CS 4780/6780: Fundamentals of Data Science **Spring 2019**

Topic 6: Linear Regression

Instructor: Daniel L. Pimentel-Alarcón

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.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_D x_D, \tag{6.1}$$

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

$$y = \mathbf{x}^T \boldsymbol{\beta}. \tag{6.2}$$

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

6.2 Finding β

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 $\boldsymbol{\beta}$ such that

$$\mathrm{y_i} \;=\; \mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\beta} + \epsilon_\mathrm{i} \qquad \text{for every } \mathrm{i} = 1, \ldots, \mathrm{N},$$

where ϵ_i is a small error. Letting $\mathbf{y} = [\mathrm{y}_1 \;\; \mathrm{y}_2 \;\; \cdots \;\; \mathrm{y}_\mathrm{N}]^\mathsf{T}$, $\mathbf{X}^\mathsf{T} = [\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \cdots \;\; \mathbf{x}_\mathrm{N}]$, and $\boldsymbol{\epsilon} = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_\mathrm{N}]^\mathsf{T}$, we can rewrite this in matrix form as:

$$\mathbf{y} \;=\; \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{6.3}$$

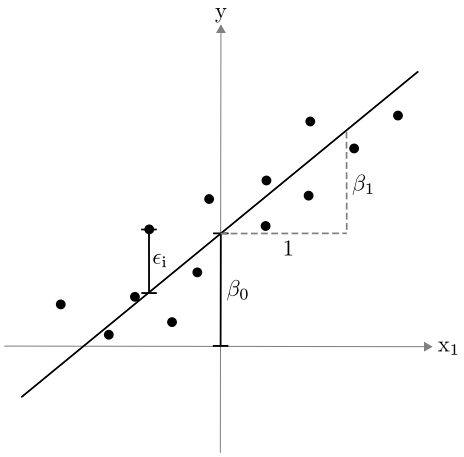
Our goal can then be rephrased as finding the $\boldsymbol{\beta}$ that minimizes the size of $\boldsymbol{\epsilon}$, or equivalently, the difference between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\beta}$, i.e.,

$$\boldsymbol{\beta}^\star \;=\; \underset{\boldsymbol{\beta}\in\mathbb{R}^{\mathrm{D}+1}}{\arg\min} \; \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{6.4}$$

Recall that

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{y_i} - \mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\beta})^2,$$

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



6.2.1 Solving (6.4)

In order to solve (6.4), notice that:

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

At this point we can take the derivative with respect to $\boldsymbol{\beta}$. Notice that $\boldsymbol{\beta}$ 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. $\boldsymbol{\beta}$ is:

$$2\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\mathsf{T}\mathbf{y}$$

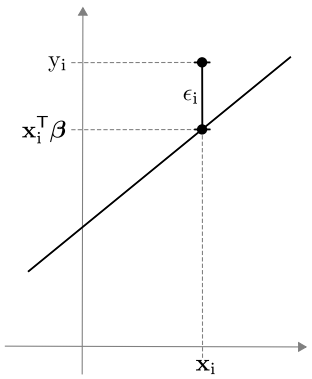
Setting this to zero and solving for $\boldsymbol{\beta}$, we have that

$$\boldsymbol{\beta}^\star \;=\; (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y},$$

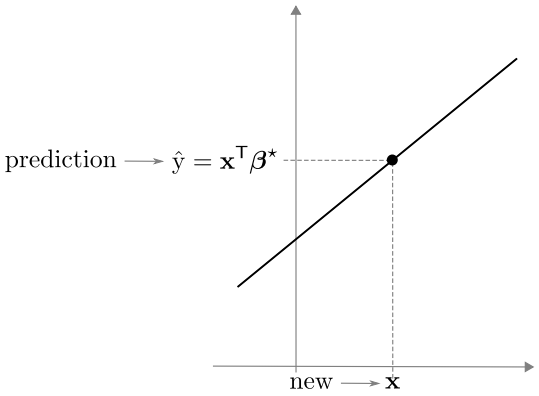
which is tightly related to the projector operator onto span$\{\mathbf{X}\}$.

6.3 We've found $\beta^\star$. 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 $\mathrm{y_i}$ as a function its corresponding $\mathbf{x}_\mathrm{i}$. Such *line* is determined by $\boldsymbol{\beta}$. Given $\mathbf{x}_\mathrm{i}$, the linear prediction of $\mathrm{y_i}$ is given by $\mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\beta}$, and ϵ_i is the *error* between the prediction $\mathbf{x}_\mathrm{i}^\mathsf{T}\boldsymbol{\beta}$ and the truth, $\mathrm{y_i}$:



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



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 $\boldsymbol{\beta}^\star = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. 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 $\mathbf{x}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$.

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.