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

|  |  |
| --- | --- |
| Factor | N (%) |
| Female gender | 686 (52.7) |
| Adult age | 882 (67.7) |
| Living in Forest |  |
| Micro filaria infected |  |
| High micro-filarial load |  |
| Eyes affected |  |

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.

|  |  |  |
| --- | --- | --- |
|  | Micro filaria infected |  |
| Gender | Yes | No |
| Male | 426 | 190 |
| Female | 396 | 290 |

OR (female) = 0.61 (0.48 – 0.77)

|  |  |  |
| --- | --- | --- |
|  | Micro filaria infected |  |
| Age | Yes | No |
| 1-19 |  |  |
| 20+ |  |  |

OR (20+) =

|  |  |  |
| --- | --- | --- |
|  | Micro filaria infected |  |
| Residence | Yes | No |
| Savanna |  |  |
| Forest |  |  |

OR (forest) =

1. As a next step, use the ‘cc’ command from the ‘Epistats’ package to confirmm 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?
2. 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.
3. 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?
4. 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’?
5. Does logistic regression confirm your findings on “adult” and “female” as potential confounders?

Forest:  
Crude OR = 2.41 (1.90 – 3.06)  
OR adjusted for sex = 2.40 (1.90 – 3.03)  
OR adjusted for adult = 3.27 (2.51 – 4.27)  
OR adjusted for both = 3.26 (2.50 – 4.27)

|  |  |  |  |
| --- | --- | --- | --- |
| Factor | N cases (%) | Crude OR (95% CI) | Adjusted OR (95% CI) |
| Residence in forest |  | 2.41 (1.90 – 3.06) | 3.26 (2.50 – 4.27) |
| Age +20 |  | 6.26 (4.86 – 9.10) | 8.03 (6.10 – 10.66) |
| Male |  |  | 1.87 (1.44 – 2.44) |

The odds of having a microfilarial infection are 3.26 times as high for those who reside in the forest compared to those in the savanna, if age and sex are held constant.

1. What happens if you add an additional term in the ‘CCInter’ command, e.g. if you type: ‘CCInter(Onch1302, "mf","forest","adult","female")’?
2. 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?
3. 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)?
4. 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.

Optional:

1. Try to do the same with ‘lesions’ or ‘highload’ as outcome.