Data Exercise in R

HeaDS Data Lab

10 June, 2022

Introduction:

If you don't have your own data to work on, you can practice your newly acquired R skills, using one of the available datasets located in the **Exercises folder**. Please see below for a brief description of each dataset:

Name	Description	Variables	Observations
HealthPrimaryCare	Here	29	453
CrohnsDisease	Here	9	117
StrokeData	Here	12	5110

The sections which follow serve as inspiration for what you could do with the dataset at hand, but you do not need to adhere to these, it is completely up to you what you would like to try out!

Read in the Dataset:

When you have chosen a dataset for analysis, see summary above, the first thing is to read it into R. You can do this the usual way using the `read_excel()` function from `readxl`, as all datasets we have are in the form of excel sheets. Name your dataset something logical.

If you are using your own dataset you will have to figure out which read function you need to get your data into R. There is a function for reading in *almost* any type of text file, `read.csv()`, `read.table()`, `read.delim()`, etc. You can either google your way to an answer or ask one of the course instructors.

Exploratory Analysis

Before making any statistical models, it is necessary to have a good understanding of ones dataset.
So, you could start by looking at:

- How many variables and observations are there in the data? Which data types are the variables.
- Are there any missing values (`NA`), and how many?
- Make some preliminary plots of your data, i.e. these do not have to be beautiful, but they should be informative to you. You could try histograms, boxplots, barplots or something else you would like.
- Looks at your plots, do you appear to have any outlier data points in your dataset?
- Get some summary statistics for your variables, i.e. means, medians, sd, sums (for numerical types) and counts per group (for categorical or factor types).

Scientific Question

Now that you have got en overview of your dataset and the variables in it, try to think about and formulate a simple scientific hypothesis/question which you would like to explore.

Lets use an example of a dataset containing information on patients with breast cancer. Lets say that we follow these patients from diagnosis, though treatment, and 10 years after treatment ends, as follow-up:

Age	BC.subtype	Treatment	Survival	Stage	Smoker	Familial.BC
60	luminal	chemotherapy	alive	2	no	no
51	basal	chemotherapy	alive	3	no	yes
58	luminal	immunotherapy	dead	4	yes	no
...	...	...	...	...	...	...

We need to decide on:

- What is our **response/outcome variable (y)**. Using the example dataset above this might be patient survival or patient subgroup (BC.subtype).
  Q: Which variable in your chosen dataset will you use as response variable? and what type is it, e.g. a categorical (two or more discrete types of outcome) or a continuous variable?
- Decide on our **predictive/explanatory variable (x)**. Which variable in the dataset is predictive of the outcome, in the example above, we might think treatment is predictive of patient survival? Could there be other variables of interest, i.e. covariates, which are predictive of patient survival in addition to treatment?
  Q: Which predictive variables (x) do you have in your datatset?

Data Wrangling

After familiarizing yourself with your dataset and deciding on which variables may be of interest to include in an analysis, you might consider weather you need to do any data wrangling/management before defining your statistical model. Here we mean things like:

- Do you need all variables in the dataset? Maybe some are redundant or uninteresting so you could remove them from the tibble.
- If you have missing values for some observations, do you want to filter out these rows before continuing analysis? There could also be other filtering criteria to thing about, removing rows with $>$ than some number of zeros or some boundary value.
- Maybe you want to make new columns combining some existing ones? An example of this, could be as we did with the downloads dataset where you converted sizes from bytes to megabytes and added the result as a new column to your tibble.
- Do you want to rename the observations of a variables to something more suited for modeling? An examples of this could be converting a categorical variable like status of smoking (noted as *smoker* or *non-smoker*) to integer values (*1* or *0*).

Modelling

You have a *tidy* dataset ready for analysis. You have decided on your response and predictive variables and you know which datatypes they are. The latter is important as this will determine which type of model you can make.

In the statistics exercise we tried out the `lm()` function to do one-way ANOVA. Use this function again to make a simple statistical model.
If the outcome variable you are interested in is continuous, then you will be doing linear regression, however, if the outcome is categorical, you are performing ANOVA (if you only have two groups, this will effectively be a t-test). **N.B** the function and annotation in R is the same for these models. Have a look a the statics exercise/presentation and the pseudo code below:

```
model1 <- lm(resp ~ pred1, data=name_of_dataset)
```

- Begin with a simple model, just one predictive variable.
- Test whether your data fulfills the model assumptions by making residual plots and evaluating these, just like in the presentation and exercise:

```
par(mfrow=c(2,2)) # makes room for 4=2x2 plots.
plot(model1)
```

- If model assumptions are not appropriate could you perform some sort of transformation to help alleviate the problem?
- Get some summary statistics from your model. Can you extract the confidence intervals of your estimate? HINT: `summary()`, `confint()`.
- Try making a model with more than one predictive variable. Remake the residual plots for you new more complex model.
- Which model is best? How could you test this? HINT: `drop1()`, `anova()`.