Data challenge & SHS: Logistic regression and linear model

Julie Josse, Gaël Varoquaux, and Bénédicte Colnet

February 2021

Abstract

In this tutorial, you will perform a logistic regression with `R`. This is the first exercice and we will do it together in class. At the end you can find an exercice with a simple linear regression you should be able to do alone at home (solutions will be given later).

Contents

Credits for this lab: **An Introduction to Statistical Learning: With Applications in R** book from Garet James, Daniela Witten, Trevor Hastie, Robert Tibshirani (in particular for the exercice on Logistic Regression and stock market) The exercice on linear model comes from Imke Mayer's labs. Thanks to them.

Logistic regression: stock market data

In this part we use the `Smarket` data, which is part of the `ISLR` library. This data set consists of percentage returns for the S&P 500 stock index over 1250 days, from the beginning of 2001 until the end of 2005.

The S&P 500,or simply the S&P, is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States. It is one of the most commonly followed equity indices. (I guess we can compare it with the French CAC 40)

Therefore you have 1250 observations on the following 9 variables.

`Year` The year that the observation was recorded

`Lag1` Percentage return for previous day

`Lag2` Percentage return for 2 days previous

`Lag3` Percentage return for 3 days previous

`Lag4` Percentage return for 4 days previous

`Lag5` Percentage return for 5 days previous

`Volume` The number of shares traded

`Today` The percentage return on the date in question

`Direction` A factor with levels Down and Up indicating whether the market had a positive or negative return on a given day

Question 1: Data exploration

Load the library `ISLR` and inspect the data set. Do you see a link between returns? For example you can also look at correlation. What can you say on the volume of shares traded over year?

Question 2: Logistic regression

Fit a logistic regression model in order to predict `Direction` using all the other available variables.

For this you can use `glm()`, a class of models that includes logistic regression.

Interpret the result. What is the coefficient that is the most linked to the outcome according to this model?

Question 3: Prediction

You can use the `predict()` function to perform prediction that the market will go up given other values. Remember that it corresponds to the quantity:

$$\mathbb{P}(Direction = Up | Lag1, \ldots, Volume)$$

If no data set is supplied to the `predict()` function, then the probabilities are computed for the training data used to fit the logistic regression model.

After doing this prediction, you will have the probability of having $Y = 1$. Now, create a confusion matrix with the function `table()` to determine how many observations were correctly or incorrectly classified.

Conclude on this model efficacy. What would you do to better assess this model efficacy?

Question 4: ROC curves

The prediction performed before is by default made with a cutoff at 0.5. But maybe another threshold would help to have a better performance. Using the library `pROC`, screen for the best cutoff. The function you will use is the function `roc()`.

Exploratory data analysis and simple regression

The database

The data are stored in the file 'bea-2006.csv'. It contains information about the economies of the 366 metropolitan statistical areas" (cities) of the US in 2006. In particular, it lists, for each city:

- the population,
- the total value of all goods and services produced for sale in the city that year per person (per capita gross metropolitan product", pcgmp),
- and the share of economic output coming from *four* selected industries.

Question 1: load data

Load the data and perform a summary analysis.

Question 2: data exploration

Produce histogram of population (density and the histogram with "bar") and the box plot of the pgmp column.

Tips: Don't hesitate to do an histogram without the outliers.

Question 3: GMP and population

Make a bivariate plot for per-capita GMP as a function of population. Describe the relationship in words. You can also try with $log(pop)$.

Question 4: simple linear model

Considering your previous plot, run the `lm()` function on the data and add the regression to the previous plot. Will you use `pop` or `log(pop)`? (would the last one still be a linear model?)

You can comment the result.

Question 5

Bonus question: could you do `log(pcgmp)` as a linear function of `pop` (or `log(pop)`)?