8.5.0 Linear Regression

Sometimes we are interested in obtaining a simple model that explains the relationship between two or more variables. For example, suppose that we are interested in studying the relationship between the income of parents and the income of their children in a certain country. In general, many factors can impact the income of a person. Nevertheless, we suspect that children from wealthier families generally tend to become wealthier when they grow up. Here, we can consider two variables:

1. The family income can be defined as the average income of parents at a certain period.
2. The child income can be defined as his/her average income at a certain period (e.g, age).

To examine the relationship between the two variables, we collect some data

$$(x_i, y_i), \quad \text{for } i = 1, 2, \cdots, n,$$

where y_i is the average income of the ith child, and x_i is the average income of his/her parents. We are often interested in finding a simple model. A **linear** model is probably the simplest model that we can define, where we write

$$y_i \approx \beta_0 + \beta_1 x_i.$$

Of course, there are other factors that impact each child's future income, so we might write

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i is modeled as a random variable. More specifically, if we approximate a child's future income by $\hat{y}_i = \beta_0 + \beta_1 x_i$, then ϵ_i indicates the error in our approximation. The goal here is to obtain the best values of β_0 and β_1 that result in the smallest errors. In other words, we would like to draw a "line" in the $x - y$ plane that best fits our data points. The line

$$\hat{y} = \beta_0 + \beta_1 x$$

is called the **regression line**. Figure 8.11 shows the regression line.

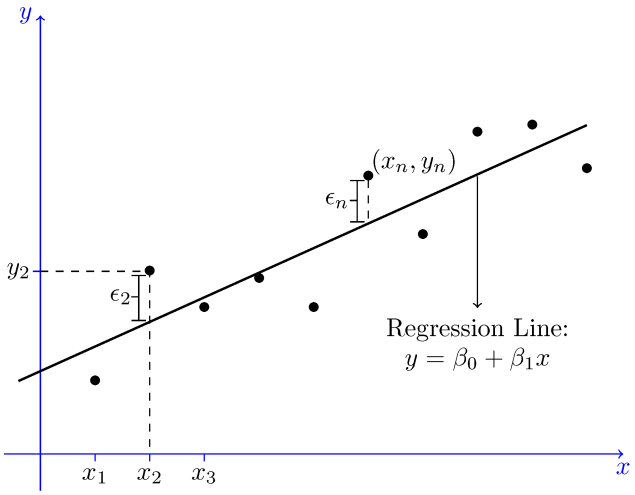

Figure 8.11 - Regression line is the line that best represents the data points (x_i, y_i).

We may summarize our model as

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

Note that since ϵ is a random variable, Y is also a random variable. The variable x is called the **predictor** or the **explanatory variable**, and the random variable Y is called the **response** variable. That is, here, we use x to predict/estimate Y.