Exercise 01: Primers

Rohit Koonireddy

24 September, 2023

This exercise will have you working on an Excel file. Your task will be to tidy the data and then read it into Renku's Rstudio.

Download data from Groenigen et al., 2014, containing soil organic matter content data from a meta analysis of CO2 experiments (available on Moodle). Open the file in Excel and navigate to the tab 'Database S1'. You will find a short description in the top-left cell: "Database S1. Overview of CO2 enrichment studies reporting soil C contents that were used in our analysis.". The main issue with this dataset is that .xlsx files are not easily readable into R without an extra package. In addition, even after saving the tab 'Database S1' as a CSV file, the table you get is not **machine-readable** into a data frame that we can work with in R. The way the data is organised into cells does not follow the structure of a dataframe and is not tidy. Recall the **tidy data** rules from the 01_primers.Rmd tutorial.

```
### packages that need to be installed
#install.packages("foreign")
#install.packages("readr")
#install.packages("readxl")
#install.packages("dplyr")
#install.packages("Hmisc")
#install.packages("fastDummies")
#install.packages("tidyr")
#install.packages("vtable")
#install.packages("stargazer")
#install.packages("wooldridge")
#install.packages("gapminder")
#install.packages("ggplot2")
#install.packages("gganimate")
#install.packages("gifski")
#install.packages("av")
#install.packages("sandwich")
#install.packages("lmtest")
#install.packages("fixest")
#install.packages("broom")
#install.packages("modelsummary")
#install.packages("AER")
#install.packages("car")
#install.packages("openxlsx")

library("foreign");
library("readr");
library("readxl");
library("dplyr");
library("Hmisc");
library("fastDummies");
```

```
library("tidyr");
library("vtable");
library("stargazer");
library("wooldridge");
library("gapminder");
library("ggplot2");
library("gganimate");
library("gifski");
library("av");
library("sandwich");
library("lmtest");
library("fixest");
library("broom");
library("modelsummary");
library("AER");
library("car");
library("openxlsx");
```

Your task is to:

1. Manually manipulate the .xlsx file to make it tidy.

2. Save the data as a .csv file (comma-separated-values).

3. Read the .csv file into RStudio.

```
# enter your solution here
setwd("Exercise 1") #chage it to directory wherever your cleaned file is located
sampleData <- read.xlsx("groenigen14sci_cleaned.xlsx");
write.csv(sampleData, file = "groenigen14sci_cleaned.csv", row.names = FALSE);
sampleDataCsv <- read.csv("groenigen14sci_cleaned.csv");
head(sampleDataCsv)
```

```
##                     Experiment                          Citation   Depth
## 1 ArizonaFACE - wheat - high N              Leavitt et al. 2001 0-15 cm
## 2 ArizonaFACE - wheat - high N              Leavitt et al. 2001 0-15 cm
## 3 ArizonaFACE - wheat - high N              Leavitt et al. 2001 0-15 cm
## 4 ArizonaFACE - wheat - high N              Leavitt et al. 2001 0-15 cm
## 5                  Biosphere 2 Trueman and  Gonzalez-Meler 2005 0-25 cm
## 6                  Biosphere 2 Trueman and  Gonzalez-Meler 2005 0-25 cm
##    Sample.date Time..years. ambient.CO2..mean..g.C.m.2
## 1    nov. 1995          0.0                    1292.840
## 2     may 1996          0.5                    1453.080
## 3    dec. 1996          1.0                    1312.560
## 4     may 1997          1.5                    1246.000
## 5 average 1999          0.0                    5863.000
## 6 average 2000          1.0                    5255.571
##   increased.CO2..mean..g.C.m.2 ambient.CO2..number. increased.CO2..number.
## 1                    1172.400                    4                      4
## 2                    1350.360                    4                      4
## 3                    1176.920                    4                      4
## 4                    1472.080                    4                      4
## 5                    6391.434                    1                      1
## 6                    4698.132                    1                      1
##   Description.of.data.source
## 1                    fig. 3
```

```
## 2                     fig. 3
## 3                     fig. 3
## 4                     fig. 3
## 5                     fig. 3
## 6                     fig. 3
##                                                 Value.treatment
## 1                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 2                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 3                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 4                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 5 bulk density: 0.91 g/cm3 for ambient, 0.97 for 800 ppm (table 2)
## 6 bulk density: 0.91 g/cm3 for ambient, 0.97 for 800 ppm (table 2)
```

4. Calculate the logarithmic response ratio as the logarithm of the ratio of soil C contents at elevated CO2 divided by soil C contents at ambient CO2, for each data point (experiment and sample date).

```
colnames(sampleDataCsv)
```

```
##  [1] "Experiment"                   "Citation"
##  [3] "Depth"                        "Sample.date"
##  [5] "Time..years."                 "ambient.CO2..mean..g.C.m.2"
##  [7] "increased.CO2..mean..g.C.m.2" "ambient.CO2..number."
##  [9] "increased.CO2..number."       "Description.of.data.source"
## [11] "Value.treatment"
```

```
# enter your solution here
sampleDataCsv$log_response_ratio <-
  log(sampleDataCsv$"increased.CO2..mean..g.C.m.2" / sampleDataCsv$"ambient.CO2..mean..g.C.m.2")
head(sampleDataCsv)
```

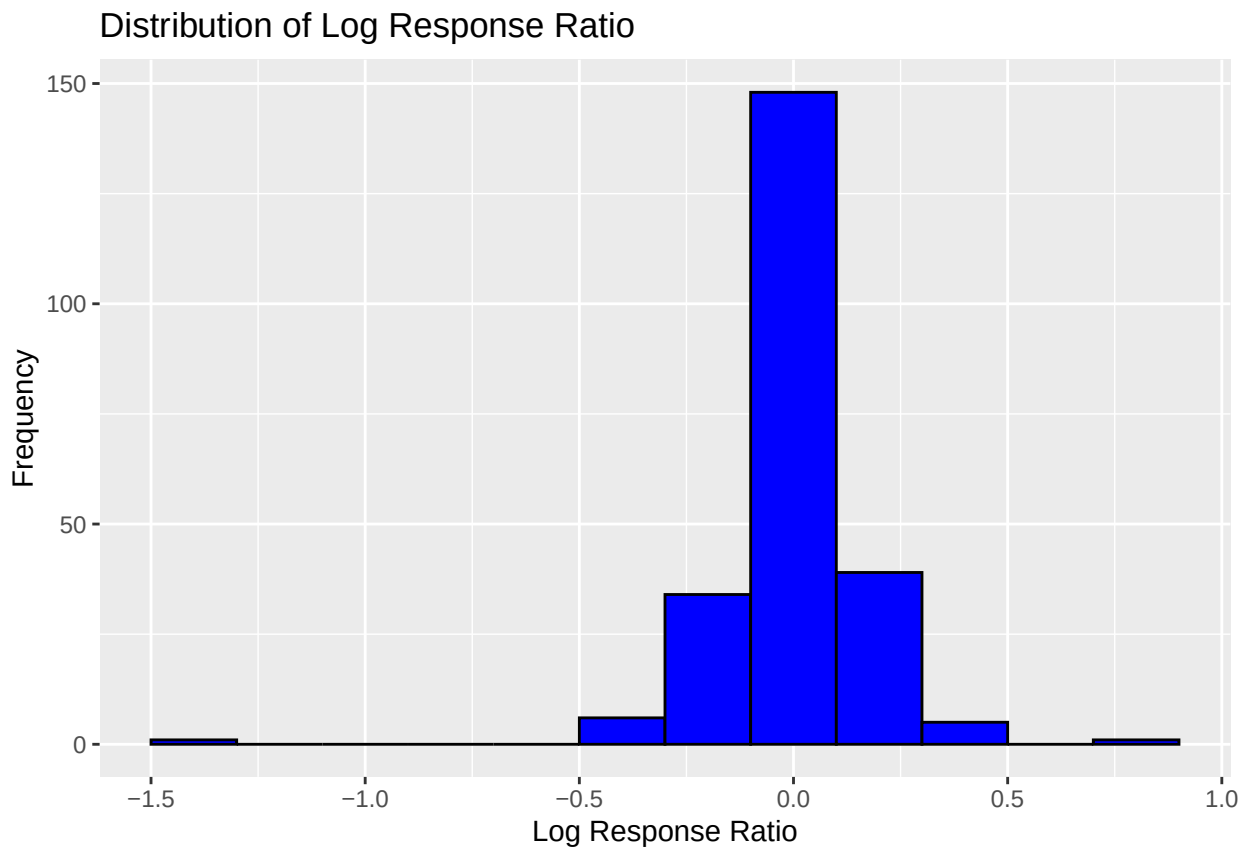
```
##                    Experiment                        Citation   Depth
## 1 ArizonaFACE - wheat - high N             Leavitt et al. 2001 0-15 cm
## 2 ArizonaFACE - wheat - high N             Leavitt et al. 2001 0-15 cm
## 3 ArizonaFACE - wheat - high N             Leavitt et al. 2001 0-15 cm
## 4 ArizonaFACE - wheat - high N             Leavitt et al. 2001 0-15 cm
## 5                 Biosphere 2 Trueman and  Gonzalez-Meler 2005 0-25 cm
## 6                 Biosphere 2 Trueman and  Gonzalez-Meler 2005 0-25 cm
##    Sample.date Time..years. ambient.CO2..mean..g.C.m.2
## 1    nov. 1995          0.0                   1292.840
## 2     may 1996          0.5                   1453.080
## 3    dec. 1996          1.0                   1312.560
## 4     may 1997          1.5                   1246.000
## 5 average 1999          0.0                   5863.000
## 6 average 2000          1.0                   5255.571
##   increased.CO2..mean..g.C.m.2 ambient.CO2..number. increased.CO2..number.
## 1                    1172.400                    4                       4
## 2                    1350.360                    4                       4
## 3                    1176.920                    4                       4
## 4                    1472.080                    4                       4
## 5                    6391.434                    1                       1
## 6                    4698.132                    1                       1
##   Description.of.data.source
## 1                     fig. 3
## 2                     fig. 3
## 3                     fig. 3
```

```
## 4                     fig. 3
## 5                     fig. 3
## 6                     fig. 3
##                                                 Value.treatment
## 1                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 2                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 3                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 4                        assumed BD= 1.48 g/cm3 (Post et al. 1988)
## 5 bulk density: 0.91 g/cm3 for ambient, 0.97 for 800 ppm (table 2)
## 6 bulk density: 0.91 g/cm3 for ambient, 0.97 for 800 ppm (table 2)
##   log_response_ratio
## 1        -0.09778842
## 2        -0.07331422
## 3        -0.10907857
## 4         0.16673795
## 5         0.08629727
## 6        -0.11212385
```

5. Visualise the distribution of the response ratio and save the plot as a .pdf file.

```r
# enter your solution here
response_ratio_plot <- ggplot(sampleDataCsv, aes(x = log_response_ratio)) +
  geom_histogram(binwidth = 0.2, fill = "blue", color = "black") +
  labs(title = "Distribution of Log Response Ratio",
       x = "Log Response Ratio",
       y = "Frequency")

# Print the plot
print(response_ratio_plot)
```




```
# Save the plot as a .pdf file
ggsave("response_ratio_distribution.pdf", plot = response_ratio_plot)
```

6. Export the new data frame in csv file.

```
# enter your solution here
print("saving final file")
```

```
## [1] "saving final file"
```

```
write.csv(sampleDataCsv, file = "final_data.csv", row.names = FALSE)
```