Week 2.6 Practical: Correlations and Linear Regressions

ADS2

Semester 2, 2023/24

Learning Objectives

After completing this problem set you will be able to:

- perform correlation and regression analysis in R
- understand the differences between these analyses and how to interpret their different outputs

This **Practical** contains a description of the two datasets used, together with examples of how to handle the data followed by **Tasks**.

Within the R Markdown file (*.Rmd*) you can find additional lines of code that were used to create the datasets as well as functions and data processing that you could find useful.

This part is compulsory and you are expected to complete the following exercises during the two hours allocated to the Practical session.

COVID-19

In this exercise you will be able to explore real data from COVID-19 cases and vaccinations. The script will download and name the datasets but in case of problems, the links are the following:

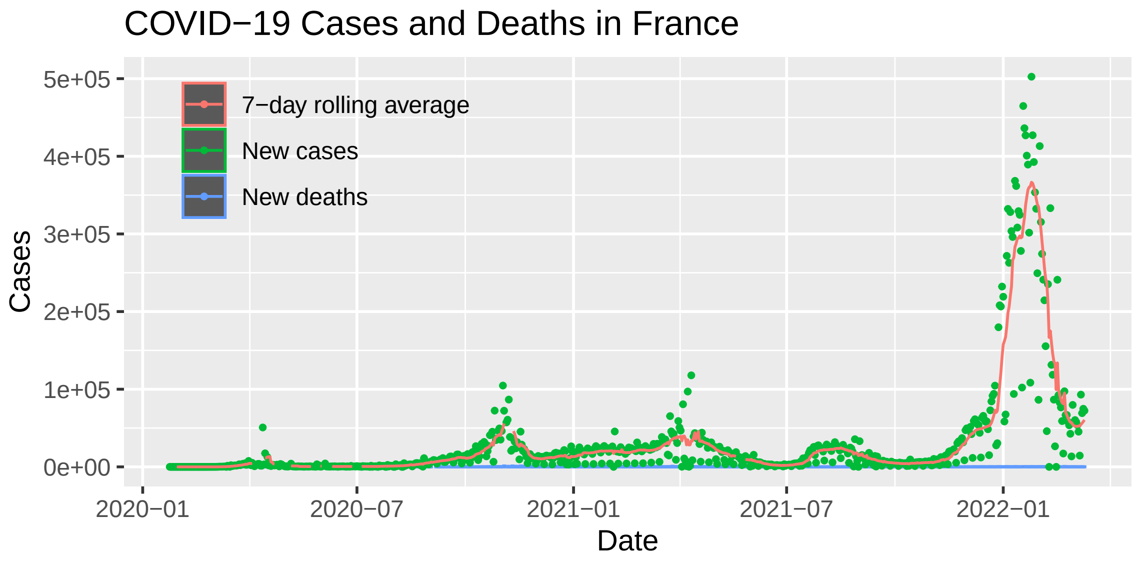
Dataset: https://github.com/hugocarlos/covid-19-data/blob/master/public/data/owid-covid-data.csv
Explanation on headers: https://github.com/hugocarlos/covid-19-data/blob/master/public/data/owid-covid-codebook.csv

```r
suppressMessages(library(ggplot2))
suppressMessages(library(dplyr))
```

```r
# Function that generates a vector with x-days rolling average for a given country and
# a given variable
# USAGE: generate_rolling_avg(subcovid, "France", "new_cases", 7)
# one_country = "France"; one_variable = "new_cases"; days = 7
generate_rolling_avg <- function(subcovid, one_country, one_variable, days = 7){
  range_days_in_one_country <- range(subcovid$date[which(subcovid$location == one_country)])
  # Identifying the dates present in subcovid
  dates_included <- seq(range_days_in_one_country[1], range_days_in_one_country[2],
                        by = "days")
  # Calculating 7-day rolling mean
  variable_means <- sapply(dates_included[-(1:6)], function(end_of_the_week){
    # end_of_the_week <- dates_included[7]
    x_days_cases <- sapply(-6:0, function(y){
      # y <- -6
      subcovid[which(subcovid$location == one_country & subcovid$date == (end_of_the_week + y)),
               one_variable]
    })
    mean(x_days_cases)
  })
  variable_means_df <- data.frame(Dates = dates_included[-(1:6)],
                                  new_variable_avg = variable_means)
}
```

The dataset includes daily data on new COVID-19 cases and deaths. This data can be used to explore the incidence of COVID cases in France during the two years of the pandemic:

```
trying <- try(covid <- read.csv("owid-covid-data.txt", header = TRUE))
if(is(trying, "try-error")){
  download.file(url = paste0("https://github.com/hugocarlos/covid-19-data/blob/master/",
                             "public/data/owid-covid-data.csv?raw=true"),
                             destfile = "owid-covid-data.txt")
  covid <- read.csv("owid-covid-data.txt", header = TRUE)
}
# Selecting some columns
subcovid <- covid %>%
  select(iso_code, location, date, new_cases, new_deaths, new_cases_per_million,
         total_cases_per_million, new_vaccinations, people_fully_vaccinated,
         aged_65_older, aged_70_older, gdp_per_capita, extreme_poverty,
         cardiovasc_death_rate, diabetes_prevalence, life_expectancy,
         human_development_index)

# To date format
subcovid$date <- as.Date(subcovid$date)

# Setting one country
one_country <- "France"

# Calculating the 7-days window average for new cases of COVID-19
cases_means_df <- generate_rolling_avg(subcovid, one_country, "new_cases", 7)

# Merging cases_means_df to subcovid
subcovid$new_cases_avg <- NA
for(i in 1:nrow(cases_means_df)){
  # i <- 1
  subcovid$new_cases_avg[which(subcovid$location == one_country &
                               subcovid$date == cases_means_df$Dates[i])] <-
    cases_means_df$new_variable_avg[i]
}

# Plot
ggplot() +
  geom_bar(stat = "identity",
           aes(x = subcovid$date[which(subcovid$location == one_country)],
               y = subcovid$new_deaths[which(subcovid$location == one_country)],
               colour = "New deaths")) +
  geom_point(aes(x = subcovid$date[which(subcovid$location == one_country)],
                 y = subcovid$new_cases[which(subcovid$location == one_country)],
                 colour = "New cases"), size = 0.7) +
  geom_line(aes(x = subcovid$date[which(subcovid$location == one_country)],
                y = subcovid$new_cases_avg[which(subcovid$location == one_country)],
                colour = "7-day rolling average")) +
  labs(x = "Date", y = "Cases") +
  ggtitle(paste0("COVID-19 Cases and Deaths in ", one_country)) +
  theme(legend.position = c(0.2, 0.8),
        legend.title = element_blank(),
        legend.background = element_blank(),
        legend.key = element_rect(fill = NULL, color = NULL))
```

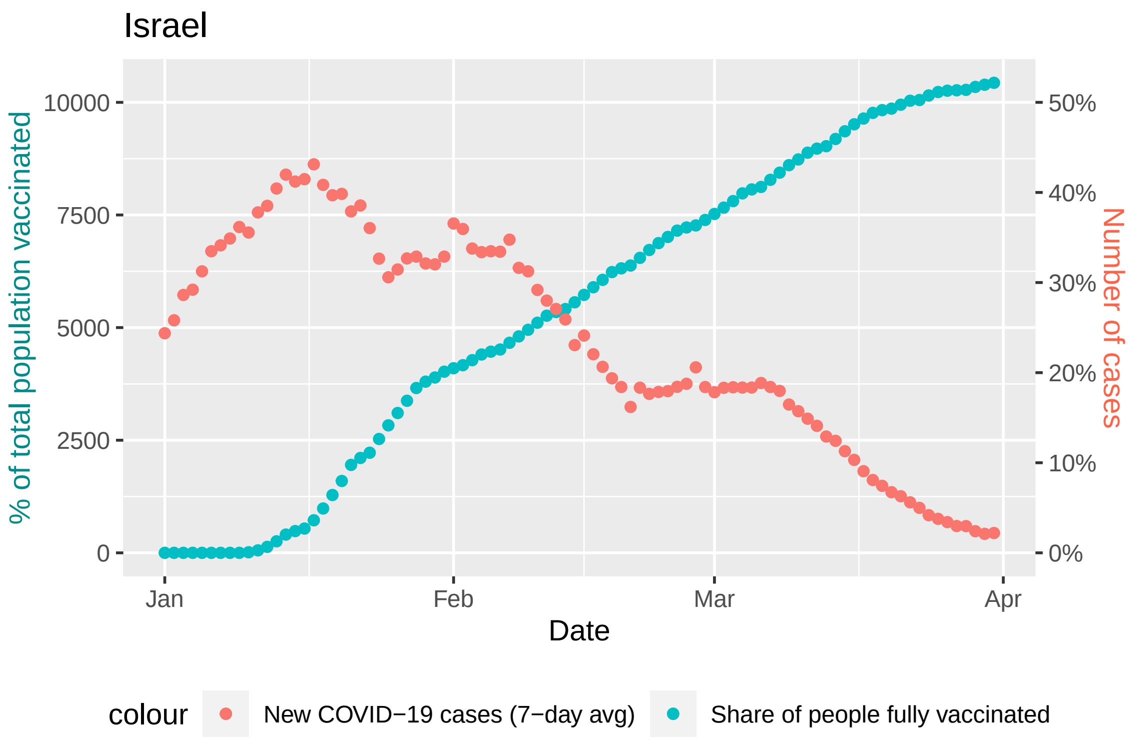


The vaccination data can be included with the number of cases, using a 7-day rolling average, data from Israel from January to March 2021 is plotted:

```
# Finding all the days from 2021, as they probably contain vaccination data
dates_from_2021 <- seq(as.Date("2021-01-01"), as.Date("2021-12-31"), by = "days")

one_country <- "Israel"
# Re-calculating the vector with 7-day rolling average of new COVID-19 cases
cases_means_df <- generate_rolling_avg(subcovid, one_country, "new_cases", 7)
```

```
# Attaching the file with the population
trying <- try(population <- read.csv("WBpopulation.csv", header = TRUE, sep = "\t"))
if(is(trying, "try-error")){
  download.file(url = paste0("https://raw.githubusercontent.com/hugocarlos/public_scripts/",
                             "master/teaching/WBpopulation.csv"),
                             destfile = "WBpopulation.csv")
  population <- read.csv("WBpopulation.csv", header = TRUE, sep = "\t")
}

# Calculating the percentage of the population fully vaccinated
Israel_population <- population$X2020[which(population$Country.Name == one_country)]
```

```
subcovid$share_fully_vaccinated <- subcovid$people_fully_vaccinated * 100 / Israel_population

# Merging cases_means_df to subcovid
subcovid$new_cases_avg <- NA
for(i in 1:nrow(cases_means_df)){
  # i <- 1
  subcovid$new_cases_avg[which(subcovid$location == one_country &
                               subcovid$date == cases_means_df$Dates[i])] <-
    cases_means_df$new_variable_avg[i]
}

subcovid %>%
  filter(location == one_country) %>%
    filter(date >= dates_from_2021[1] & date < as.Date("2021-04-01")) %>%
  ggplot() +
  geom_point(aes(x = date, y = share_fully_vaccinated * 200,
                 colour = "Share of people fully vaccinated")) +
  geom_point(aes(x = date, y = new_cases_avg, colour = "New COVID-19 cases (7-day avg)")) +
  scale_y_continuous(name = "% of total population vaccinated",
                     sec.axis = sec_axis(~./200, name = "Number of cases",
                                         labels = function(b){
                                           paste0(b, "%")
                                          })) +
  xlab("Date") +
  theme(axis.title.y = element_text(color = "cyan4"),
        axis.title.y.right = element_text(color = "tomato"),
        legend.position = "bottom") +
  ggtitle(paste0(one_country))
```



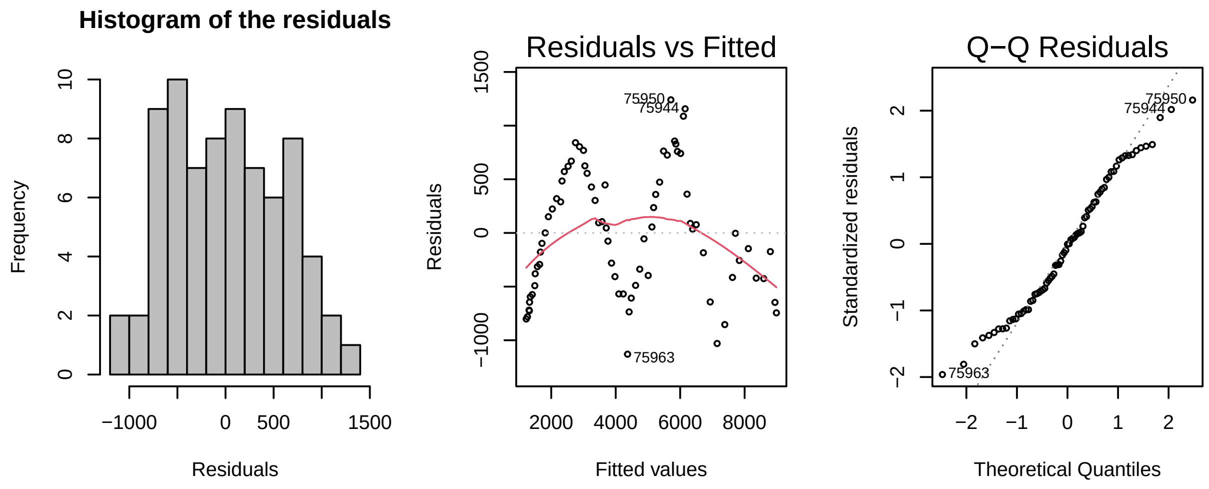
The previous plot suggests that there might be an effect on vaccination after the first half of January 2021, lasting until the end of April. Can the number of new COVID-19 cases be explained by the vaccination progress?

```r
covid_onecountry <- subcovid[which(subcovid$location == "Israel" &
                                     subcovid$date >= as.Date("2021-01-15") &
                                     subcovid$date < as.Date("2021-03-31")), ]
# Calculating the Correlation Coefficient
cor(covid_onecountry$new_cases_avg,
    covid_onecountry$share_fully_vaccinated,
    use = "complete.obs")
```

```
## [1] -0.9692676
```

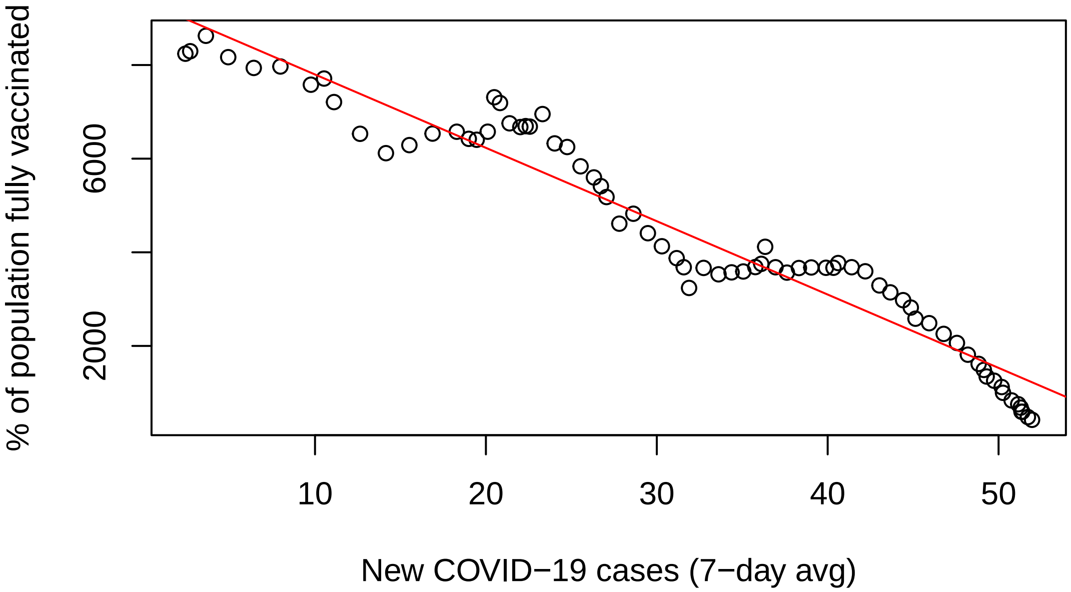
If we want to calculate the simple linear regression of cases as the result of vaccinations, we need to verify that the assumptions of the Least Squares are satisfied.

```r
lm_Israel <- lm(new_cases_avg ~ share_fully_vaccinated, covid_onecountry)
par(mfrow = c(1, 3))
hist(residuals(lm_Israel), breaks = 15, col = "gray",
     main = "Histogram of the residuals", xlab = "Residuals", cex = 0.6)
plot(lm_Israel, which = c(1, 2), cex = 0.6)
```



The Residuals vs Fitted values plot reveals what we saw in the cases plot: there are short ranges where the linearity is partially lost. For this time, we will proceed to see the linear regression.

```r
plot(x = covid_onecountry$share_fully_vaccinated, y = covid_onecountry$new_cases_avg,
     xlab = "New COVID-19 cases (7-day avg)",
     ylab = "% of population fully vaccinated")
abline(lm_Israel, col = "red")
```


% of population fully vaccinated
2000
6000
10
20
30
40
50
New COVID−19 cases (7−day avg)

Tasks

- Do you consider this correlation good enough to infer that the vaccination in Israel had an effect on the number of new infections during the first three months of 2021?
- What other factors would you add to the linear regression to improve the fit?
- Can you find another country (or countries) where the correlation between the share of the population that have received one or the full number of doses closely correlates with the recent number of new COVID-19 infections?
- Consider a restricted period of time (around Winter or around Summer), not all countries might be equally affected by the change of the seasons.
- Choose one country and develop a method that uses local linear regressions to assess when during the COVID-19 epidemic it is observed an increase, a decrease or no change in the number of new cases. The output could be a table showing that for some periods of time (for example from October to mid November) there was an increase or a decrease in cases.

Previous versions by author: Hugo Samano, Dmytro Shytikov, Zhaoyuan Fang, Jingyuan Chen

Last update by DJ MacGregor in 2024