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 predicting poverty, 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
```

To start building statistical models that predict poverty, we need to proceed in the following way. We will predict poverty in a simple setting, and predict poverty on the basis of educational attainment and democracy scores. We will start from scratch, meaning that we ourselves will look for relevant data sources, put them together, and build some simple statistical models. Finally, we will check our solutions against standard benchmarks, and will attempt to interpret them. We will also set a seed to make answers comparable:

```
set.seed(2)
```

You have to hand in a **pdf** knitted `.Rmarkdown` document on Blackboard before the deadline. The basis of the document is the present document, which you will fill in yourself with your own code snippets, answers to open questions, and code output (figures, tables, etc.).

2. Obtain the required poverty data

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 x: 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 ratio at national poverty lines (% of population)*", the proportion of the population with daily per capita income below the national poverty line.

```
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.

```
poverty_data <- wb_data("SI.POV.NAHC")
```

Now, 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)
```

Question x: Remove the `NA` observations from the dataset. Hint: you can use `na.omit`, but you have to `select` only a couple of columns.

```
poverty_data <- poverty_data %>%
  select(1:5) %>%
  na.omit()
```

Question x: For which countries is this data available? Print a vector of the first 10 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() %>% head(10)
```

```
##  [1] "Afghanistan" "Albania"     "Algeria"     "Angola"      "Argentina"
##  [6] "Armenia"     "Austria"     "Azerbaijan"  "Bangladesh"  "Belarus"
```

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()
```

For convenience, we might want to rename our poverty variable. One way to do this is as follows:

```
poverty_data <- poverty_data %>%
  rename(poverty = `SI.POV.NAHC`)
```

The `rename` function follows the `new = old` syntax, which you can verify by typing `?rename` in the console.

3. Obtain the required democracy data

Next, we will download datasets related to varieties of democracy. This data is contained in the Github package `democracyData`. The package is not available on the official R package repository CRAN, but it a custom package which is downloadable from somebody's github repository. To download packages from Github, we need the `devtools` package, and then, we can install the package using the `install_github()` function.

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

After installing the package, we want to load a dataset from the package. To get a sense of what the data looks like, you can have a look here. The particular dataset we will use is called the `anckar` dataset, and contains data used in Anckar and Fredriksson (2018).

```
anckar <- democracyData::anckar
```

Have a look at the Anckar and Fredriksson dataset we just downloaded:

```
head(anckar)
```

```
## # A tibble: 6 x 13
##   anckar~1 ancka~2 abbre~3  year democ~4 monar~5 regim~6 regim~7 popel~8 exten~9
##   <chr>      <dbl> <chr>   <dbl>   <dbl>   <dbl> <fct>   <fct>     <dbl> <chr>
## 1 UNITED ~       2 USA      1800       1       0 Presid~ Presid~       1 United~
## 2 UNITED ~       2 USA      1801       1       0 Presid~ Presid~       0 United~
## 3 UNITED ~       2 USA      1802       1       0 Presid~ Presid~       0 United~
## 4 UNITED ~       2 USA      1803       1       0 Presid~ Presid~       0 United~
## 5 UNITED ~       2 USA      1804       1       0 Presid~ Presid~       0 United~
## 6 UNITED ~       2 USA      1805       1       0 Presid~ Presid~       1 United~
## # ... with 3 more variables: GWn <dbl>, cown <int>, in_GW_system <lgl>, and
## #   abbreviated variable names 1: anckar_country, 2: anckar_ccode,
## #   3: abbreviation, 4: democracy, 5: monarchy, 6: regimebroadcat,
## #   7: regimenarrowcat, 8: popelection, 9: extended_country_name
```

The dataset contains a couple of indicators whether a country is a democracy at a given point in time as judged by the authors mentioned above. There is also some miscellaneous information. However, if we compare this dataset to the `poverty_data` dataset, we notice that some countries have different names. It will therefore be difficult to bring them together. Fortunately, the `democracyData` dataset contains a function called `country_year_coder`, which allows us to look up country code. This is where the `iso2c` (or `iso3c`) variable comes in handy.

```
anckar <- democracyData::country_year_coder(anckar,
                                            country_col = anckar_country,
                                            date_col = year,
                                            code_col = anckar_ccode,
                                            include_in_output = "iso2c")
```

As you may have noticed, by executing this command, we have supplemented our `data.frame` with an `iso2c` column. This, in turn, allows us to match (better terminology: merge) the observations according to the observations in the `poverty_data` dataset.

But before that, we remove variables we do not need: we want to select the variables `iso2c`, `year`, and `democracy`, and merge them to the `poverty_data` dataframe.

Question x: Select these variables and write them to a `data.frame` called `to_be_merged`.

```
to_be_merged <- anckar %>% select(iso2c, year, democracy)
```

Question x: Use the `left_join` command to merge these two variables to the `poverty_data` dataframe. Read `?left_join` to correctly specify the `by` argument in this function. Make sure to match on two variables: the `iso2c` codes and the years.

```
data <- left_join(poverty_data, to_be_merged, by = c("iso2c" = "iso2c", "date" = "year"))
```

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

$$\text{povyesno}_i = \begin{cases} 1 & \text{if Poverty } > \text{median(Poverty)} \\ 0 & \text{otherwise} \end{cases}$$

You can easily do so using the `if_else` function from `dplyr`. Check `?if_else` for its syntax.

```
data <- data %>%
  mutate(povyesno = if_else(poverty > median(poverty, na.rm = TRUE), 1, 0))
```

4. Obtain the required educational data

Next, we will want to integrate information about educational attainment. Fortunately, the *World Bank* also collects many different kinds of data on educational attainment. For an impression, have a look at:

```
educ_search <- wbstats::wb_search("educational attainment")
```

For this setting, we'll take the variable `SE.PRM.CUAT.ZS`, which is the fraction of the population aged 25 and over that at least complete primary education in a particular country in a particular year.

Question x: Download this variable and store it in a dataset called `education` using the function `wb_data`.

```
education <- wb_data('SE.PRM.CUAT.ZS')
```

Question x: Select the first five variables and then omit the `NA` observations from this dataset. Also rename the variable to `educ`.

```
education <- education %>% select(1:5) %>% na.omit() %>% rename(educ = `SE.PRM.CUAT.ZS`)
```

Question x: Now use `left_join` again to merge your already assembled `data` with the `education` data:

```
data <- left_join(data, education, by = c("iso2c" = "iso2c", "date" = "date"))
```

There are a couple of unnecessary variables in your dataset, and the merges have caused some `NA`s. They can be removed using the `select` function in the following way:

```
data <- data %>%
  select(-iso3c.x, -country.x, -iso3c.y, -country.y) %>%
  na.omit()
```

Now we are ready to train a statistical model.

5. Building a linear regression model

Now, we want to predict poverty by means of two factors: (i) historical educational investment, and (ii) democracy. We have just undertaken various efforts to collect all these data in the data.frame called `data`. Next, we'll proceed to estimate a statistical model to predict poverty. In particular, we will estimate the following model:

$$\text{poverty}_i = \alpha + \beta_1 \cdot \text{democracy}_i + \beta_2 \cdot \text{education}_i + \epsilon_i$$

This model estimates the coefficients α, β_1 and β_2. Do you think β_1 and β_2 will be greater than zero, smaller than zero, or zero, and why?

It turns out that this linear model can be solved using the gradient descent algorithm. This means that we are describing a loss function, randomly initialize parameters, and then compute the gradient, and move the parameters in the direction *opposite* the largest increase.

In this case, the loss function we use is:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

where $\hat{y}_i = \alpha + \beta_1 \cdot \text{democracy}_i + \beta_2 \cdot \text{education}_i$, the predicted value for y_i, given values of education and democracy, and parameter values α, β_1, β_2. It is our job to set the parameters such that our loss function is minimized. To do so, we employ gradient descent. This involves taken the derivative of the loss function with respect to the parameters α, β_1, β_2. Let's focus on β_1, then you can do the remainder by yourself.

$$\frac{\partial L}{\partial \beta_1} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \beta_1} = -\frac{1}{n}\sum_{i=1}^{n} 2(y_i - \hat{y}) \cdot x_i$$

In practice, people often leave out the -2 constant term, and write the gradient as follows:

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i) \cdot x_i$$

..so that the following updating rule for the β-coefficients can be determined:

$$\beta^i + 1 = \beta^i - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i) \cdot x_i$$

Below is an implementation of such an algorithm. You do not have to change much, but you have to implement the gradients for β_1 and β_2 yourself.

```
gradient_descent <- function(x, y, learn_rate, conv_threshold, n, max_iter) {
  beta1 <- runif(1, 0, 1)
  beta2 <- runif(1, 0, 1)
  alpha <- runif(1, 0, 1)
  yhat <- beta1 * x[,1] + beta2 * x[,2] + alpha
  MSE <- sum((y - yhat) ^ 2) / n
  converged = F
  iterations = 0
  while(converged == F) {
    ## Implement the gradient descent algorithm
    ## Complete the derivatives in the three lines below
    beta1_new <- beta1 - learn_rate * ((1 / n) * (sum((yhat - y) * x[,1])))
    beta2_new <- beta2 - learn_rate * ((1 / n) * sum((yhat - y) * x[,2]))
    alpha_new <- alpha - learn_rate * ((1 / n) * (sum(yhat - y)))

    beta1 <- beta1_new
```

```
    beta2 <- beta2_new
    alpha <- alpha_new

    yhat <- beta1 * x[,1] + beta2 * x[,2] + alpha
    MSE_new <- sum((y - yhat) ^ 2) / n

    if(MSE - MSE_new <= conv_threshold) {
      converged = T
      return(paste("Converged. Optimal intercept:", alpha, "Optimal slopes:", "\n",
                   beta1,  beta2))
    }
    iterations = iterations + 1
    if(iterations > max_iter) {
      converged = T
      return(paste("Optimal intercept:", alpha, "Optimal slopes:", "\n",
                   beta1, beta2))
    }
  }
}
```

Next, we'll run the algorithm. You'll get an output containing three numbers, the estimated $\alpha, \beta_1, and \beta_2$, corresponding to a constant term, `education`, `democracy` respectively.

```
linear_gd <- gradient_descent(cbind(data$educ, data$democracy),
                  data$poverty,
                  learn_rate = 0.0001,
                  conv_threshold = 0.001,
                  n = length(data$poverty),
                  2500000)
```

Question x: Print the three numbers you got in your RMarkdown document. Note: the algorithm might take a couple of minutes to run.

6. Building a logistic regression model

Finally, we'll analyze the same data using a logistic regression model. This time, however, we will analyze not the `poverty` variable, but a dichotomized version of it: `povyesno`, which you created before. A key element in a logistic regression model is the so-called *sigmoid* function. The sigmoid function $S(x)$ is defined as follows:

$$S(x) = \frac{1}{1 + e^{-x}}$$

Question x: Implement the sigmoid function as a function in R

```
sigmoid <- function(x){
  1/(1+exp(-x))
}
```

Question x: Differentiate the sigmoid function with respect to x. Rewrite this derivative to show that it is equal to $S(x)(1 - S(x))$ using the definition of $S(x)$ given above.

Question x: Implement the derivative of the sigmoid function in R.

```r
sigmoid_derivative <- function(x){

  sigmoid(x) * (1-sigmoid(x))

}
```

A nice thing about this sigmoid function is that its outcomes always fall between 0 and 1. We'll use the sigmoid function to build a logistic regression model. In particular, instead of a simple argument x, we are predicting the outcome, poverty, on the basis of the same two independent variables as before, and two corresponding coefficients. Our sigmoid function would thus look like this:

$$\text{Poverty}_i = \frac{1}{1 + e^{-[\alpha + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{stab}_i]}}$$

Question x: Use the chain rule from calculus to identify the derivatives of the above particular sigmoid function with respect to education$_i$ and stability$_i$ and α. Hint: use the result from above, and use that:

$$\frac{\partial S(.)}{\partial \text{education}_i} = \frac{\partial S(x)}{\partial x} \cdot \frac{\partial x}{\partial \text{education}_i}$$

Write the results in an equation delineated by `$$ $$`, or take a picture with your phone, upload it and insert it in the document by typing `![](\link_to_picture.png)` in the document.

In the next few steps, we want to implement these functions in our gradient descent algorithm. We can largely use the same setup as we used in the linear model, with the exception of the derivatives, because the loss function you use is slightly different.

Question x: fill the derivatives in this algorithm yourself.

```r
gradient_descent <- function(x, y, learn_rate, conv_threshold, n, max_iter) {
  beta1 <- runif(1, -0.5, 0.1)
  beta2 <- runif(1, -0.5, -0)
  alpha <- runif(1, 4, 5)

  yhat <- sigmoid(alpha + beta1 * x[,1] + beta2 * x[,2])
  loss <- -(1 / n) * sum(y * log(yhat) + (1-y)* log(1-yhat))

  converged = F
  iterations = 0

  while(converged == F) {
    ## Complete the derivatives in the three lines below
    beta1_new <- beta1 - learn_rate * ((1 / n) * (sum((yhat - y) * x[,1])))
    beta2_new <- beta2 - learn_rate * ((1 / n) * sum((yhat - y) * x[,2]))
    alpha_new <- alpha - learn_rate * ((1 / n) * (sum(yhat - y)))

    beta1 <- beta1_new
    beta2 <- beta2_new
    alpha <- alpha_new

    yhat <- sigmoid(alpha + beta1 * x[,1] + beta2 * x[,2])
```

```
    loss_new <- -(1 / n) * sum(y * log(yhat) + (1-y)* log(1-yhat))

    if(loss - loss_new <= conv_threshold) {
      converged = T
      return(paste("Converged. Optimal intercept:", alpha,
                   "Optimal slopes:", beta1, "\n", beta2))
    }
    iterations = iterations + 1
    if(iterations > max_iter) {
      converged = T
      return(paste("Optimal intercept:", alpha,
                   "Optimal slopes:", beta1, "\n", beta2))
    }
  }
}
```

Now we can use this algorithm to run our gradient descent algorithm, this time for a logistic regression model. You can expect it to take about a minute:

```
logit_gd <- gradient_descent(cbind(data$educ, data$democracy), data$povyesno,
                 learn_rate = 0.001,
                 conv_threshold = 0.01,
                 n = length(data$povyesno),
                 max_iter = 1.5e6)
```

7. Evaluate your output and compare it to standard solutions

In this section, we'll think about the output of our analysis and the interpretation of our results. Both linear regression and logistic regression can also be performed using an analytical solution for the minimum loss. For linear regression, this function is implemented in R in the `lm` function. Check the syntax for the `lm` method using `?lm`.

Question x: Verify that the solution from the analytical model (`lm(y ~ x1 + x2, data = data)`) is identical to the solution you obtained when performing gradient descent.

```
linear_model <- lm(poverty ~ educ + democracy, data = data)
linear_model
```

```
##
## Call:
## lm(formula = poverty ~ educ + democracy, data = data)
##
## Coefficients:
## (Intercept)         educ    democracy
##     52.1045      -0.4018       4.9591
```

Question x: Verify that the solution from the analytical logistic regression model `glm(y ~ x1 + x2, data = data, family = "binomial")` is identical to the solution when performing gradient descent.

```
logit_model <- glm(povyesno ~ educ + democracy, data = data, family = "binomial")
logit_model
```

```
##
## Call:  glm(formula = povyesno ~ educ + democracy, family = "binomial",
##     data = data)
##
## Coefficients:
## (Intercept)         educ    democracy
##    5.864405    -0.074700    -0.001083
##
## Degrees of Freedom: 250 Total (i.e. Null);  248 Residual
## Null Deviance:       335.8
## Residual Deviance: 272.6     AIC: 278.6
```

Question x: you can also report a summary of the models estimated using the `modelsummary` package and function. Report a summary of the model below. Add stars indicating statistical significance to your table.

```
modelsummary::modelsummary(list(
  "Poverty" = linear_model,
  "Poverty (Yes/No)" = logit_model),
  stars=T)
```

Finally, we also want to use what we've learned about poverty.

Question x: Give an interpretation of the linear model. What would happen to poverty when a country becomes more democratic, and what would happen to poverty if a country had invested more in education? Is the effect large? Explain.

Question x: Give an interpretation of the logistic regression model. What would happen the likelihood to be poor when a country becomes more democratic, and what would happen if a country had invested more in education? Is the effect large? Explain.

Question x: Generate the predicted values from the linear model. Find the *highest* predicted value. For which country is this (show the `iso2c` code) And what is the predicted value? Use code to find all of this information. Hint: use `predict` and `which.max`, and then look at `data`.

```
number <- predict(linear_model) %>% which.max()
value <- predict(linear_model) %>% max()

value
```

```
## [1] 51.90993
```

```
data[number,]
```

```
## # A tibble: 1 x 6
##   iso2c  date poverty democracy povyesno  educ
##   <chr> <dbl>   <dbl>     <dbl>    <dbl> <dbl>
## 1 ML     2016    46.8         1        1  12.8
```

The End