Linear Regression

Cheikh Mbacké BEYE

2020-08-30

Table of Contents

## ENVIRONNEMENT DE TRAVAIL

library(dplyr)  
library(tidyr)  
library(ggplot2)  
library(fst)  
library(readr)

Linear regression is a very simple approach for supervised learning. In particular, linear regression is useful tool for predicting a quantitative response. The data set used in this part is the *Advertising* data. It describes sales( in thousand of units) for a particular product as function of advertising budgets (in thousand of dollars) for *TV*, *Radio* and *Newspaper* media. Suppose that our role as data scientist is the answer the following questions based in *Advertising* data.

* 1 is there a relationship between advertising budget and sales? Our first goal should be to determine whether the data provide evidence of an association between a devrtising expenditure and sales. If the evidence is weak, the one might argue that no money should be spent on adevtising.
* 2 How stong is the relationship between advertising budget and sales? Assuming that there is a relationship between aadvertising budget and sales, we would like to know how strong is the relationship. In other words, given a certaiin advertising budget, can we predict sales accurately?
* 3 Which media contibutes to sales? Do all three media: *TV*, *Radio* and *Newspaper* contribute to sales or do just one or two of the media contribute? To answer the question, we must find a way to separate out the invidual effects of each medium when we have spent money on all three media.
* 4 How accurately can we estimate the effect of each of medium on sales? For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase?
* 5 How accurately can we predict future sales? For any given level of television, radio or newspaper advertising, what is our prediction for sales and what is the accuracy of this prediction?
* 6 Is the ralation linear ? If there is approximatively a straight-line relationship between adevertising expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still possible to transform the predictor or the response so that linear regression can be used.
* 7 Is there a synergy among the advertising media? Perhaps spending $50,000 on television advertising and $50,000 on radio advertising results in more sales than allocating $100,000 to either tekevision or radio individualy. This situation is known as *interaction* effect.

## SIMPLE LINEAR REGRESSION

Simple regression lives up its name: it’s a very straightforward approach for predicting a quantitative response *Y* on the basis of a single predictor variable *X*. It assumes that there is a linear relationship between *X* and *Y*. Mathematically, we can write this linear relationship as follow

and are unknown: represents the intercept and the slope in the linear model. These values are known as the model parameters or coefficients. Once we have used our training data to produce estimates and for the model coefficients, we can predict future value of *Y* on the basisof a particular value, *x* of *X* by computing

The main purpose is the computation of the straigntline which is as close as possible of data points. The must common approach involves the *least square* criterion. Let be the prediction for *Y* based on the th value of *X* . Then reprsents the th residual; this is the difference between the obseved response value and the th response value predicted by our linear model. The Residual Sum of Squares (RSS) is defined as follow

The least square approach choses and to minimize the RSS. We can easily demonstrate that the minimizers are defined as follow

### Importations des données et traitement

Advertising <- read\_csv('Advertising.csv')  
str(Advertising)

tibble [200 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
 $ X1 : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...  
 $ TV : num [1:200] 230.1 44.5 17.2 151.5 180.8 ...  
 $ Radio : num [1:200] 37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...  
 $ Newspaper: num [1:200] 69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...  
 $ Sales : num [1:200] 22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...  
 - attr(\*, "spec")=  
 .. cols(  
 .. X1 = col\_double(),  
 .. TV = col\_double(),  
 .. Radio = col\_double(),  
 .. Newspaper = col\_double(),  
 .. Sales = col\_double()  
 .. )

skimr::skim(Advertising)

Data summary

|  |  |
| --- | --- |
| Name | Advertising |
| Number of rows | 200 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| X1 | 0 | 1 | 100.50 | 57.88 | 1.0 | 50.75 | 100.50 | 150.25 | 200.0 | ▇▇▇▇▇ |
| TV | 0 | 1 | 147.04 | 85.85 | 0.7 | 74.38 | 149.75 | 218.82 | 296.4 | ▇▆▆▇▆ |
| Radio | 0 | 1 | 23.26 | 14.85 | 0.0 | 9.97 | 22.90 | 36.52 | 49.6 | ▇▆▆▆▆ |
| Newspaper | 0 | 1 | 30.55 | 21.78 | 0.3 | 12.75 | 25.75 | 45.10 | 114.0 | ▇▆▃▁▁ |
| Sales | 0 | 1 | 14.02 | 5.22 | 1.6 | 10.38 | 12.90 | 17.40 | 27.0 | ▁▇▇▅▂ |

par(mfrow=c(1,3))  
with(Advertising,  
 plot(Sales,TV,pch=19)  
 )  
with(Advertising,  
 plot(Sales,Radio,pch=19)  
 )  
with(Advertising,  
 plot(Sales,Newspaper,pch=19)  
 )

