Regressão Linear Simples

Parte 2 – Na visão de Escolha de Modelos

Se Y for uma variável aleatória, então

o modelo $Y = y$ não

poderá produzir valores

de Y, com precisão.

Modelo Simples

$$Y = \mu + \varepsilon$$

$\mu = \mathbb{E}(Y)$: constante

$\mathbb{E}(\varepsilon) = 0$

$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y) = \sigma_y^2$

Se $Y \sim N(\mu, \sigma_y^2)$

então basta sabermos μ e σ_y^2 para que o modelo apresentado tenha uma definição completa.

$Y = \mu + \varepsilon$

$\mu = \mathbb{E}(Y)$: constante

$\varepsilon \sim N(0; \sigma_y^2)$

Quem seria um estimador para μ?

Como μ é esperança de Y

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Quem seria um estimador para σ_y^2?

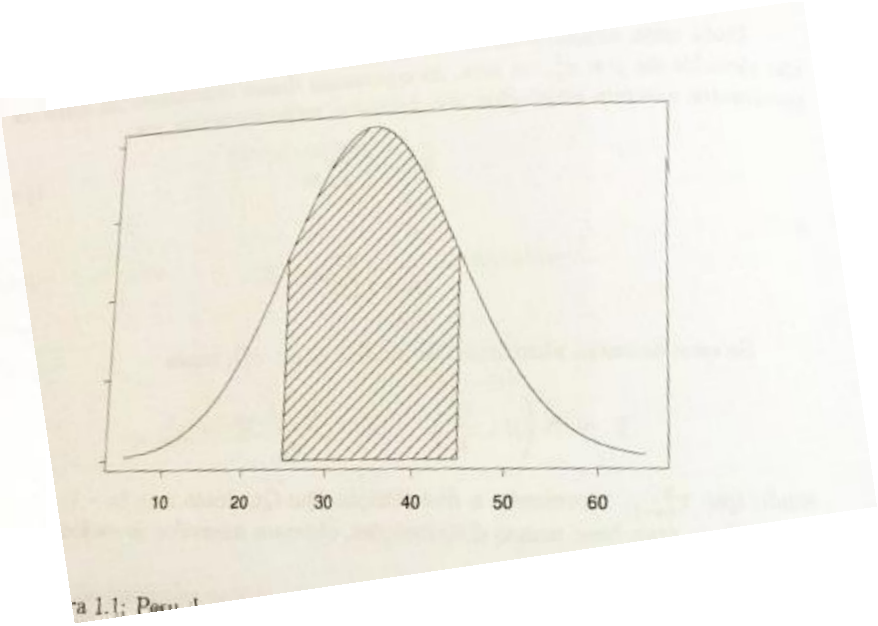
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Exemplo 1:

Exemplo 1.1

Y : peso

O peso de meninas de 7 a 11 anos de uma certa comunidade é a variável aleatória de interesse. Suponha que esta variável seja normal com média 35 kg e variância 100 kg^2. Assim, num sorteio, onde cada menina tenha a mesma chance de ser escolhida, com probabilidade 0,68, observamos um peso na faixa [25 ; 45]. Por outro ângulo, podemos dizer que aproximadamente 68% das meninas têm pesos neste intervalo. A Figura 1.1 ilustra este modelo de probabilidade.



a 1.1: Per

$$P(\mu - \sigma \leq Y \leq \mu + \sigma) = 68\%$$

$$P(35 - 10 \leq Y \leq 35 + 10) = 68\%$$

$$P(25 \leq Y \leq 45) = 68\%$$

68% das meninas tem peso entre 25 e 45 kg.

Modelo de Regressão Linear simples

E se agora, Y dependesse de X ?

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

$\mathbf{E}(\epsilon) = 0 \quad Var(\epsilon) = \sigma^2$

Exemplo 1.2

No Exemplo 1.1 supomos que o peso de meninas, de 7 a 11 anos, de uma certa comunidade é normal com esperança 35 kg e variância 100 kg^2. Vamos usar o fato de que a altura é altamente relacionada ao peso. Assim, o peso de meninas de determinada altura está mais concentrado em uma faixa específica. Vamos supor que o peso, vinculado à altura, seja normal com esperança dependendo do valor da altura e com variância 36 kg^2. Por exemplo, esperanças 30 kg, 36 kg e 40 kg, correspondentes às alturas 1,35 m, 1,40 m e 1,50 m. Temos os seguintes resultados:

Altura (m)	faixa de pesos (kg) com probabilidade 0,68	faixa de pesos (kg) com probabilidade 0,95
1,35	[24 ; 36]	[18 ; 42]
1,40	[30 ; 42]	[24 ; 48]
1,50	[34 ; 46]	[28 ; 52]

Peso: Y
Altura: x

O peso tem distribuição Normal de média E(Y|x), que não é uma esperança condicional, e variância igual a 36, que não depende de x

Quando estratificamos pela altura é natural que a dispersão seja menor

Comparando os intervalos de probabilidade 0,68 com os valores apresentados no Exemplo 1.1, vemos que agora há mais precisão: intervalos específicos e mais concentrados. A Figura 1.3 ilustra estes modelos de probabilidade.

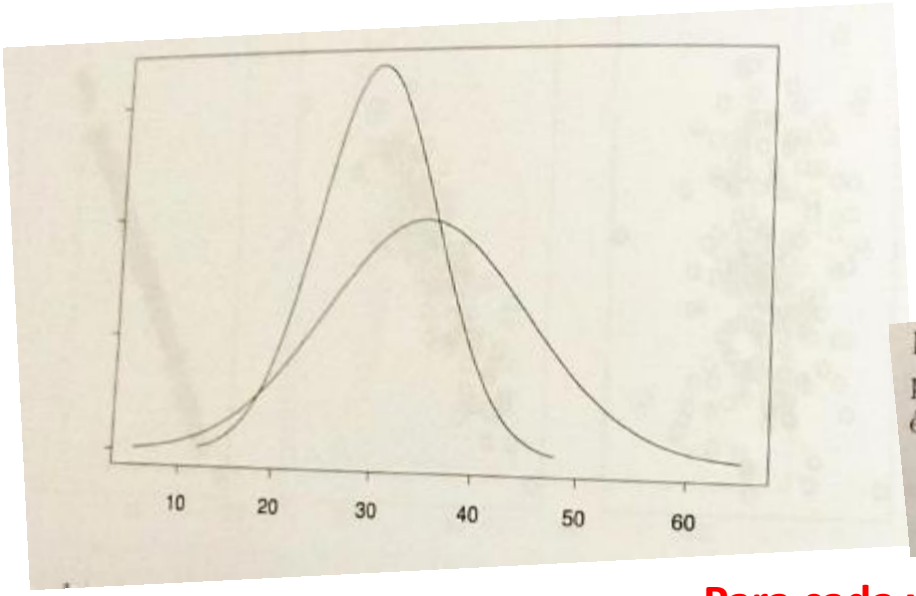


Figura 1.3: Distribuição do peso de meninas – a densidade mais dispersa é a densidade de pesos em geral; a densidade mais concentrada é a densidade de pesos de meninas com 1,35 m de altura.

Para cada x temos uma Esperança diferente de Y, e por conseguinte, uma nova distribuição.

Escolha do Modelo

Para termos certeza que de que um modelo de Regressão Linear Simples é o melhor modelo para os nossos dados, podemos fazer um teste de hipóteses para o parâmetro β_1 :

- A Hipótese Nula é representada por: $H_0: \beta_1 = 0$
- A Hipótese Alternativa é representada por: $H_1: \beta_1 \neq 0$

Observar que não temos o objetivo de destruir uma teoria pré-estabelecida (representada por H_0), mas verificar se Y depende mesmo da variável preditiva x (X).

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

versus the *alternative hypothesis*

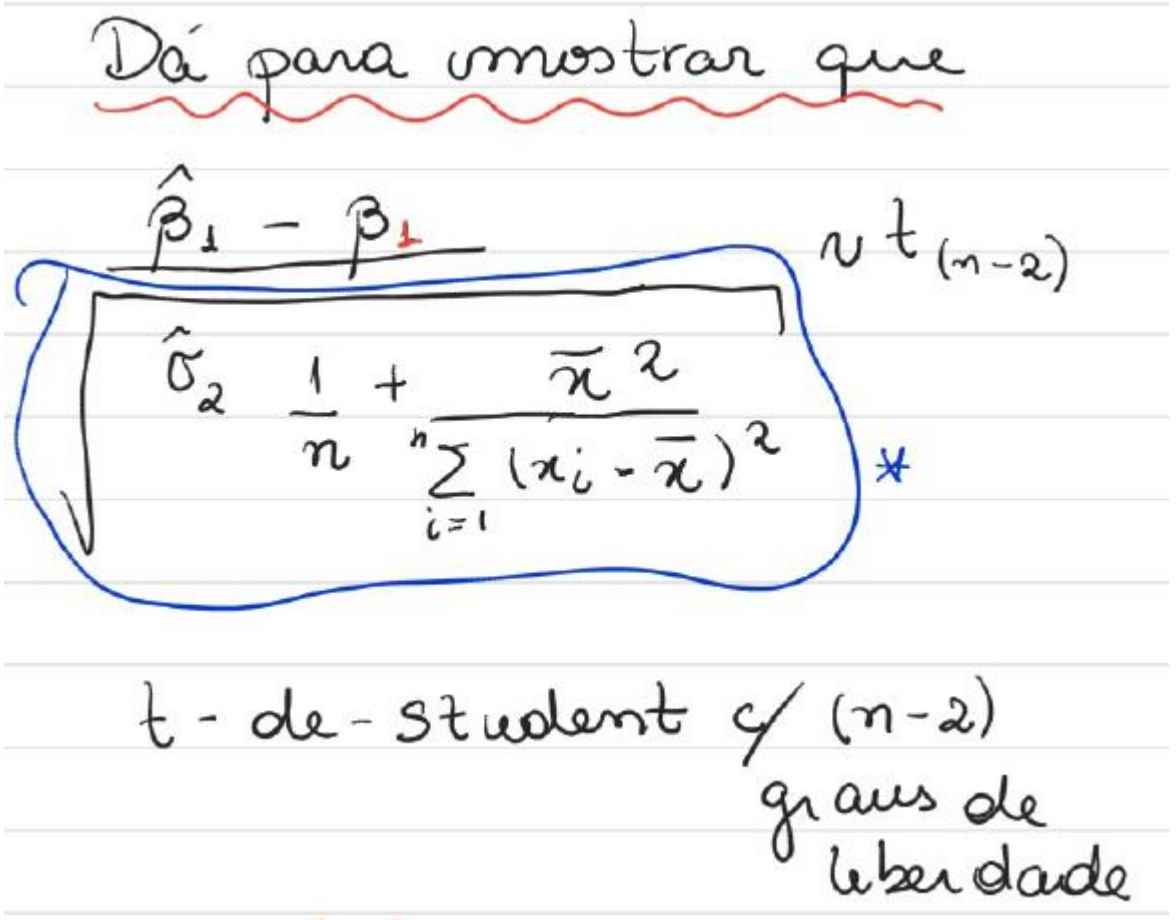
$$H_a : \text{There is some relationship between } X \text{ and } Y.$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0,$$



Estatística usada para o teste de hipóteses

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

α: nível de significância

$$\alpha = \mathbb{P}(\hat{\beta}_1 \neq 0 \mid H_0)$$

$$= \mathbb{P}(t_c^1 \leq \hat{\beta}_1 \leq t_c^2 \mid \beta_1 = 0)$$

t-de-Student