

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

Morten Hjorth-Jensen<sup>1,2</sup>

Department of Physics, University of Oslo<sup>1</sup>

Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University<sup>2</sup>

Oct 24, 2017

© 1999-2017, Morten Hjorth-Jensen. Released under CC Attribution-NonCommercial 4.0 license

## 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

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

where  $\epsilon_i$  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

$$y_0 = \beta_0 + \beta_1 x_0^1 + \beta_2 x_0^2 + \dots + \beta_{n-1} x_0^{n-1} + \epsilon_0$$

$$y_1 = \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + \dots + \beta_{n-1} x_1^{n-1} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2^1 + \beta_2 x_2^2 + \dots + \beta_{n-1} x_2^{n-1} + \epsilon_2$$

.....

$$y_{n-1} = \beta_0 + \beta_1 x_{n-1}^1 + \beta_2 x_{n-1}^2 + \dots + \beta_{n-1} x_{n-1}^{n-1} + \epsilon_{n-1}.$$

## Rewriting the fitting procedure as a linear algebra problem, follows

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

$$\hat{X} = \begin{bmatrix} 1 & x_0^1 & x_0^2 & \dots & x_0^{n-1} \\ 1 & x_1^1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2^1 & x_2^2 & \dots & x_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n-1}^1 & x_{n-1}^2 & \dots & x_{n-1}^{n-1} \end{bmatrix}$$

we can rewrite our equations as

$$\hat{y} = \hat{X} \hat{\beta} + \hat{\epsilon}.$$

## 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

$$y_0 = \beta_0 x_{00} + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_{n-1} x_{0,n-1} + \epsilon_0$$

$$y_1 = \beta_0 x_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_{n-1} x_{1,n-1} + \epsilon_1$$

$$y_2 = \beta_0 x_{20} + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_{n-1} x_{2,n-1} + \epsilon_2$$

.....

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{n-1} x_{i,n-1} + \epsilon_i$$

.....

$$y_{n-1} = \beta_0 x_{n-1,0} + \beta_1 x_{n-1,1} + \beta_2 x_{n-1,2} + \dots + \beta_{n-1} x_{n-1,n-1} + \epsilon_{n-1}.$$

## Generalizing the fitting procedure as a linear algebra problem

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

$$\hat{X} = \begin{bmatrix} x_{00} & x_{01} & x_{02} & \dots & \dots & x_{0,n-1} \\ x_{10} & x_{11} & x_{12} & \dots & \dots & x_{1,n-1} \\ x_{20} & x_{21} & x_{22} & \dots & \dots & x_{2,n-1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n-1,0} & x_{n-1,1} & x_{n-1,2} & \dots & \dots & x_{n-1,n-1} \end{bmatrix}$$

and without loss of generality we rewrite again our equations as

$$\hat{y} = \hat{X}\hat{\beta} + \hat{\epsilon}.$$

The left-hand side of this equation forms known. Our error vector  $\hat{\epsilon}$  and the parameter vector  $\hat{\beta}$  are our unknown quantities. How can we obtain the optimal set of  $\beta_i$  values?

## Optimizing our parameters

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

$$y_0 = \beta_0 x_{00} + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_{n-1} x_{0,n-1} + \epsilon_0$$

$$y_1 = \beta_0 x_{10} + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_{n-1} x_{1,n-1} + \epsilon_1$$

$$y_2 = \beta_0 x_{20} + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_{n-1} x_{2,n-1} + \epsilon_1$$

.....

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{n-1} x_{i,n-1} + \epsilon_i$$

.....

$$y_{n-1} = \beta_0 x_{n-1,0} + \beta_1 x_{n-1,1} + \beta_2 x_{n-1,2} + \dots + \beta_{n-1} x_{n-1,n-1} + \epsilon_{n-1}.$$

## Optimizing our parameters, more details

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

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

and in order to find the optimal parameters  $\beta_i$  instead of solving the above linear algebra problem, we define a function which gives a measure of the spread between the values  $y_i$  (which represent hopefully the exact values) and the parametrized values  $\hat{y}_i$ , namely

$$Q(\hat{\beta}) = \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 = (\hat{y} - \hat{y})^T (\hat{y} - \hat{y}),$$

or using the matrix  $\hat{X}$  as

$$Q(\hat{\beta}) = (\hat{y} - \hat{X}\hat{\beta})^T (\hat{y} - \hat{X}\hat{\beta}).$$

## Interpretations and optimizing our parameters

The function

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

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

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

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

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

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

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

## Interpretations and optimizing our parameters

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

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

and with

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

we have

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

meaning that the solution for  $\hat{\beta}$  is the one which minimizes the residuals. Later we will link this with the maximum likelihood approach.

## The $\chi^2$ function

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.

Introducing the standard deviation  $\sigma_i$  for each measurement  $y_i$ , we define now the  $\chi^2$  function as

$$\chi^2(\hat{\beta}) = \sum_{i=0}^{n-1} \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} = (\hat{y} - \hat{\hat{y}})^T \frac{1}{\hat{\Sigma}^2} (\hat{y} - \hat{\hat{y}}),$$

where the matrix  $\hat{\Sigma}$  is a diagonal matrix with  $\sigma_i$  as matrix elements.

## The $\chi^2$ function

In order to find the parameters  $\beta_j$  we will then minimize the spread of  $\chi^2(\hat{\beta})$  by requiring

$$\frac{\partial \chi^2(\hat{\beta})}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \left[ \sum_{i=0}^{n-1} \left( \frac{y_i - \beta_0 x_{i,0} - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \dots - \beta_{n-1} x_{i,n-1}}{\sigma_i} \right)^2 \right]$$

which results in

$$\frac{\partial \chi^2(\hat{\beta})}{\partial \beta_j} = -2 \left[ \sum_{i=0}^{n-1} \frac{x_{ij}}{\sigma_i} \left( \frac{y_i - \beta_0 x_{i,0} - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \dots - \beta_{n-1} x_{i,n-1}}{\sigma_i} \right) \right]$$

or in a matrix-vector form as

$$\frac{\partial \chi^2(\hat{\beta})}{\partial \hat{\beta}} = 0 = \hat{A}^T (\hat{b} - \hat{A}\hat{\beta}).$$

where we have defined the matrix  $\hat{A} = \hat{X}/\hat{\Sigma}$  with matrix elements  $a_{ij} = x_{ij}/\sigma_i$  and the vector  $\hat{b}$  with elements  $b_i = y_i/\sigma_i$ .

## The $\chi^2$ function

We can rewrite

$$\frac{\partial \chi^2(\hat{\beta})}{\partial \hat{\beta}} = 0 = \hat{A}^T (\hat{b} - \hat{A}\hat{\beta}),$$

as

$$\hat{A}^T \hat{b} = \hat{A}^T \hat{A} \hat{\beta},$$

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

$$\hat{\beta} = (\hat{A}^T \hat{A})^{-1} \hat{A}^T \hat{b}.$$

## The $\chi^2$ function

If we then introduce the matrix

$$\hat{H} = \hat{A}^T \hat{A},$$

we have then the following expression for the parameters  $\beta_j$  (the matrix elements of  $\hat{H}$  are  $h_{ij}$ )

$$\beta_j = \sum_{k=0}^{p-1} h_{jk} \sum_{i=0}^{n-1} \frac{y_i x_{ik}}{\sigma_i} = \sum_{k=0}^{p-1} h_{jk} \sum_{i=0}^{n-1} b_i a_{ik}$$

We state without proof the expression for the uncertainty in the parameters  $\beta_j$  as

$$\sigma^2(\beta_j) = \sum_{i=0}^{n-1} \sigma_i^2 \left( \frac{\partial \beta_j}{\partial y_i} \right)^2,$$

resulting in

$$\sigma^2(\beta_j) = \sum_{i=0}^{n-1} \sigma_i^2 \left( \frac{\partial \beta_j}{\partial y_i} \right)^2$$

## The $\chi^2$ function

The first step here is to approximate the function  $y$  with a first-order polynomial, that is we write

$$y = y(x) \rightarrow y(x_i) \approx \beta_0 + \beta_1 x_i.$$

By computing the derivatives of  $\chi^2$  with respect to  $\beta_0$  and  $\beta_1$  show that these are given by

$$\frac{\partial \chi^2(\hat{\beta})}{\partial \beta_0} = -2 \left[ \sum_{i=0}^1 \left( \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma_i^2} \right) \right] = 0,$$

and

$$\frac{\partial \chi^2(\hat{\beta})}{\partial \beta_1} = -2 \left[ \sum_{i=0}^1 x_i \left( \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma_i^2} \right) \right] = 0.$$

## The $\chi^2$ function

We define then

$$\gamma = \sum_{i=0}^1 \frac{1}{\sigma_i^2},$$

$$\gamma_x = \sum_{i=0}^1 \frac{x_i}{\sigma_i^2},$$

$$\gamma_y = \sum_{i=0}^1 \left( \frac{y_i}{\sigma_i^2} \right),$$

$$\gamma_{xx} = \sum_{i=0}^1 \frac{x_i x_i}{\sigma_i^2},$$

$$\gamma_{xy} = \sum_{i=0}^1 \frac{y_i x_i}{\sigma_i^2},$$

and show that

$$\beta_0 = \frac{\gamma_{xx} \gamma_y - \gamma_x \gamma_{xy}}{\gamma_{xx} \gamma - \gamma_x^2},$$

## The singular value decomposition

How can we use the singular value decomposition to find the parameters  $\beta_j$ ? More details will come. We first note that a general  $m \times n$  matrix  $\hat{A}$  can be written in terms of a diagonal matrix  $\hat{\Sigma}$  of dimensionality  $n \times n$  and two orthogonal matrices  $\hat{U}$  and  $\hat{V}$ , where the first has dimensionality  $m \times m$  and the last dimensionality  $n \times n$ . We have then

$$\hat{A} = \hat{U} \hat{\Sigma} \hat{V}$$