TMA4315 Generalized linear models

Module 3: BINARY REGRESSION

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

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 stragegy: cohort, but also case-control data good for logit model

Aim of binary regression

Two aims

1. Construct a model to help understand the relationship between a "success probability" and one or several explanatory variables. The response measurements are binary (present/absent, true/false, healthy/diseased).
2. Use the model for estimation and prediction of success probabilites.

Two running examples: mortality of beetles and probability of respiratory infant disease.

Example: Dose response of beetles

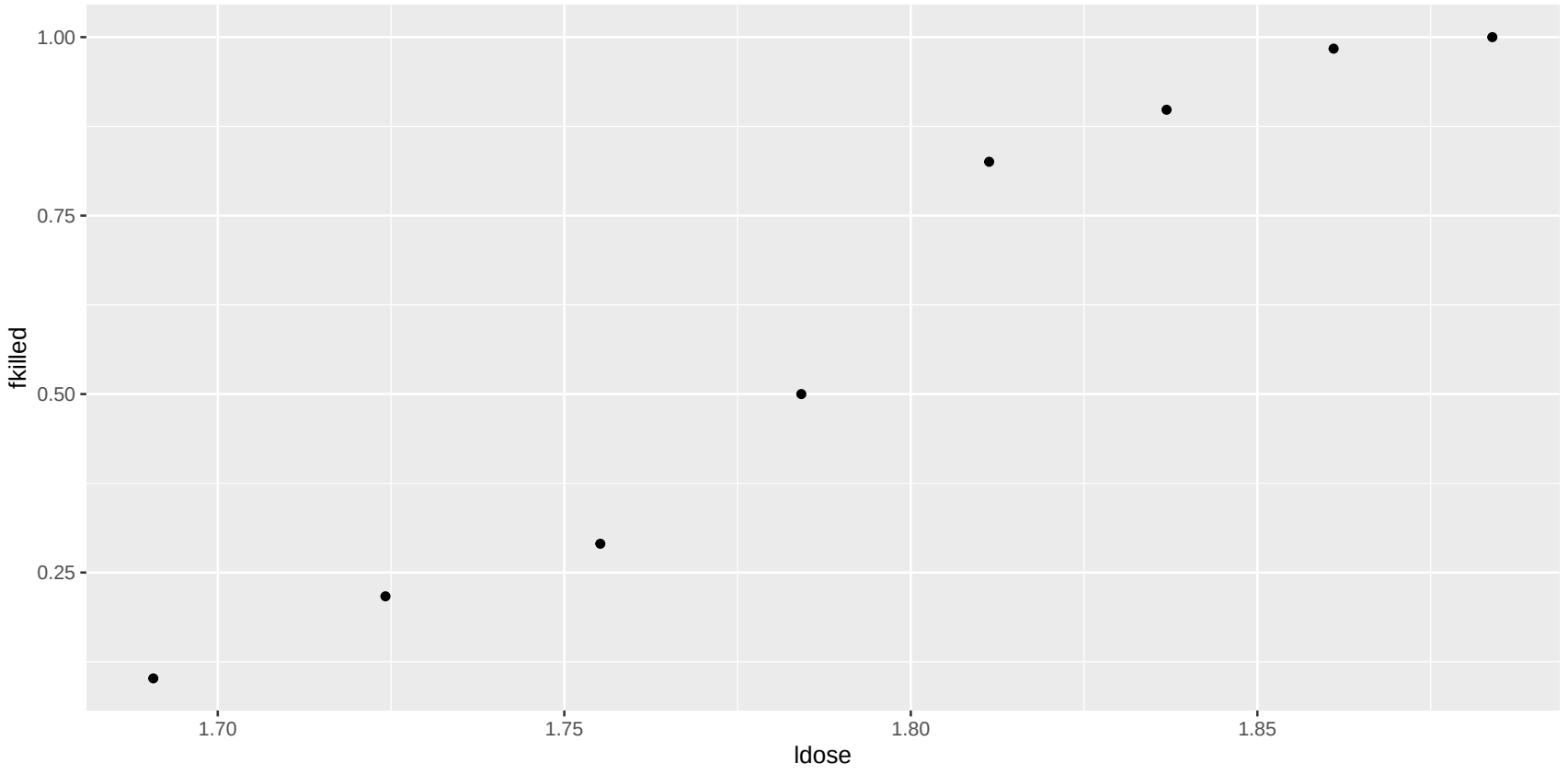
A total of 481 beetles were exposed to 8 different concentration of CS_2 (data on log10-dose).
For each beetle is was recorded if the beetle was alive or killed at the given concentration.
Data for beetle i: $Y_i = 0$ if beetle i was alive and $Y_i = 1$ if it was killed, and x_i is then the log10-dose beetle i was given.

The table below shows the 8 values of the log10-dose against the number of beetles alive and killed. The plot shows log10-dose on the horizontal axis and fraction of beetles killed (killed/total) for each log10-dose.

```
library(investr)
# from aggregated to individual data (because these data w
ldose=rep(round(beetle$ldose, 2), beetle$n)
y=NULL; for (i in 1:8) y=c(y,rep(0,beetle$n[i]-beetle$y[i])
beetleds=data.frame("killed"=y,"ldose"=ldose)
knitr::kable(table(beetleds), digits = 2)
```

	1.69	1.72	1.76	1.78	1.81	1.84	1.86	1.88
0	53	47	44	28	11	6	1	0
1	6	13	18	28	52	53	61	60

```r
# plot from aggregated data
frac=beetle$y/beetle$n
dss=data.frame(fkilled=frac,ldose=beetle$ldose)
ggplot(dss,aes(ldose,fkilled))+
  geom_point()
```



Q:

a. What might be the effect (mathematical function) of the log10-dose on the probability of killing a beetle?
b. How can this curve be part of a regression model?

How to model a binary response?

In multiple linear regression we have

1. Random component: Distribution of response: $Y_i \sim N(\mu_i, \sigma^2)$, where μ_i is *parameter of interest* and σ^2 is *nuisance*.
2. Systematic component: Linear predictor: $\eta_i = \mathbf{x}_i^T \beta$. Here $\mathbf{x}_i$ is our fixed (not random) p-dimensional column vector of covariates (intercept included).
3. Link: Connection between the linear predictor and the mean (parameter of interest): $\mu_i = \eta_i$.

- It would not make sense to fit the continuous linear regression to Y_i when $Y_i = \{0, 1\}$ - since Y_i is not a continuous random variable, and Y_i is not normal.
- So, we need to change 1. We keep 2. And, we make 3. more general.

Binary regression

1. $Y_i \sim \text{bin}(n_i, \pi_i)$.

First look at $n_i = 1$ (i.e. a Bernoulli distribution).
Our parameter of interest is π_i which is the mean $\text{E}(Y_i) = \mu_i = \pi_i$.
For a generalized linear model (GLM) we require that the distribution of the response is an exponential family. We have seen in M1 that the binomial distribution is an exponential family.

Linear Predictor

$\eta_i = \mathbf{x}_i^T \beta.$

Link Function

3. Relationships between the mean $\mu_i = \pi_i$ and the linear predictor η_i:

$$g(\mu_i) = \eta_i$$

and the inverse of the link function, called the *response function*, and denoted by

$$h(\eta_i) = g^{-1}(\eta_i) = \mu_i$$

We thus also have to require that the link function is monotone, and we will soon see that we also need to require that it is twice differential.

Response function for binary regression

Based on selecting a cumulative distribution function (cdf) as the response function.

The cdf will always be within [0,1], and the cdf is monotone - which will help us to interpret results.

The most popular response functions are:

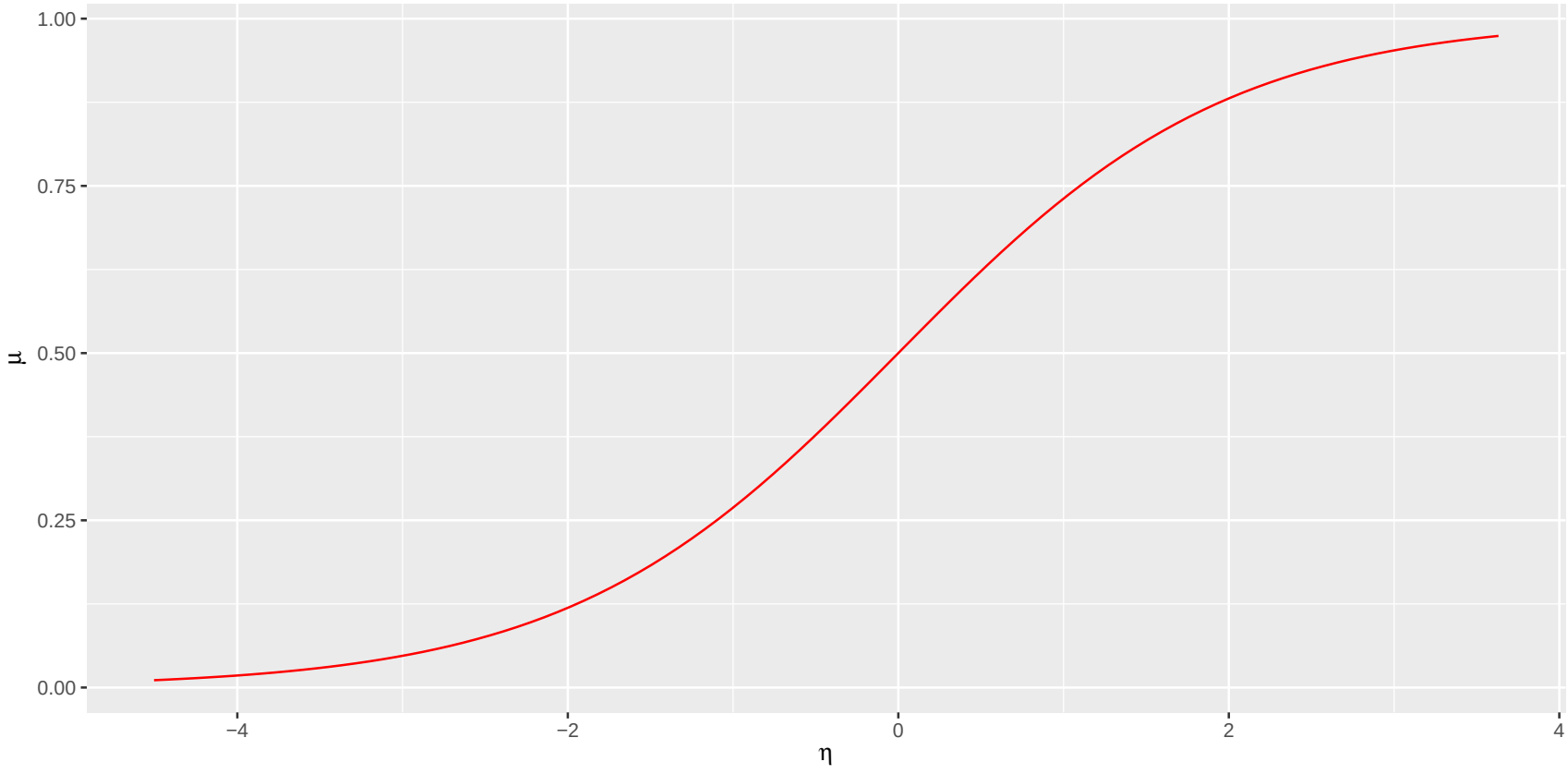
- *logistic cdf* (with corresponding *logit* link function) referred to as the *logit model*,
- *normal cdf* - (with corresponding *probit* link function) referred to as the *probit model* ,
- the *extreme minimum-value cdf* (with corresponding *complementary log-log* link function) referred to as the *complementary log-log model*.

In this module we focus on the logit model.

The logit model aka logistic regression

In the beetle example we have a simple linear predictor:
$\eta_i = \beta_0 + \beta_1 x_i$ where x_i is the log10-dose for beetle i.
Assume that β_0 = -60.1 and β_1 = 33.9. (These values are estimates from our data, and we will see later how to find these estimates using maximum likelihood estimation.)

Below the response function is plotted for η_i =-60.1+33.9x_i.



Q: Explain to your neighbour what is on the x- and y-axis of this plot. Where are the observed log10-doses in this graph?

Link and reponse function

The logit model is based on the logistic cdf as the response function, given as

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

or alternatively as the link function (the inverse of the response function)

$$g(\mu_i) = h^{-1}(\mu_i) = \ln(\frac{\mu_i}{1 - \mu_i}) = \ln(\frac{\pi_i}{1 - \pi_i})$$

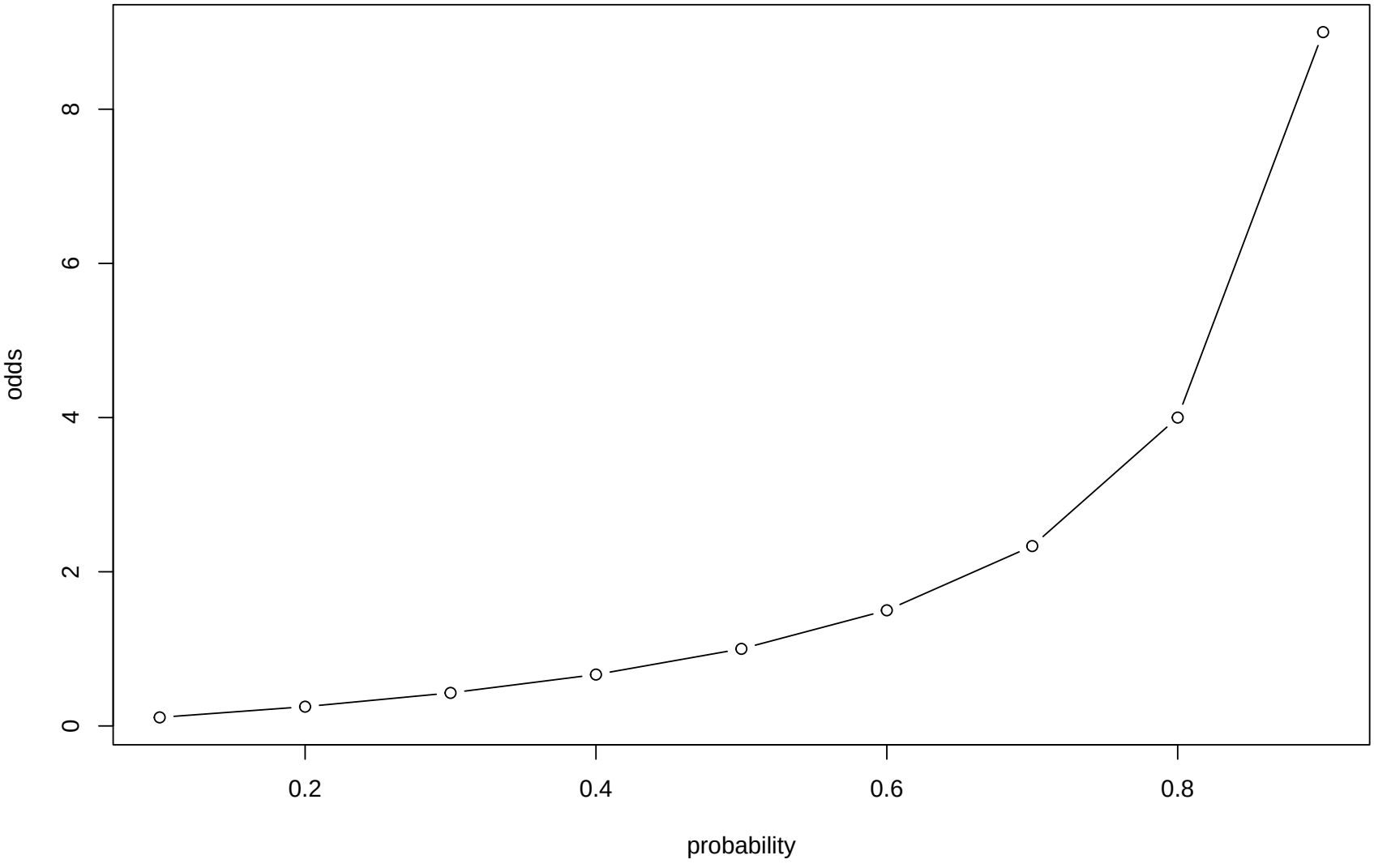
Hands-on: show this for yourself.

Interpreting the logit model

If the value of the linear predictor η_i changes to $\eta_i + 1$ the probability π increases non-linearly from $\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ to $\frac{\exp(\eta_i+1)}{1+\exp(\eta_i+1)}$, as shown in the graph above.

Before we go further: do you know about the odds? The ratio $\frac{P(Y_i=1)}{P(Y_i=0)} = \frac{\pi_i}{1-\pi_1}$ is called the *odds*. If $\pi_i = \frac{1}{2}$ then the odds is 1, and if $\pi_i = \frac{1}{4}$ then the odds is $\frac{1}{3}$. We may make a table for probability vs. odds in R:

pivec	0.10	0.20	0.30	0.40	0.5	0.6	0.70	0.8	0.9
odds	0.11	0.25	0.43	0.67	1.0	1.5	2.33	4.0	9.0



Odds may be seen to be a better scale than probability to represent chance, and is used in betting. In addition, odds are unbounded above.

We look at the link function (inverse of the response function). Let us assume that our linear predictor has k covariates present

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

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

$$\eta_i = \ln(\frac{\pi_i}{1 - \pi_i})$$

$$\ln(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdots \exp(\beta_k x_{ik})$$

We have a *multiplicative model* for the odds.

So, what if we increase x_{1i} to $x_{1i} + 1$?

If the covariate x_{1i} increases by one unit (while all other covariates are kept fixed) then the odds is multiplied by $\exp(\beta_1)$:

$$\begin{aligned}
\frac{P(Y_i = 1 \mid x_{i1} + 1)}{P(Y_i = 0) \mid x_{i1} + 1)} &= \exp(\beta_0) \cdot \exp(\beta_1(x_{i1} + 1)) \cdots \exp(\beta_k x_{ik}) \\
&= \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \exp(\beta_1) \cdots \exp(\beta_k x_{ik}) \\
&= \frac{P(Y_i = 1 \mid x_{i1})}{P(Y_i = 0 \mid x_{i1})} \cdot \exp(\beta_1)
\end{aligned}$$

This means that if x_{i1} increases by 1 then: if $\beta_1 < 0$ we get a decrease in the odds, if $\beta_1 = 0$ no change, and if $\beta_1 > 0$ we have an increase. In the logit model $\exp(\beta_1)$ can be easier to interpret than β_1.

To Sum Up

For the linear predictor we interpret effects in the same way as for the linear model (in Module 2), then we transform this linear effect in η into a nonlinear effect for $\pi = \frac{\exp(\eta)}{1+\exp(\eta)}$, and use the odds to interpret changes.

Q: If x_{i1} increases by 1 ANDβ_1 is small, what is the relationship between the change in the odds, the change in the log odds and the change in the probability?

Infant respiratory disease: interpretating parameter estimates

This example is taken from Faraway (2006): "Extending the linear model with R"

We select a sample of newborn babies (girls and boys) where the parents had decided on the method of feeding (bottle, breast, breast with some supplement), and then monitored the babies during their first year to see if they developed infant respiratory disease (the event we want to model).

We fit a logistic regression to the data, and focus on the parameter estimates.

```r
library(faraway)
data(babyfood)
# babyfood
xtabs(disease/(disease+nondisease)~sex+food,babyfood)
```

```
##        food
## sex          Bottle     Breast      Suppl
##   Boy  0.16812227 0.09514170 0.12925170
##   Girl 0.12500000 0.06681034 0.12598425
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.61	0.11	-14.35	0.00
sexGirl	-0.31	0.14	-2.22	0.03
foodBreast	-0.67	0.15	-4.37	0.00
foodSuppl	-0.17	0.21	-0.84	0.40

Questions

Observe that the two factors by default is coded with dummy variable coding, and that `sexBoy` is the reference category for sex and `foodBottle` the reference category for feeding method.

1: Explain how to interpret the `Estimate` for `sexGirl`, `foodBreast` and `foodSuppl`.

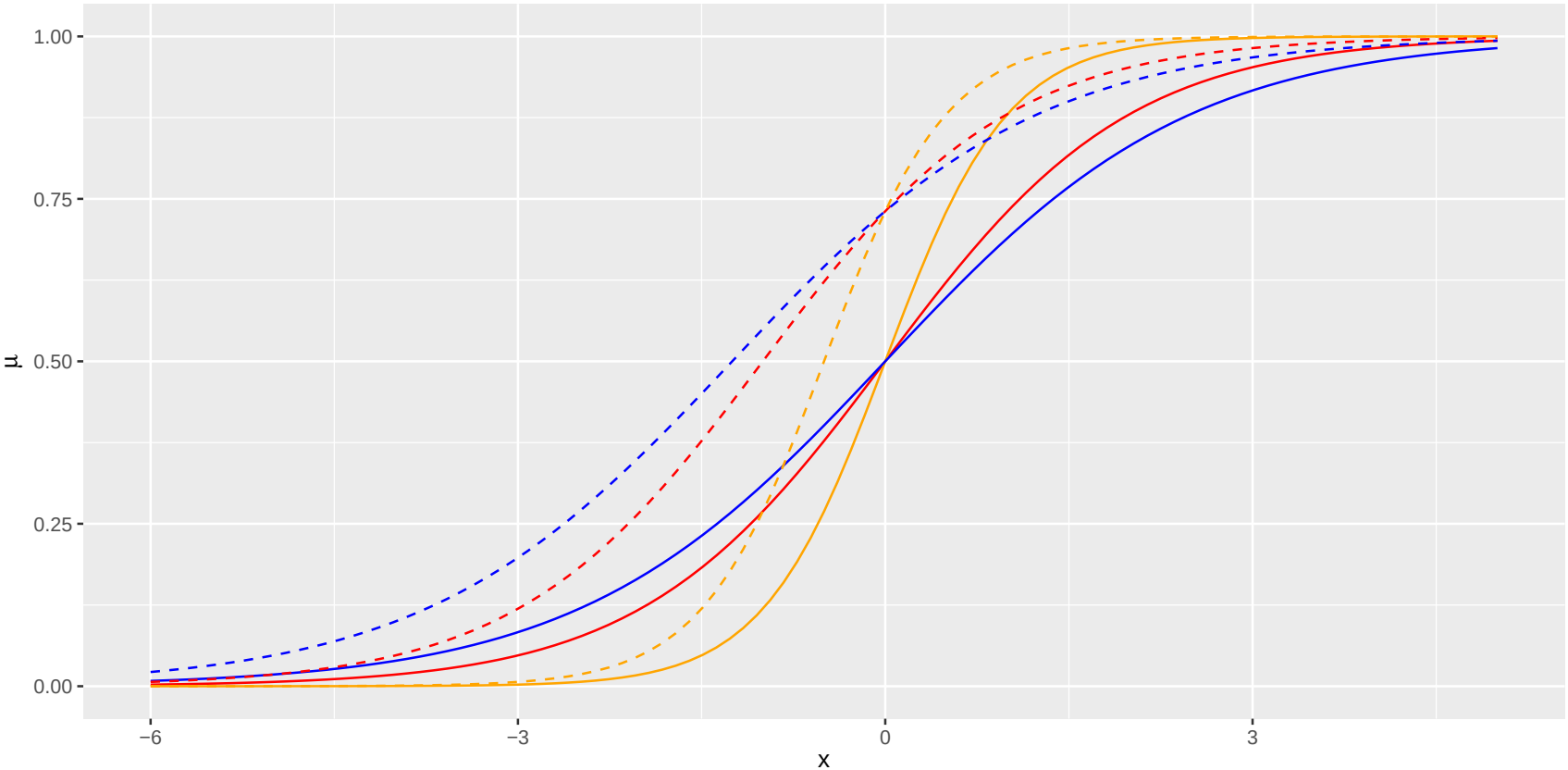
2: What are the 6 values given by the call to `predict`? What is the least favourable combination of sex and method of feeding? And the most favourable?

```
print(predict(fit,type="response"), digits=2) #gives predi
```

```
##     1     2     3     4     5     6
## 0.166 0.144 0.093 0.127 0.109 0.069
```

More response function plots for the logit model

The response function as a function of the covariate x and not of η. Solid lines: $\beta_0 = 0$ and β_1 is 0.8 (blue), 1 (red) and 2 (orange), and dashed lines with $\beta_0 = 1$.



Grouped vs. individual data

So far we have only mentioned individual data.

However, in both the examples we have looked at some covariate vectors are *identical* (rows in the design matrix are identical). We call these unique combinations of covariates *covariate patterns*, and say we have *grouped* data.

disease	nondisease	sex	food
77	381	Boy	Bottle
19	128	Boy	Suppl
47	447	Boy	Breast
48	336	Girl	Bottle
16	111	Girl	Suppl
31	433	Girl	Breast

Here we have 6 groups of covariate patterns. The first group has covariates `Boy` and `Bottle`, there are 77+381= 458 babies with this combination and 77 of these got the disease.

We prefer to group data if possible. Grouping is good because then data can be kept in a condensed form, it will speed up computations and makes model diagnosis easier (than for individual data).

For the grouped data we still have a binomial distribution, and possible generalization is to let

- $n_j\bar{Y}_j$ be the number of successes in group j,
- which means that $\bar{Y}_j = \frac{1}{n_j}\sum Y_i$ where the sum is over all i in group j.

Further

$$n_j\bar{Y}_j \sim \text{bin}(n_j, \pi_j)$$

such that $\text{E}(n_j\bar{Y}_j) = n_j\pi_j$ and $\text{Var}(n_j\bar{Y}_j) = n_j\pi_j(1-\pi_j)$, and $\text{E}(\bar{Y}_j) = \pi_j$ and $\text{Var}(\bar{Y}_j) = \frac{1}{n_j}\pi_j(1-\pi_j)$

We then keep the linear predictor, and the link function is still $\eta_j = \ln(\frac{\pi_j}{1-\pi_j})$. That is, we do not model the mean $n_j\pi_j$ but π_j directly.

Likelihood and derivations thereof

Our parameter of interest is the vector β of regression coefficients, and we have no nuicance parameters (because the variance is related directly to the π_j and n_j is known).

We would like to estimate β from maximizing the likelihood, but we will soon see that we have no closed form solution. First we look at the likelihood, the log-likelihood and first and second derivatives thereof.

For simplicity we do the derivations for the case where $n_i = 1$, but then include the results for the case where we have G covariate patterns with n_j observations of each pattern.

Assumptions

1. $Y_i \sim \text{bin}(n_i = 1, \pi_i)$, and $\text{E}(Y_i) = \mu_i = \pi_i$, and $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$.
2. Linear predictor: $\eta_i = \mathbf{x}_i^T \beta$.
3. Logit link

$$\eta_i = \ln(\frac{\pi_i}{1 - \pi_i}) = g(\mu_i)$$

and (inverse thereof) logistic response function

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

We will also need:

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

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.

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} f(y_i; \beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

Loglikelihood $l(\beta)$

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \ln L_i(\beta) = \sum_{i=1}^{n} l_i(\beta) \tag{1}$$

$$= \sum_{i=1}^{n} [y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)] \tag{2}$$

$$= \sum_{i=1}^{n} [y_i \ln(\frac{\pi_i}{1 - \pi_i}) + \ln(1 - \pi_i)] \tag{3}$$

Observe that the log-likelihood is a sum of invidual contributions for each observation pair i.

The log-likelihood is now expressed as a function of π_i, but we want to make this a function of β and the connection between π_i and β goes through η_i. We have that $\pi = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ and in our log-likelihood we need

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

and

$$\ln(\frac{\pi_i}{1-\pi_1}) = \eta_i$$

(the last is our logit link function).

Then we get:

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

which is now our function of η_i.

Finally, since $\eta_i = \mathbf{x}_i^T \beta$,

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

Q: What does the graph of l look like as a function of β?

If we look at the beetle example we only have one covariate (in addition to the intercept) - so this means that we have $\beta = (\beta_0, \beta_1)$. Plotting the log-likelihood (for the beetle data set) will be one of the tasks for the interactive lecture.

But, next we take partial derivatives, and then we will (instead of using this formula) look at $l_i(\beta) = l_i(\eta_i(\beta))$ and use the chain rule.

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.
Solving $s(\beta) = 0$ wil give us our MLEs

We will need the following:

Chain rule: $\frac{df(u(x))}{du} = \frac{df}{du} \cdot \frac{du}{dx}$,

Product rule: $(u \cdot v)' = u' \cdot v + u \cdot v'$,

Fraction rule: $(\frac{u}{v})' = \frac{u' \cdot v - u \cdot v'}{v^2}$,

$\frac{d \ln(x)}{dx} = \frac{1}{x}$, $\frac{d \exp(x)}{dx} = \exp(x)$ and $\frac{d(\frac{1}{x})}{dx} = -\frac{1}{x^2}$.

Partial derivatives of scalar wrt a vector $\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{b}} = \mathbf{a}$

and later we will also need $\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{b}^T} = (\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{b}})^T = \mathbf{a}^T$.

Here we go:

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

Again, observe that the score function is a sum of individual contributions for each observation pair i.

We will use the chain rule to calculate $s_i(\beta)$.

$$s_i(\beta) = \frac{\partial l_i(\beta)}{\partial \beta} = \frac{\partial l_i(\beta)}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta} = \frac{\partial [y_i \eta_i - \ln(1 + \exp(\eta_i))]}{\partial \eta_i} \cdot \frac{\partial [\mathbf{x}_i^T \beta]}{\partial \beta}$$

$$s_i(\beta) = (y_i - \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}) \cdot \mathbf{x}_i = (y_i - \pi_i)\mathbf{x}_i$$

The score function is given as:

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

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

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

Next week we will see how we can do that using the Newton-Raphson or Fisher Scoring iterative methods, but first we will work on finding the mean and covariance matrix of the score vector - and the derivatives of the score vector (the Hessian, which is minus the observed Fisher matrix).

Remark: in Module 5 we will see that the general formula for GLMs is:

$$s(\beta) = \sum_{i=1}^{n} [\frac{y_i - \mu_i}{\text{Var}(Y_i)} \mathbf{x}_i \frac{\partial \mu_i}{\partial \eta_i}] = \sum_{i=1}^{n} [\frac{y_i - \mu_i}{\text{Var}(Y_i)} \mathbf{x}_i h'(\eta_i)] = \mathbf{X}^T \mathbf{D} \Sigma^{-1} (\mathbf{y} - \mu)$$

where $\mathbf{X}$ is the $n \times p$ design matrix,
$\mathbf{D} = \text{diag}(h'(\eta_1), h'(\eta_2), \ldots, h'(\eta_n))$ is a diagonal matrix with the derivatives of the response function evaluated at each observation. Further, $\Sigma = \text{diag}(\text{Var}(Y_1), \text{Var}(Y_2), \ldots, \text{Var}(Y_n))$ is a diagonal matrix with the variance for each response, and $\mathbf{y}$ is the observed $n \times 1$ vector of responses and μ is the $n \times 1$ vector of individual expectations $\mu_i = \text{E}(Y_i) = h(\eta_i)$.