

Topic 6: Linear Regression

GO GREEN. AVOID PRINTING, OR PRINT DOUBLE-SIDED.

6.1 Introduction

One of the most elemental problems in data science can be summarized as predicting the value of a variable of interest y as a function of other variables x_1, \dots, x_D . For example:

- Predicting my glucose level (variable of interest) as a function of my height, weight, age, and gender (other variables).
- Predicting stock prices (variable of interest) as a function of the market state (other variables).
- Predicting the activation level of a gene that determines a disease, like cancer (variable of interest) as a function of other genes' activation levels (other variables); this is often known as genomics wide association studies (GWAS).
- Predicting magnitude of solar flares (variable of interest) as a function of solar images (other variables).

The main idea behind linear regression is to write y as a linear combination of x_1, \dots, x_D , i.e.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_D x_D, \quad (6.1)$$

where β_0 is essentially an *offset*, and $\beta_1, \beta_2, \dots, \beta_D$ are the weights of each variable. For instance, in our glucose example, (6.1) is essentially saying:

$$\text{glucose level} = \beta_0 + \beta_1 \text{height} + \beta_2 \text{weight} + \beta_3 \text{age} + \beta_4 \text{gender}.$$

Notice that by letting $\mathbf{x} = [1 \ x_1 \ x_2 \ \dots \ x_D]^T$ and $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_D]^T$ we can rewrite (6.1) in vector form as

$$y = \mathbf{x}^T \boldsymbol{\beta}. \quad (6.2)$$

The goal is to determine the weights vector $\boldsymbol{\beta}$ that best explains y as a function of \mathbf{x} .

6.2 Finding $\boldsymbol{\beta}$

In order to find the coefficient vector $\boldsymbol{\beta}$ that best explains y as a function of \mathbf{x} we use *training data*. The main idea is to observe *training pairs* $\{\mathbf{x}_i, y_i\}_{i=1}^N$, and find the vector $\boldsymbol{\beta}$ such that $y_i \approx \mathbf{x}_i^T \boldsymbol{\beta}$ for every $i = 1, \dots, N$.

More precisely, we want to find β such that

$$y_i = \mathbf{x}_i^\top \beta + \epsilon_i \quad \text{for every } i = 1, \dots, N,$$

where ϵ_i is a small error. Letting $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_N]^\top$, $\mathbf{X}^\top = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_N]$, and $\boldsymbol{\epsilon} = [\epsilon_1 \ \epsilon_2 \ \dots \ \epsilon_N]^\top$, we can rewrite this in matrix form as:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}. \quad (6.3)$$

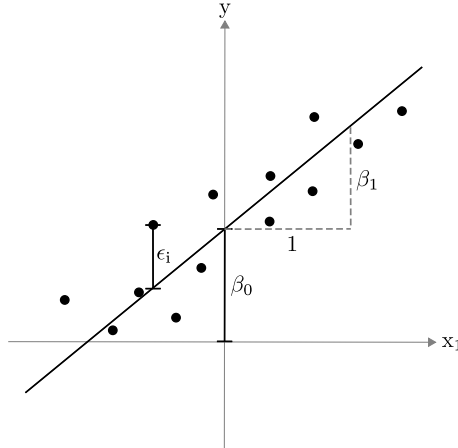
Our goal can then be rephrased as finding the β that minimizes the size of $\boldsymbol{\epsilon}$, or equivalently, the difference between \mathbf{y} and $\mathbf{X}\beta$, i.e.,

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^{D+1}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2. \quad (6.4)$$

Recall that

$$\|\mathbf{y} - \mathbf{X}\beta\|_2^2 = \sum_{i=1}^N y_i - \mathbf{x}_i^\top \beta,$$

so (6.4) is minimizing the error of the entire sample. Intuitively, β^* is the *line* that best explains the y_i 's as function of the \mathbf{x}_i 's. This is illustrated in the following figure, where each point represents a pair (\mathbf{x}_i, y_i) :



6.2.1 Solving (6.4)

In order to solve (6.4), notice that:

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 &= (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \\ &= \mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta. \end{aligned}$$

At this point we can take the derivative with respect to β . Notice that β is a vector, so taking derivative is a bit tricky. To learn more about how to take derivatives w.r.t. vectors and matrices see *Old and new matrix algebra useful for statistics* by Thomas P. Minka. Using the tricks therein, we know that the derivative w.r.t. β is:

$$2\mathbf{X}^\top \mathbf{X}\beta - 2\mathbf{X}^\top \mathbf{y}$$

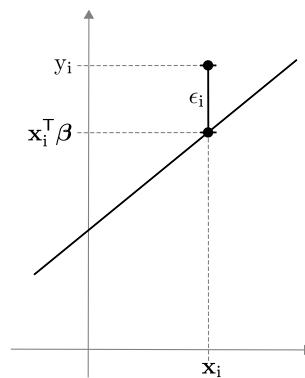
Setting this to zero and solving for β , we have that

$$\beta^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

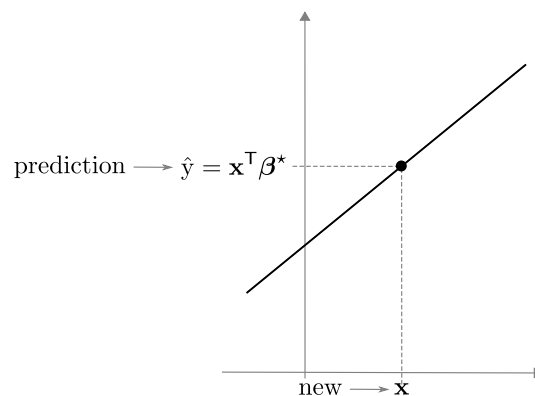
which is tightly related to the projector operator onto $\text{span}\{\mathbf{X}\}$.

6.3 We've found β^* . Then what?

Recall that our ultimate goal is to predict y as a function of \mathbf{x} . To this end, linear regression aims to find the *line* that explains each y_i as a function its corresponding \mathbf{x}_i . Such *line* is determined by β . Given \mathbf{x}_i , the linear prediction of y_i is given by $\mathbf{x}_i^T \beta$, and ϵ_i is the *error* between the prediction $\mathbf{x}_i^T \beta$ and the truth, y_i :



Once we have found β^* , we have found the line that *best* explains each y_i in our training data as a function of its corresponding \mathbf{x}_i . Substituting β^* , we see that the best linear prediction of each y_i is given by $\mathbf{x}_i^T \beta^* = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Extrapolating this, given a new vector \mathbf{x} (for which we do not know y), we can predict its y as $\mathbf{x}^T \beta^* = \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$:



6.4 A Top to Bottom Example

Scientists have discovered a new deadly disease with 90% mortality rate. However, if detected in time, people can take preventive measures to reduce mortality rate to only 5%. The problem is that the test to determine

a person's risk to develop this new disease is extremely expensive. To overcome this problem, you decide to run an experiment. You will test N people, and record their results in a data vector \mathbf{y} . In addition you will store other D indicators about these N people (such as height, weight, and gender) in a data matrix \mathbf{X} . Using this information you will try to find a coefficient vector $\boldsymbol{\beta}$ that best explains \mathbf{y} as a function of \mathbf{X} . In words, you will try to predict a person's risk of developing the disease as a function of their other indicators.

Since you took CS 4780/6780, you know that the best $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Using this information, you are now able to predict the chance that *every* person in the world develops the disease as follows. For each person whose risk is unknown, collect the D indicators in a vector \mathbf{x} (which can be easily and cheaply done), and compute her risk as $\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.

This is precisely the approach that is currently done in genomic wide association studies (GWAS) to predict a person's risk to develop cancer and other diseases, where the D indicators include people's genome, in addition to their demographics.