# Advanced Statistical Methods in Epidemiology

## Logistic Regression Exercise 3: MODELLING

##### DATABASE onch1302.csv

This exercise is based on the ‘onch1302.csv’ dataset, which you have seen before. 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

This time you are asked to make two logistic regression models, first using classical model selection, then change-in-estimate model selection for the effect of area of residence (forest or savannah). Please answer the questions below.

**Part 1:**

Question 1: Which are the steps you will go through for classical model selection?

* Describe study population
* Explore univariate associations with disease
* Select variables significant at p = 0.10 to build saturated model
* Remove variables one by one, starting with the highest p-value and always perform likelihood ratio testing to check whether complex model is not significantly more precise, until you reach a variable that cannot be left out.
* Check for interactions

setwd("xxx/Datasets\_csv")

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

head(onch1302)

Question 2: Please start by making a table describing the study population. Remember that to get tables with percentages you can use the “proportions” or “prop.table” function around a table (both functions are the same).

onch1302$female <- ifelse(onch1302$sex==0, 0,1)

onch1302$forest <- ifelse(onch1302$area==0, 0,1)

with(onch1302, table(mf))

with(onch1302, table(forest))

with(onch1302, table(female))

with(onch1302, table(agegrp))

with(onch1302, table(lesions))

with(onch1302, proportions(table(mf)))

with(onch1302, proportions(table(forest)))

with(onch1302, proportions(table(female)))

with(onch1302, proportions(table(agegrp)))

with(onch1302, proportions(table(lesions)))

Table 1: Description of study population

|  |  |
| --- | --- |
| **Factor (n=1302)** | **Number (%)** |
| Female gender | 686(52.7) |
| Age group 5-9 | 202(15.5) |
| 10-19 | 218(16.7) |
| 20-39 | 424(32.6) |
| 40+ | 458(35.2) |
| Infected with micro filaria | 822(63.1) |
| Living in forest | 754(57.9) |
| Eye lesions | 201(15.4) |

Question 3: Next step is to explore the bivariate associations with the outcome variable, microfilarial infection. Please explore these associations using bivariate logistic regression models and construct an appropriate table. Though odds ratios with 95% confidence intervals show which coefficients are significant, please add a column with p-values from the likelihood ratio test so we can decide which factors to test in our multivariate model.

with(onch1302, table(forest, mf))

with(onch1302, table(female, mf))

with(onch1302, table(agegrp, mf))

with(onch1302, table(lesions, mf))

cases <-subset(onch1302, mf==1)

controls <-subset(onch1302, mf==0)

with(cases, proportions(table(forest)))

with(cases, proportions(table(female)))

with(cases, proportions(table(agegrp)))

with(cases, proportions(table(lesions)))

with(controls, proportions(table(forest)))

with(controls, proportions(table(female)))

with(controls, proportions(table(agegrp)))

with(controls, proportions(table(lesions)))

OR\_female <- (0.4817518/(1-0.4817518))/(0.6041667/(1-0.6041667))

OR\_female

Mod1 <- glm(mf ~ forest , family=binomial, data=onch1302)

Mod2 <- glm(mf ~ female , family=binomial, data=onch1302)

Mod3 <- glm(mf ~ factor(agegrp) , family=binomial, data=onch1302)

Mod4 <- glm(mf ~ lesions , family=binomial, data=onch1302)

summary(Mod1)

summary(Mod2)

summary(Mod3)

summary(Mod4)

exp(coef(Mod1))

exp(confint(Mod1))

exp(coef(Mod2))

exp(confint(Mod2))

exp(coef(Mod3))

exp(confint(Mod3))

exp(coef(Mod4))

exp(confint(Mod4))

anova(Mod1, test="Chisq")

anova(Mod2, test="Chisq")

anova(Mod3, test="Chisq")

anova(Mod4, test="Chisq")

Table 2: Bivariate associations with infection:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Factor (n=1302)** | **Cases  (% exposed)** | **Non-cases**  **(% exposed)** | **OR (95% CI)** | **p-value** |
| Female gender | 396 (48.2) | 290 (60.4) | 0.61 (0.48-0.76) | <0.0001 |
| Age group: 5-9 years | 46 (5.6) | 156 (32.5) | Ref. | <0.0001 |
| 10-19 years | 99 (12.0) | 119 (24.8) | 2.82 (1.86-4.33) |
| 20-39 year | 299 (36.4) | 125 (26.0) | 8.11(5.54-12.08) |
| 40 years and above | 378 (46.0) | 80 (16.7) | 16.02 (10.74-24.31) |
| Living in forest | 541(65.8) | 213(44.4) | 2.41(1.92-3.04) | <0.0001 |
| Eye lesions | 182 (22.1) | 19 (4.0) | 6.90 (4.35-11.59) | <0.0001 |

Question 4: Which variables will you include in the multivariate model? Should eye lesions be included?

All variables are clearly statistically significant. I would not recommend including eye lesions because it is a result of the disease rather than a risk factor.

Question 5: Fit the model with ‘age group’ as factor, ‘forest’ and ‘female’, which is now the weakest factor (highest p-value)? Does dropping it make the model significantly less precise? Present the final model in a table.

Mod5<- glm(mf ~ forest+female+factor(agegrp) , family=binomial, data=onch1302)

summary(Mod5)

exp(coef(Mod5))

exp(confint(Mod5))

Mod5a<- glm(mf ~ forest+factor(agegrp) , family=binomial, data=onch1302)

anova(Mod5, Mod5a, test="Chisq")

Table 3: Associations with infections, multivariate

|  |  |
| --- | --- |
| **Factor (n=1302)** | **OR (95% CI)** |
| Female gender | 0.56 (0.43-0.73) |
| Age group: 5-9 years | Ref |
| 10-19 years | 2.57 (1.66-4.00) |
| 20-39 year | 10.46 (6.98-15.96) |
| 40 years and above | 17.66 (11.64-27.28) |
| Living in forest | 3.07 (2.35-4.04) |

Question 6: Which are the three potential interactions? Does any of them significantly improve the model? If so, make a table with the odds ratio’s in which you have taken into account the interaction.

Mod6<- glm(mf ~ forest\*female+factor(agegrp) , family=binomial, data=onch1302)

Mod7<- glm(mf ~ forest+female\*factor(agegrp) , family=binomial, data=onch1302)

Mod8<- glm(mf ~ forest\*factor(agegrp)+female , family=binomial, data=onch1302)

anova(Mod5, Mod6, test="Chisq")

anova(Mod5, Mod7, test="Chisq")

anova(Mod5, Mod8, test="Chisq")

women <-subset(onch1302, female==1)

men <-subset(onch1302, female==0)

Mod9<- glm(mf ~ forest+factor(agegrp) , family=binomial, data=women)

Mod10<- glm(mf ~ forest+factor(agegrp) , family=binomial, data=men)

summary(Mod9)

exp(coef(Mod9))

exp(confint(Mod9))

summary(Mod10)

exp(coef(Mod10))

exp(confint(Mod10))

Table 4: Multivariate associations, final model:

|  |  |  |
| --- | --- | --- |
| **Factor (n=1302)** | **Women (n=686)**  **OR (95% CI)** | **Men (N=616)**  **OR (95% CI)** |
| Age group: 5-9 years | Ref | Ref |
| 10-19 years | 2.14 (1.13-4.11) | 3.19 (1.76-5.91) |
| 20-39 year | 10.48 (5.97-19.14) | 11.56 (6.42-21.50) |
| 40 years and above | 23.12 (12.63-44.08) | 13.84 (7.87-25.07) |
| Living in forest | 4.15 (2.86-6.12) | 2.14 (1.44-3.20) |

We could still explore whether a linear trend can be assumed in age group, there is a consistent increase both in the male and the female sub group but for now let’s stop here. We could also have done a better check for interaction by age group by recoding it into a binary variable. We’ll do all that in the second part.

**Part 2:**

Question 7: Now that you have made the final predictive model, let’s continue with an explanatory model for the effect of living in the forest versus the savannah. Which factors should we consider? We will follow these steps:

*1. Fit a model containing only the primary exposure variable of interest; write down the effect size, which is the crude OR.*

*2. construct models with two variables, including the primary exposure and another exposure; identify confounders by comparing crude with adjusted ORs*

*3. construct saturated model containing all exposure variables which resulted in a 10% change in effect size of the primary exposure (ref. step 2)*

*4. eliminate exposure variables one after the other, but do not eliminate established confounders; retain exposure variable if removing it resulted in a 10% change in effect size of the primary exposure (ref. step 3)*

*5. Explore interaction between primary exposure and other exposure variables; neglect other types of interaction*

Question 8: Present a table with results of logistic regressions on forest alone, forest + female and forest + age group. Did the odds ratios of forest change as a result of this control for confounding?

Mod11<- glm(mf ~ forest, family=binomial, data=onch1302)

Mod12<- glm(mf ~ forest + female, family=binomial, data=onch1302)

Mod13<- glm(mf ~ forest +factor(agegrp), family=binomial, data=onch1302)

summary(Mod11)

exp(coef(Mod11))

summary(Mod12)

exp(coef(Mod12))

summary(Mod13)

exp(coef(Mod13))

|  |  |  |
| --- | --- | --- |
| **Factor** | **Co-factor** | **OR** |
| Living in forest |  | 2.41 |
| Living in forest | Female gender | 2.40 |
| Living in forest | Age group | 3.08 |

Question 9: Gender is not a confounder, but age group is. Normally this would be the end of our procedure. However, we would also like to explore interaction effects between the primary exposure and each of the confounders. This is best done by recoding age group to two levels, creating a variable ‘adult’ set to ‘FALSE’ if age is 0-19 years and ‘TRUE’ if age is 20 years or above. Is ‘adult’ or ‘female’ an effect modifier in the association between ‘forest’ and disease?

Female is out because it is not a confounder. Adult is not a significant effect modifier, p = 0.10.

onch1302$adult <- ifelse(onch1302$agegrp <= 1, 0,1)

table(onch1302$agegrp, onch1302$adult)

Mod14<- glm(mf ~ forest +factor(adult), family=binomial, data=onch1302)

Mod15<- glm(mf ~ forest\*factor(adult), family=binomial, data=onch1302)

summary(Mod14)

summary(Mod15)

anova(Mod15, Mod14, test="Chisq")