Exercises day 7

Basic Statistics for health researchers 2025

March 17, 2025

Warming up

Before starting the exercise below, learn from the R-demo of Lecture 7 (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 "Birth Weight" data, collected from 189 women who came to the Baystate Medical Centre, Massachusetts, in 1986.

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.

Question 2 (data preparation)

1. Use the function `table()` to look at the frequencies that were observed for the variables indicating the number of:

   - previous premature births of the mother (`ptl`)
   - medical consultations during the first trimester of pregnancy (`fvt`)

   Are some values rarely observed? **Hint:** you can simply use this R code.

```
table(BW$ptl)
table(BW$fvt)
```

2. Create and add to the data the new variables:

   - `ptl2` that takes the value "`1+`" if the mother has had at least one premature birth before, "`0`" otherwise.
   - `fvt2` that takes the value "`0`", "`1`" or "`2+`", if the mother has had 0, 1 or at least 2 medical consultations during the first trimester.

   Hint: you can use the following R code:

```
BW$fvt2 <- factor(as.numeric(BW$fvt>0) + as.numeric(BW$fvt>1),
                  levels=c(0,1,2),
                  labels=c("0","1","2+"))
BW$ptl2 <- factor(as.numeric(BW$ptl>0),
                  levels=c(0,1),
                  labels=c("0","1+"))
```

3. Make the variables `race` and `smoke` become factor variables and give appropriate labels. **Hint:** you can use the following R code:

```
BW$race <- factor(BW$race,
                  levels=c(1,2,3),
                  labels=c("white","black","other"))
BW$smoke <- factor(BW$smoke,
                   levels=c(0,1),
                   labels=c("no","yes"))
```

4. Update the variables `bwt` and `lwt` such that they indicate the weight of the baby and mother in kg (note: 1 kg=2.205 pounds). **Hint:** you can use the following R code:

```
BW$bwt <- BW$bwt/1000
BW$lwt <- BW$lwt/2.205
```

Question 3

We assume that we are primarily interested in comparing the birth weight of infant born from smokers and non-smokers women.

1. Make an appropriate "Table 1" to compare the distribution of the following variables between mothers who smoke and those who do not: `age`, `lwt`, `race`, `ui`, `ht`, `ptl2`, `fvt2`. **Hint:** use the `univariateTable()` function as examplified in the R-demo and lecture slides.

2. Are the two groups similar with respect to all these variables? If no, for which variables do you observe "substantial" differences?

Question 4

We assume that background clinical knowledge suggest that the following variables are expected to influence the birth weight of the infant: `age`, `lwt`, `ui`, `ht`, `ptl2`, `fvt2`. In addition, background knowledge suggest that access to and quality of health care and education can be associated with the birth weight of the child. In the USA, those aspects have been observed to be strongly correlated with racial or ethnic groups. Hence, in addition to the above listed variables, we will further adjust using the variable `race` in our analysis.

1. Estimate a multiple linear model to compare the average birth weight of children born from smoking and nonsmoking mothers, whom mothers are otherwise similar with respect to the above listed variables. We will assume that subject matter knowledge justifies the choice of including one unique interaction term: between smoking status and age. It is further assumed that `age` can enter in the model as a quantitative variable. **Hint:** you can use the following R code:

   ```
   lm(bwt∼smoke*age + lwt + race + ui + ht + ptl2 + fvt2,data=BW)
   ```

2. Take the time to interpret the results, especially those most relvant for the main research question about smoking.

3. Do the results confirm that modeling an interaction between smoking status and age seemed relevant?

4. Refit the model after removing 27 to all ages and 55 kg to all mother weights. Take the time to interpret the results again. What is now the interpretation of the intercept? **Hint:** you can first use the following R code to define new variables

   ```
   BW$age27 <- BW$age-27
   BW$lwt55 <- BW$lwt-55
   ```

   Then, you can refit the model using the variables `age27` and `lwt55` instead of `age` and `lwt`.

5. What is the estimated birth weight difference between two babies, one born from a smoker, the other one from a non-smoker, both born from mothers being 27 years old and similar with respect to all other variables included in the model? Provide a confidence interval.

6. Same question for mothers being 20 and 30 years old (instead of 27). Overall, what do you conclude from this model, about the association between smoking and birth weight?

Question 5

1. Does the model fit "suggest" that there is a difference in birth weight between babies born from mothers from different racial groups, who are otherwise similar with respect to smoking, age and all other variables included in the model?

2. Perform "all-pairwise" comparisons and adjust for multiple comparisons. **Hint:** you can copy-paste the corresponding R code example of the R-demo and only change the name of the fitted model (`lm4` in the R-demo) and the name of the variable which defines the group to compare (`Country` in the R-demo). Between which groups do you observe significant differences?

3. Based on the results of all-pairwise comparisons, which p-value can you report along with the conclusion that the data suggest an association between birth weight and race (adjusted for age, smoking etc, i.e., all other variables included in the model)?

4. Just for pedagogical purpose and practice, compute an alternative p-value to test the association between birth weight and race (adjusted for age, smoking etc) using a F-test. Is the result similar?

Question 6

Important: if you do not have enough time or if you are not sufficiently comfortable with R programming, **do not write any R code for this question**. Instead, just think and answer the question of item 3, using the plot provided at the last page. This plot has been produced following the instructions of items 1 and 2.

1. Plot the observations of birth weights against the age of the mother. Use two colors to distinguish between observations from smoking and nonsmoking mothers.

2. Add two lines to display the estimated mean birth weight as a function of the age of a mother and her smoking status, assuming that in any case the mother:

   - weighs 60 kg
   - is white
   - has history of hypertension
   - has no uterine irritability
   - has had no medical consultation during the first trimester of pregnancy
   - has never had a premature birth before.

3. How would the two lines change if we were to change one or some of the above assumption about the mother?

4. **(Additional, only if tyou have enough time)** Add a graphical representation of confidence intervals and prediction intervals for smokers.

Exercise B **(additional, if time allows)**

For this exercise we will work with the "Brain" data, collected from 20 mouse litters. In this exercise, we want to investigate how the litter size and body weight of infant mice relate to the weight of their brain. The aim of this exercise is to illustrate how different results and interpretations can be obtained from different models, especially between univariate and multiple models.

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. How many observations do we have per litter size?

Question 2

1. Plot the observations of brain weights against those of litter sizes.
2. Estimate a linear model to describe how the average brain weight relates to the litter size.
3. Add the regression line to the plot.
4. What can you conclude from the estimated slope and corresponding confidence interval?

Question 3

Why would it be unreasonable to describe the association between brain weights and litter size using a correlation coefficient?

Question 4

Same question as the Question 2, but consider now body weight instead of litter size.

Question 5

1. Estimate a **multiple** linear model for the average brain weights using the two variables litter size and brain weight. Do not model an interaction.
2. What can you conclude from the estimated parameters and corresponding confidence intervals?
3. At first sight, are the results surprising, as compared to those of question 2? What could possibly explain the difference between the results of this question and question 2?

Question 6

Same question as the Question 2, but consider now body weight instead of brain weight.

Question 7

Based on the complementary results from Question 6, can you now better support your explanations provided in your answer to question 5?

Question 8

Same question as the Question 2, but consider now the ratio between brain weight and body weight instead of brain weight. Do the results also support your explanations provided in your answer to question 5?

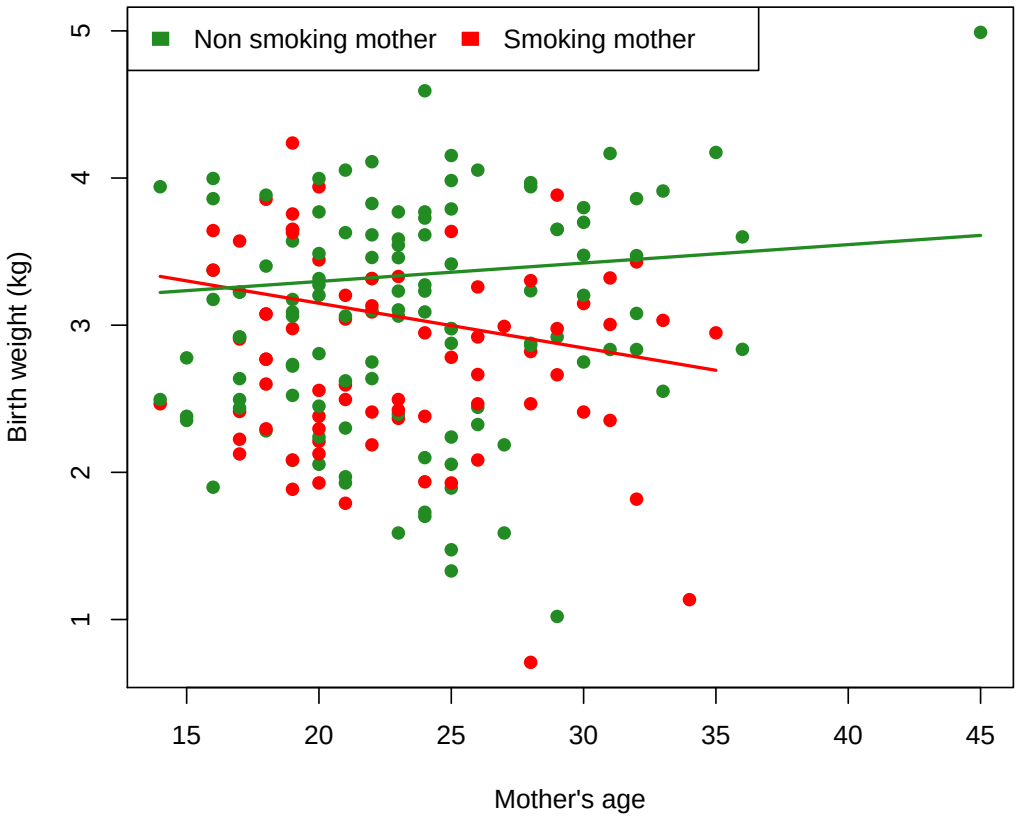

Figure 1: Plot produced following the intructions of items 1 & 2, Question 6, Exercise A.