Logistic Regression

Practical in R - instructions

1. Getting started

- Log into www.ucl.ac.uk/moodle
- Download the file **logregression_data.dta** to your space on the workstation
- Start *RStudio*, and on "Environment" click on "Import Dataset" and choose "From Stata…". Just browse your dataset file from your folder and click "import".

2. Examine data

- names(logregression_data)
- summary(logregression_data)

- Our response variable will be diabetes
  - table(logregression_data$diabetes)
  - prop.table(table(logregression_data$diabetes))
  - mean(logregression_data$diabetes)
  - sd(logregression_data$diabetes)
  - summary(logregression_data$diabetes)
  - prop.table(table(logregression_data$diabetes, logregression_data$sex),2)

    # In case it is too hard to understand this table, you can add a label to the categories of variables diabetes and sex by typing:
    - logregression_data$diabetes <- factor(logregression_data$diabetes, levels = c(0,1), labels = c("No", "Yes"))
    - logregression_data$sex <- factor(logregression_data$sex, levels = c(0,1), labels = c("Men", "Women"))
  - prop.table(table(logregression_data$diabetes, logregression_data$ag16g10),2)

QUESTION 1 - What are your preliminary ideas about the relationship between diabetes and sex and diabetes and ag16g10?

3. LOGISTIC REGRESSION

- Use logistic regression to examine the association between diabetes and ag16g10 and interpret the OR

- model1 <- glm(diabetes ~ factor(ag16g10),
  data = logregression_data,
  family = binomial(link = "logit"))
- summary(model1)
- exp(cbind(odds=coef(model1), confint(model1)))

QUESTION 2 - How can you interpret the odds ratio of those aged 45-54? And of those aged 75+?

- Use logistic regression to examine the association between **diabetes** and **sex** and interpret the OR
  - sex <- factor(sex)
  - sex <- relevel(sex, ref="0")
  - model2 <- glm(diabetes ~ sex,
    data = logregression_data,
    family = binomial(link = "logit"))
  - summary(model2)
  - exp(cbind(odds=coef(model2), confint(model2)))

4. MULTIVARIATE ANALYSES

Exercise 4.1

Run a logistic regression for **diabetes** (outcome variable) and **sex**, **ag16g10** (age groups), **ethnici** (ethnicity) as independent variables. Explain the results of the model and test whether ethnicity should be kept into the model or not.

- model3 <- glm(diabetes ~ sex + factor(ag16g10),
  data = subset(logregression_data, ethnici>0),
  family = binomial(link = "logit"))
- summary(model3)
- exp(cbind(odds=coef(model3), confint(model3)))

- model4<- glm(diabetes ~ sex + factor(ag16g10) + factor(ethnici),
  data = logregression_data,
  family = binomial(link = "logit"))
- summary(model4)

- exp(cbind(odds=coef(model4), confint(model4)))
- install.packages("lmtest")
- library(lmtest)
- lrtest(model3, model4)

Exercise 4.2

Run a logistic regression for the association between obesity (**bmi30**) as outcome and social class (**sclass2**), smoking status (**smoker**), alcohol consumption (**overlim**), adjusted for **age** and **separately** for men and women.

- model5<- glm(bmi30 ~ sclass2 + factor(smoker) + overlim + age,
  data = subset(logregression_data, sex==0),
  family = binomial(link = "logit"))
- summary(model5)
- Ask for odds ratio: exp(cbind(odds=coef(model5), confint(model5)))

- model6<- glm(bmi30 ~ sclass2 + factor(smoker) + overlim + age,
  data = subset(logregression_data, sex==1),
  family = binomial(link = "logit"))
- summary(model6)
- exp(cbind(odds=coef(model6), confint(model6)))

5. INTERACTION TERMS

Exercise 5.1

- Fit a logistic regression model for the effects of **sex**, **agegr** (indicates whether age is above or below 50) and their interaction on the odds of CVD
  - model7 <- glm(cvddef1 ~ sex + agegr + sex:agegr, data = logregression_data, family = binomial(link = "logit"))
  - summary(model7)
  - exp(cbind(odds=coef(model7), confint(model7)))

- Write down the following odds ratios from the output:

  1) The odds ratio for the effect of sex (women versus men) at the baseline value of age (≤50):

2) The odds ratio for the effect of age at the baseline value of sex (men):

3) The interaction term between sex and age:

4) The estimated odds ratio for women vs men among those not at the baseline of age (aged 51+):

5) The estimated odds ratio for the effect of age (51+ vs ≤50) among women:

Summarise your results:

Exercise 5.2

- Fit a logistic regression model for CVD with an interaction term between sex and physical activity (**adt30gp**). Check the variables first, then interpret the results of each of the odds ratios obtained. Finally do a likelihood ratio test to see if there is an effect modification and based on the result run the appropriate model.

  - model8 <- glm(cvddef1 ~ factor(sex) + factor(adt30gp) + factor(sex):factor(adt30gp), data = logregression_data, family = binomial(link = "logit"))
  - summary(model8)
  - exp(cbind(odds=coef(model8), confint(model8)))

  - model9 <- glm(cvddef1 ~ factor(sex) + factor(adt30gp), data = logregression_data, family = binomial(link = "logit"))
  - summary(model9)
  - exp(cbind(odds=coef(model9), confint(model9)))

  - lrtest(model8, model9)

  - model10 <- glm(cvddef1 ~ factor(sex), data = subset(logregression_data, adt30gp==1), family = binomial(link = "logit"))
  - summary(model10)
  - exp(cbind(odds=coef(model10), confint(model10)))

  - model11 <- glm(cvddef1 ~ factor(sex), data = subset(logregression_data, adt30gp==2), family = binomial(link = "logit"))
  - summary(model11)
  - exp(cbind(odds=coef(model11), confint(model11)))

- model12 <- glm(cvddef1 ~ factor(sex), data = subset(logregression_data, adt30gp==3), family = binomial(link = "logit"))
- summary(model12)
- exp(cbind(odds=coef(model12), confint(model12)))

OPTIONAL EXERCISE

- In the model for exercise 1 we want to test whether there is an effect modification between sex and alcohol and sex and smoking – i.e. whether sex modifies the effect of alcohol, and the effect of smoking. To do that we include in the model 2 interaction terms as follows:
  - model13 <- glm(bmi30 ~ age + sclass2 + factor(sex) + overlim + factor(smoker) + factor(sex):overlim + factor(sex):factor(smoker), data = logregression_data, family = binomial(link = "logit"))
  - summary(model13)
  - exp(cbind(odds=coef(model13), confint(model13)))

- Do you get a warning about collinearity and **sex**? Why is this?

- *Is the interaction term between sex and alcohol significant? What do you conclude?

- *Is the interaction term between sex and smoking significant? How would you assess the overall significance of the interaction?

- *Run the LR test to check for the overall significance of the interaction term between sex and smoking status. Discuss the result.