Introduction to General and Generalized Linear Models

Generalized Linear Models - IIIb

Henrik Madsen

March 18, 2012

Examples – Overdispersion and Offset!

- Germination of Orobanche (overdispersion)
- Accident rates (offset)

- Some comments

Germination of Orobanche

Binomial distribution

Modelling overdispersion

Diagnostics

Germination of Orobanche *

Orobanche is a genus of parasitic plants without chlorophyll that grows on the roots of flowering plants. An experiment was made where a bach of seeds of the species Orobanche aegyptiaca was brushed onto a plate containing an extract prepared from the roots of either a bean or a cucumber plant. The number of seeds that germinated was then recorded. Two varieties of Orobanche aegyptiaca namely O.a. 75 and O.a. 73 were used in the experiment.

O. aegyptiaca 75				*O. aegyptiaca* 73			
Bean		Cucumber		Bean		Cucumber	
y	n	y	n	y	n	y	n
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

*Modelling binary data, David Collett

Data

```
> dat<-read.table('seeds.dat',header=T)
> head(dat)
  variety root  y  n
1       1    1 10 39
2       1    1 23 62
3       1    1 23 81
4       1    1 26 51
5       1    1 17 39
6       1    2  5  6
> str(dat)
'data.frame': 21 obs. of  4 variables:
 $ variety: int  1 1 1 1 1 1 1 1 1 1 ...
 $ root   : int  1 1 1 1 1 2 2 2 2 2 ...
 $ y      : int  10 23 23 26 17 5 53 55 32 46 ...
 $ n      : int  39 62 81 51 39 6 74 72 51 79 ...
```

The model

We shall assume that the number of seeds that germinated y_i in each independent experiment followers a binomial distribution:

$$y_i \sim \text{Bin}(n_i, p_i) \ , \ \text{where}$$
$$\text{logit}(p_i) = \mu + \alpha(\text{root}_i) + \beta(\text{variety}_i) + \gamma(\text{root}_i, \text{variety}_i)$$

Model fitting

```
> dat$variety<-as.factor(dat$variety)
> dat$root<-as.factor(dat$root)
> dat$resp<-cbind(dat$y,(dat$n-dat$y))
> fit1<-glm(resp~variety*root,
+           family=binomial(link=logit),
+           data=dat)
> fit1

Call:  glm(formula = resp ~ variety * root, family = binomial(link = logit),
    data = dat)

Coefficients:
   (Intercept)        variety2           root2  variety2:root2
       -0.5582          0.1459          1.3182         -0.7781

Degrees of Freedom: 20 Total (i.e. Null);  17 Residual
Null Deviance:     98.72
Residual Deviance: 33.28  AIC: 117.9
```

Deviance table

From the output we can make a table:

Source	f	Deviance	Mean deviance
Model $\mathcal{H}_M$	3	65.44	21.81
Residual (Error)	17	33.28	1.96
Corrected total	20	98.72	4.94

The p-value for the test for model sufficiency

```
> pval<-1-pchisq(33.28,17)
> pval
[1] 0.01038509
```

Overdispersion?

The deviance is to big. Possible reasons are:

- Incorrect linear predictor
- Incorrect link function
- Outliers
- Influential observations
- Incorrect choose of distribution

To check this we need to look at the residuals! If all the above looks ok the reason might be over-dispersion.

Overdispersion

- In the case of over-dispersion the variance is larger than expected for the given distribution.
- When data are *overdispersed*, a *dispersion parameter*, σ^2, should be included in the model.
- We use $\mathrm{Var}[Y_i] = \sigma^2 V(\mu_i)/w_i$ with σ^2 denoting the overdispersion.
- Including a dispersion parameter does not affect the estimation of the mean value parameters $\boldsymbol{\beta}$.
- Including a dispersion parameter does affect the standard errors of $\boldsymbol{\beta}$.
- The distribution of the test statistics will be influenced.

The dispersion parameter

Approximate moment estimate for the dispersion parameter

It is common practice to use the residual deviance $\mathrm{D}(\boldsymbol{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ as basis for the estimation of σ^2 and use the result that $\mathrm{D}(\boldsymbol{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ is approximately distributed as $\sigma^2\chi^2(n-k)$. It then follows that

$$\widehat{\sigma}^2_{dev} = \frac{\mathrm{D}(\boldsymbol{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{n-k}$$

is asymptotically unbiased for σ^2.

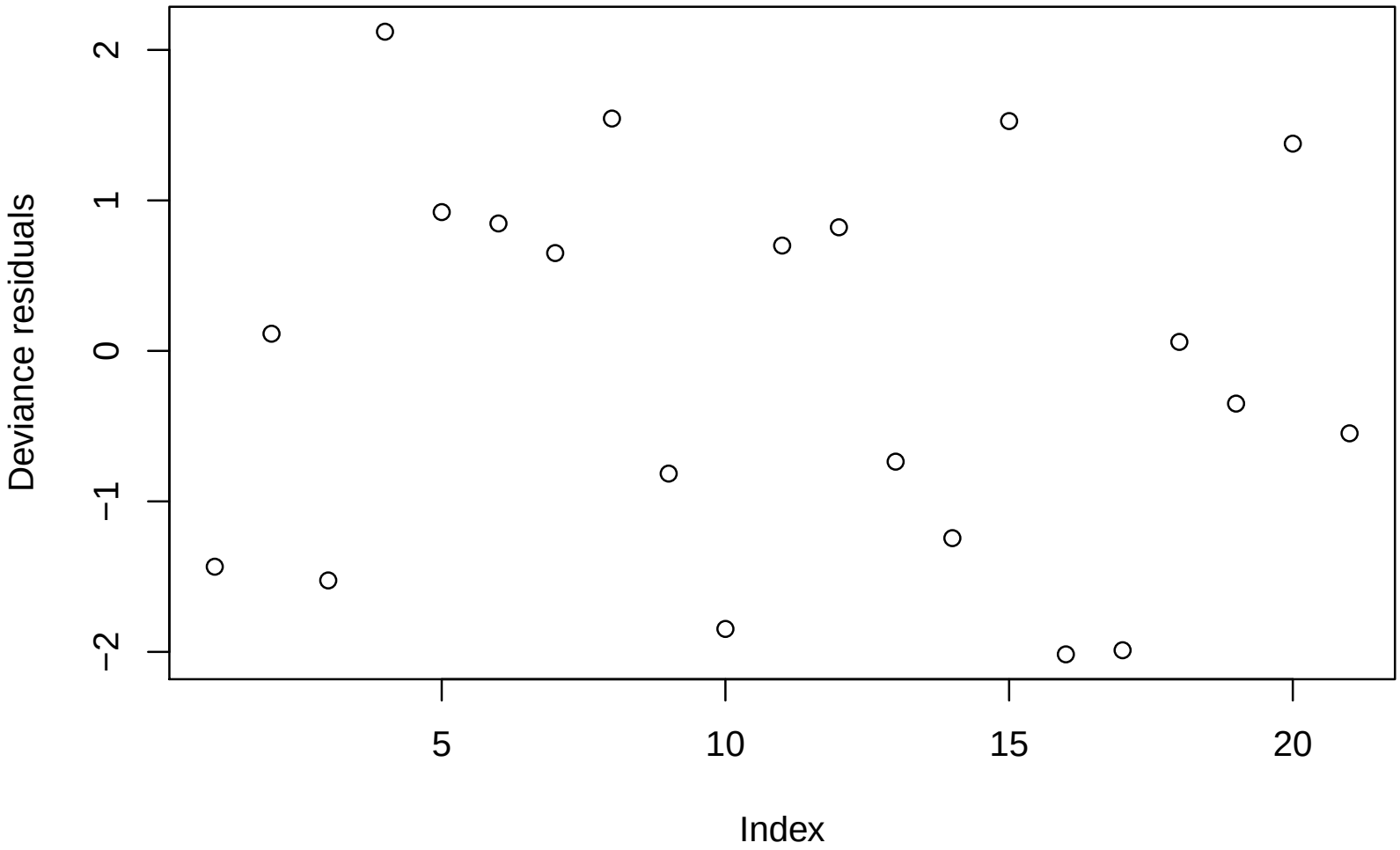
Alternatively, one would utilize the corresponding Pearson goodness of fit statistic

$$X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

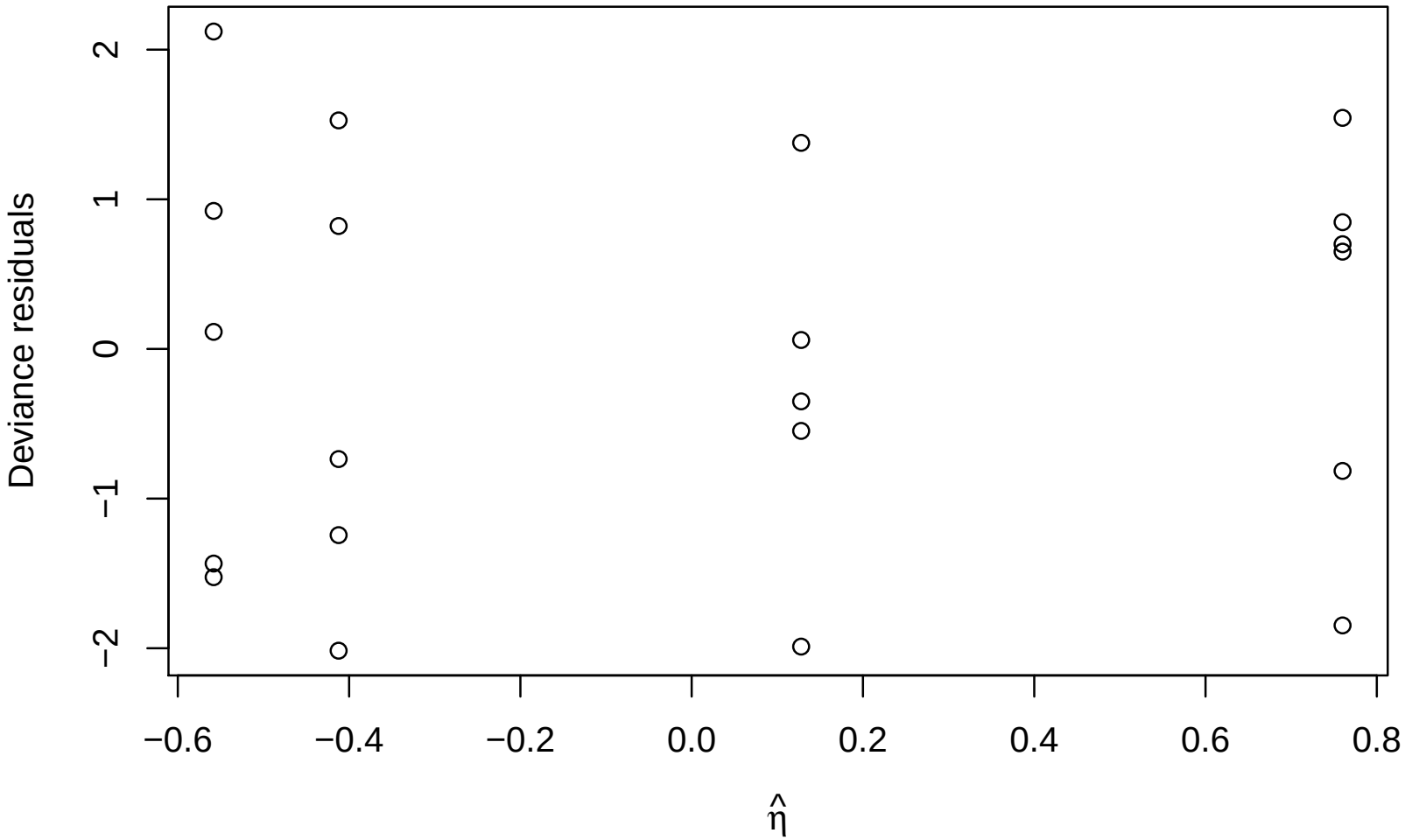
which likewise follows a $\sigma^2\chi^2(n-k)$-distribution, and use the estimator

$$\widehat{\sigma}^2_{Pears} = \frac{X^2}{n-k}.$$

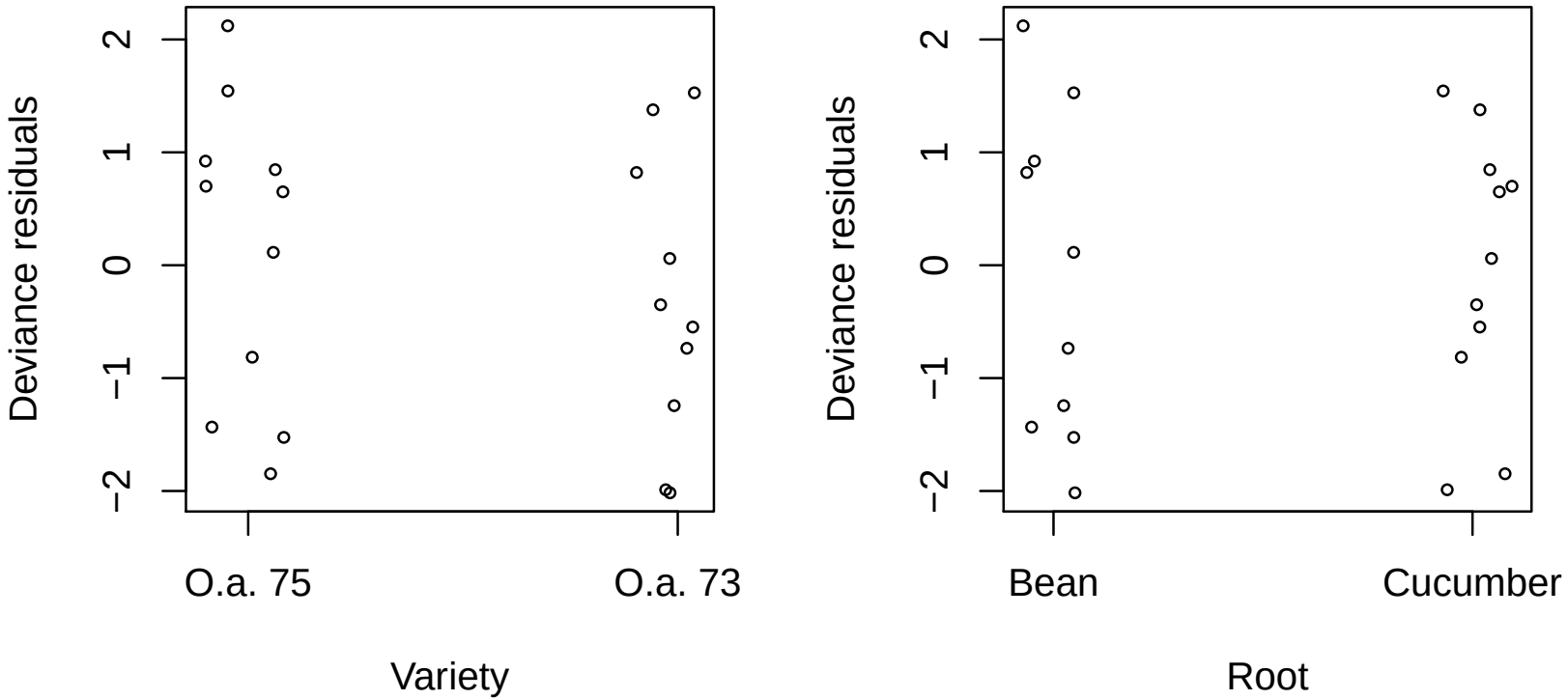
```
> resDev<-residuals(fit1,type='deviance') # Deviance residuals
> plot(resDev, ylab="Deviance residuals")
```



```
> plot(predict(fit1),resDev,xlab=(expression(hat(eta))),
+      ylab="Deviance residuals")
```



```
> par(mfrow=c(1,2))
> plot(jitter(as.numeric(dat$variety),amount=0.1), resDev, xlab='Variety',
+      ylab="Deviance residuals", cex=0.6, axes=FALSE)
> box()
> axis(1,label=c('O.a. 75','O.a. 73'),at=c(1,2))
> axis(2)
> plot(jitter(as.numeric(dat$root),amount=0.1), resDev, xlab='Root',
+      ylab="Deviance residuals", cex=0.6, axes=FALSE)
> box()
> axis(1,label=c('Bean','Cucumber'),at=c(1,2))
> axis(2)
```



Possible reasons for overdispersion

Nothing in the plots is shows an indication that the model is not reasonable. We conclude that the big residual deviance is because of overdispersion.

In binomial models overdispersion can often be explained by variation between the response probabilities or correlation between the binary responses. In this case it might because of:

- The batches of seeds of particular spices germinated in a particular root extract are not homogeneous.
- The batches were not germinated under similar experimental conditions.
- When a seed in a particular batch germinates a chemical is released that promotes germination in the remaining seeds of the batch.

Overdispersion - some facts

- The residual deviance cannot be used as a goodness of fit in the case of overdispersion.
- In the case of overdispersion an F-test should be used in stead of the χ^2 test. The test is not exact in contrast to the Gaussian case.
- When fitting a model to overdispersed data in R we use
  `family = quasibinomial` for binomial data and
  `family = quasipoisson` for Poisson data.
- The families differ from the binomial and poisson families only in that the dispersion parameter is not fixed at one, so they can model over-dispersion.

Fit of model with overdispersion

```
> fit2<-glm(resp~variety*root,family=quasibinomial,data=dat)
> summary(fit2)

Call:
glm(formula = resp ~ variety * root, family = quasibinomial,
    data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.5582     0.1720  -3.246  0.00475 **
variety2        0.1459     0.3045   0.479  0.63789
root2           1.3182     0.2422   5.444 4.38e-05 ***
variety2:root2 -0.7781     0.4181  -1.861  0.08014 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.861832)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
```

Compare to summary of standard model (wrong here)

```
> # JUST TO COMPARE THIS MODEL IS CONSIDERED WRONG HERE
> summary(fit1)

Call:
glm(formula = resp ~ variety * root, family = binomial(link = logit),
    data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01617  -1.24398   0.05995   0.84695   2.12123

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.5582     0.1260  -4.429 9.46e-06 ***
variety2        0.1459     0.2232   0.654   0.5132
root2           1.3182     0.1775   7.428 1.10e-13 ***
variety2:root2 -0.7781     0.3064  -2.539   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98.719  on 20  degrees of freedom
Residual deviance: 33.278  on 17  degrees of freedom
```

Model reduction

Note that the standard errors shown in the summary output are bigger than without the overdispersion - multiplied with $\sigma = \sqrt{1.8618}$

```
> fit2<-glm(resp~variety*root,family=quasibinomial,data=dat)
> drop1(fit2, test="F")

Single term deletions

Model:
resp ~ variety * root
             Df Deviance F value  Pr(>F)
<none>            33.278
variety:root  1   39.686  3.2736 0.08812 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model reduction

```
> fit3<-glm(resp~variety+root,family=quasibinomial,data=dat)
> drop1(fit3, test="F")

Single term deletions

Model:
resp ~ variety + root
        Df Deviance F value    Pr(>F)
<none>       39.686
variety  1   42.751  1.3902    0.2537
root     1   96.175 25.6214 8.124e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model reduction

```
> fit4<-glm(resp~root,family=quasibinomial,data=dat)
> drop1(fit4, test="F")

Single term deletions

Model:
resp ~ root
       Df Deviance F value    Pr(>F)
<none>      42.751
root    1   98.719  24.874 8.176e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model results

```
> par<-coef(fit4)
> par

(Intercept)       root2
 -0.5121761   1.0574031

> std<-sqrt(diag(vcov(fit4)))
> std

(Intercept)       root2
  0.1531186   0.2118211

> par+std%o%c(lower=-1,upper=1)*qt(0.975,19)

                lower      upper
(Intercept) -0.8326570 -0.1916952
root2        0.6140564  1.5007498

> confint.default(fit4)  # same as above but with quantile qnorm(0.975)

                2.5 %     97.5 %
(Intercept) -0.8122830 -0.2120691
root2        0.6422414  1.4725649
```

Model results

Probability of germination is $\frac{e^{-0.512}}{1+e^{-0.512}} \approx 37\%$ on bean roots.

Probability of germination is $\frac{e^{-0.512+1.0574}}{1+e^{-0.512+1.0574}} \approx 63\%$ on cucumber roots.

The odds ratio becomes:

$$\frac{\text{odds}(\text{Germination}|\text{Cucumber})}{\text{odds}(\text{Germination}|\text{Bean})} \approx 2.88$$

with confidence interval from 1.9 to 4.4.

Consider The model

Will still assume that the number of seeds that germinated y_i in each independent experiment followers a binomial distribution:

$$y_i \sim \text{Bin}(n_i, p_i) \text{ , where}$$
$$\text{logit}(p_i) = \mu + \alpha(\text{root}_i) + \beta(\text{variety}_i) + \gamma(\text{root}_i, \text{variety}_i) + B_i$$

Where $B_i \sim N(0, \sigma^2)$

Notice B_i is unobserved

In some sense this model does exactly what we need.

Can we even handle such a model? Yes! Wait for next chapter...

Accident rates

Poisson distribution

Rate data

Use of offset

Accident rates [†]

Events that may be assumed to follow a Poisson distribution are sometimes recorded on units of different size. For example number of crimes recorded in a number of cities depends on the size of the city. Data of this type are called *rate data*.

If we denote the measure of size with t, we can model this type of data as:

$$\log\left(\frac{\mu}{t}\right) = \boldsymbol{X}\boldsymbol{\beta}$$

and then

$$\log(\mu) = \log(t) + \boldsymbol{X}\boldsymbol{\beta}$$

[†] Generalized linear models, Ulf Olsson

Accident rates

The data are accidents rates for elderly drivers, subdivided by sex. For each sex, the number of person years (in thousands) are also given.

	Females	Males
No. of accidents	175	320
No. of person years	17.30	21.40

We can model these data using Poisson distribution and a log link and using number of person years as offset.

Fitting the model

```
> fit1<-glm(y~offset(log(years))+sex,family=poisson,data=dat)

> anova(fit1,test='Chisq')
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     1     17.852
sex   1   17.852          0  1.155e-14 2.388e-05
```

We can see from the output that sex is significant.

Parameter estimates - relative accident rate

```
> summary(fit1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.31408    0.07559  30.612  < 2e-16
sex2         0.39085    0.09402   4.157 3.22e-05

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.7852e+01  on 1  degrees of freedom
Residual deviance: 1.1546e-14  on 0  degrees of freedom
```

Using the output we can calculate the ratio as

```
> exp(0.3908)
[1] 1.478163
```

The conclusion is that the risk of having an accident is 1.478 times bigger for males than for females.

Some comments

Residual deviance as goodness of fit - binomial/binary data

- When $\sum_i n_i$ is reasonable large the χ^2-approximation of the residual deviance is usually good and the residual deviance can be used as a goodness of fit.
- The approximation is not particularly good if some of the binomial denominators n_i are very small and the fitted probabilities under the current model are near zero or unity.
- In the special case when n_i, for all i, is equal to 1, that is the data is binary, the deviance is not even approximately distributed as χ^2 and the deviance can not be used as a goodness of fit.

More comments...

- In a binomial setup where all n_i are big the standardized deviance residuals should be closed to Gaussian. The normal probability plot can be used to check this.
- In a Poisson setup where the counts are big the standardized deviance residuals should be closed to Gaussian. The normal probability plot can be used to check this.
- In a binomial setup where x_i (number of successes) are very small in some of the groups numerical problems sometimes occur in the estimation. This is often seen in very large standard errors of the parameter estimates.