Aula 30/09/2020

Tópicos dicutidos

- *Statistical learning should not be viewed as a series of black boxes.*

- **No single approach will perform well in all possible applications.**

- **Without understanding all of the cogs (engrenagens ou mecanismos) inside the box, or the interaction between those cogs, it is impossible to select the best box.**

- **Hence, we have attempted to carefully describe the model, intuition, assumptions, and trade-offs behind each of the methods that we consider.**

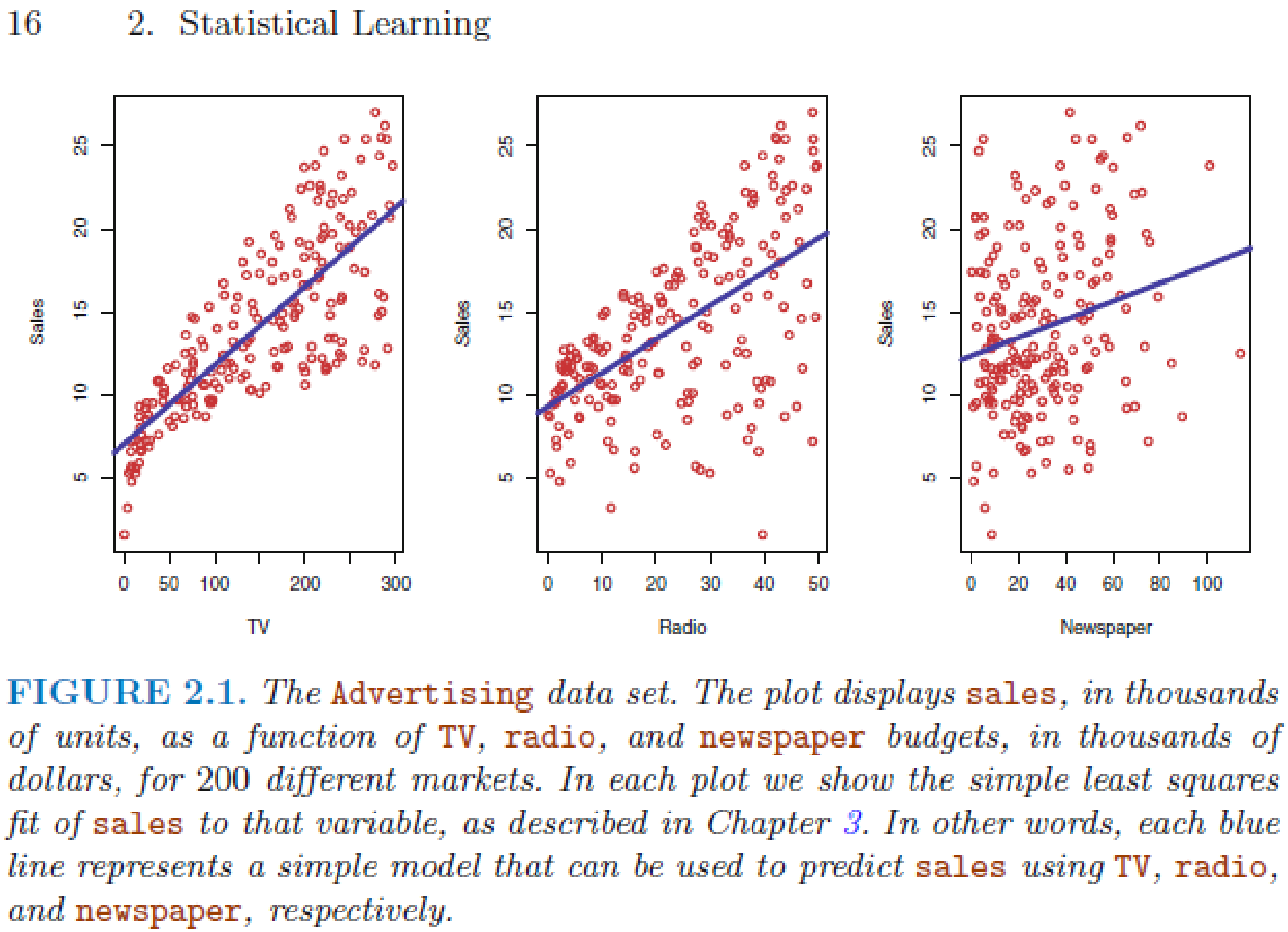
Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. In this setting, $\hat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

Exemplo 1



16 2. Statistical Learning

FIGURE 2.1. *The* Advertising *data set. The plot displays* sales*, in thousands of units, as a function of* TV*,* radio*, and* newspaper *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* sales *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* sales *using* TV*,* radio*, and* newspaper*, respectively.*

Exemplo 2

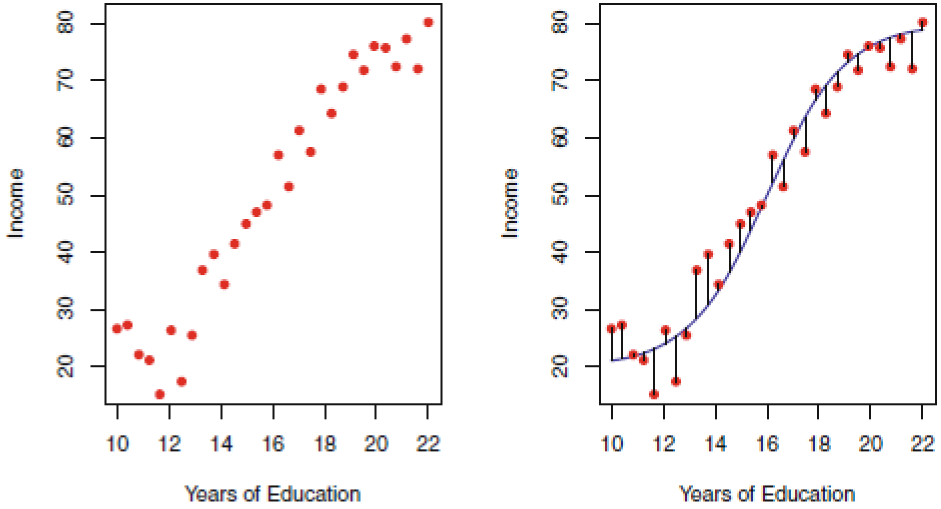
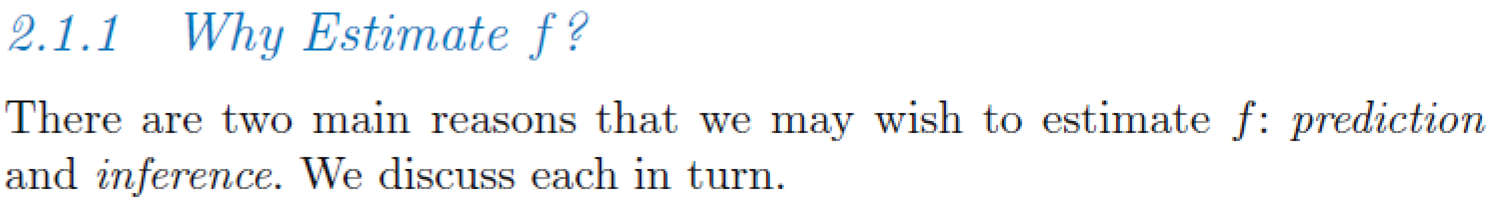


FIGURE 2.2. *The* Income *data set.* Left: *The red dots are the observed values of* income *(in tens of thousands of dollars) and* years of education *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* income *and* years of education*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

In essence, statistical learning refers to a set of approaches for estimating f.

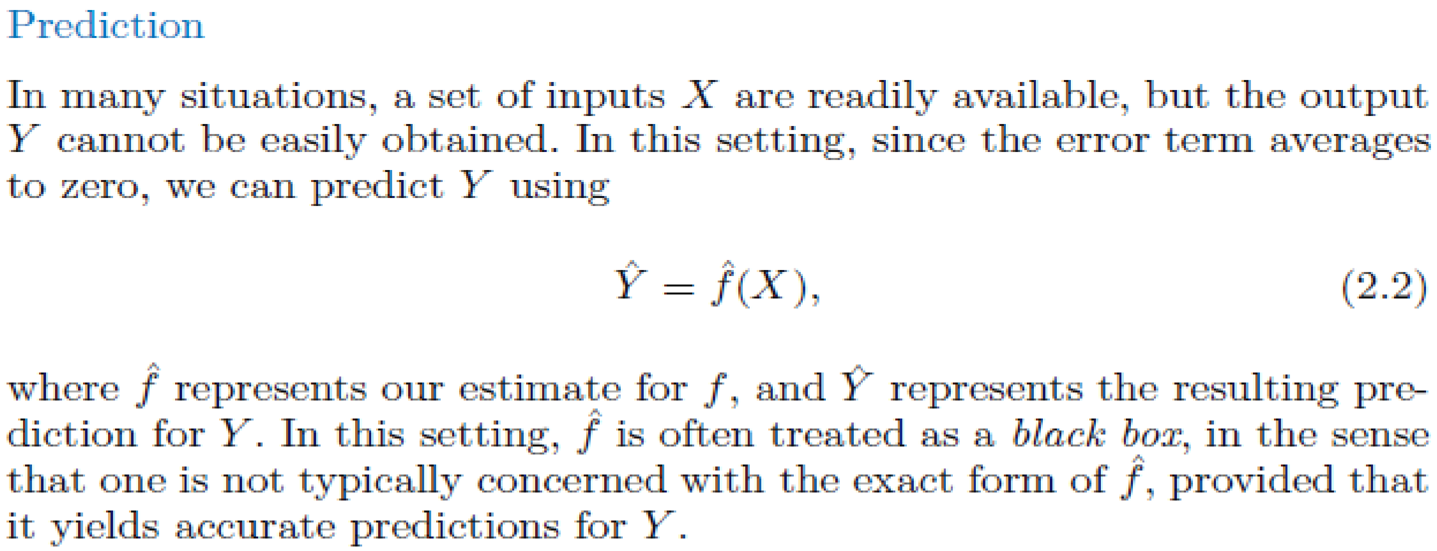


2.1.1 *Why Estimate f?*

There are two main reasons that we may wish to estimate f: *prediction* and *inference*. We discuss each in turn.

Predição

Inferência



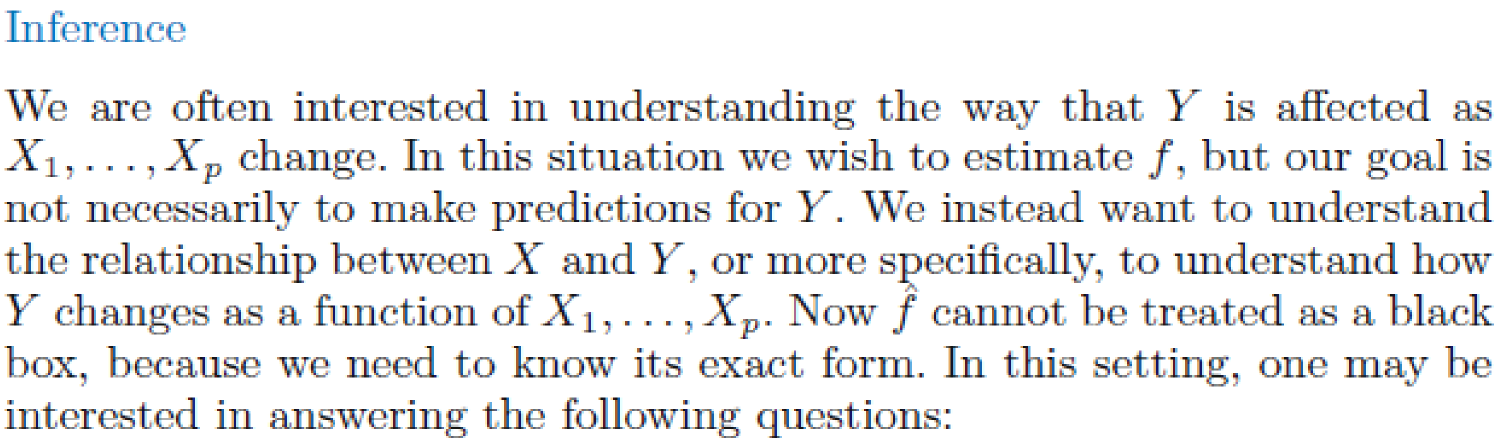
Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. In this setting, $\hat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

- Temos acesso aos dados de entrada (inputs X) mas é difícil obter os dados de saída (the output Y)
- A forma exata de f não é tão importante, desde que forneça valores de Y com certa precisão.
- No exemplo 2 é mostrada uma f, mas é conhecida por que os dados foram obtidos por uma simulação que usou a própria f



Inference

We are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. We instead want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, \ldots, X_p$. Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in answering the following questions:

- Quais são de fato as variáveis preditoras (input) que interferem de fato no valor de Y (reposta/output) ?
- Como as variáveis preditoras interferem em Y: positiva ou negativamente?
- Qual de fato é a relação das variáveis preditoras e a resposta Y: é uma relação linear?