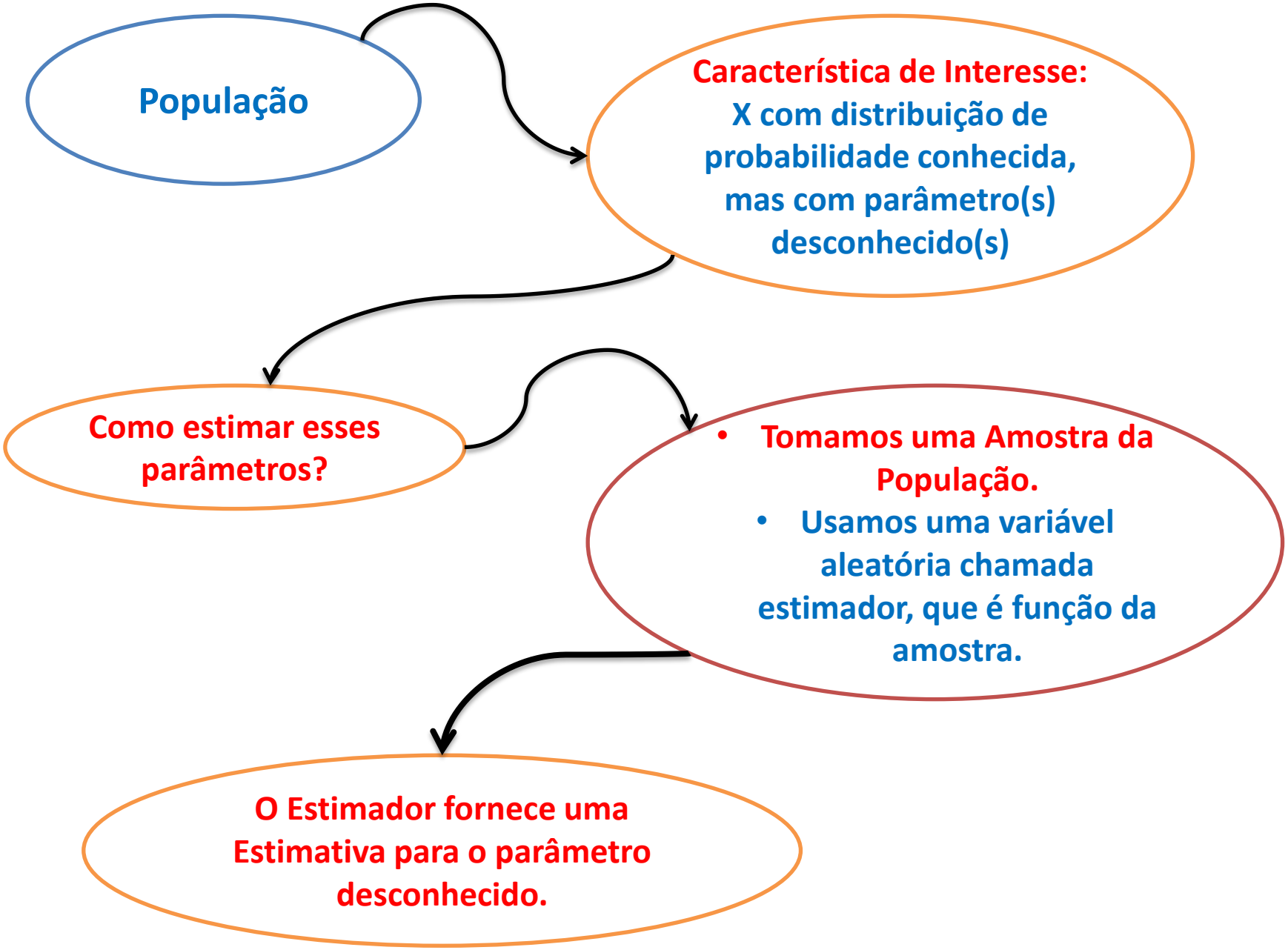
Aula Regressão Linear

Parte 1


População
Característica de Interesse:
X com distribuição de probabilidade conhecida, mas com parâmetro(s) desconhecido(s)
Como estimar esses parâmetros?
• Tomamos uma Amostra da População.
• Usamos uma variável aleatória chamada estimador, que é função da amostra.
O Estimador fornece uma Estimativa para o parâmetro desconhecido.

Discutimos nas aulas passadas: $Y = f(X) + \epsilon$

E agora temos: $Y = \beta_0 + \beta_1 X + \epsilon.$

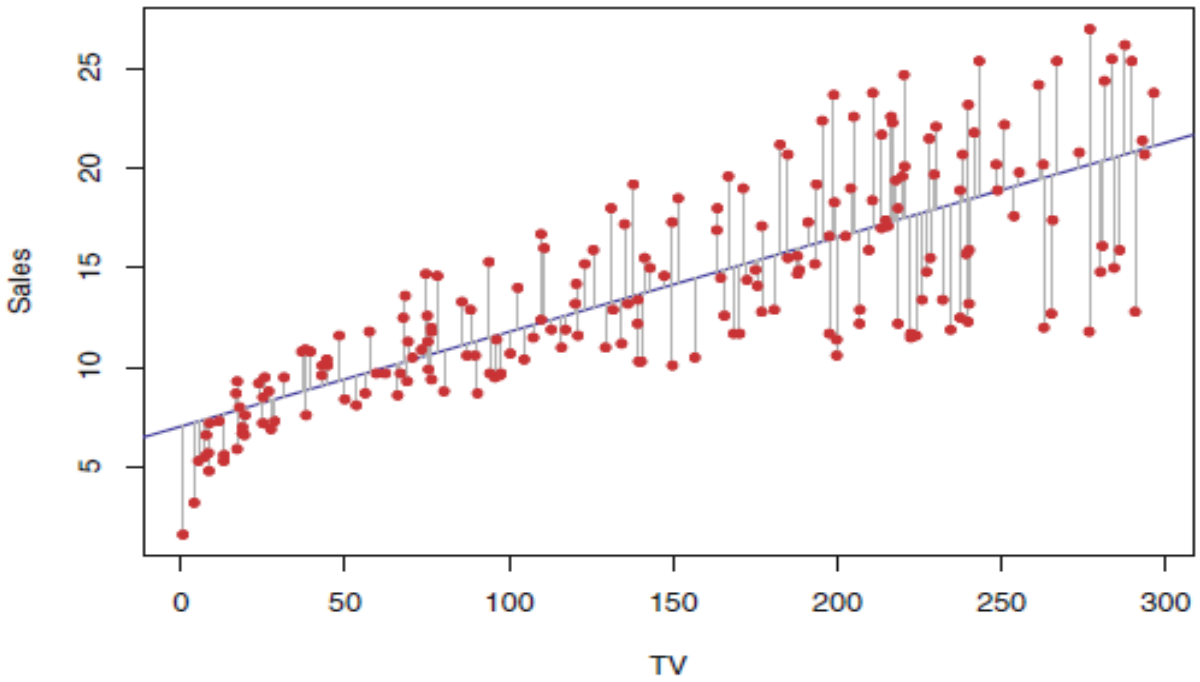


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

$$Y \approx \beta_0 + \beta_1 X.$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

Usar o conjunto de treinamento para estimar os parâmetros desconhecidos.

E poder fazer previsões, dado um valor de x, encontramos um valor de y

TV: gastos com propaganda na televisão.

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

Dada um certo valor de "TV" podemos predizer a quantidade de vendas do produto.

O método do mínimos quadrados fornecem estimativas para os parâmetros desconhecidos da função linear proposta.

Temos um conjunto de dados:

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Minimizando a soma dos resíduos ao quadrado, em relação aos dois parâmetros desconhecidos, temos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

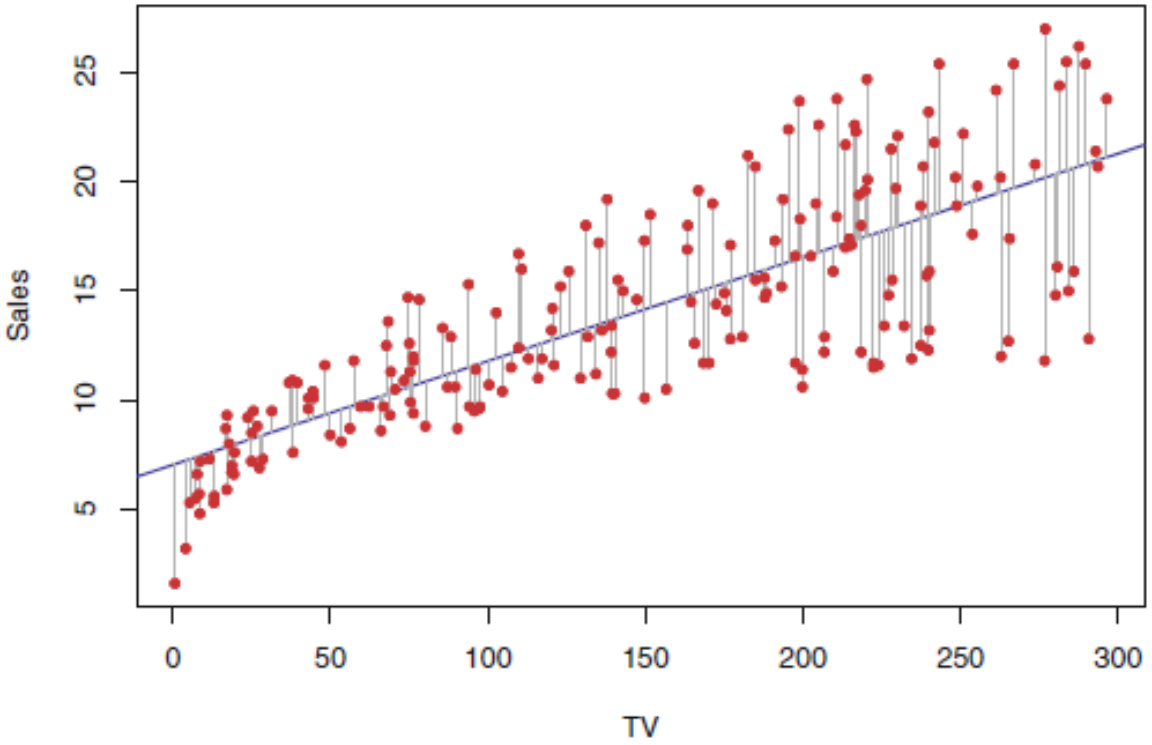


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$ ⇒ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$

$\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475.$

Figura que mostra um exemplo de aplicação do Cálculo II em Inferência Estatística

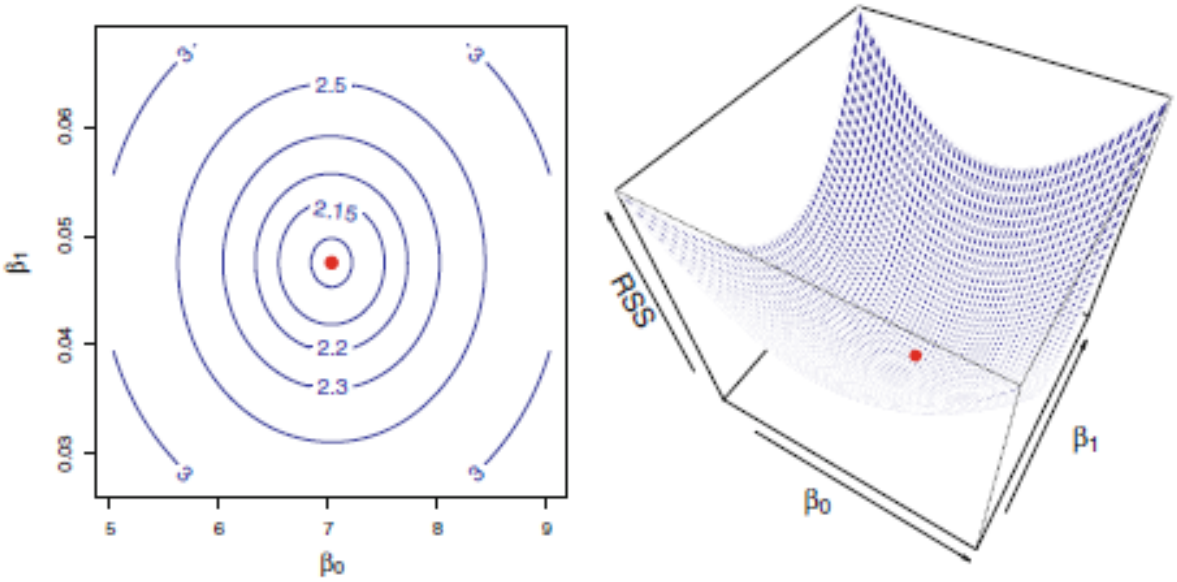


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* `Advertising` *data, using* `sales` *as the response and* `TV` *as the predictor. The red dots correspond to the least squares estimates* $\hat{\beta}_0$ *and* $\hat{\beta}_1$*, given by (3.4).*

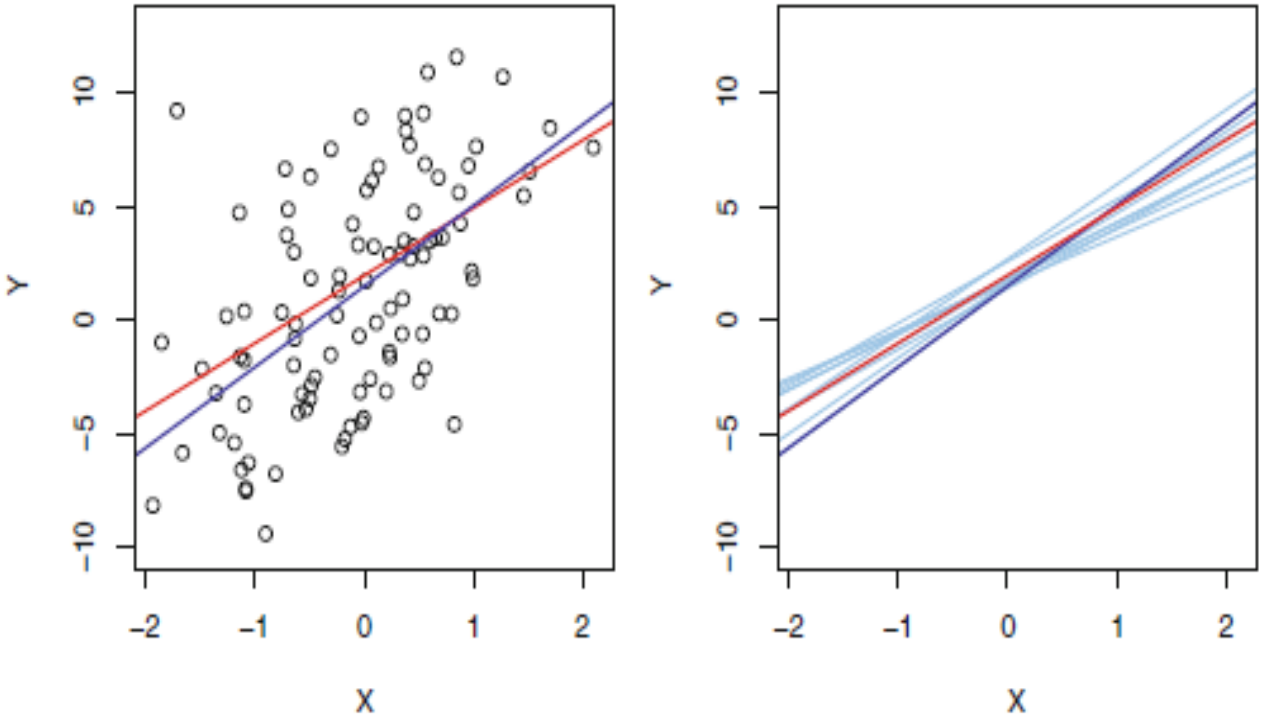


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Foram criados 100 valores aleatórios de X. E 100 valores de Y foram criados a partir de:

$$Y = 2 + 3X + \epsilon,$$

Linha vermelha: relação verdadeira.