# Multi variable analysis, Logistic regression

## Exercise 1: LOGISTIC REGRESSION

**Prevalence study of onchocerciasis in Sierra Leone.**

The exercise is based on onch1302.csv.

Onchocerciasis (commonly known as River Blindness) is a chronic filarial disease found in sub-Saharan Africa and some parts of Central and South America. An onchocerciasis project was set up in 1982 in the Bo district of Sierra Leone. The aims of the project were to study epidemiological, clinical, immunological and entomological aspects of the disease. Prevalence surveys were undertaken in villages selected on the basis of potential high endemicity, being situated on or near rivers which are the breeding sites for the *Simulium damnosum* blackfly. Of the twelve villages included in the present dataset, five were situated in the south and east of the country in the `forest' zone and the other seven were in the `savannah' zone of the country. A census was taken of each village, and all villagers over the age of five years were asked to participate in the study. Coverage was over 90% in all but one of the selected villages. Diagnosis was made by taking a skin-snip, and clinical and an ocular examination were also performed. The file ONCH1302 contains data for all 1,302 subjects.

**Database onch1302:**

* area Area of residence: 0=savanna, 1=forest
* sex 0=male, 1=female
* agegrp Age group 0=5-9 1=10-19 2=20-39 3=40+
* mf microfilarial infection: 0=no, 1=yes
* mfload number of microfilariae in skin snip from iliac crest: 0=none 1=1-9 2=10-49 3=50+
* lesions Presence of eye lesions: 0=no, 1=yes

1. Upload the dataset in R. To make the naming a bit more logical, copy the ‘sex’ variable into a new variable called ‘female’ and the ‘area’ variable into one called ‘forest’. For ‘mf’ and ‘lesions’ you may use the variables as is. Next, recode variables with more than 2 levels (‘agegrp’ and ‘mfload’) to binary variables. For ‘agegrp’ use ‘adult’ and code as ‘TRUE’ for those aged 20 and older, ‘FALSE’ for those under 20 years of age. For ‘mfload’ you can recode to ‘highload’ taking values 0 and 1 together as ‘FALSE’ and values 2 and 3 together as ‘TRUE’.

Onch1302 <- read.table("onch1302.csv")

Onch1302$female <- Onch1302$sex

Onch1302$forest <- Onch1302$area

Onch1302$adult <- Onch1302$agegrp > 1

Onch1302$highload <- Onch1302$mfload > 1

1. Get an overview of the data by univariable analysis. Construct a table showing numbers and frequencies of each of the variables.

table(Onch1302$female)

proportions(table(Onch1302$female))

table(Onch1302$forest)

proportions(table(Onch1302$forest))

table(Onch1302$highload)

proportions(table(Onch1302$highload))

table(Onch1302$lesions)

proportions(table(Onch1302$lesions))

table(Onch1302$mf)

proportions(table(Onch1302$mf))

Table 1. Overview of study population

|  |  |
| --- | --- |
| **Factor (n = 1302)** | **Number (%)** |
| Female gender | 686 (52.7) |
| Adult age | 883 (67.7) |
| Living in Forest | 754 (57.9) |
| Micro filaria infected | 822 (63.1) |
| High micro-filarial load | 455 (35.0) |
| Eyes affected | 201 (15.4) |

1. Make three 2x2 tables to explore the association between mf infection and the three exposures, ‘female’, ‘adult’ and ‘forest’ and manually compute the odds ratios between exposed and unexposed.

table(Onch1302$female, Onch1302$mf)

table(Onch1302$forest, Onch1302$mf)

table(Onch1302$female, Onch1302$forest)

|  |  |  |
| --- | --- | --- |
|  | MF infected | |
| Adult | Yes | No |
| Yes | 677 | 205 |
| No | 145 | 275 |

OR = 6.3

|  |  |  |
| --- | --- | --- |
|  | MF infected | |
| Female | Yes | No |
| Yes | 396 | 426 |
| No | 290 | 190 |

OR = 0.6

|  |  |  |
| --- | --- | --- |
|  | MF infected | |
| Forest | Yes | No |
| Yes | 541 | 213 |
| No | 281 | 267 |

OR = 2.4

1. As a next step, use the ‘cc’ command from the ‘Epistats’ package to confirm the associations between ‘mf’ and exposures ‘female’, ‘adult’ and ‘forest’ and add the 95% confidence intervals. Are these associations statistically significant? Who are more at risk, men or women?

Loading the Epistats package:

library(EpiStats)

Using the ‘cc’ command to calculate odds ratios:

cc(Onch1302, "mf", "female")

cc(Onch1302, "mf", "adult")

cc(Onch1302, "mf", "forest")

Table 2. Univariate associations with microfilarial infection

|  |  |  |
| --- | --- | --- |
| **Factor (n = 1302)** | **Number (%)** | **OR (95% CI)** |
| Female gender | 686 (52.7) | 0.6 (0.5-0.8) |
| Adult age | 883 (67.7) | 6.3 (4.8-8.2) |
| Living in Forest | 754 (57.9) | 2.4 (1.9-3.1) |

1. Now compare results from table-based analyses with results from logistic regression. Are they consistent? For each of the three models note the OR and it’s 95% CI.

GLM.1 <- glm(mf ~ female, family=binomial, data=Onch1302)

summary(GLM.1)

exp(coef(GLM.1))

exp(confint(GLM.1))

GLM.2 <- glm(mf ~ adult, family=binomial, data=Onch1302)

summary(GLM.2)

exp(coef(GLM.2))

exp(confint(GLM.2))

GLM.3 <- glm(mf ~ forest, family=binomial, data=Onch1302)

summary(GLM.3)

exp(coef(GLM.3))

exp(confint(GLM.3))

Table 3. Univariate associations with microfilarial infection with logistic regression

|  |  |  |
| --- | --- | --- |
| **Factor (n = 1302)** | **Number (%)** | **OR (95% CI)** |
| Female gender | 686 (52.7) | 0.6 (0.5-0.8) |
| Adult age | 883 (67.7) | 6.3 (4.9-8.1) |
| Living in Forest | 754 (57.9) | 2.4 (1.9-3.0) |

1. The only exposure that we could do something about in an intervention is ‘living in the forest’, do you think age or gender could be confounders in the association between ‘living in the forest’ and being infected with micro filaria?   
     
   If it is permanent residence, then probably not. But if adult men spend part of the year in the forest for professional reasons, both age and gender could be associated with living in the forest and both of them are also associated with the outcome, micro-filarial infection. They do not seem to be in the causal pathway.
2. Use ‘CCInter’ from the ‘Epistats’ package to test for confounding. Is there confounding or interaction in the association between ‘mf’ and ‘forest’ by either ‘female’ or ‘adult’?

CCInter(Onch1302, "mf","forest","female")

CCInter(Onch1302, "mf","forest","adult")

Female: No interaction (though p =0.09), no confounding,

Adult: No interaction (p=0.10) but confounding, MHOR = 3.2

1. Does logistic regression confirm your findings on “adult” and “female” as potential confounders?

Yes, if you add ‘adult’ to the model the OR for ‘mf’ becomes 3.3, if you add female it remains 2.4

GLM.4 <- glm(mf ~ forest + adult, family=binomial, data=Onch1302)

summary(GLM.4)

exp(coef(GLM.4))

GLM.5 <- glm(mf ~ forest + female, family=binomial, data=Onch1302)

summary(GLM.5)

exp(coef(GLM.5))

1. What happens if you add an additional term in the ‘CCInter’ command, e.g. if you type: ‘CCInter(Onch1302, "mf","forest","adult","female")’?

CCInter(Onch1302, "mf","forest","adult")

CCInter(Onch1302, "mf","forest","adult","female")

The resulting MH OR is just the same and if you look at the output you will see that factor ‘female’ has just not been taken into account.

1. Now try to fit a logistic regression model with ‘mf’ as outcome variable and ‘forest’, ‘adult’ and ‘female’ as predictors. What happens now to the odds ratio of ‘forest’ in comparison to the model with only ‘forest’ and ‘adult’ as predictors?

GLM.6 <- glm(mf ~ forest + adult + female, family=binomial, data=Onch1302)

summary(GLM.6)

exp(coef(GLM.6)) # Exponentiated coefficients ("odds ratios")  
  
There is a minimal change, from 3.2577809 to 3.2653865, this is no surprise because we already knew that ‘female’ was not a confounder.

1. Please check that the null deviance (the -2LLR of the null model) is 1714.1 on 1301 degrees of freedom. For the model with ‘forest’ and ‘adult’ it is 1416.7 on 1299 degrees of freedom. For the model with ‘forest’, ‘adult’ and ‘female’ it is 1394.1 on 1298 degrees of freedom. So the difference between the last two models is 1416.7-1394.1, which is equal to 22.6, on 1299-1298, i.e. 1 degree of freedom. Is this a significant difference, please check a chi square table or in Excel (formula CHIDIST(22.6,1)?

Yes, this is highly significant, with 1 degree of freedom and a chi square of 22.6, p = 0.000002.

1. So apparently the model with the three terms, ‘forest’, ‘adult’ and ‘female’ is significantly better than the model with only ‘forest’ and ‘adult’. The easier way to check this is to do a likelihood ratio test in R. First run the complex model with three terms, then run the simple model with two terms and next use the anova() function to make the comparison.

GLM.7 <- glm(mf ~ forest + adult+female, family=binomial, data=Onch1302)

GLM.8 <- glm(mf ~ forest + adult, family=binomial, data=Onch1302)

anova(GLM.8, GLM.7, test="Chisq")

Indeed the difference is 22.562 with 1 degree of freedom and the p-value is 0.000002035. Depending on which of the two models you select first it will turn out positive or negative but the chi square and p-value will not change.

What is important to ensure is that the simple model is ‘nested’ in the complex model and that the numbers of observations are the same. The latter may present a problem if your additional variable has some missing values while the other variable does not have missing values. These will be left out when you do the complex model but used again when you do the simple model which makes the two non-comparable. Fortunately R Commander will notice this and will provide a warning message ‘models were not all fitted to the same size of dataset’.

If you want to try this out for yourself, make a new dataset with just the subset of rows where ‘female’ is not missing. Then run the two models again with this dataset and try to do the LR-testing.

Onch1302\_complete <- subset(Onch1302, !is.na(Onch1302$female))

GLM.9 <- glm(mf ~ forest + adult+female, family=binomial, data=Onch1302\_complete)

GLM.10 <- glm(mf ~ forest + adult, family=binomial, data=Onch1302\_complete)

anova(GLM.10, GLM.9, test="Chisq")