Exercises day 6

Basic Statistics for health researchers

March 12 2025

Warming up

Before starting the exercise below, learn from the R-demo of Lecture 6 (available from the course webpage):

1. Read and run the code.
2. Check that the output matches the results presented on the slides.
3. Do not hesitate to add your own comments into the script.

Exercise A

For this exercise we will work with the "Myocardial Infarction (MI)" data, which come from a case-control study (available from the course webpage). We will assume that the aim of this observational study and data analysis is to shed light on the following research question:

- Is the use of oral contraceptives associated with an increased risk of myocardial infarction (MI)?

We further assume that we want to take into account the following background knowledge for the data analysis. Previous studies suggest that:

- smoking is associated with an increased risk of MI.
- oral contraceptives could increase the risk of MI differently for smokers, non-smokers and former smokers.
- the risk of MI naturally increases with age. For instance, it is expected to be different in the three age groups 15-39, 40-55 and 55 or above.

Part I

In this first part, we proceed to preliminary analyses, to "explore" the data and practice with "simple" examples. In the second part of the exercise, we will proceed to the main analysis of the data, which aims to shed light on the research question and lead to the main conclusions.

Question 1

Read the description of the data and load the data into R. Visualize the first lines of the data and a summary of the data. Which variables have missing data?

Question 2

1. Create and add to the data a factor variable `Smoke` which explicitly indicates whether each woman is a current smoker, former smoker or has never smoke. How many women are there in each group? **Hint:** you can use this code:

```
MI$Smoke <- factor(MI$tobacco,
                   levels=c(1,2,3),
                   labels=c("never smoked",
                            "current smoker",
                            "former smoker"))
```

2. Estimate a "simple" (univariate) logistic model to investigate the association between myocardial infarction (`mi`) and smoking status (`Smoke`). Interpret the results.

3. Compute a frequency table to see how many women:

   - who had never smoked have experienced MI.
   - who had never smoked have **not** experienced MI.
   - who were current smokers have experienced MI.
   - who were current smokers have **not** experienced MI.

   Hint: you can use the `table()` function as seen last week, e.g. like this,:

```
table(Smoke=MI$Smoke,MI=MI$mi)
```

4. Explain how you can deduce the estimated value of one of the model parameters from these four numbers. Compute this estimated value "by hand" using these four numbers and check that you indeed find the same result.
   Hint: remember the "simple" formula $OR = (a \cdot d)/(b \cdot c)$ seen last week, when discussing 2 by 2 tables.

5. Make all-pairwise comparison and adjust for multiple testing (because of the three comparisons). Overall, do the result show a significant association between MI and smoking status? **Hint:** use the `glht()` function of the package `multcomp` as exemplified in the R-demo.

6. Which group(s), if any, have significantly different risks of MI? Report the corresponding estimated odds ratio(s) with (adjusted) confidence interval(s) and p-value(s).

Question 3

1. Estimate a "simple" (univariate) logistic model to investigate the association between myocardial infarction (`mi`) and use of oral contraceptives (`oc`).

2. Interpret the results and write a conclusion sentence which includes the estimated odds ratio and the corresponding 95% confidence interval and p-value.

Question 4

1. Compute a frequency (2 by 3) table to compare the proportions of women who are current smoker, former smoker or who have never smoked among those who use oral contraceptives and those who do not use them.
   Hint: you can use the `table()` and `prop.table()` functions. With the `prop.table()` function, you can use the option `margin=1` or `margin=2`, if you want proportions that sum up to one in rows or columns, as already seen last week.

2. Explain why this comparison is interesting to better understand the results of the previous question.

3. Estimate a (multiple) logistic model to model the risk of myocardial infarction (`mi`) using the two variables corresponding to smoking status (`Smoke`) and use oral contraceptives (`oc`). Do not model an interaction. Interpret the results.

4. Why does the above model seem inappropriate for the aim of the data analysis? How could you improve it? **Hint:** re-read what previous studies suggest (see first page).

Question 5

1. Estimate a (multiple) logistic model to model the risk of myocardial infarction (`mi`) using the two variables corresponding to smoking status (`Smoke`) and oral contraceptives (`oc`). Model an interaction between the two variables. Interpret the results.

2. Do the results suggest that the use of oral contraceptives is **differently** associated with MI depending on the smoking status?

Question 6

1. Make some boxplots to compare the distribution of ages in the six groups of women defined by all possible combinations of smoking status and use of oral contraceptives.
   Hint: you can first define a new variable with the `interaction` function and then use the `boxplot` function as shown below. The idea is that new variable will indicate in which of the six groups each woman belongs.

```
MI$group <- interaction(MI$oc,MI$Smoke)
boxplot(MI$age~interaction(MI$oc,MI$Smoke),
```

```
xlab="Group",
ylab="Age",
col=rep(c("forestgreen","red"),3))
```

2. Does the comparison of the boxplots suggest that compared women are similar with respect to age, when comparing women using oral contraceptives to those who do not, both having the same smoking habbits? What consequences does it have on the interpretation of the results to the previous question? **Hint:** re-read what previous studies suggest.

Part II

We now proceed to the main analysis of the data, which aims to shed light on the research question and lead to the main conclusions.

Question 7

Create and add to the data a categorical variable `AgeGroup` to indicate in which of these three groups each woman belongs: 15-39, 40-55 and 55 or above. How many women to you observe in each group?
Hint: you can use the `cut()` function as exemplified in the R-demo or shown below, and then use the `table()` function to see the number of women we observe in each group.

```
MI$AgeGroup <- cut(MI$age,
                   breaks=c(15,40,55,100),
                   include.lowest=TRUE)
```

Question 8

1. Estimate a (multiple) logistic model to model the risk of myocardial infarction (`mi`) using the three variables corresponding to smoking status (`Smoke`), use of oral contraceptives (`oc`) and age group (`AgeGroup`). Model an **interaction** between smoking status (`Smoke`) and use of oral contraceptives (`oc`). Interpret the results.

2. Why does the model seems reasonable for the aim of the main analysis.

3. Write down the conclusions sentences (corresponding to the research question) and cite the relevant estimated odds ratios, confidence intervals and p-values.

Part III

When the main results of a statistical analysis rely on some modeling assumptions, it might be interesting to investigate how "robust" (i.e. stable or sensitive) the most important results are to small changes in these modeling assumptions. This is particularly interesting when some of the modeling assumptions are, to some extent, arbitrary. The additional analysis that are made to answer the same research questions under such (slightly) different modeling assumptions are called "sensitivity analyses".

Question 9

Perform a sensitivity analysis by changing the way the variable age enters the model:

1. Use the variable age as a continuous variable (assuming a linear effect of age) and refit the model.
2. Do the main conclusions change?