Data regression: simple linear regression

Lluís Garrido – lluis.garrido@ub.edu

November 2016

Abstract

This laboratory is focused on the principles of data regression. We will focus on some of the well known methods and on the importance of the error funcion used to perform the regression. Notice that this lab is still focused on unconstrained optimization.

1 Introduction

Linear regression is an approach to model the relationship between two continuous, quantitative, variables. One variable, denoted x, is regarded as the predictor or independent variable. The other variable, denoted as y, is regarded as the response or dependent variable.

Take a look at figure 1. We may see a set of data points in blue. In this example the objective is to approximate the data points using a simple linear regression, i.e. we only have one predictor variable.

What is the best fitting line? Let $\{x_i, y_i\}$ with $i = 1 \dots m$ be the data points. Let $\hat{y}_i$ be the fitted value for unit i. The equation for the best fitting line is $\hat{y}_i = b_0 + b_1 x_i$, where b_0 and b_1 are the parameters to be estimated. The parameters b_0 and b_1 may be computed in such a way that the prediction error (or residual error), $e_i = \hat{y}_i - y_i$, is minimized. A line that fits de data "best" will be one for which the m prediction errors are as small as possible in some overall sense. Let us see some ways to achieve this goal.

2 Least squares

The least squares criterion just says "minimize the sum of the squared prediction errors.", that is, we need to find the values b_0 and b_1 that minimize

$$Q = \frac{1}{2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{2}\sum_{i=1}^{m} e_i^2 \tag{1}$$

How can the previous expression be solved? We may used the gradient descent method, for instance, to obtain the optimal values for b_0 and b_1. Observe, however, that the previous expression is a quadratic (linear) expression and

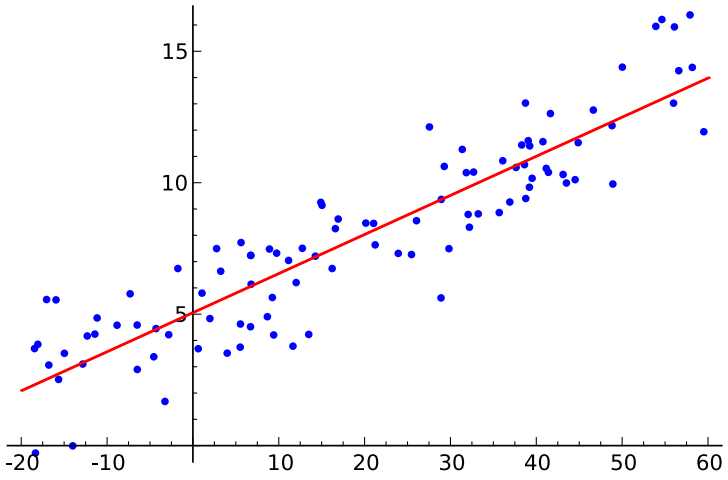

Figure 1: Simple linear regression example. Data points are marked in blue and the objective is to model the points with a linear model. Obtained from https://en.wikipedia.org/wiki/Linear_regression.

thus its solution could be obtained in just one step of the Newton method. In other words, the previous expression has a closed solution and this is one of the reasons why the least squares method is commonly used.

Let us briefly describe how the closed solution can be computed (it is not obtained through application of the Newton method). Recall that the residual error, $e_i = y_i - \hat{y}_i$, is computed for $i = 1 \dots m$. That is,

$$\begin{aligned} b_0\, x_1 + b_1 - y_1 &= e_1 \\ b_0\, x_2 + b_1 - y_2 &= e_2 \\ \vdots &\quad \vdots \\ b_0\, x_m + b_1 - y_m &= e_m \end{aligned}$$

The previous expression can be compactly written as

$$\underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}}_{\mathbf{b}} - \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}}_{\mathbf{e}}$$

Equation (1) can be rewritten as

$$Q = \frac{1}{2} \|\mathbf{e}\|^2 = \frac{1}{2} \|\mathbf{A}\,\mathbf{b} - \mathbf{y}\|^2$$

The solution to this least squares problem is known to be

$$\mathbf{b} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$$

For the particular problem we are dealing with, $\mathbf{b} = (b_0, b_1)^T$, matrix $\mathbf{A}^T\mathbf{A}$ has size 2×2, no matter how many data points we have, and thus its inverse is easy to compute.

Exercises

You are proposed to use the Anscombe's dataset (https://en.wikipedia.org/wiki/Anscombe's_quartet) to perform the tests. It is a relatively simple dataset since the number of elements of each dataset is rather low. We propose you to perform the next experiments:

1. Graph each of the datasets and see how the samples are distributed. Making a plot of the dataset is commonly an important step to visually analyze the samples. You are recommended to perform it previous to any automatic analysis.

2. For each of the dataset, compute the parameters $(b_0, b_1)^T$ using the least squares method. For that purpose you may use the closed solution. Plot the obtained line and see if it correctly fits the dataset. You will see that the least squares method is sensible to outliers. An outlier is a sample that markedly deviates from the other observations in the set.

3 Robust functions

The least squares method has the advantage of having a closed solution. However, the least squares method is known to be sensitive to outliers. The reason is due to the fact that the least squares method minimizes the squared error which may be very large for an outlier. In order to minimize the squared error for the outliers the value of the model, $\mathbf{b} = (b_0, b_1)^T$, has to be "adapted accordingly" and thus it may have nothing to do with what we are looking for.

We need to modify equation (1) so as to make it more robust to outliers. Without entering into exhaustive details, a way to proceed is to minimize

$$\sum_{i=1}^{m} \rho(e_i)$$

where $\rho(u)$ is a robust error function. For the least squared method $\rho(u) = \frac{1}{2}u^2$, but one may take other functions such as the Huber function, which is defined to be[1]

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & |u| \leq c \\ \frac{1}{2}c(2|u| - c) & |u| > c \end{cases} \tag{2}$$

[1] If you are interested in more information on this issue you may begin with the reference Steward, C. "Robust parameter estimation in computer vision", SIAM Review, Vol. 41, No. 3, pp. 513–537, 1999.

Observe that the Huber function is the least squared function if $|u| \leq c$. For $|u| > c$ it is a function that does not square the error value and thus the outlier won't be given the same "importance" that is given by the least squares error.

Exercises

Assume, for simplicity, that $c = 1$ is taken for the Huber function

1. Plot the least squares function, $\rho(u) = \frac{1}{2}u^2$, and compare it with the Huber function, equation (2), in order to see the "importance" that is given to each prediction error u. You may, for instance, plot the function $\rho(u)$ for $|u| \leq 10$.

2. Compute the parameters $(b_0, b_1)^T$ using the Huber function. For that issue you will need to use the Newton method (recall that you'll need to use the gradient method if the Hessian is not positive definite). Plot the obtained line and and compare the result with the result you obtained with the least squares method.

3. If you want to perform more experiments, you may use other robust functions such as the Cauchy function,

$$\rho(u) = \frac{b^2}{2} \log \left[1 + \left(\frac{u}{b} \right)^2 \right]$$

   The latter function is even more robust to outliers than the Huber function.

Report

You are asked to deliver an *individual* report of the two sections. Just explain each of the steps you have followed. If you want to include some parts of code, please include it within the report. Do not include it as separate files. You may just deliver the Python notebook if you want.

The deadline to deliver this report is November 22nd at 3 p.m. (15h).