Assignment 2: Linear and Logistic Regression
Introduction to Applied Data Science
2022-2023

Bas Machielsen
a.h.machielsen@uu.nl

April 2023

Assignment 2: Linear and Logistic Regression to Predict Poverty

1. Preparation

In this assignment, you will build your own linear and logistic regression models, using the gradient descent algorithms we talked about in the lecture. Similarly to the previous assignment, you will fill up the code chunks left empty in this document, and you will interpret them in this document. To start with, please replace my name and e-mail address with yours. Then, remove the lines:

```
output:
  pdf_document:
    includes:
      in_header: "preamble.tex"
    latex_engine: xelatex
```

from the document, and replace them with:

```
output: pdf_document
```

2. Obtain the required data

Next, we are going to make use of the `wbstats` package, which you have seen before. If you haven't done so already, you can install the package with `install.packages('wbstats')`. Make sure not to put this in your Rmarkdown document, as R will then attempt to install this package every time you knit your document.

The `wbstats` package allows you to navigate the World Bank database, and download datasets without having to visit the World Bank website, download the data into a spreadsheet, and subsequently load it into R. With the help of this package, we just download the data into R right away.

Question 1: Read this so-called vignette to find out how to navigate the World Bank data using the `wbstats` package. Download the variable that measures the poverty headcount index, the proportion of the population with daily per capita income (in 2011 PPP) below the poverty line at $1,90.

```
library(wbstats)
all_poverty_variables <- wbstats::wb_search("poverty")

View(all_poverty_variables)
```

Change `eval = FALSE` to `eval = TRUE` below once you've found the right answer.

In addition, we will also load the `tidyverse` package, as we're going to do some data wrangling to glue various pieces of data together:

```
library(tidyverse)
```

```
poverty_data <- wb_data("1.0.HCount.1.90usd")
```

Question 2: For which countries is this data available? Print a vector of the unique country names. Do not type a string variable of all the country names yourself, but use code to extract this.

```
poverty_data %>% select(country) %>% pull() %>% unique()
```

```
##  [1] "Argentina"          "Bolivia"            "Brazil"
##  [4] "Chile"              "Colombia"           "Costa Rica"
##  [7] "Dominican Republic" "Ecuador"            "Guatemala"
## [10] "Honduras"           "Mexico"             "Nicaragua"
## [13] "Panama"             "Peru"               "Paraguay"
## [16] "El Salvador"        "Uruguay"            "Venezuela, RB"
```

Next, save the *unique* `iso2c` and `iso3c` codes in two separate vectors:

```
iso2c <- unique(poverty_data %>% select(iso2c)) %>% pull()

iso3c <- unique(poverty_data %>% select(iso3c)) %>% pull()
```

Next, we will download datasets related to varieties of democracy.

```
#install.packages("devtools")
devtools::install_github("xmarquez/vdem")
```

Question x: Find out what the median is of the $1.90 poverty head count variable in the dataset you downloaded. Make a new dummy variable `povyesno`, meaning:

$$\text{povyesno}_i = \begin{cases} 1 & \text{if 1.0.HCount.1.90usd } > \text{median(1.0.HCount.1.90usd)} \\ 0 & \text{otherwise} \end{cases}$$