Generalized linear models

January 5, 2017

Generalized linear models encompass a large class of models which can be applied in situations where the response variable is not normally distributed as assumed in linear models (including multiple regression, analysis of variance, and analysis of covariance).

1 Binomial response

Suppose that the response variable is binomially distribud, that is, each observation Y is based on say, n independent Bernoulli trials, and the probability of a particular event has probability p for all trials on which observation Y is based. We want to model how the probability p depends on explanatory variables of interesting, either numerical variables or factors.

Rather than working with the probability p,

$$0 \leq p \leq 1, \tag{1}$$

we may work with the odds of the event $p/(1-p)$. For example, if $p = 0.9$ the corresponding odds is 9, a probability $p = 1/2$ gives an odds of 1, and if $p = 0.1$ the odds is 0.11. In general, the odds of an event will be a non-negative quantity with no upper bound, that is, $0 \leq p/(1-p) \leq +\infty$. Taking the log of the odds $\ln p/(1-p)$ known as the logit transformed probability $\operatorname{logit} p$, we thus get a quantity which can take any positive or negative real value,

$$-\infty \leq \operatorname{logit} p \leq +\infty \tag{2}$$

We then assume that

$$\operatorname{logit} p = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\eta} \tag{3}$$

where the right hands side η is the so called linear predictor of the model. Solving for p we see that this ensures that

$$p = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}} \tag{4}$$

always takes a meaningful value between 0 and 1 without any additional constraints on the parameters.

In the same way as for linear models, the linear predictor may involve factors encoded using dummy variables.

To make the notation less cumbersome, it should be noted that we have omitted indices i indicating different observations on Y, p, n and the explanatory variables $x_1, x_2, \ldots, x_k$ above.

2 Deviance

Suppose that we have data of the form given in Table 1. The likelihood function for a generalized linear model for these data assuming that Y_i is binomially distributed with parameters n_i and

y_i	n_i	x_i
5	10	-.2
2	5	-.1
7	20	0
9	10	.1
8	12	.2

Table 1: Example data set

p_i where $\operatorname{logit} p_i = \beta_0 + \beta_1 x_i$ becomes

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \\ &= \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \right)^{y_i} \left(1 - \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \right)^{n_i - y_i}, \end{aligned} \tag{5}$$

that is, a function of the observed data and the unknown parameters β_0 and β_1.

In gereral, the parameters of a generalized linear model are estimated by maximising the likelihood function (using a numerical method known as iterated reweighted least squares, IRLS). As we add more explantory variables to a model, the observed maximum likelihood will always increase until the number of parameters p (regression coefficients) equals the number of observations n.

Based on the observed maximum log likelihood of a fitted model we define the deviance D as

$$D = 2(\ln L_{\text{saturated}} - \ln L) \tag{6}$$

where $\ln L$ is the maximum log likelihood of the fitted model and $\ln L_{\text{full}}$ is the maximum log likelihood of the so called saturated model, that is, a model having as many paramters p as there are observations n. The deviance measures how well a given model fits the data and plays a role similar to the residual sum of squares of a linear model.

2.1 Testing goodness-of-fit

In contrast to linear models in which the response is normal with an uknown variance σ^2 which must be estimated from the data, the variance of a binomially distributed response variable is given, once the mean is specified. This makes it possible to assess the goodness-of-fit of a given model.

If a given model H_0 fitted to the data is true, it follows that the deviance D is approximately chi-square distributed with $n - p$ degrees of freedom. If the observed deviance is sufficiently large we may reject the this null hypothesis and conclude that the model does not fit the data.

Consider the following example for Dalgaard, p. 232.

```
> no.yes <- c("No","Yes")
> smoking <- gl(2,1,8,no.yes)
> obesity <- gl(2,2,8,no.yes)
> snoring <- gl(2,4,8,no.yes)
> n.tot <- c(60,17,8,2,187,85,51,23)
> n.hyp <- c(5,2,1,0,35,13,15,8)
> data.frame(smoking,obesity,snoring,n.tot,n.hyp)
  smoking obesity snoring n.tot n.hyp
1      No      No      No    60     5
2     Yes      No      No    17     2
```

```
3      No     Yes      No     8     1
4     Yes     Yes      No     2     0
5      No      No     Yes   187    35
6     Yes      No     Yes    85    13
7      No     Yes     Yes    51    15
8     Yes     Yes     Yes    23     8
> prop <- n.hyp/n.tot
> hyp.mod <- glm(prop ~ smoking + obesity + snoring, fam=binomial(), weight=n.tot)
> summary(hyp.mod)

Call:
glm(formula = prop ~ smoking + obesity + snoring, family = binomial(),
    weights = n.tot)

Deviance Residuals:
       1         2         3         4         5         6         7         8
-0.04344   0.54145  -0.25476  -0.80051   0.19759  -0.46602  -0.21262   0.56231

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537

Number of Fisher Scoring iterations: 4
```

The deviance D of this model (the "Residual deviance") is 1.61 with $n - p = 8 - 4 = 4$ degrees of freedom. In this case the observed deviance is well below it's expected value of 4 under H_0 as well as the upper 0.05-quantile of the chi-square distribution (the critical value)

```
> qchisq(0.05,df=4,lower.tail=F)
[1] 9.487729
```

so in this case we can not reject the null hypothesis, that is, we have no evidence that the fitted model is wrong. The p-value for the test of goodness-of-fit becomes

```
> pchisq(1.61,df=4,lower.tail=F)
[1] 0.8069937
```

2.2 Tests between alternative models

In addition to testing the goodness-of-fit of any given model, the deviance is used in tests between different nested alternatives. Just like the residual sum of squares, the deviance always decreases as we add more terms to a model. Suppose the H_1 is an extension of some model H_0

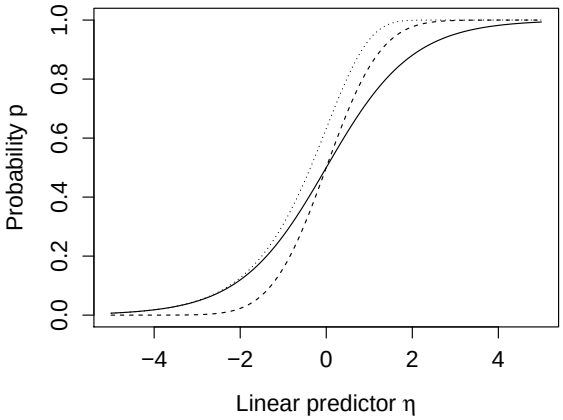

Figure 1: Relationship between the probability p of the modelled event and the linear predictor η for the logit (solid line), probit (dashed line) and cloglog (dotted line) choice of link function.

obtained by adding an extra explanatory variable and let p_1 and p_0 be the number of parameters estimated under H_1 and H_0. Then the change in deviance

$$D_0 - D_1 \tag{7}$$

is approximately chi-square with $p_1 - p_0$ degrees of freedom under H_0. If the observed change in deviance is sufficiently large we reject H_0 in favour of H_1.

Tests of this kind can be obtained as follows

```
> drop1(hyp.mod,test="Chisq")
Single term deletions

Model:
prop ~ smoking + obesity + snoring
        Df Deviance    AIC    LRT Pr(Chi)
<none>       1.6184 34.537
smoking  1   1.6781 32.597 0.0597 0.80694
obesity  1   7.2750 38.194 5.6566 0.01739 *
snoring  1   7.2963 38.215 5.6779 0.01718 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For factors with only two levels, the same hypothesis is being tested here as in the output from `summary( )`; the difference between the P values are a result of the somewhat different approximations used. For factors with more than two levels, we always need to used `drop1( )`.

3 Other link functions for binomial data

The logit link function used above has the advantage that the parameters of the models can be easily interpreted in terms of oddsratios. However, depending on the context, other link functions (see Fig. 1) may be prefereable.

3.1 Probit link

Consider the logistic regression on p. 240 where the probability that girls between the age of 8 and 20 years has had their first menstruation is modelled as a function of age x using the logistic regression model

$$\operatorname{logit} p = \beta_0 + \beta_1 x. \tag{8}$$

This model can be used to estimate the mean age at first menstuation $x_0 = -\beta_0/\beta_1$. Clearly, the age at which the first menstruation occurs must have some distribution in the population and and from the slope of the relationship between p and x it should be possible to estimate the standard deviation of this distribution.

Suppose that the time T at which the first menstruation occurs has a normal distribution with mean μ and standard deviation σ. Then, at a given age x, the probability p that the first menstruation has already occured is

$$p = P(T \leq x) \tag{9}$$

Now, since $T \sim N(\mu, \sigma^2)$ it follows that $(T - \mu)/\sigma$ has a standard normal distribution and we can rewrite (9) as

$$p = P\left(\frac{T-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \phi\left(\frac{x-\mu}{\sigma}\right) \tag{10}$$

where ϕ is the cumulative standard normal density (denoted G in Løvås). This equation, specifying an alternative relationship between p and age x to (8), can alternatively be written as

$$\phi^{-1}(p) = \frac{x-\mu}{\sigma} \tag{11}$$

or, after reparameterization, as

$$\operatorname{probit} p = \beta_0 + \beta_1 x \tag{12}$$

where the probit link function is the inverse of the cumulative standard normal density ϕ and the new parameters, the regression coefficients

$$\beta_0 = -\frac{\mu}{\sigma}, \qquad \beta_1 = \frac{1}{\sigma}. \tag{13}$$

This model can be fitted in R by specifying the alternative probit link function as follows.

```
> menmod <- glm(menarche ~ age,binomial(link="probit"))
> summary(menmod)

Call:
glm(formula = menarche ~ age, family = binomial(link = "probit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.32986  -0.15223   0.00028   0.07228   2.48281

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.37033    1.06346  -10.69   <2e-16 ***
age           0.86233    0.08106   10.64   <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: 719.39  on 518  degrees of freedom
Residual deviance: 197.39  on 517  degrees of freedom
AIC: 201.39

Number of Fisher Scoring iterations: 8
```

Interestingly, the new choice of link function gives a better fitted indicated by the slightly smaller residual deviance. Having estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, estimates of μ and σ are found by solving (13) with respect to μ and σ which yields the estimators

$$\hat{\mu} = -\frac{\hat{\beta}_0}{\hat{\beta}_1}, \qquad \hat{\sigma} = \frac{1}{\hat{\beta}_1}. \tag{14}$$

Thus, the estimated standard deviation of the age at first menstruation is $\hat{\sigma} = 1/0.8623 = 1.15$ years. Standard errors of $\hat{\mu}$ and $\hat{\sigma}$ can be computed using the method in handout 3, see assignment 7.

3.2 Complementary log-log link

Suppose that we observe if given individuals have died during given time intervals of different lengths t and we want to model how the probability of dying depends on explanatory variables $x_1, x_2, \ldots, x_k$ of interest as well as the length t of the time intervals. If the rate of mortality λ of a given individual is constant with respect to time, then the life span of a given indidivual T will follow an exponential distribution with parameter λ and the probability that an individual has died is

$$p = P(T \leq t) = 1 - e^{-\lambda t}. \tag{15}$$

Our interest is in how the explanatory variables affects the rate of mortality λ. The rate of mortality of a given individual is necessarily non-negative, that is,

$$0 \leq \lambda \leq +\infty. \tag{16}$$

This suggest that it is reasonable to assume that the $\ln \lambda$ depends linearly on the linear predictor η involving the different explanatory variables, that is,

$$\ln \lambda = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\eta}. \tag{17}$$

This assumption implies that the different covariates have a multiplicative effect on the rate of mortality λ since

$$\lambda = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}. \tag{18}$$

For example, the rate of mortality may may be 20% higher for one of the sexes.

Subtitution of (18) into (15) the relationship between p and the linear predictor η becomes

$$p = 1 - e^{-e^{\beta_0 + \cdots + \beta_k x_k} t} \tag{19}$$

which can be rewritten as

$$\operatorname{cloglog} p = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \ln t \tag{20}$$

where

$$\operatorname{cloglog} p = \ln(-\ln(1-p)), \tag{21}$$

is the so called complementary log log link function (see Fig. 1).

In contrast to the probit and logit link functions, because the cloglog link function is not symmetric, the model is changed and not only reparameterized if we choose to model the probability of survival rather than death.

The final term on the right hand side of (20), the log of the length of the time intervals, is a so called offset variable in the model which accounts for how the probability p necessarily increases in a certain way with the length t of the time interval. For instance, for small p, $\operatorname{cloglog} p \approx \ln p$ which implies that probability of death p as it should becomes directly proportional to the length of the the time interval t. Offset terms can be thought of as explanatory variables for which the regression coefficient is known to be exactly one a priori. To include an offset term in a model use the additional `offset` argument in the call to the `glm` function when fitting the model in R.

Suppose that we mark 10 female and 10 male newborn roedeer fawns with radio collars and that we relocate each individual after either $t = 7$ og $t = 14$ days by means of radio tracking. At this point we determine if each individual has been lost to fox predation (indicated by a response $y = 1$) or if each individual is still alive (indicated by the response $y = 0$). This data can then by represented by the following data frame

```
> fawndata
   y    sex  t
1  1   male  7
2  0   male  7
3  0   male  7
4  0   male  7
5  1   male  7
6  1   male 14
7  1   male 14
8  1   male 14
9  1   male 14
10 1   male 14
11 1 female  7
12 0 female  7
13 0 female  7
14 0 female  7
15 0 female  7
16 0 female 14
17 0 female 14
18 1 female 14
19 0 female 14
20 1 female 14
```

A model based on the cloglog link using sex as an explantory categorical variable (a factor) and using $\ln t$ as an offset can now be fitted as follows.

```
> fawnmod <- glm(y~sex,family=binomial(link="cloglog"),offset=log(t))
> summary(fawnmod)

Call:
glm(formula = y ~ sex, family = binomial(link = "cloglog"), offset = log(t))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.38076  -0.98701  -0.06535   0.67213   1.75029
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9938     0.4075  -4.893 9.94e-07 ***
sexfemale    -1.3646     0.7094  -1.924   0.0544 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.017  on 19  degrees of freedom
Residual deviance: 21.022  on 18  degrees of freedom
AIC: 25.022

Number of Fisher Scoring iterations: 5
```

With a linear predictor involving one factor and one offset variable, the model can now be written in mathematical notation as

$$\operatorname{cloglog} p = \mu + \alpha_i + \ln t \tag{22}$$

where μ is the "intercept" of the model and α_i is the effect of the i'th level of factor i. The corresponding relationship between the mortality rate λ and the linear predictor is

$$\lambda = e^{\mu+\alpha_i}. \tag{23}$$

Since we are comparing only two groups only two parameters can be estimated. By default, α_1 is set equal to zero. The estimate of the mortality rate for the first sex (males) is thus only

$$e^{\hat{\mu}} = e^{-1.99} = 0.136 \tag{24}$$

(per day) whereas for the second sex (females), the rate of mortality is

$$e^{\hat{\mu}+\hat{\alpha}_2} = e^{-1.99-1.36} = 0.0351, \tag{25}$$

that is, a reduced mortality rate by a factor of $e^{-1.36} = 0.257$ relative to the first sex (males).

Thus, if we somewhat naively assume the rate of mortality remains constant also beyond the length of the time interval of 14 days, the expected lifespans become 7.31 and 28.5 days for males and females, respectively. For a real world example of this form of data see Aanes & Andersen (1996).

4 Infinite parameter estimates as a result of linear separation

The following example illustrates a quite common phenomena when the response is binomial and we have only a moderate amount of data and many possible explanatory variables.

Suppose we have the following simple data set.

```
x <- 1:10
y <- c(0,0,0,0,1,1,1,1,1,1)
plot(x,y)
```

If we fit a logistic regression to these data, we get the following warning messages.

```
> mod <- glm(y~x,fam=binomial)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

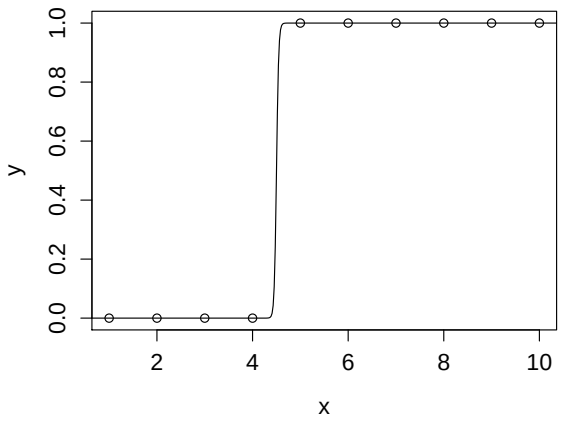

Figure 2: Linear separation in a model with one numerical explanatory variable.

The summary of the model gives us the following parameter estimates.

```
> summary(mod)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -200.37  265802.23  -0.001        1
x               44.52   58511.58   0.001        1
    Null deviance: 1.3460e+01  on 9  degrees of freedom
Residual deviance: 8.6042e-10  on 8  degrees of freedom
AIC: 4
```

The estimate of the regression coefficient for for the effect of the explanatory x and in particular its standard error and associated tests looks suspect. Fig. 2 shows the predicted values based on the fitted model together with the observed data.

```
xx <- seq(0,1,len=100)
pp <- predict(mod,newdata=data.frame(x=xx),type="response")
lines(xx,pp)
```

This shows that the estimated model fits the data very well, in fact, almost perfectly as indicated by a residual deviance which is almost exactly zero.

This phenomena arise because the response y is always 1 for all values of x above a certain threshold and otherwise always zero. This implies that the likelihood (the probability of the observed y's) is maximised when slope and intercept of the model goes to infinity and minus infinity in such a way that the resulting predicted values coinsides with the observations.

The summary, however, suggests that the effect of the x on y is non-significant. This test, however, is based on the estimated standard error of the slope which is computed from the curvature of the log-likelihood surface at the maximum likelihood. Since the the log-likelihood in this cases essentially takes the form of a flat ridge extending into infinity (Fig. 3), the standard error is greatly inflated and the test breaks down. If we instead tests $H_0 : \text{logit}\, p = \beta_0$ against $H_1 : \text{logit}\, p = \beta_0 + \beta_1 x$ using `drop1` we get a highly significant test result

```
> drop1(mod,test="Chisq")
Single term deletions

Model:
```

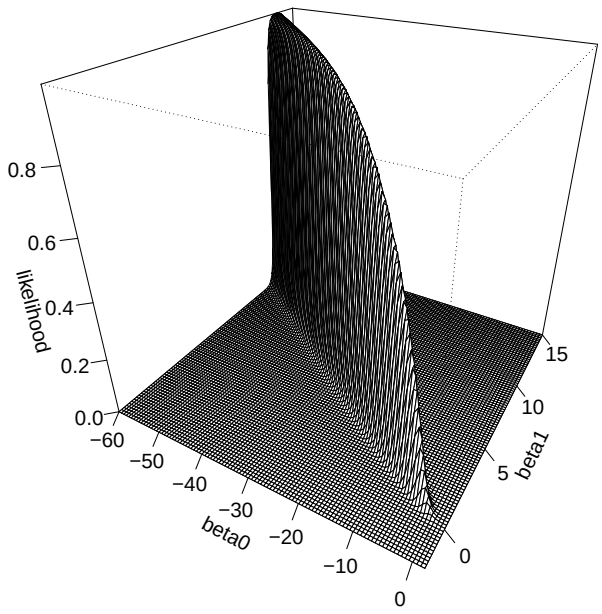

Figure 3: A plot of the likelihood function for the logistic regression model for data with linear separation

```
y ~ x
       Df Deviance    AIC    LRT   Pr(Chi)
<none>       0.000  4.000
x       1   13.460 15.460 13.460 0.0002437 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This test rests on the fact that that the change in deviance is approximately chi-square distributed under the null hypothesis. The observed data, strongly supporting H_1, does not invalidate this approximation for the distribution of the change in deviance under H_0.

The above phenomena can arise in more complex models with many explanatory variables. Consider the following data set.

```
> x1 <- rep(1:10,2)
> x2 <- rep(c(0,1),c(10,10))
> y <- c(1,1,1,0,0,0,0,0,0,0,
+        1,1,1,1,1,1,0,0,0,0)
> data.frame(x1,x2,y)
   x1 x2 y
1   1  0 1
2   2  0 1
3   3  0 1
4   4  0 0
5   5  0 0
6   6  0 0
7   7  0 0
```

```
8   8  0 0
9   9  0 0
10 10  0 0
11  1  1 1
12  2  1 1
13  3  1 1
14  4  1 1
15  5  1 1
16  6  1 1
17  7  1 0
18  8  1 0
19  9  1 0
20 10  1 0
```

Here x_2 can represent an dummy variable encoding of a factor with two levels. If we first fit the model using either x_1 or x_2 as the only explanatory variable nothing exceptional happens. However, if including the both x_1 and x_2 in the same model, we again get the same warning messages as in the simple example above aswell as suspiciously large parameter estimates and associated standard errors.

```
> mod2 <- glm(y ~ x1 + x2, family=binomial())
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(mod2)

Call:
glm(formula = y ~ x1 + x2, family = binomial())

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    155.07  140546.83   0.001        1
x1             -44.29   39066.66  -0.001        1
x2             132.79  122010.82   0.001        1
    Null deviance: 2.7526e+01  on 19  degrees of freedom
Residual deviance: 1.9320e-09  on 17  degrees of freedom
AIC: 6
```

The fitted model is shown in Fig. 4. Again, the estimated model fits almost perfectly to the observed data, in this case because there exist a linear combination of the explanatory variables

$$L = c_1 x_1 + c_2 x_2 \tag{26}$$

such that $y = 1$ for all values of $L > L_0$ and $y = 0$ for all values of $L < L_0$. Again, the likelihood is maximised when both slopes in the regression goes to infinity.

Consider a final example obtained by changing the reponse as follows

```
y <- c(1,1,0,1,0,1,0,0,0,0,
       1,1,1,1,1,1,1,1,1,1)
> data.frame(x1,x2,y)
   x1 x2 y
1   1  0 1
2   2  0 1
3   3  0 0
```

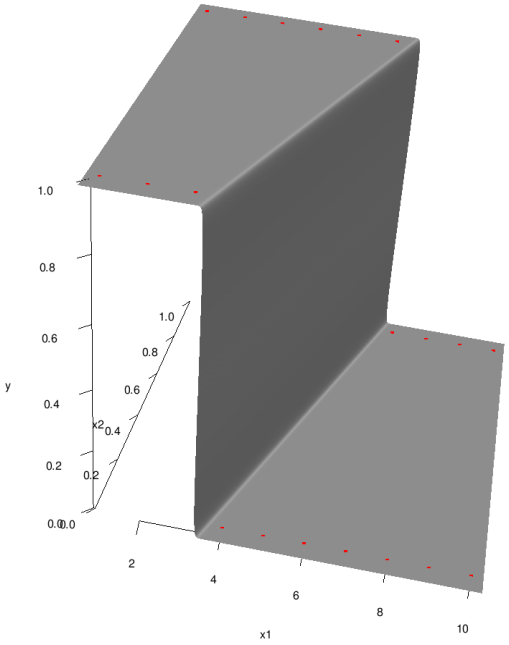

Figure 4: Linear separation in a model with two explanatory variables.

```
4    4  0 1
5    5  0 0
6    6  0 1
7    7  0 0
8    8  0 0
9    9  0 0
10 10  0 0
11  1  1 1
12  2  1 1
13  3  1 1
14  4  1 1
15  5  1 1
16  6  1 1
17  7  1 1
18  8  1 1
19  9  1 1
20 10  1 1
```

Now we always observe a response $y = 1$ for the second level of the factor (when the dummy variable $x_2 = 1$). Fitting a logistic regression to these data we get

```
> mod2 <- glm(y ~ x1 + x2, family=binomial())
> summary(mod2)

Call:
glm(formula = y ~ x1 + x2, family = binomial())

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.9265     2.0601   1.421   0.1554
```

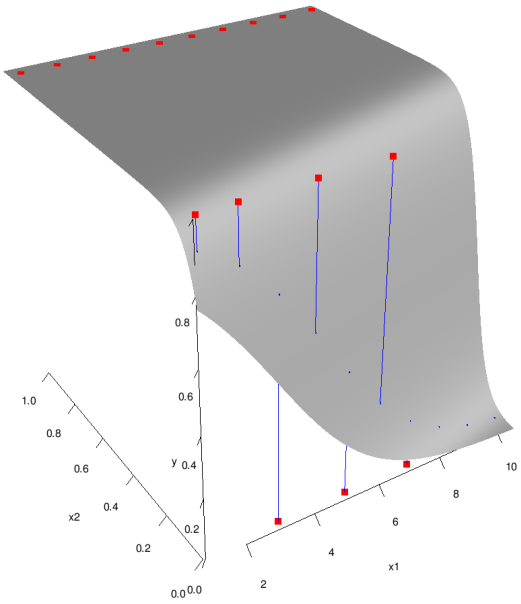

Figure 5: An example for which the response y is always 1 for one level (indicated by the dummy variable x_2) of a factor included as an explanatory variable in the model.

```
x1            -0.6622     0.4001  -1.655   0.0979 .
x2            22.3725  4800.9802   0.005   0.9963

    Null deviance: 24.4346  on 19  degrees of freedom
Residual deviance:  8.6202  on 17  degrees of freedom
AIC: 14.620
```

Fig. 5 shows a plot of the observed and predicted values as function of x_1 and x_2. In this case we can trust the estimate of the regression coefficient β_1 for the effect of x_1 but not that of x_2 since the maximum likelihood estimate of β_2 becomes infinite.

To conclude, in cases where linear separation occurs, tests between different nested alternatives may still be valid, but estimates of some of the parameters can no longer be trusted and more data needs to be collected. An alternative approach is to assume that the effects associated with different levels of a factor comes from some a common distribibution. This leads to so called generalized linear mixed models (Pinheiro and Bates, 2009)

5 Poisson response

See Dalgaard, ch. 15.

6 Overdispersion

6.1 Processes generating over- and underdispersion

The assumption that the response is binomial with parameters n and p implies the variance of the response variable is $\operatorname{Var} Y = np(1-p)$. Having specified how model for how p depends on the explanatory variables of interest, for example,

$$\operatorname{logit} p = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{27}$$

the variance of response variable is also given. The expression for the variance of a binomially distributed variable follows from the assumption that the individual trials in the Bernoulli trials are independent. If $Y \sim \text{bin}(n, p)$, we can write Y as sum

$$Y = I_1 + I_2 + \cdots + I_n \tag{28}$$

where the I_i''s are variables indicating success or failure in each Bernoulli trial. It then follows that

$$\text{Var}\, Y = \sum_{i=1}^{n} \text{Var}\, I_i + 2 \sum_{i<j} \text{Cov}(I_i, I_j). \tag{29}$$

If the trials in the Bernoulli sequence are independent, all covariances are zero and (29) simplifies to

$$\text{Var}\, Y = n\, \text{Var}\, I_i = np(1-p). \tag{30}$$

If the Bernoulli trials are not independent, however, the variance of Y may either increase or decrease. For example, if we consider the survival of n nestlings, competition between different individuals for limited resources provided by the parents may create a negative dependency between their individual survival. This would make the above covariances negative and deflate the variance of the total number of nestlings Y leaving the nest. Similarly, positive covariances between the I_i's, as a result of cooperation, say improved vigilance, might inflate the variance in the number of offspring surviving a given time interval after leaving the nest. In humans, there is a weak positive correlation between the sex of children in the same family creating a slight degree of over-dispersion in family sex-ratio (Lindsey and Altham, 2002).

Similar complications may arise also for Poisson processes. Just like individual trials in a Bernoulli sequence are assumed to be independent, a Poisson process has the property that the number of occurances during disjoint subintervals are assumed to be indepenent. The total number of occurances Y during a time interval of length t is then Poisson distributed with expectation $EY = \lambda t$ where λ is the rate of the process. The assumption of independence between disjoint subintervals implies that the variance $\text{Var}\, Y$ has to be equal to the expected value λt.

Suppose that we count the number of individuals of minke whale seen from a ship moving a along a given transect at constant speed. As a result of flocking behaviour, the number of individuals seen during disjoint adjacent subintervals will not be independent, since we are more likely to observe another individual shortly after a given observation. This leads to an increase in the variance of the total number seen during a time interval of length t beyond the Poisson variance λt.

Similarly, negative dependencies may arise reducing the variance if each occurance involves a certain handling time. For instance, in many species, giving birth involves pregnancy and parantal care during which a new child cannot be conceived. This will deflate the variance in the total number of offspring conceived during a time interval of length t below the Poisson variance. Modern contraception will of course further complicate this process.

Finally, overdispersion may arise as a result of heterogeneity between different individuals not accounted for by the fitted model. Suppose that we study the number of offspring produced by different parents. Each parent produces a maximum number of n offspring and each offspring survive with probability p. Conditional on p, the variance in number of surviving offspring Y of a given parent is then

$$\text{Var}(Y|p) = np(1-p), \tag{31}$$

and the conditional expectation is

$$E(Y|p) = np \tag{32}$$

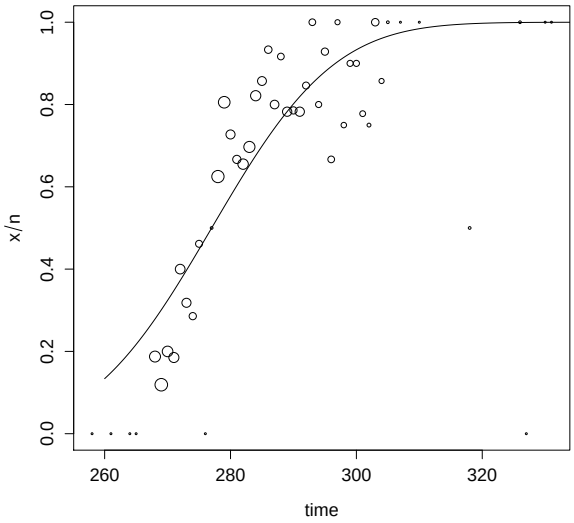

Figure 6: Proportion of female moose having ovulated at different times (number of days since January 1.

Suppose that that the mean value of p among the different parents, Ep, is p_0 and that the variance among parents $\operatorname{Var} p = \phi p_0(1-p_0)$ and . From the law of total variance it follows that

$$\begin{aligned} \operatorname{Var} Y &= E \operatorname{Var}(Y|p) + \operatorname{Var} E(Y|p) \\ &= Enp(1-p) + \operatorname{Var} np \\ &= np_0(1-p_0)[1+(n-1)\phi], \end{aligned} \tag{33}$$

that is, the variance is inflated by a factor $\varphi = 1 + (n-1)\phi$ relative to what we might expect from the simple binomial model. From this equation it can be also seen that a single Bernoulli variable cannot be over-dispersed (the case of $n = 1$).

6.2 Detecting overdispersion

A large deviance may be an indication of overdispersion. Testing if overdispersion is present is based on the fact that the deviance of the model has a chi-square distribution with $n-p$ degrees of freedom under the null hypothesis that there is no overdispersion. This null hypothesis may then be rejected if the observed deviance of the model is greater than the upper α-quantile of the chi-square distribution (the critical value of the test).

This test of overdispersion thus follows the same procedure as the test for the goodness-of-fit of the model. A positive test result may thus be an indication of overdispersion but it may also indicate that the functional relationship between the response and the explanatory variable assumed in the model is wrong.

This is illustrated by problem 3 in assignment 7 concerning how the proportion of individual female moose having ovulated increase over time during the rut (Fig. 6). The summary of the model,

```
> summary(mod)

Call:
glm(formula = prop ~ time, family = binomial(link = "probit"),
    weights = n)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8703  -1.0580   0.0004   0.5604   3.2028

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.057365   1.642587  -10.99   <2e-16 ***
time          0.065188   0.005852   11.14   <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: 254.607  on 48  degrees of freedom
Residual deviance:  90.165  on 47  degrees of freedom
AIC: 189.29
```

shows that the observed deviance is much greater than it's expected value under H_0 equal to 47, aswell as the critical value of the test

```
> qchisq(.95,df=47)
[1] 64.00111
```

and we can thus reject H_0. However, the bad fit of the model is not due to overdispersion arising from dependencies between different individual but is due the fact that the functional relationship between time and the probability p is wrong as can be seen by the systematic pattern in the residuals. As we shall see later, an alternative model which fits the data almost perfectly, can be formulated (assignment 11).

6.3 Accounting for overdispersion

Suppose that the variance of the repsonse variable of a generalized linear model is greater the binomial or Poisson variance. If we ignore this and use a model based on the assumption that the variance is exactly binomial or Poisson, we will underestimate the standard errors of the parameter estimates. Ignoring overdispersion will also lead to a higher probability of type I error - incorrectly rejecting true null hypotheses.

The simplest way of accounting for overdispersion it is assume that the variance of the response variable is inflated by an unknown scale parameter φ such that the variance is

$$\operatorname{Var} Y = \varphi np(1-p) \tag{34}$$

for models with a binomial response variables and

$$\operatorname{Var} Y = \varphi\mu \tag{35}$$

for Poisson response. One estimator of the scale parameter ψ is then the observed deviance divided by the residual degrees of freedom of the model,

$$\hat{\varphi} = \frac{D}{n-p}. \tag{36}$$

Note the similarity with estimator $s^2 = \mathrm{SSD}_{res}/(n-k)$ of σ^2 in linear models with normal response. In practice, a slightly different estimator based on th sum of squared Pearson residuals,

$$\hat{\varphi} = \frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{y_i/n - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)/n}}\right)^2, \tag{37}$$

for a binomial response, is used instead (see Venables and Ripley, 2009 for details).

Having introduced the unknown scale parameter φ in the model, the change in deviance but now scaled by the parameter φ,

$$\frac{D_0 - D_1}{\varphi} \tag{38}$$

is again approximately chi-square distribututed with $p_1 - p_0$ degrees of freedom. In pratice, since φ must be estimated, the test statistic

$$\frac{(D_0 - D_1)/(p_1 - p_0)}{\hat{\varphi}} \tag{39}$$

which is approximately F-distributed with $p_1 - p_0$ and $n - p_1$ degrees of freedom. Again, note the similarity to the corresponding test for linear models.

To fit a model with the additional scale parameter φ use `family=quasibinomial( )` or `family=quasipoisson( )` when fitting the model with the `glm( )` function. F-tests between different model alternatives are obtained using `drop1( )` or `add1( )` with the argument `test="F"`.

7 Ordinal response (optional material)

7.1 Probit regression revisited