# V. Statistical analysis in R (exercises)

Data Science Lab, University of Copenhagen

14 March, 2022

Except for the last two parts, this exercise is about analysis with so-called *Gaussian linear models*. This is the class of models where data points are assumed to be independent, Gaussian (i.e. with a normal distribution) and with the same variance. All such models are fitted with the `lm` function in R, like so:
`lm(y~x, data = my.data)`
The term *regression* is usually used for models where all explanatory variables are numerical, whereas the term *analysis of variance* (ANOVA) is usually used for models where all explanatory variables are categorical (i.e. factors). However, predictors of both types can be included in the same Gaussian linear model. In the end, we are also concerned with logistic regression models (for binary outcomes e.g. cases vs. controls) and linear mixed models (for data with a block structure).

The purpose of this exercise is to show how to fit and work with statistical models in R. This means that the analyses are not necessarily those that should be used in a real-life analysis of the data.

Start with the **Data** part. The later parts are *almost* independent of each other, so you can choose the parts that are most appropriate for you. However, not all statistical concepts are studied in all the parts present.

### Data

The dataset to be used is about risk factors associated with low infant birth weight. The data are available in the R-package **MASS** (part of base R, so it is installed automatically) as a data frame called birthwt. If nothing else is mentioned, then use the infants's birth weight, `bwt`, as outcome (response).

1. Use the commands below to load the MASS package and get the help page for the data frame. The help page will appear in the lower right window of RStudio. Read about the data; in particular about the variables `age`, `lwt`, `ftv`, `smoke`, and `bwt`.

```
library(MASS)
?birthwt
```

2. `birthwt` is not a tibble (the modern way to organize data), but we can make it one:

```
library(tidyverse)
birthData <- as_tibble(birthwt)
birthData
```

3. Mother's smoking habits (`smoke`) is coded as numerical variable. Make it a factor (categorical variable).

4. The `ftv` is a numerical variable, but we would also like a categorical version of it, with groups corresponding to zero visits, one visit, and two or more visits, respectively. Try the following commands, and check the results:

```
table(birthData$ftv)
birthData <- mutate(birthData, ftvFac = factor(ftv))
birthData <- mutate(birthData, visits = fct_collapse(ftvFac, Never="0", Once="1", other_level="MoreThan
summary(birthData)
```

5. Make groupwise boxplots of `bwt` by `visits`, i.e., boxplots of birht weight for those women who did not see their doctor, those who saw their doctor once, and those who saw their doctor more than once during the first trimester - all boxplots in one graph. Do the same thing for `smoke` (instead of `visits`).

6. You may also make groupwise boxplots for each combination of `smoke` and `visits`. You add a combination term with a `:`, e.g. `smoke:visits`. What is your initial impression about potential associations?

7. Make a scatterplot with `lwt` on the $x$-axis and `bwt` on the $y$-axis. Modify the plot such that points are coloured according to visits.
Modify it further such that the symbol types are different for smokers and non-smokers. This is done by adding `pch=smoke`.

### Regression

The term *regression* is usually used for models where all explanatory variables are numerical.

If you are not too familiar with linear regression, you can have a look at what the `summary()` output means here and for understanding regression `plot()`, have a look here.

8. Fit a linear regression model with `bwt` as response and `lwt` as covariate, and identify the estimates for the intercept and for the slope. Find the 95% confidence interval for the slope parameter by using the `confint()` function.

9. Just by calling `plot()` on your regression model you made you can carry out model validation. Does the model seem appropriate for the data?

10. Fit the multiple linear regression model where you include `lwt` as well as `age` and `ftv` (numerical variables) as covariates. Identify the parameter estimates, and consider what their interpretation is.

11. Try the following commands, and see if you can figure out what the outcome means. You should replace `reg2` with the name of the model object from the previous question.

```
newData <- data.frame(lwt=100, age=25, ftv=0)
newData
predict(reg2, newData)
predict(reg2, newData, interval="prediction")
```

12. Fit the multiple linear regression model again, but now only using data from mothers with a weight less than 160 pounds (`lwt` $< 160$). *Hint:* Use the `filter` function and change the `data` argument in the `lm` command.

### ANOVA

The term *analysis of variance* (ANOVA) is usually used for models where all explanatory variables are categorical (factors). It is important that you have coded `smoke` as a factor, cf. Question 3.
First, install and load the R-package `emmeans`.

```
library(emmeans)
```

13. Fit the oneway ANOVA (with `lm`) where the expected value of `bwt` is allowed to differ between smokers and non-smokers. Find the estimated birth weight for infants from smokers as well as non-smokers. Is there significant difference between smokers and non-smokers when it comes to infant's birth weight? *Hints:* Use `summary()` and the `emmeans()` function from the `emmeans` R-package. **N.B** make sure you know what needs to go into the function, use ?.

14. Fit the oneway ANOVA where you use `visits` as the explanatory variable. Find the estimated birth weight for each group, and make the pairwise comparisons between the groups. Furthermore, carry out the $F$-test for the overall comparison of the three levels.
*Hint* `drop1(, test="F")`. What is the conclusion?

15. Now, consider both `visits` and `smoke` as explanatory variables. Since they are both categorical variables (factors), the relevant model is a *twoway* ANOVA. Fit the twoway ANOVA model *without interaction*, and make sure you understand the estimates:

```
twoway1 <- lm(bwt ~ visits + smoke, data=birthData)
summary(twoway1)
```

16. Fit the twoway ANOVA model *with interaction* (use the command below). Then use `anova` and/or `drop1` to test if the interaction between visits and smoking habits is statistically significant.

```
twoway2 <- lm(bwt ~ visits * smoke, data=birthData)
```

17. You should still use `twoway2` in for this question. Use `emmeans` to compute the expected birth weight of infants for smokers and non-smokers, respectively, on average over the three levels of `visits`.

### Models with numerical as well as categorical predictors

Predictors of any type can be included in Gaussian linear models, still using `lm`.

18. Fit a model where `lwt` (numeric) and `smoke` (factor) are included as predictors in an additive way:

```
model1 <- lm(bwt ~ lwt + smoke, data=birthData)
summary(model1)
```

What is the interpretation of the estimates?

19. Fit a model with *interaction* between the two predictors, replacing `+` by a `*`:

```
model2 <- lm(bwt ~ lwt * smoke, data=birthData)
summary(model2)
```

What is the interpretation of the estimates? Is there evidence in the data that the effect of mother's weight on infant's weight differs between smokers and non-smokers? *Hint*: What does the last question have to do with interaction? Use `anova` on appropriate models.

20. Fit the model with additive effects of mothers's weight, smoking status, age, and visit status, and carry out model validation. Give an interpretation of the estimate associated to `smoke`. Does smoking affect the weight of infants?

### Logistic regression

The variable `low` is 1 if birth weight is less than 2500 g, and 0 if birth weight is larger than 2500g, see the plot

```
ggplot(birthData, aes(x=bwt, y=low)) + geom_point()
```

Consider for a moment the situation where the actual birth weight (`bwt`) was not registered, such that `low` was the only information on the child. Hence, the outcome is binary (`low` has two values), and the relevant analysis would be a *logistic regression* where the probablity $Pr(low = 1)$ is described in terms of predictors. For example, we may consider a model with mothers's weight, smoking status, age, and visit status as predictors, just like in the previous question.

21. Fit the model, and consider the estimates:

```
logreg1 <- glm(low ~ lwt + smoke + age + visits, data=birthData, family="binomial")
summary(logreg1)
```

Does this model give evidence for an effect of smoking on the weight of infants? Compare the signs of the estimates from `model3` (Gaussian model in the solution for Question 18) and `logreg1` (logistic regression model). Can you explain 'what happens'?

### Linear mixed models

22. Assume for a moment that the 189 births took place at 19 different medical centers with 10 births at each center, except for one center with only nine births. This is not the case, so we have to generate a center variable artificially. You should of course never invent such an artificial structure for a real dataset! Anyway, let's do it like this:

```
set.seed(123)
center <- sample(rep(1:19, each=10)[1:189])
center
birthData <- mutate(birthData,center=factor(center))
```

The `rep` command repeats the number from 1 to 19, 10 times each. We only need the first 189 numbers. The `sample` changes the order of the 189 numbers at random. The `set.seed` command has the effect that you get the same sample each time you run the commands. The last line includes the new variable in the original dataset as a categorical variable.

23. The `center` variable would typically be included in the model as a *random effect*. Gaussian models with both fixed and random effects are called *linear mixed models*, and are fitted with the `lmer` function from the `lme4` package. Run the code below and identify relevant estimates. Remember that `lme4` must be installed before the commands below can be used.

```
library(lme4)
lmm1 <- lmer(bwt ~ lwt + smoke + age + visits + (1|center), data=birthData)
summary(lmm1)
```