# Advanced Statistical Methods in Epidemiology

## Exercise 2: LOGISTIC REGRESSION: INTERACTION AND DUMMIES

The first questions are based on the mwanza.csv dataset (list of variables below). The question which interests us most in this part of the exercise is whether and how education is associated with HIV infection.

1. Explore how education and age are associated with HIV infection; the list of variables is presented overleaf. Create two factor variables, ‘age1\_f’ and ‘ed\_f’, based on ‘age1’ and ‘ed’ respectively.

mwanza <- read.table("mwanza.csv")

age1\_f <- factor(age1, labels=c("15-19","20-24","25-29","30-34","35-44","45-54"))

mwanza$ed\_f <- factor(mwanza$ed, labels=c("0","1-3","4-6","7+"))

table(mwanza$age1\_f)

table(mwanza$ed\_f)

table(mwanza$age1\_f, mwanza$case)

proportions(table(mwanza$age1\_f, mwanza$case), margin=1)

table(mwanza$ed\_f, mwanza$case)

proportions(table(mwanza$ed\_f, mwanza$case), margin=1)

|  |  |  |  |
| --- | --- | --- | --- |
| Age group | Case | Control | OR (95% CI) |
| 15-19 | 13 (11.9) | 96(88.1) | Ref |
| 20-24 | 57(34.5) | 108(65.5) | 3.9 (2.1-7.8) |
| 25-29 | 39(31.7) | 84(68.3) | 3.4 (1.8-7.1) |
| 30-34 | 33(28.0) | 85(72.0) | 2.9 (1.4-6.0) |
| 35-44 | 30(21.9) | 107(78.1) | 2.1 (1.0-4.3) |
| 45-54 | 17(15.3) | 94(84.7) | 1.3 (0.6-3.0) |

|  |  |  |  |
| --- | --- | --- | --- |
| Years of education | Case | Control | OR (95% CI) |
| 0 | 49 (15.7) | 263 (84.3) | Ref. |
| 1-3 | 24 (32.0) | 51 (68.0) | 2.5 (1.4-4.5) |
| 4-6 | 110 (30.1) | 255 (69.9) | 2.3 (1.6-3.4) |
| 7+ | 6 (54.5) | 5 (45.5) | 6.4 (1.9-23.1) |

We can see that all age groups have higher odds to be infected than age group ‘15-19’ (but not all significantly so), and that educated women have significantly higher odds to be infected than uneducated women.

1. Now we are interested in a possible interaction between these two variables. First we should reduce the number of categories for both of them. Transform “ed” into a binary variable called “ed2”, with should take value 1 for “having been at school for at least 1 year” and the value 0 otherwise. Check whether recoding has worked. The new variable “age2” should be coded as follows: 1=15‑19, 2=20‑29, 3=30‑44, 4=45 and more. In your own interest: always check that your recoding has worked.  
     
   Recoding of education (to binary) and age group (from 4 to 6 groups) variables:

ed2 <- mwanza$ed > 1

mwanza$age2[mwanza$age1 == 1] <- 1

mwanza$age2[mwanza$age1 == 2 | mwanza$age1 == 3] <- 2

mwanza$age2[mwanza$age1 == 4 | mwanza$age1 == 5] <- 3

mwanza$age2[mwanza$age1 == 6] <- 4

To check whether the recoding worked:

table(mwanza$ed, mwanza$ed2)

table(mwanza$age1, mwanza$age2)

1. Use logistic regression to look for interaction between education and age.

GLM.5 <- glm(case ~ ed2 + factor(age2), family=binomial, data=mwanza)

summary(GLM.5)

exp(coef(GLM.5))

exp(confint(GLM.5))

|  |  |
| --- | --- |
| Age group | OR (95% CI) |
| 15-19 | Ref |
| 20-29 | 3.8 (2.1-7.5) |
| 29-44 | 3.2 (1.7-6.4) |
| 45-54 | 2.4 (1.0-5.5) |
| Education | 2.4 (1.6-3.6) |

GLM.6 <- glm(case ~ ed2 \* factor(age2), family=binomial, data=mwanza)

anova(GLM.5, GLM.6, test="Chisq")

Adding an interaction term did not significantly improve the model (p=0.09). You have treated age as a categorical (dummy) variable. What happens if you treat age (variable ‘age2’) as numerical? The model would be simpler (principle of parsimony) but is it not significantly less precise?

1. Combining levels of exposure often increases the power of the likelihood ratio test to show interaction. Therefore:   
   a) dichotomise age in a new variable “young”, where 15-19 years is coded as ‘TRUE’ and 20 years is coded as ‘FALSE’. Compare to ‘age2’ to check that it worked well.  
   b) test for interaction, using logistic regression  
   c) estimate appropriate odds ratios for the association between education and HIV infection  
   d) interpret these OR:

mwanza$young <- mwanza$age1 == 1

GLM.9 <- glm(case ~ ed2 + young, family=binomial, data=mwanza)

summary(GLM.9)

GLM.10 <- glm(case ~ ed2 \* young, family=binomial, data=mwanza)

summary(GLM.10)

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

Now the interaction is significant, p= 0.01776. Testing for interaction is more sensitive if we dichotomize, i.e. use two groups only.

This is the output for the odds ratios with confidence intervals. There are 4 groups, young uneducated, young educated, older uneducated and older uneducated. The baseline now are older women without education. So we can calculate odds ratios for each group.

exp(confint(GLM.10))

Exponentiated Coefficients and Confidence Bounds

Estimate 2.5 % 97.5 %

(Intercept) 0.1836735 0.13200733 0.2496751

ed2 3.0610396 2.10056832 4.5268821

young 1.2098765 0.33774768 3.4251394

ed2:young 0.1696256 0.04620559 0.7138024

|  |  |
| --- | --- |
| Category | OR |
| Young uneducated | 1.2098765 |
| Young educated | 1.2098765\*3.0610396\*0.1696256= 0.6282050 |
| Older uneducated | 1.0 (ref) |
| Older educated | 3.0610396 |

So for older women education is a risk factor (OR = 3.06), for younger women it is protective,   
OR 0.6282050 /1.2098765 = 0.5192307.

You can also split the dataset in two subsets, young and older and assess the effect of education there. Please check whether your odds ratios match.

young\_women <- subset(mwanza, subset=young==1)

older\_women <- subset(mwanza, subset=young==0)

summary(young\_women)

GLM.11 <- glm(case ~ ed2, family=binomial, data=young\_women)

summary(GLM.11)

exp(coef(GLM.11))

GLM.12 <- glm(case ~ ed2, family=binomial, data=older\_women)

summary(GLM.12)

exp(coef(GLM.12))

Young:

> exp(coef(GLM.10))

(Intercept) ed2

0.2222222 0.5192308

Older:

> exp(coef(GLM.11))

(Intercept) ed2

0.1836735 3.0610396

**mwanza.csv**

Case control study of risk factors for HIV in women, Mwanza Tanzania

As part of a prospective study of the impact of STD control on the incidence of HIV infection in Mwanza, Tanzania, a baseline survey of HIV prevalence was carried out in 12 communities. All seropositive women (15 years and above) were revisited and, where possible) interviewed about potential risk factors for HIV infection using a standard questionnaire. In addition to interviewing HIV +ve women, a random sample of HIV -ve women were selected from the population lists prepared during the baseline survey and these women were also revisited and, where possible, interviewed. No matching of controls with cases was performed.

idno identity number

comp community 1-12

case 0 = control 1 = case

age1 age group: 1=15-19 2=20-24 3=25-29 4=30-34 5=35-44 6=45-54

ed education: 1=none/adult only 2=1-3 yrs 3=4-6 yrs 4=7+ yrs

eth ethnic group: 1=Sukuma 2=Mkara 3=other 9=missing

rel religion: 1=Moslim 2=Catholic 3=Protestant 4=other 9=missing

msta marital status: 1=currently married 2=divorced/widowed 3=never married  
 9=missing

bld blood transfusion in last 5 years: 1=no 2=yes 9=missing

inj injections in past 1 year: 1=none 2=1 3=2-4 4=5-9 5=10+ 9=missing

skin skin incisions or tattoos: 1=no 2=yes 9=missing

fsex age at first sex: 1=<15 2=15-19 3=20+ 4=never 9=missing

npa number of sexual partners ever: 1=0-1 2=2-4 3=5-9 4=10-19  
 5=20-49 6=50+ 9=missing

pa1 sex partners in last year: 1=none 2=1 3=2 4=3-4 5=5+  
 9=missing

usedc ever used a condom: 1=no 2=yes 9=missing

ud genital ulcer or discharge in past year: 1=no 2=yes 9=missing

ark perceived risk of HIV/AIDS: 1=none/slight 2=quite likely  
 3=very likely/already infected 4=don't know

srk perceived risk of STDs: 1=none/slight 2=quite likely  
 3=very likely/already infected 4=don't know