. $Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$
2. $\eta_j = x_i^T \beta$
3. $\mu_i = \eta_i$ (identity response function and link function)

Likelihood:

$$L(\beta) = \left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)\right)$$

Loglikelihood:

$$l(\beta) = \ln L(\beta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$$

Since $(y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) = Y^TY - 2Y^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta$, then

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = -\frac{1}{2\sigma^2}(2\mathbf{X}^T\mathbf{X}\beta - 2\mathbf{X}^TY) = \frac{1}{\sigma^2}(\mathbf{X}^TY - \mathbf{X}^T\mathbf{X}\beta)$$

and $s(\widehat{\beta}) = 0$ gives $\mathbf{X}^TY - \mathbf{X}^T\mathbf{X}\beta = 0$ which can be solved on closed form giving $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY$. So, no need for iterative methods.

Finally, observed Fisher information matrix.

$$H(\beta) = \frac{\partial s(\beta)}{\partial \beta^T} = -\frac{\partial}{\partial \beta_T}(\frac{1}{\sigma^2}\mathbf{X}^T Y - \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\beta) = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}$$

which is independent on β, and also we see that $F(\beta) = \mathsf{E}(H(\beta)) = H(\beta)$ since no random variables are present. The identity link is also the canonical link. Finally, the (asymptotic) covariance of the ML estimate is $F^{-1}(\widehat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ which we know as $\mathsf{Cov}(\widehat{\beta})$.

Exponential family - and canonical link

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

- 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 $\mathrm{Var}(Y_i) = b''(\theta_i) \cdot \frac{\phi}{w_i}$.

In Module 1 we found that the binomial distribution $Y_i \sim \text{bin}(1, \pi_i)$ is an exponential family (derivation from Module 1: https://www.math.ntnu.no/emner/TMA4315/2017h/Module1ExponentialFamily.pdf)

and that

- $\theta_i = \ln(\frac{\pi_i}{1-\pi_i})$ is the canonical parameter
- $\phi = 1$, no nuisance
- $w_i = 1$
- $b(\theta_i) = \ln(1 + \exp(\theta_i))$

Recall that in a GLM we choose a link function g, linking the linear predictor and the mean: $\eta_i = g(\mu_i)$. For the logit model we had that $\eta_i = \ln(\frac{\pi_i}{1-\pi_i})$.

Now (new to us) - every exponential family has a unique *canonical link function* such that

$$\theta_i = \eta_i$$

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

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

to have a canonical link.

Q: Is the logit link the canonical link for the binary model?

A:

Yes, since $\theta_i = \ln(\frac{\pi_i}{1-\pi_i}) = g(\pi_i)$ then the logit link is the canonical link for the binary regression.

##Properties of a GLM with canonical link

1. The log-likelihod is always concave so that the ML estimated is always unique (given that it exists).
2. The observed Fisher information matrix $H(\beta)$ *equals* the expected Fisher information matrix $F(\beta)$. That is,

$$-\frac{\partial^2 l}{\partial \beta \beta^T} = \mathsf{E}(-\frac{\partial^2 l}{\partial \beta \beta^T})$$

Proving this is beyond the scope of this course.

Parameter estimation - in practise

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

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

We have that the score function for the logit model is:

$$s(\beta) = \sum_{j=1}^{G} \mathbf{x}_j (y_j - n_j \pi_j)$$