Miniproject: Logistic regression - Poisson regression - Gamma regression - GLM

Ulrich Halekoh and Søren Højsgaard Created: November 7, 2011

Contents

TOPICS: Logistic regression

1 Shuttle

On January, 28, 1986, the shuttle Challenger (flight 61-I) exploded shortly after the start. Later it was found that 'a combustion gas leak' caused the catastrophe. The leakage may by attributed to the failure of one or several of six O-rings, which task was to seal the field joints of the motor.

The night before the event engineers discussed the problem in a three hour telephone conference. In the conference a plot of the failure of O-rings in former missions were discussed, but only data for flights with failures were used, the flights without failures were considered not to contribute valuable information. Later the Rogers Commission noted that this was a failure of a proper data analysis.

```
library(LiSciData)
data(shuttleOrings)
```

1. Plot the number of damages against the temperature. What seems to be the conclusion on the effect of temperature, if one disregards the flights where no damages were observed?

2. Fit a logistic regression model using `temp` and `pres` as covariates.

3. Use a model where the temperature is the only covariate. Predict the probability of the damage of at least one O-ring at the temperature of 31 Fahrenheit (=(31-32)*5/9 ^{0}C=-0.6 ^{0}C.) (Remember, if π is the probability of a ring to fail, the probability of at least one ring to fail is $1-(1-\pi)^6$.) Provide a confidence interval.

4. Fit models with the probit and the cloglog-link. Predict again the above probability.

2 Mendel's law and primulae

TOPICS: Test for binomial proportions

A total of 560 plants were classified by type of their leafs (flat or crimped) and the type of their eye (normal or Primrose Queen). (In the picture `http://www.primulaworld.com/`

PWWeb/gallery/slides/sinensisTM3.html the eye is the yellow area round the mouth of the corolla tube). According to simple Mendelian law one would expect a 3:1 ratio for the dominant to the recessive characteristic for each part of the plant. For leafs one would expect 420 plants with flat leaves and 140 plants with crimped leaves. For the eye one would would expect 420 plants normal and 120 plants Primrose Queen.

	Leaves		
Eye	Flat	Crimped	Total
Normal	328	77	405
Primrose Queen	122	33	155
Total	450	110	560

Concentrate first on the leaves where 450 are flat and 110 crimped. Describe the distribution of the observations, formulate the hypothesis and test it. Repeat the analysis for the eyes.

3 A simple example on different presentation of tabulated data

In a study on the ratio of born boys to girls one observed 10 families with 1 child and 17 families with 2 children and recorded the number of boys. Answer the question whether boys are equally frequent then girls in a family.

How would you test the supposition? The experimenter tabulated the data in three different ways. Which tabulation do you think is most convenient for your analysis?

Table 1: Number of families with 1 or 2 children classified after their number of boys.

	Number of boys			Total
number children	0	1	2	
1	9	1	-	10
2	2	10	5	17

Solution:

We use Table 2. We assume that in each family $i = 1, \ldots, 27$ of size n_i the number of boys is binomially distributed with probability π,

$$Y_i \sim bin(n_i, \pi)$$

The number of families give the frequency how often a certain number of boys is observed in a specific family size. We use this as a frequency in the model fit.

Table 2: Number of families with 1 or 2 children classified after their number of boys and girls.

boys	girls	number children	number families
0	1	1	9
0	2	2	2
1	0	1	1
1	1	2	10
2	0	2	5

Table 3: Number of families with 1 or 2 children classified after their number of boys and girls.

number boys	number girls		
	0	1	2
0	-	9	2
1	1	10	-
2	2	-	-

```
dboys<-data.frame(boys=c(0,0,1,1,2),girls=c(1,2,0,1,0),nfam=c(9,2,1,10,5))
```

Fit

```
m<-glm(cbind(boys,girls)~1,data=dboys,family=binomial,weights=nfam)
```

We generate now a data set, where each family is individually represented, e.g. the 2 families with 0 boy and 1 girl are represented by 2 observations (0,1) and (0,1). The data set will have 10+17=27 rows, one for each family.

```
explode<-function(dat,number){
 #explodes data by copying each row in data with the entry i  number
 # number is a column of dat!
 numb<-dat[,number]
 dat<-dat[! (is.na(numb) | numb==0),]
 numb<-dat[,number]
 dat$inde<-1:nrow(dat)
 numb<-dat[,number]
 indec<-data.frame(inde=rep(dat$inde,numb))
 datexploded<-merge(dat,indec,by='inde')
 }
dboysE<-explode(dboys,'nfam')
```

An alternative solution is therefore

```
m.a<-glm(cbind(boys,girls)~1,data=dboysE,family=binomial)
```

4 Low birth weight in infants

1. The data set `birthwt` contains information about 189 births at a US hospital.

```
library(MASS)
data(birthwt)
```

2. Use the help page for the data frame to find out what variables the data set contain.

```
help(birthwt)
```

3. Run the example code of the data frame. This creates the data frame `bwt` that aggregate levels of some of the variables of `birthwt`.

```
example(birthwt)
```

4. We treat 'low' as the response-variable and wish to find the effect of the other variable on the weight of the children. What type of model will we use?

5. Examine the effect of the independent variables on the birthweight. Which can be left out your model.

6. Make a residual analysis of your fitted model.

5 Damage of carrots by carrot larvae

Wheatley and Freeman (1982) analysed an experiment on the impact of two different insecticides, applied in different soil depths, on the damage of carrots by carrot fly larvae. Additionally to the two insecticides two control experiments without applied insecticides were performed. The experiment was repeated three times.

From each plot on which a treatment had been applied a number of carrots were chosen, washed and classified as damaged or not.

Reading data and printing the data

```
data(carrotfly,package='LiSciData')
```

`dami` are the number of damaged carrots in replicate `i` and `exami` are the corresponding analysed carrots.

1.
   - The data are given in wide format (all observations for one `insecticide` × `depth` combinations are in one row. Transform the data such that in each row we have the two observations (number of damaged and examined carrots) for

each replicate. Take care to create a variable `replicate` that identifies the replicates. (The example code of the carrotfly-data may be of help (you need to load the package `LiSciData`).

Solution:

```
library(LiSciData)
example(carrotfly)
```

```
crrtfl data(carrotfly)

crrtfl #reshape the data into long fromat
crrtfl carrot<-reshape(carrotfly,direction='long',varying=list(c('dam1','dam2','dam3'),
crrtfl                              c('exam1','exam2','exam3')),
crrtfl                         v.names=c('dam','exam'),
crrtfl                         times =c(1,2,3),timevar='replicate')

crrtfl #removing the id variable
crrtfl carrot<-subset(carrot,select=-id)
```

- Make the variables `replicate` a factor.

Solution:

```
carrot<-transform(carrot,replicate=factor(replicate))
```

- Create a factor variable `alltreat` has as levels the combinations of the levels of `insecticide` and `depth`. `alltreat` represents thus a factor variable with 11 levels representing 11 considering all combinations of insecticide and depth as 11 different treatments. Use the function `interaction` to perform the task. The function `interaction` generates a factor with levels that consists of all possible combinations of the levels of `depth` and `insecticide` . It will therefore also create level-combinations that are not presented in the data, e.g. the combination `insecticide='control', depth=1`. To get rid of these combinations use the `drop=T` argument.

Solution:

```
carrot<-transform(carrot,alltreat=interaction(depth,insecticide,drop=T))
```

2. Formulate a model and fit a model, where you analyse the impact of replicate and the 11 different treatments on the damage of carrots.

Solution: We assume that the data are binomially distributed.

```
g<-glm(cbind(dam,exam-dam)~replicate+alltreat,data=carrot,family=binomial)
```

3. Are there signs of overdispersion?

Solution:

```
X2<-sum(residuals(g,type='pearson')^2)/g$df.residual
X2
```

```
[1] 5.362905
```

4. Test whether the treatments have an impact on the number of damaged roots.

Solution:

```
g.quasi<-glm(cbind(dam,exam-dam)~replicate+alltreat,data=carrot,family=quasibinomial)
anova(g.quasi,test='F')
```

```
Analysis of Deviance Table

Model: quasibinomial, link: logit

Response: cbind(dam, exam - dam)

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                         35    1824.41
replicate  2    64.69        33    1759.73  6.031  0.007839 **
alltreat  10  1636.75        23     122.97 30.520 6.417e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

or

```
g.quasi.0<-glm(cbind(dam,exam-dam)~replicate,data=carrot,family=quasibinomial)
anova(g.quasi,g.quasi.0,test='F')
```

```
Analysis of Deviance Table

Model 1: cbind(dam, exam - dam) ~ replicate + alltreat
Model 2: cbind(dam, exam - dam) ~ replicate
  Resid. Df Resid. Dev  Df Deviance     F    Pr(>F)
1        23     122.97
2        33    1759.73 -10  -1636.8 30.52 6.417e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Test whether the effect of the insecticides are different for the depth 1.

6. Test the null-hypothesis: The effects of the insecticides are not different on any level of depth?

Solution:

This hypothesis is formulated by several contrasts to be zero at the same time. If the model is

$$logit(\pi) = \mu + \alpha_{id} + \beta_r$$

then the hypothesis of no insecticide effect is

$$\alpha_{diazonon,d} - \alpha_{disulfoton,d} = 0$$

for all $d = 1, 2, 5, 5, 10, 25$. (Why is $d = 0$ not included?) You can use the `esticon` function of `doBy` package to test this hypothesis: The parameter estimates are

```
coef(g.quasi)
```

```
          (Intercept)            replicate2            replicate3
            1.6341642             0.5031748             0.5863955
   alltreat1.diazinon  alltreat2.5.diazinon    alltreat5.diazinon
           -1.1202164            -2.1065815            -3.6135831
  alltreat10.diazinon   alltreat25.diazinon  alltreat1.disulfoton
           -4.0302763            -1.9878067            -1.4707183
alltreat2.5.disulfoton  alltreat5.disulfoton alltreat10.disulfoton
           -1.6703448            -2.2139604            -2.8947688
 alltreat25.disulfoton
           -1.1735287
```

the contrast vector for the difference

$$\alpha_{diazonon,d} - \alpha_{disulfoton,d} = 0$$

would be

```
lamb1<-lamb2<-lamb5<-lamb10<-lamb25<-0*coef(g.quasi)
lamb1[c('alltreat1.diazinon','alltreat1.disulfoton')]<-c(-1,1)
lamb2[c('alltreat2.5.diazinon','alltreat2.5.disulfoton')]<-c(-1,1)
lamb5[c('alltreat5.diazinon','alltreat5.disulfoton')]<-c(-1,1)
lamb10[c('alltreat10.diazinon','alltreat10.disulfoton')]<-c(-1,1)
lamb25[c('alltreat25.diazinon','alltreat25.disulfoton')]<-c(-1,1)
lamb<-rbind(lamb1,lamb2,lamb5,lamb10,lamb25)
```

and using the esticon function with the matrix *lamb* as contrast matrix:

```
esticon(g.quasi,cm=lamb,joint.test=T)
```

```
   X2.stat DF   Pr(>|X^2|)
1 37.95923  5 3.845064e-07
```

7. Assume now that the transformed data logit($dam/exam$) are normally distributed. Fit a model with the same covariates as above and compare the parameter estimates and their standard errors.

6 Esophageal cancer

1. The data set `esoph` contains records from a case-control study of esophageal cancer. For 88 combinations of age , alcohol consumption and tobacco consumption the number of cases and controls are given.

```
data(esoph)
```

2. Make some descriptive plots of data.
3. Formulate a model for your data.
4. Fit the model with interactions between all three predictors.
5. Eliminate factors to simplify the model as far as is possible.
6. All three factors are ordered. Convert the factors to numerical variables using the `recodevar` of the package `doBy`.
7. Can the model be simplified using the numerical representation of the predictors?
8. Analyse the residuals of your final model.

7 Discoveries

1. The data `discoveries` lists the number of great inventions and scientific discoveries in each of the years between 1860 to 1959.

```
data(discoveries)
```

2. Plot the data.
3. Add a smooth line to the plot.
4. How would you model data. Write down the systematic and the random part of your model.
5. Has the discovery rate remained constant over time?

8 Bacteria

We reanalyze data we used already in the last project day. In an experiment on nutrients on bacterial growth, bacteria were grown in different media with the nutrients sucrose and leucine added. After fours days the numbers of bacteria, `density`, were counted.

```
library(LiSciData)
data(bactsucrose)
```

1. Assume the responses `density` to be normally distributed. Assume the predictors `day`, `sucrose` and `leucine` as factors. Consider a model which is additive in `day` and with an interaction between the other two factors. Which link function would the Box-Cox approach suggest?

2. Fit the corresponding model using the identified link assuming the data to be normally distributed.

3. We want to identify an appropriate variance function. Take the absolute residuals `abs(res)` (on the response scale) from the above model and the predicted values `fit` (on the response scale) and fit the linear model

$$E(\log(|res|)) = \phi + \lambda \log(fit)$$

   Twice the slope of the fitted equation will give the exponent λ in the variance function

$$V(\mu) = \phi\mu^{\lambda}$$

4. Use the estimated $2 \cdot \lambda$ to propose a different distribution for the observations. Fit this new model.

5. Check whether you still need the interaction between sucrose and leucine.

6. Fit two models with the `log` link assuming either normally or gamma distributed data. Use AIC and cross-validation to decide which model to choose.

Solution:

```
bactsucrose<-transform(bactsucrose,y=density/max(density),day=factor(day),sucrose=factor(sucrose),leucine=factor(leucine))
library(MASS)
boxcox(y~day+sucrose+leucine+sucrose:leucine,data=bactsucrose)
m.normal.log<- glm(density~day+sucrose+leucine+sucrose:leucine,data=bactsucrose,family=gaussian(link=log))
res<-residuals(m.normal.log,type='response')
fit<-predict(m.normal.log,type='response')
g<-glm(log(abs(res))~log(fit))
m.normal.log<- glm(density~day+sucrose+leucine+sucrose:leucine,data=bactsucrose,family=gaussian(link=log))
m.gamma.log<-glm(density~day+sucrose+leucine+sucrose:leucine,data=bactsucrose,family=Gamma(link=log))
```

```
library(boot)
AIC(m.normal.log)
AIC(m.gamma.log)
cv.glm(bactsucrose,m.normal.log)$delta
cv.glm(bactsucrose,m.gamma.log)$delta
```

9 Criminal teens and Poisson distribution

Sometimes the Poisson distribution can explain a phenomenon just as a random phenomenon, making more intricate argumentation superfluous. In an article from 17. December 2007, Politiken wrote **Violent teens live in small villages - not the large city** *It was beautiful in the countryside. In old days. Today the small villages as Farsø, Bredebro, Aarup and Holsted are on the top if the number of sentences for violence for persons aged between 15 and 17 is counted. If the number of young persons that are involved in violence is related to the number of teens in their community, then the largest towns are ranked very low with Copenhagen at number 81 and Aarhus as number 125 of the 272 communities in 2006. An explanation could be that the small communities have the cheapest accommodations.* A sociologist confronted with these data commented, that the findings were surprising.

We want to find out whether the data provide evidence for something surprising going on or whether a simple assumption of equal rates based on Poisson counts would explain the data.

1. First we make a little theoretical exercise. Assume we have 3 councils with the following population of teens:

   ```
   pop<-rep(c(10,100,1000),each=c(1,1,1))
   ```

   We assume further that each community has the same problem with violent teens, i.e. we assume that all towns have the same rate of violent teens. Let us assume the violence-rate is 3%, e.g. 3 violent per 100 teens. We assume that the number of violent teens is Poisson distributed and generate for each town the number of violent teens according to this assumption A single replication would be coded as:

   ```
   b<-rpois(length(pop),lambda=0.03*pop)
   ```

   We repeat this 1000 times

   ```
   set.seed(89)
   b<-matrix(NA,length(pop),1000)
   for (i in 1:1000) {
   ```

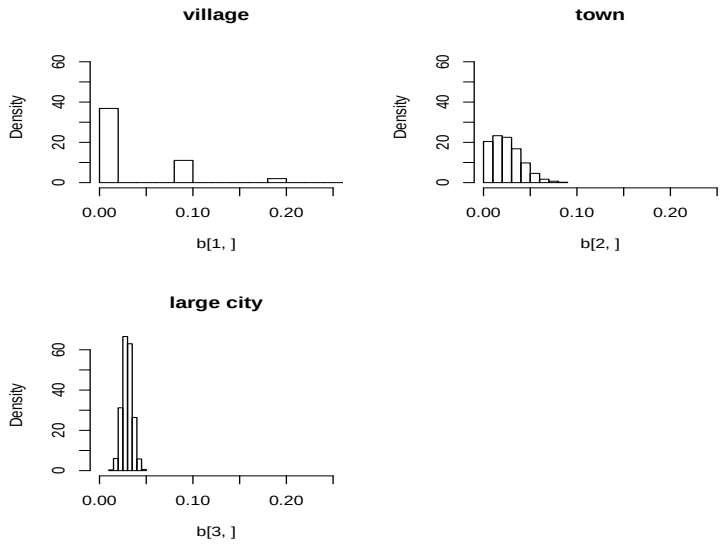
```
b[,i]<-rpois(length(pop),lambda=0.03*pop)
}
```

Now we calculate for each town the proportion of violent teens for each of our replication

```
b<-b/pop
```

We plot the distribution of the proportion for each town

```
par(mfrow=c(2,2))
 hist(b[1,],xlim=c(0,0.25),ylim=c(0,65),main='village',probability=TRUE)
 hist(b[2,],xlim=c(0,0.25),ylim=c(0,65),main='town',probability=TRUE)
 hist(b[3,],xlim=c(0,0.25),ylim=c(0,65),main='large city',probability=TRUE)
```



Looking at the plots, which type of city (i.e. village, town, large city) do you think will most probably be

(a) among those with the smallest proportions of violent teens,

(b) among those with the largest proportions of violent teens,

(c) among those with a median amount of violent teens?

Relate your conclusion to that of the article that the violent live in small but not in large cities.

2. We analyze now the data. We consider only the non-aggregated data, i.e. we do not use the data for the 'Amt's and for 'Hele landet'. Additionally we exclude all data with a ratio of 0 because for these data we cannot calculate the number of teens aged between 15 and 17 in a city.

   Preparation of the data, the rows for 'Amt' and 'Hele-landet' are deleted and the number of teens aged 15 to 17 in each town is calculated.

```
v<-get(data(poissonviolence,package='LiSciData'))
v$pop<-with(v,number/ratio*1000)
is.amt<-grep('Amt',v$kommune)
v<-v[-is.amt,]
v<-subset(v, !(kommune %in% c('Hele_landet')))
```

(a) Fit now two alternative models to the data
   i. A simple Poisson model to the data, assuming a common violence rate for each community. Consider the necessity to account for overdispersion.
   ii. A model, where the violence rate is proportional to the size of the town

(b) Are the two models statistically different?

(c) What is the likely consequence that we left out the towns where the ratio of violent teens had been estimated to zero?

Solution: A proposal solution to the model fitting

```
M0<-glm(number~offset(log(pop)),data=v,family=poisson)
X2pearson<-function(m) sum(residuals(m,type='pearson')^2)/m$df.residual
X2pearson(M0)
```

```
[1] 1.774907
```

```
M1<-glm(number~offset(log(pop)),data=v,family=quasipoisson)
M2<-glm(number~log(pop)+offset(log(pop)),data=v,family=quasipoisson)
```

10 Survial data

The data in set `whitebloodcell` of the package `LiSciData` contain the survival time (in weeks) of patients after the diagnosis of leukemia. Also the logarithm to the base 10 of the counts of the white blood cells at the time of diagnoses is given.

Survival times can sometimes be assumed to be exponentially distributed.

The exponential distribution is a special case of the gamma distribution with the dispersion parameter (the reciprocal of the shape parameter) equal to 1. The density if the Gamma distribution is

$$f(y) = \frac{1}{s^a \Gamma(a)} y^{a-1} e^{-\frac{y}{s}}$$

with $s > 0$ the scale parameter, $a > 0$ the shape parameter.

Setting $a = 1$ and defining the rate $\lambda = 1/s$ we get the most common form of the density for the exponential distribution

$$f(y) = \lambda e^{-\lambda y}.$$

1. What is the expectation and the variance of an exponential distribution. (HINT: Refer to the lecture notes for the Gamma distribution)

Solution: The Gamma distribution has expectation and variance

$$\mathbb{E}(Y) = a \cdot s \quad \mathbb{V}\text{ar}(Y) = a \cdot s^2$$

With $a = 1$ and $s = \frac{1}{\lambda}$ we have for an exponentially distributed random variable Y

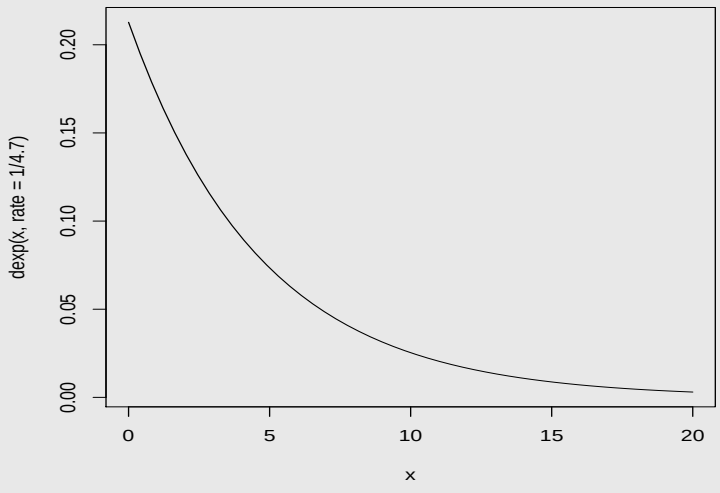
$$\mathbb{E}(Y) = 1/\lambda \quad \mathbb{V}\text{ar}(Y) = 1/\lambda^2$$

2. Draw the density of the exponential distribution with the expectation of 4.7. (HINT: use the R-function `dexp`).

Solution:

```
x<-seq(0,20,l=50)
```

```
plot(x,dexp(x,rate=1/4.7),type='l')
```



3. Assume that the expected survival times may be described by

$$\mathbb{E}(Y_i|x_i) = \exp(\alpha + \gamma x_i)$$

where x_i are the $\log_{10}$ counts of the white blood cells.

Fit an appropriate GLM (write down the formula for the linear predictor) and give an estimate for the parameters and their standard errors.

Be aware of the following: The estimation of the parameters of the linear predictor of a GLM is independent of the estimate for the dispersion. This is reflected in R by the fact that a `glm` object has no dispersion attribute. But if you you use the `summary` function, the dispersion parameter is either estimated by default via the Pearson's X^2 or you may specify it.

Because we assume that our data are exponentially distributed, you must choose the dispersion equal to 1 (remember that the dispersion is the reciprocal of the shape parameter: $\phi = 1/a$)!

Solution: The model is linear on the log-scale, we use therefore the `log`-link:

$$\log(\mathbb{E}(Y_i|x_i)) = \alpha + \gamma x_i$$

```
library(LiSciData)
data(whitebloodcell)
g<-glm(time~logcount,data=whitebloodcell,family=Gamma(link=log))
```

ATTENTION: We must explicitely specify the `log`-link for the Gamma family, because R assumes by default the canonical link for the Gamma distribution, i.e. the inverse function.

```
summary(g,dispersion=1)$coef
```

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept)  8.343253   1.608815  5.185960 2.149046e-07
logcount    -1.080825   0.389811 -2.772689 5.559521e-03
```

4. What is the expected life time for patients with a $\log_{10}$ cell count of 4.5? (HINT: Use the `predict` function).

Solution:

```
predict(g,newdata=data.frame(logcount=4.5),type='response')
```

```
       1
32.44487
```

5. Provide a confidence interval for this expected lifetime. (HINT: Use the `se` argument of the `predict` function and set the `dispersion` argument to 1).

Solution:

```
e<-predict(g,newdata=data.frame(logcount=4.5),dispersion=1,se=TRUE)
```

```
ci.log<-e$fit+ 1.96* e$se.fit *c(-1,1)
```

```
[1] 2.906002 4.053082
```

```
ci<-exp(ci.log)
```

```
[1] 18.28355 57.57466
```

If we repeat the calculation of the confidecne interval, but incorporate the dispersion into its calculation, we get a narrower interval (Note that I do NOT set the value of dispersion)

```
e<-predict(g,newdata=data.frame(logcount=4.5),se=TRUE)
ci.log<-e$fit+ 1.96* e$se.fit *c(-1,1)
ci<-exp(ci.log)
```

```
ci<-exp(ci.log)
```

```
[1] 18.57472 56.67214
```

11 Insurance claims

Install the CRAN-package `faraway`. In the `faraway`-package you find the data set `motorins` which contain the number of car insurance claims for 1797 groups of cars in Sweden in 1977.

```
data(motorins,package='faraway')
```

We want examine the effect of Zone, Make and on the number of claims.

1. Formulate a model to solve the problem. Which variable could be used as an offset in this model? Check your model.

```
g.quasi<-glm(cbind(dam,exam-dam)~replicate+alltreat,data=carrot,family=quasibinomial)
carrot$p<-carrot$dam/carrot$exam
g<-glm(log(p/(1-p))~replicate+alltreat,data=carrot,family=gaussian)
```

2. Compare your previous analysis to an analysis where you assume that the data a log-normally distributed.

   (a) Fit a normal linear model for the log-transformed data (i.e. assume that the survival times are log-normally distributed)

Solution:

```
g.lognormal<-lm(log(time)~logcount,data=whitebloodcell)
```

   (b) Based on your model, estimate the mean life time of patients with a $\log_{10}$ cell count of 4.5?. (HINT: predict first the expected log-survival time, and transform this to the original scale using the formula given in the lecture notes.)

   What is your conclusion from comparing the two ways of analysis for these data?

Solution:

```
exp(
 predict(g.lognormal,newdata=data.frame(logcount=4.5))+
 0.5*summary(g.lognormal)$sigma^2)
```

```
       1
28.29349
```

12 Weight loss under training

TOPICS: Generalized linear models; gee

Twenty men participated in a 'waist loss' program where there initial weight and their weight after completion of the program were measured. The question is, whether the program had an effect.

```
library(LiSciData)
data(waistloss)
```

1. Plot the weights against the before/ after factor combing the points for each man.

Solution:

```
library(lattice)
waistL<-reshape(waistloss,direction='long',varying=list(c('before','after')),v.names='weight',time=c('before','after'),
 timevar='time')
waistL$timenum<-with(waistL,ifelse(time=='before',-1,1))
xyplot(weight~timenum,groups=id,type='l',data=waistL)
```

2. Formulate and fit a model where you assume that all observations are independent.
3. Formulate and fit an appropriate model to the data taking the correlation between the repeated measurements into account.
4. Do you think that the working correlation `exchangeable` and `ar1` could yield different results?
5. Compare the estimates and standard errors of the two previous model fits.
6. Analyze the difference of the weights for each man. Compare the results to your previous analysis. The test you perform here is also known as a 'paired t-test'. (You can try to use the R-function `t.test`. The result should be the same as for the analysis of the differences).

Solution:

```
library(geepack)
waistL<-with(waistL,waistL[order(id,time),])
m.un<-glm(weight~time,data=waistL)
m.gee<-geeglm(weight~time,data=waistL,id=id,corstr='exchangeable')
waistloss$diff<-with(waistloss,before-after)
m.diff<-glm(diff~1,data=waistloss)
```

13 Sitka Spruce trees

TOPICS: Generalized linear models

In a study on the effect of ozone on the growth of trees, 54 sitka spruce trees were grown in an atmosphere enriched with ozone, and 25 trees were grown under normal conditions. The 79 trees of the same age were randomly assigned to the two groups at the beginning of the growth experiment.

In 1989, the second year of growth, the size of the trees was measured at roughly monthly intervals.

The data set `sitka89` in our data library `LiSciData` contains the size of the trees measured as $\log(h * d^2)$ where h is the height of a tree and d its diameter, the time of measurement in days after January, 1st, 1988, an identification number for each tree and the indicator of the treatment.

1. Plot the growth curves for all trees and the mean of the growth curves for the ozone and the control group
2. Propose and fit a model assuming that all observations are independent. Use the time-variable as a factor.
3. Fit a model taking into account the repeated measurements on the same tree.
4. Compare the parameter estimates and their standard errors of the two fits.
5. Test the null-hypothesis that there is no effects of ozone.
6. Provide the effect-estimates for each tree.
7. Make some residual plots.

14 Influence of method of anaesthetic gas application.

TOPICS: Generalized linear Models

In a study the effect of two different airway-masks (classical FM and experimental LMA) on the post-operative experience of sore throat was analyzed. Data from surgeries with 10 consultants (anaethesists) were collected on different patients.

The data are available as as

```
data(soreThroat,package='LiSciData')
```

1. Plot the proportion of patients experiencing sore throat against the type of airway mask.
2. Formulate and fit a model to the data assuming independence of the observations.
3. Formulate and fit a random intercept model taking account of the correlation between observations from the same consultant.
4. Provide parameter estimates and their confidence intervals.
5. Test the null- hypothesis that the two airway masks are equivalent.
6. Give the estimates for the consultant effects.

15 Damage of carrots by carrot fly larvae

TOPICS: Generalized linear models; link function; variance function

In a cross-sectional study the relation between plasma retinol and person characteristics and dietary factors were collected. Low level of plasma retinol is suspected ti be related to increase risk of cancer.

The data `plasmaRetinol` (removing the one observations with zero plasma concentration):

```
v<-get(data(plasmaRetinol,package='LiSciData'))
v<-subset(v,betaplasma>0)
```

```
v<-transform(v,sex=factor(sex,labels=c('M','F')),
 smokstat=factor(smokstat,labels=c('nev','for','cur')),
 vituse=factor(vituse))
names(v)[names(v)=='beteadiet']<-'betadiet'
```

1. Fit the following three models to the response `betaplasma` and the independent factors `vituse`,`sexp`,`age`,`kcal`,`fat`,`fat`, `alcohol`,`cholestorolbetadiet`. Decide which variables should be factors and which variables should be covariates.

   (a) response is normally distributed, link is the identity

   (b) response is normally distributed, link is the logarithm,

   (c) response is Gamma distributed, link is the logarithm

   Which model seems to fit best?

```
m.normal.id<-glm(betaplasma~vituse+sex +age + kcal+fat+fiber+alcohol+cholesterol+betadiet,data=v)
#
m.normal.log<-glm(betaplasma~vituse+sex +age + kcal+fat+fiber+alcohol+cholesterol+betadiet,data=v,family=gaussian(link=
#
m.gamma.log<-glm(betaplasma~vituse+sex +age + kcal+fat+fiber+I(fiber^2)+alcohol+cholesterol+betadiet,data=v,family=Gamm
```

2. In the third model, make q-q-plots of the Pearson and the deviance residuals, What do you observe?

3. Fit a generalized additive model (using the package `mgcv`), using smooth terms for all covariates. Would you need to enhance your model?

```
library(mgcv)
mm<-gam(betaplasma~vituse+sex +s(age) + s(kcal)+s(fat)+s(fiber)+s(alcohol)+
 s(cholesterol)+s(betadiet),
        data=v)
```

16 Fruitfly

TOPICS: Generalized linear models; link function; variance function

In study on the impact of sexual activity on lift-time of male fruit-flies, three groups with different degrees of sexual activity were measured. Additionally the thorax-length of males was measured because it is known to have an impact on longevity.

```
data(fruitfly,package='faraway')
sapply(fruitfly,class)
```

```
   thorax longevity  activity
"numeric" "integer"  "factor"
```

Formulate and fit an appropriate model to the data.

1. Start with a simple normal modal using the identity link
2. use a model where the variance function is quadratic in the mean.
3. use the preceding model bu use now a log-link. (Because `longevity` is a measure of lifetime, one often analysis the logarithm of such positive data).

Use residual analysis and the AIC-criterion to find an appropriate model.

What would the conclusion be if one would not use the thorax-length in the analysis?

17 Bacteria

In an experiment on nutrients on bacterial growth, bacteria were grown in different media with the nutrients sucrose and leucione add3ed. After fours days the number of bacteria `density`were counted.

```
v<-get(data(bactsucrose,package='LiSciData'))
v$y<-with(v,density/quantile(density,0.9))
```

Fit following models

1. Assume the responses to be normally distributed and use a identity link. Fit the maximal model.

2. Take the absolute residuals `abs(res)=` from the above model and the predicted values `fit` and fit the linear model

   `lm(log(abs(res))~log(fit))`

   Twice the slope of the fitted equation will give the exponent λ in the variance function

   $$V(\mu) = \phi\mu^{\lambda}$$

   Use the estimated λ to improve your model-fit.

3. Check whether you still need the interaction between sucrose and leucine.

4. Try in your model the `log` and the `identity`-link. Use AIC and cross-validation to decide which model to choose.

```
m.gamma.log<-glm(density~day+sucrose+leucine,data=v,family=Gamma(link=log))
m.gamma.id<-glm(density~day+sucrose+leucine,data=v,family=Gamma(link=identity))
```

```
library(boot)
extractAIC(m.gamma.id)
```

```
[1]    4.000 2074.988
```

```
extractAIC(m.gamma.log)
```

```
[1]    4.000 2043.355
```

```
cv.glm(v,m.gamma.id)$delta
```

```
[1] 1.013426e+19 1.010748e+19
```

```
cv.glm(v,m.gamma.log)$delta
```

```
[1] 6.007113e+18 5.904827e+18
```

18 Hodkin

TOPICS: Generalized linear models; link function; variance function

```
library(LiSciData)
data(hodkin)
v<-reshape(hodkin,direction='long',varying=list(c('hodk','nonhodk')),v.names='number',timevar='hodkin',times=c('yes','non'))
v$hodkin<-factor(v$hodkin)
```

1. Formulate and fit an appropriate model.Check for overdispersion. Does the q-q-plot of the standardized residuals tell you something about overdispersion?

```
m<-glm(number~hodkin,data=v,family=poisson)
m.ov<-glm(number~hodkin,data=v,family=quasipoisson)
```

2. Fit a model assuming the negative binomial distribution. Use the `glm.nb` function of the package `MASS`

```
library(MASS)
m.neg<-glm.nb(number~hodkin,data=v)
```

3. Fit a model assuming the data to be normally distributed and us the log-ling. Compare the coefficients, the predicted values and the AIC's of all your fitted models.

4. Fitting the model with the normal assumption, is there a difference in AIC using the log or the identity link?