Computer Lab Practical n. 4

Today's task

Examples of use of logistic regression.

How to start

• Access the course web page on Moodle (/Biology). Download and save the three data files beetle.dat, titanic.rda and snails.dat.

How carbon sulfide does no good to beetles

• Beetles were exposed to different doses of carbon sulfide (CS_2) for 5 hours. At the end of the period, the status, alive or dead, was assessed for each beetle.
• Load the data. Note: *beetle.dat* is a text file, hence we need to assign a name to the dataset when we load it into R.

```
beetle <- read.table(file="~/beetle.dat", header=TRUE)
```

The variable CS2 represents the dose level and the variable D whether the beetle is dead or alive.
• Inspect the data and scroll through them. *Do you already notice something?*
• Though only 7 different dose levels were tested (with about 60 beetles allocated to each dose level), we can nonetheless estimate the probability of death for an arbitrary level in the following way.

– Start off by adding a new variable to the dataset $\Rightarrow$ Convert the numerical variable *CS2* to factor.

```
beetle$fCS2 <- as.factor(beetle$CS2)
```

– χ^2 test for independence.

```
# Frequency table
freqtab <- table(beetle$fCS2,beetle$D)

# Row percentage (margin=1)
round(prop.table(freqtab, margin = 1), digits=3)

# Pearson's Chi-squared test
chisq.test(freqtab, correct=FALSE)
```

Question: What do you notice?

• Let us now fit a logistic regression model to estimate how the probability of death varies as the exposure dose varies. The expression $\vartheta(x)$ defines the "probability that a beetle dies when $CS2 = x$" as

defined below.

$$\vartheta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \text{or, equivalently,} \quad \log\left(\frac{\vartheta(x)}{1 - \vartheta(x)}\right) = \beta_0 + \beta_1 x$$

To estimate β_0 and β_1 run the following.

```
fit <- glm(D ~ CS2, data=beetle, family = binomial(link=logit))
summary(fit)
```

Your findings?

- The fitted probability is

$$\text{Prob}(D = Yes|CS2 = x) = \frac{\exp(-58.44 + 32.97x)}{1 + \exp(-58.44 + 32.97x)}.$$

You may plot it by specifying the following instruction in your Script window.
`curve(exp(-58.44+32.97*x)/(1+exp(-58.44+32.97*x)),from=1.6,to=2,xlab='CS2 dose')`

- Package inserts of medications or operating manuals for toxic material often report statements such as *"The LD50 of insects (or Gunea pigs) is xx"* where xx represents a number. *LD* stands for Lethal Dose and *LD50* represents the dose at which the probability of dying is 0.5.
Question: How could you use your fitted model to estimate the LD50?

Do you like snails?

- Load the data of snails.dat file you downloaded.

```
dt_snail <- read.table("~/snails.dat",header = T)
```

- Inspect the data. The experiment and the measured variables (as specified in the data file) are as follows. Groups of "N" snails of two different "Species" were kept for "Exposure" weeks in an environment with constant temperature fixed at "Temp" and relative humidity fixed at "Rel.Hum". After the exposure period, the number of dead snails ("Deaths") was counted. For instance, row 27 of the dataset tells us that 5 snails died out of 20 of species A which were kept for 3 weeks at a temperature of 20 degrees and relative humidity of 60%.
- Let us fit a logistic regression model to estimate the probability of death as a function of temperature, humidity, species and exposure period. We now have to account that each row of the data set represents a group of snails, not a single snail.

```
fit <- glm(formula = cbind(Deaths, N - Deaths) ~ Exposure + Rel.Hum + Temp + Species,
           family = "binomial", data = dt_snail)
summary(fit)
```

The cbind instruction in the above model formula creates an object which contains both the number of dead snails (Deaths) and the number of snails which are still alive at the end of the exposure period (N−Deaths), and this for every experimental condition (that is, for each combination of exposure time, temperature, humidity and species).
Assignment: Comment upon the results. In particular, assuming that the snails in question are

bred to be eaten and that they roughly have the same gastronomic value, would you suggest raising those of species A or B? And, in what kind of environment?

OPTIONAL

"Women and children first!"

• Load the titanic.rda data file.
Note: this is an *rda* object and you can load it into the workspace as it is.

```
load("~/titanic.rda")
```

• Inspect the dataset. It contains information on survival of the Titanic passengers on its first and only (you know why!) crossing of the Atlantic ocean. Apart from their names, look for the following variables:
Survived (*Yes*/No): states whether the passenger survived or not;
Type (*child*/*female*/*male*): depends on the age of the passenger (less than 10 years old/female older than 10 years/male older than 10 years);
Class (first/second/third): the class in which the passenger travelled.
The file does not include all passengers because age and/or sex was not known for some passengers.
• The data were collected by the *British Board of Trade* which concluded that the old seafaring policy of "Women and children first!" had been followed with no class discrimination.
Let us estimate a logistic regression model for the probability of surviving as a function of Type and Class of the passenger. The formula needs to be

$$Survived \sim Type + Class.$$

```
fit <- glm(Survived ~ Type + Class, family=binomial(logit), data=titanic)
summary(fit)
```

The fitted model can be written as

$$\text{Prob}(Survived) = \frac{e^{\eta}}{1+e^{\eta}} \quad \text{with} \quad \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4,$$

where X_1, X_2, X_3 e X_4 amount to 1 if the passenger was a woman, a man, travelled in second class or in third class, respectively. (To see when they amount to 1 check the (T.*something*) part attached to the "name" of the coefficient.)
Assignment: Comment upon your results. In particular, answer the following questions:
(a) which passenger group had the highest probability of surviving?
(b) which passenger group experienced the highest mortality?
(c) do the parameter estimates support the statement that preference was given to women and children?
(d) more to children or more to women?
(e) do you agree with the conclusions of the British Board of Trade *that the second and third class passenger were not discriminated in their access to the (few) available lifeboats?*
• In this case, there is no numerical variable; all variables are categorical. You may verify your conclusion by using the following simple "descriptive" method.
Use Survived as row variable, Class as column variable and Type as control variable. This table, called a *contingency table*, summarizes the number of passengers which were on board of the Titanic per

possible combination of the levels of the three variables.

```
# table(row.var, column.var, control.var)
table(titanic$Survived, titanic$Class, titanic$Type)
```

Question: Does this table tell the same story as the fitted regression model? (Forgetting about the p-values.)

How to check hypotheses on more than one parameter

• The p-values reported in the table of coefficients verify the null hypothesis that the particular "β" they refer to is zero. However, sometimes we may be interested in verifying hypotheses which refer to several parameters.
So, for instance, with respect to the model fitted to the Titanic passenger data, to verify whether "the class of the passenger did not determine her/his survival" amounts to testing the following system of hypotheses:

$$H_0 : \beta_3 = 0 \text{ and } \beta_4 = 0 \quad \text{against} \quad H_1 : \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0.$$

• To verify a hypothesis of the above type (= more than one zero "β") in RStudio we have to proceed as follows.

1. Fit the model without the variable(s) which correspond to the β's which need be tested.
   In our case, fit the model by only using Type as covariate.
   ```
   fit_1 <- glm(Survived ~ Type , family="binomial", data=titanic)
   summary(fit_1)
   ```

2. Compare two models
   ```
   anova(fit,fit_1, test="Chisq")
   ```

• We may also test for linear combinations of the parameters. For instance, we may ask ourselves whether the treatment reserved to the second class passengers on board of the Titanic was different from the one reserved to third class passengers, that is, whether

$$H_0 : \beta_3 - \beta_4 = 0 \quad \text{against} \quad H_1 : \beta_3 - \beta_4 \neq 0.$$

To verify this hypothesis in RStudio run the following.

1. Test linear hypotheses of the form

   $$H_0 : a_0\beta_0 + a_2\beta_1 + \cdot + a_k\beta_k = b$$

   where $a_0, \ldots, a_k$ e b are arbitrary constants to be input using the dialog windows that just popped up. In our case $k = 4$; the linear combination of interest is $a_0 = a_1 = a_2 = 0$, $a_3 = 1$, $a_4 = -1$ e $b = 0$.
   ```
   library(car)
   .Hypothesis <- matrix(c(0,0,0,1,-1), 1, 5, byrow=TRUE)
   .RHS <- c(0)
   linearHypothesis(fit_1, .Hypothesis, rhs=.RHS, test="Chisq")
   ```

   Your comments?

2. Try and use the same procedure to test further hypotheses of the same type. You will see how flexible it is.