CAPÍTULO 6. REGRESSÃO LINEAR SIMPLES

6.1. O MODELO

O modelo de regressão linear simples para n observações pode ser escrito como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ para } i = 1,\ldots, n \qquad (6.1)$$

- Linear: significa que o modelo (6.1) é linear em seus parâmetros β_0 e β_1.

  Efetivamente, $E(y_i) = \beta_0 + \beta_1 x_i$ é linear nos parâmetros.

- Simples: indica que existe somente uma variável regressora x para predizer a resposta y,

- y_i e ε_i são variáveis aleatórias e os valores de x_i são constantes conhecidas ⇒ podemos usar os mesmos valores de $x_1, \ldots, x_n$ em novas amostragens.

Suposições adicionais:

1. $E(\varepsilon_i) = 0$ (ou $E(y_i) = \beta_0 + \beta_1 x_i$) para todo $i = 1,\ldots, n$
2. $var(\varepsilon_i) = \sigma^2$ (ou $var(y_i) = \sigma^2$) para todo $i = 1,\ldots, n$
3. $cov(\varepsilon_i, \varepsilon_j) = 0$ (ou $cov(y_i, y_j) = 0$) para todo $i \neq j$

A suposição 1 estabelece que o modelo (6.1) é correto, que y_i depende somente de x_i e que toda outra variação em y_i é aleatória.

A suposição 2 (*homocedasticidade*) estabelece que $var(\varepsilon_i)$ é constante e não depende dos valores de x_i.

A suposição 3 indica que os ε's (ou os y's) **não são correlacionados entre si**.

A suposição de **normalidade** será incluída mais tarde na seção 6.3 e daí os y's (ou os ε's) passarão a ser **independentes**.

Qualquer uma dessas suposições pode falhar quando trabalhamos com dados reais.

Um **gráfico de dispersão dos dados** poderá revelar alguma fuga das suposições 1 e 2.

Técnicas para verificar se essas suposições estão satisfeitas serão discutidas no capítulo 9.

6.2. ESTIMAÇÃO DE β_0, β_1 E σ^2

- Amostra de n pares (x_1, y_1), (x_2, y_2),…, (x_n, y_n)
- Estimar os parâmetros β_0, β_1 e σ^2, usando o Método dos Mínimos Quadrados Ordinários (MQO), que não faz qualquer suposição sobre a distribuição dos y's.
- Buscar estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizem a soma de quadrados dos desvios $(y_i - \hat{y}_i)$, em que $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \qquad (6.2)$$

Nota: $\hat{y}_i$ estima $E(y_i)$ e não y_i, isto é, $\hat{\beta}_0 + \hat{\beta}_1 x_i$ estima $\beta_0 + \beta_1 x_i$ e não, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

- Para obter $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, diferenciamos (6.2) em relação a $\hat{\beta}_0$ e a $\hat{\beta}_1$ e igualamos os resultados a zero:

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \qquad (6.3)$$

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0 \qquad (6.4)$$

- As soluções para (6.3) e (6.4) são dadas por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (6.5)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (6.6)$$

Para verificar que $\hat{\beta}_0$ e $\hat{\beta}_1$ minimizam $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ em (6.2), devemos examinar o sinal da derivada segunda e observar que $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ não tem máximo e que a derivada primeira produz um **mínimo global**.

Exemplo 6.2. Estudantes de Estatística alegam que as tarefas de casa não ajudam a prepará-los para o exame final. Os escores do exame (y) e das tarefas (x) para os 18 alunos da classe foram:

y	x	y	x	y	x
95	96	72	89	35	0
80	77	66	47	50	30
0	0	98	90	72	59
0	0	90	93	55	77
79	78	0	18	75	74
77	64	95	86	66	67

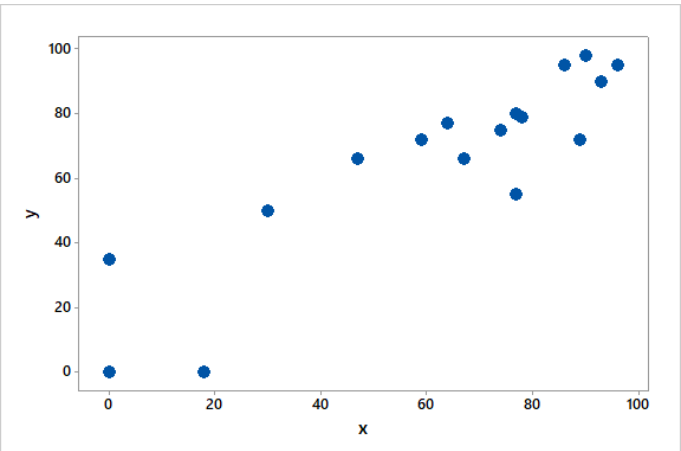

Figura 6.1.a. Gráfico de dispersão dos escores obtidos nas tarefas e na prova.

Usando (6.5) e (6.6) obtemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{811965 - 18(58{,}056)(61{,}389)}{80199 - 18(58{,}056)^2} = 0{,}8726$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 61{,}389 - 0{,}8726(58{,}056) = 10{,}73$$

A equação de predição da nota do exame em função da nota nas tarefas fica:

$$\hat{y}_i = 10{,}73 + 0{,}8726 x_i$$

O gráfico desta reta ajustada e os 18 pontos estão plotados na Figura 6.1.b.

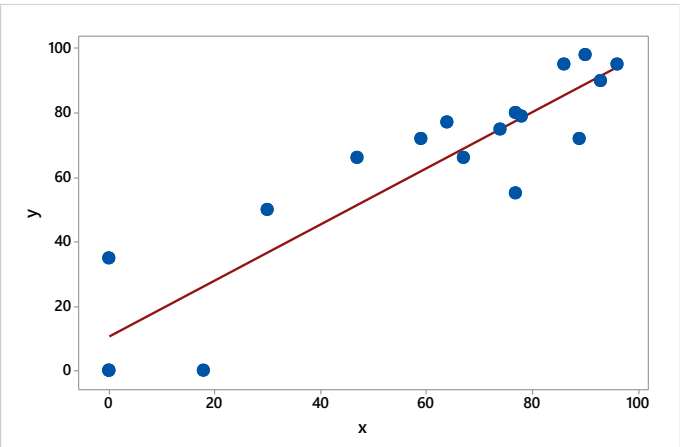

Figura 6.1.b. Gráfico de dispersão dos escores obtidos nas tarefas e na prova, com linha de regressão.

- A aparente tendência linear entre x e y no gráfico não estabelece causa e efeito entre a tarefa de casa e os resultados no exame.

- Olhando a distribuição vertical dos pontos, a suposição de que $var(\varepsilon_i) = \sigma^2$, constante, para $i = 1,\ldots, 18$, parece ser razoável.

- As suposições apresentadas na seção (6.1) não foram usadas na derivação dos estimadores de mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$.

- Se as três suposições estão satisfeitas, $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores não viesados e têm as menores variâncias dentre todos os estimadores não viesados de β_0 e β_1.

Usando as três suposições, obtemos:

- $E(\hat{\beta}_0) = \beta_0$
- $var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$
- $E(\hat{\beta}_1) = \beta_1$
- $var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sigma^2$

Assumindo que $x_1, x_2, \ldots, x_n$ permanecem os mesmos em amostras futuras, $E(\hat{\beta}_1)$ e $var(\hat{\beta}_1)$ são constantes.

- $var(\hat{\beta}_0)$ é minimizada quando $\bar{x} = 0$.
- $var(\hat{\beta}_1)$ é minimizada quando $\sum_{i=1}^{n}(x_i - \bar{x})^2$ é maximizada.

- O **Método dos Mínimos Quadrados Ordinários** não fornece um estimador para $var(y_i) = \sigma^2$, mas usando $\hat{y}_i$ como um bom estimador de $E(y_i)$, nós estimamos σ^2 por:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{SQResiduo}{n-2} \qquad (6.11)$$

- $\hat{\varepsilon}_i = (y_i - \hat{y}_i)$ é chamado de resíduo de y_i e $SQResiduo$ é a soma de quadrados residual ou soma de quadrados dos erros.

- Importante: s^2 é um estimador não viesado de σ^2.

- Intuitivamente: dividimos $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ por $(n-2)$ ao invés de $(n-1)$, porque $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ envolve dois parâmetros que precisaram ser estimados.

- $SQResiduo = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (6.13)$$

em que $\sum_{i=1}^{n}(y_i - \bar{y})^2 = SQTotal$.

Então podemos escrever que

$$SQResiduo = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 < \sum_{i=1}^{n}(y_i - \bar{y})^2 = SQTotal,$$

ou simplesmente que

$$SQResiduo < SQTotal$$

6.3. TESTE DE HIPÓTESES E INTERVALO DE CONFIANÇA PARA β_1

O maior interesse é testar a hipótese H_0: $\beta_1 = 0$, que estabelece que não existe uma relação linear entre y e x no modelo (6.1). A hipótese H_0: $\beta_1 = c$ (para $c \neq 0$) tem interesse menor.

Nota: Para construir um teste para H_0: $\beta_1 = 0$, nós precisamos admitir que $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Pode-se provar que $\hat{\beta}_1$ e s^2 têm as seguintes propriedades:

1) $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$

2) $\frac{(n-2)s^2}{\sigma^2}$ tem distribuição $\chi^2(n-2)$

3) $\hat{\beta}_1$ e s^2 são independentes.

Dessas propriedades segue por (5.34) que a estatística:

$$t = \frac{\widehat{\beta}_1}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \qquad (6.14)$$

tem distribuição t não central com $(n-2)$ graus de liberdade e parâmetro de não centralidade igual a:

$$\delta = E(\hat{\beta}_1) / \sqrt{var(\hat{\beta}_1)} = \beta_1 / \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Se $\beta_1 = 0$ (não rejeitarmos H_0) a estatística t em (6.14) tem distribuição t central com $(n-2)$ graus de liberdade.

- Para uma hipótese alternativa bilateral $H_1: \beta_1 \neq 0$, rejeitaremos $H_0: \beta_1 = 0$ se

$$| t | \geq t(\alpha/2, n-2)$$

onde $t(\alpha/2, n-2)$ é o percentil de ordem (100–α/2) da distribuição t com $n-2$ graus de liberdade e α é o nível de significância do teste.

- Alternativamente, rejeitaremos $H_0: \beta_1 = 0$ se p-valor $\leq \alpha$, em que p-valor ($p\text{-}value$) é o nível descritivo do teste ($\hat{\alpha}$), que no caso de H_1 bilateral, é calculado como

$$\hat{\alpha} = p\text{-valor} = 2P(t > t_{calc})$$

em que t_{calc} é o valor da estatística (6.14)

Os limites de um intervalo de confiança para β_1 com $100(1-\alpha)\%$ de confiança, $IC[\,\beta_1, 100(1-\alpha)\%]$, são calculados por:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \qquad (6.15)$$

6.4. COEFICIENTE DE DETERMINAÇÃO

O coeficiente de determinação R^2 é definido como

$$r^2 = \frac{SQReg}{SQTotal} = 1 - \frac{SQResiduo}{SQTotal} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (6.16)$$

Em que:

$SQReg$ é a soma de quadrados da regressão

$SQResiduo$ é a soma de quadrados residual

$SQTotal$ é a soma de quadrados total que pode ser particionada como:

$$SQTotal = SQReg + SQResiduo$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6.17)$$

- r^2 indica a proporção da variação em y que é explicada pelo modelo ou que é devida à regressão em x.

Usamos a notação r^2 em (6.16) porque o coeficiente de determinação é igual ao quadrado do coeficiente de correlação amostral entre y e x, que pode ser expresso como:

$$r(x,y) = \frac{s_{xy}}{\sqrt{s_x^2\, s_y^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (6.18)$$

onde s_{xy} é dado por (5.15).

Quando x é uma variável aleatória, $r(x, y)$ estima a correlação populacional entre as duas variáveis.

A estatística t em (6.14) usada para testar H_0: $\beta_1 = 0$, também pode ser expressa em termos de $r(x, y)$, como segue:

$$t = \frac{r(x,y)\sqrt{n-2}}{\sqrt{1 - r^2(x,y)}} \tag{6.20}$$

EXERCÍCIOS: Ver exercícios das páginas 119-120 do livro-texto.