Data Analysis and Machine Learning: Linear Regression and more Advanced Regression Analysis

Morten Hjorth-Jensen[1,2]

Department of Physics, University of Oslo[1]

Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University[2]

Oct 17, 2017

Regression analysis, overarching aims

Regression modeling deals with the description of the sampling distribution of a given random variable y varies as function of another variable or a set of such variables $\hat{x} = [x_0, x_1, \dots, x_p]^T$. The first variable is called the **dependent**, the **outcome** or the **response** variable while the set of variables $\hat{x}$ is called the independent variable, or the predictor variable or the explanatory variable.

A regression model aims at finding a likelihood function $p(y|\hat{x})$, that is the conditional distribution for y with a given $\hat{x}$. The estimation of $p(y|\hat{x})$ is made using a data set with

- n cases $i = 0, 1, 2, \dots, n-1$
- Response (dependent or outcome) variable y_i with $i = 0, 1, 2, \dots, n-1$
- p Explanatory (independent or predictor) variables $\hat{x}_i = [x_{i0}, x_{i1}, \dots, x_{ip}]$ with $i = 0, 1, 2, \dots, n-1$

The goal of the regression analysis is to extract/exploit relationship between y_i and $\hat{x}_i$ in or to infer causal dependencies,

General linear models

Before we proceed let us study a case from linear algebra where we aim at fitting a set of data $\hat{y} = [y_0, y_1, \dots, y_{n-1}]$. We could think of these data as a result of an experiment or a complicated numerical experiment. These data are functions of a series of variables $\hat{x} = [x_0, x_1, \dots, x_{n-1}]$, that is $y_i = y(x_i)$ with $i = 0, 1, 2, \dots, n-1$. The variables x_i could represent physical quantities like time, temperature, position etc. We assume that $y(x)$ is a smooth function.

Since obtaining these data points may not be trivial, we want to use these data to fit a function which can allow us to make predictions for values of y which are not in the present set. The perhaps simplest approach is to assume we can parametrize our function in terms of a polynomial of degree $n-1$ with n points, that is

$$y = y(x) \rightarrow y(x_i) = \tilde{y}_i + \epsilon_i = \sum_{j=0}^{n-1} \beta_i x_i^j + \epsilon_i,$$

where ϵ_i is the error in our approximation.

Rewriting the fitting procedure as a linear algebra problem

For every set of values y_i, x_i we have thus the corresponding set of equations

$$\begin{aligned}
y_0 &= \beta_0 + \beta_1 x_0^1 + \beta_2 x_0^2 + \dots + \beta_{n-1} x_0^{n-1} + \epsilon_0 \\
y_1 &= \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + \dots + \beta_{n-1} x_1^{n-1} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2^1 + \beta_2 x_2^2 + \dots + \beta_{n-1} x_2^{n-1} + \epsilon_2 \\
\dots & \dots \\
y_{n-1} &= \beta_0 + \beta_1 x_{n-1}^1 + \beta_2 x_{n-1}^2 + \dots + \beta_1 x_{n-1}^{n-1} + \epsilon_{n-1}.
\end{aligned}$$

Defining the vectors

$$\hat{y} = [y_0, y_1, y_2, \dots, y_{n-1}]^T,$$

$$\hat{\beta} = [\beta_0, \beta_1, \beta_2, \dots, \beta_{n-1}]^T,$$

$$\hat{\epsilon} = [\epsilon_0, \epsilon_1, \epsilon_2, \dots, \epsilon_{n-1}]^T,$$

and the matrix

Generalizing the fitting procedure as a linear algebra problem

We are obviously not limited to the above polynomial. We could replace the various powers of x with elements of Fourier series, that is, instead of x_i^j we could have $\cos(jx_i)$ or $\sin(jx_i)$, or time series or other orthogonal functions. For every set of values y_i, x_i we can then generalize the equations to

$$\begin{aligned}
y_0 &= \beta_0 x_{00} + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_{n-1} x_{0n-1} + \epsilon_0 \\
y_1 &= \beta_0 x_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_{n-1} x_{1n-1} + \epsilon_1 \\
y_2 &= \beta_0 x_{20} + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_{n-1} x_{2n-1} + \epsilon_2 \\
\dots & \dots \\
y_i &= \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{n-1} x_{in-1} + \epsilon_i \\
\dots & \dots \\
y_{n-1} &= \beta_0 x_{n-1,0} + \beta_1 x_{n-1,2} + \beta_2 x_{n-1,2} + \dots + \beta_1 x_{n-1}^{n-1,n-1} + \epsilon_{n-1}.
\end{aligned}$$

We redefine in turn the matrix $\hat{X}$ as

Optimizing our parameters

We have defined the matrix $\hat{X}$

$$\begin{aligned}
y_0 &= \beta_0 x_{00} + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_{n-1} x_{0n-1} + \epsilon_0 \\
y_1 &= \beta_0 x_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_{n-1} x_{1n-1} + \epsilon_1 \\
y_2 &= \beta_0 x_{20} + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_{n-1} x_{2n-1} + \epsilon_1 \\
\dots & \dots \\
y_i &= \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{n-1} x_{in-1} + \epsilon_1 \\
\dots & \dots \\
y_{n-1} &= \beta_0 x_{n-1,0} + \beta_1 x_{n-1,2} + \beta_2 x_{n-1,2} + \dots + \beta_1 x_{n-1,n-1} + \epsilon_{n-1}.
\end{aligned}$$

We well use this matrix to define the approximation $\hat{\tilde{y}}$ via the unknown quantity $\hat{\beta}$ as

$$\hat{\tilde{y}} = \hat{X}\hat{\beta},$$

and in order to find the optimal parameters β_i instead of solving the above linear algebra problem, we define a function which gives

Interpretations and optimizing our parameters

The function

$$
Q(\hat{\beta}) = \left(\hat{y} - \hat{X}\hat{\beta}\right)^T \left(\hat{y} - \hat{X}\hat{\beta}\right),
$$

can be linked to the variance of the quantity y_i if we interpret the latter as the mean value of for example a numerical experiment. When linking below with the maximum likelihood approach below, we will indeed interpret y_i as a mean value

$$
y_i = \langle y_i \rangle = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{n-1} x_{i,n-1} + \epsilon_i,
$$

where $\langle y_i \rangle$ is the mean value. Keep in mind also that till now we have treated y_i as the exact value. Normally, the response (dependent or outcome) variable y_i the outcome of a numerical experiment or another type of experiment and is thus only an approximation to the true value. It is then always accompanied by an error estimate, often limited to a statistical error estimate given by the standard deviation discussed earlier. In the discussion here we will treat y_i as our exact value for the response variable.

Interpretations and optimizing our parameters

We can rewrite

$$
\frac{\partial Q(\hat{\beta})}{\partial \hat{\beta}} = 0 = \hat{X}^T \left(\hat{y} - \hat{X}\hat{\beta}\right),
$$

as

$$
\hat{X}^T \hat{y} = \hat{X}^T \hat{X}\hat{\beta},
$$

and if the matrix $\hat{X}^T\hat{X}$ is invertible we have the solution

$$
\hat{\beta} = \left(\hat{X}^T\hat{X}\right)^{-1} \hat{X}^T \hat{y}.
$$

The residuals $\hat{\epsilon}$ are in turn given by

$$
\hat{\epsilon} = \hat{y} - \hat{\tilde{y}} = \hat{y} - \hat{X}\hat{\beta},
$$

and with

$$
\hat{X}^T \left(\hat{y} - \hat{X}\hat{\beta}\right) = 0,
$$

we have

$$
\hat{X}^T\hat{\epsilon} = \hat{X}^T\left(\hat{y} - \hat{X}\hat{\beta}\right) = 0,
$$

The singular value decompostion

Here we derive the equations for the SVD.