DSBD - Modelos Lineares

Slides: Cesar Taconeli, Apres.: José Padilha

03 de agosto, 2018

Aula 2 - Regressão linear simples

Definição e propriedades

- O modelo de regressão linear simples é definido por uma reta que estabelece a relação entre uma variável resposta y e uma única variável explicativa x, da seguinte forma:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{1}$$

em que β_0 é o intercepto e β_1 a inclinação da reta, e ϵ representa o erro aleatório.

- Usualmente assumimos que os erros tem média zero e variância (desconhecida) constante, isso é, $E(\epsilon) = 0$ e $Var(\epsilon) = \sigma^2$.
- Adicionalmente, vamos supor que os erros associados a diferentes observações sejam não correlacionados, o que implica $Cov(\epsilon_i, \epsilon_{i'}) = 0$.

Definição e propriedades

- Condicional a um valor observado x, a média da distribuição de y fica dada por:

$$E(y|x) = \beta_0 + \beta_1 x. \tag{2}$$

- A variância de y, condicional a x, é dada por:

$$Var(y|x) = \sigma^2. \tag{3}$$

Definição e propriedades

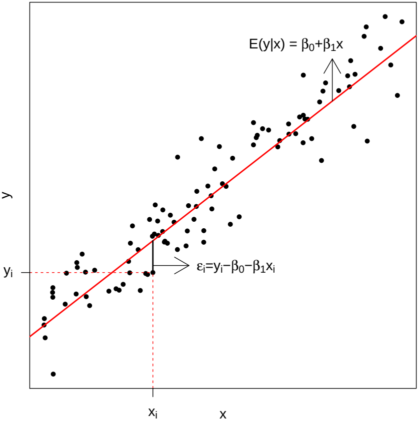

Figura 1: Regressão linear simples.

Definição e propriedades

- Interpretação dos parâmetros do modelo:
  - β_1 expressa a alteração no valor esperado de y associada ao acréscimo de uma unidade em x;
  - β_0 é o valor esperado de y quando $x = 0$ (caso $x = 0$ faça parte do suporte do problema).

Definição e propriedades

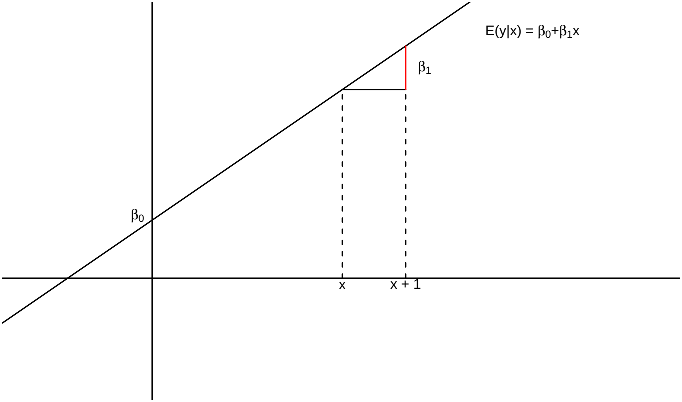

Figura 2: Interpretação dos parâmetros.

Estimação por mínimos quadrados

- A estimação de β_0 e β_1 por mínimos quadrados baseia-se em n observações para as quais se dispõe dos valores de x e y, ou seja, (x_1, y_1), (x_2, y_2),. . . , (x_n, y_n):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, ..., n. \tag{4}$$

- O método de mínimos quadrados baseia-se na determinação de β_0 e β_1 tal que a soma de quadrados dos erros, definida na sequência, seja mínima:

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{5}$$

Estimação por mínimos quadrados

Exemplo 1

Os dados a seguir referem-se às alturas de plantas (y, em centímetros) com diferentes idades (x, em semanas).

Idade (x)	1	2	3	4	5	6	7
Altura (y)	5	13	16	23	33	38	40

Estimação por mínimos quadrados

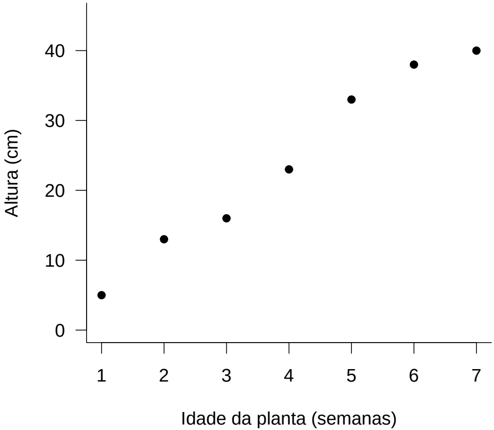

Figura 3: Gráfico de dispersão para os dados das plantas.

Estimação por mínimos quadrados

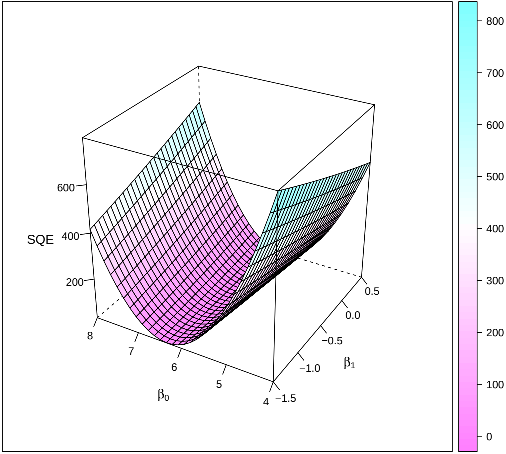

Figura 4: Ilustração da estimação por mínimos quadrados.

Estimação por mínimos quadrados

- Observando a figura 4, as estimativas de mínimos quadrados para β_0 e β_1 (denotadas por $\hat{\beta}_0$ e $\hat{\beta}_1$) correspondem aos valores de β_0 e β_1 tais que SQE seja mínimo.
- Para o presente problema, as estimativas de mínimos quadrados são dadas por $\hat{\beta}_0 = -0.57$ e $\hat{\beta}_1 = 6.14$.
- O modelo ajustado é usualmente expresso da seguinte forma:

$$\hat{y} = -0.57 + 6.14x, \tag{6}$$

em que $\hat{y}$ denota a altura predita pelo modelo para uma planta com idade x.

Estimação por mínimos quadrados

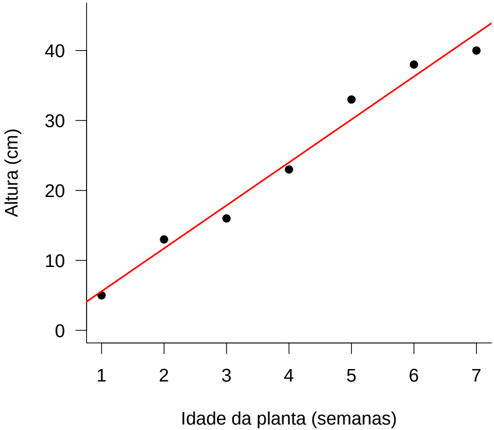

Figura 5: Gráfico de dispersão para os dados das plantas com a reta de regressão de mínimos quadrados.

Estimação por mínimos quadrados

- Os estimadores de mínimos quadrados devem satisfazer:

$$\frac{\partial S}{\partial \beta_0}\Big|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0; \tag{7}$$

$$\frac{\partial S}{\partial \beta_1}\Big|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0. \tag{8}$$

Estimação por mínimos quadrados

- A solução do sistema apresentado resulta nos seguintes estimadores de mínimos quadrados:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9}$$

e

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} y_i. \tag{10}$$

Estimação por mínimos quadrados

- O modelo de regressão linear simples ajustado pode ser representado, genericamente, da seguinte forma:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{11}$$

- A diferença entre o valor observado e o valor ajustado para uma particular observação é definido **resíduo**:

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \qquad i = 1, 2, ..., n. \tag{12}$$

- Ao contrário dos erros, resíduos podem ser calculados, e são importantes para a checagem da qualidade do ajuste.

Propriedades dos estimadores de mínimos quadrados

- Os estimadores de mínimos quadrados são combinações lineares dos $y's$;
- Os estimadores de mínimos quadrados são não viciados em relação aos respectivos parâmetros:

$$E(\hat{\beta}_0) = \beta_0; \qquad E(\hat{\beta}_1) = \beta_1. \tag{13}$$

- As variâncias de $\hat{\beta}_1$ e $\hat{\beta}_0$ são dadas, respectivamente, por:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \tag{14}$$

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right). \tag{15}$$

Propriedades dos estimadores de mínimos quadrados

Teorema de Gauss Markov

Satisfeitas as suposições assumidas para a distribuição dos erros, os estimadores de mínimos quadrados tem menor variância que quaisquer outros estimadores não viciados que sejam combinações lineares dos $y's$.

Estimação de σ^2

- A estimação de σ^2 é necessária para avaliar a precisão de $\hat{\beta}_0$ e $\hat{\beta}_1$, construir intervalos de confiança e executar testes de hipóteses.
- O estimador usual de σ^2 é baseado na soma de quadrados de resíduos:

$$SQRes = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{16}$$

- Como o valor esperado de SQRes é $(n-2)\sigma^2$, um estimador não viciado de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SQRes}{n-2} = QMRes. \tag{17}$$

- Por depender da soma de quadrados de resíduos, a especificação incorreta do modelo compromete o uso de $\hat{\sigma}^2$ na estimação de σ^2.

Regressão com dados centrados

- Uma forma alternativa de conduzir a análise de regressão é considerando os desvios da variável explicativa em torno de sua média:

$$y_i = \beta_0' + \beta_1'(x_i - \bar{x}) + \epsilon_i. \tag{18}$$

- O efeito de centrar os valores de x_i em torno de $\bar{x}$ é deslocar a origem dos $x's$ de zero para $\bar{x}$.

Regressão com dados centrados

- Como resultado, apenas o intercepto do modelo fica alterado para $\beta_0' = \beta_0 + \beta_1 \bar{x}$, em que β_0 e β_1 são os parâmetros do modelo com a variável x não centrada.

- O estimador de mínimos quadrados de β_0' fica dado por $\bar{y}$, e o estimador de β_1 não é afetado pela transformação. Portanto, o modelo ajustado fica dado por:

$$\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x}) \tag{19}$$

Testes de hipóteses e intervalos de confiança para os parâmetros do modelo

- Neste ponto teremos que assumir, adicionalmente, que os erros são normalmente distribuídos (isto é, os erros são independentes com $\epsilon \sim Normal(0, \sigma^2)$).

- A suposição de que os erros têm distribuição Normal implica $y|x \overset{ind}{\sim} Normal(\beta_0 + \beta_1 x, \sigma^2)$.

Testes de hipóteses e intervalos de confiança para os parâmetros do modelo

- Como $\hat{\beta}_1$ é uma combinação linear dos y's, decorre que também $\hat{\beta}_1$ tem distribuição Normal:

$$\hat{\beta}_1 \sim Normal\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right). \tag{20}$$

- De maneira semelhante:

$$\hat{\beta}_0 \sim Normal\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right) \tag{21}$$

Testes de hipóteses e intervalos de confiança para os parâmetros do modelo

- A distribuição conjunta dos estimadores de mínimos quadrados é dada por:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) & \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix} \right), \tag{22}$$

em que $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ e N_2 denota a distribuição Normal bivariada.

Testes de hipóteses e intervalos de confiança para β_1

- Vamos considerar o teste de que β_1 é igual a um particular valor postulado constante β_{10}:

$$H_0 : \beta_1 = \beta_{10} \ \ vs \ \ H_1 : \beta_1 \neq \beta_{10}. \tag{23}$$

- Então, sob a hipótese H_0 (ou seja, assumindo que $\beta_1 = \beta_{10}$:

$$Z = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim Normal(0, 1). \tag{24}$$

Testes de hipóteses e intervalos de confiança para β_1

- Como σ^2 geralmente é desconhecido, ele usualmente é estimado usando o seguinte estimador:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SQRes}{n-2} = QMRes. \tag{25}$$

- O estimador $\hat{\sigma}^2$ é não viciado e consistente na estimação de σ^2. Além disso, sua distribuição, sob as especificações do modelo, é dada por:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}, \tag{26}$$

em que χ^2 denota a distribuição qui-quadrado com $n-2$ graus de liberdade.

Testes de hipóteses e intervalos de confiança para β_1

- Substituindo σ^2 por $\hat{\sigma}^2$ em (24), temos:

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}, \qquad (27)$$

em que t_{n-2} representa a distribuição t-Student com $n-2$ graus de liberdade.

- Com base no resultado (30) pode-se conduzir o teste da hipótese $H_0 : \beta_1 = \beta_{10}$.

- Fixando o nível de significância em α, H_0 será rejeitada se $|t| > |t_{n-2;\alpha/2}|$, em que $t_{n-2;\alpha/2}$ é o quantil $\alpha/2$ da distribuição t_{n-2}.

Testes de hipóteses e intervalos de confiança para β_1

- O nível descritivo (valor-p) do teste fica definido por:

$$p = 2 \times P(X > |t|), \text{ em que } X \sim t_{n-2}. \tag{28}$$

- Um intervalo de confiança $100(1-\alpha)\%$ para β_1 é definido pelo par de limites:

$$\hat{\beta}_1 \mp t_{n-2;\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{29}$$

Teste da significância da regressão

- Uma importante hipótese a ser testada é $H_0 : \beta_1 = 0$ *vs* $H_0 : \beta_1 \neq 0$.

- Chamamos esse teste de **teste da significância da regressão linear simples**.

- Neste caso, a estatística do teste fica dada por:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}, , \tag{30}$$

que será rejeitada, a um nível de significância α, se $|t| > |t_{n-2;\alpha/2}|$

Teste da significância da regressão

- É importante ressaltar que a não rejeição de $H_0 : \beta_1 = 0$ permite concluir que não há relação linear entre y e x, mas não que não se tenha relação entre as variáveis.

- Além disso, ainda que H_0 seja rejeitada, isso não implica que um modelo não linear (como um polinômio, por exemplo), seja mais adequado para explicar a relação entre as variáveis.

Testes de hipóteses e intervalos de confiança para β_0

- De maneira similar, considere $H_0 : \beta_0 = \beta_{00}$ *vs* $H_1 : \beta_0 \neq \beta_{00}$ um par de hipóteses postuladas para o intercepto do modelo.
- Sob as suposições do modelo:

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim t_{n-2}, \tag{31}$$

sob a suposição de que a hipótese nula é verdadeira.

Testes de hipóteses e intervalos de confiança para β_0

- Fixando o nível de significância em α, novamente H_0 será rejeitada se $|t| > |t_{n-2;\alpha/2}|$, em que $t_{n-2;\alpha/2}$ é o quantil $\alpha/2$ da distribuição t_{n-2}.

- Um intervalo de confiança $100(1-\alpha)\%$ para β_0 é definido pelo par de limites:

$$\hat{\beta}_0 \mp t_{n-2;\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} \qquad (32)$$

Intervalo de confiança para σ^2

- Um intervalo de confiança $100(1-\alpha)\%$ para σ^2 pode ser obtido com base na distribuição qui-quadrado (χ^2):

$$\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;1-\alpha/2}} \quad ; \quad \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;\alpha/2}}, \tag{33}$$

em que $\chi^2_{n-2;\alpha/2}$ e $\chi^2_{n-2;1-\alpha/2}$ são os quantis $\alpha/2$ e $1-\alpha/2$ da distribuição qui-quadrado com $n-2$ graus de liberdade.

Intervalo de confiança para a resposta média

- Suponha que se deseja estimar a média de y para um particular valor $x = x_0$.
- A estimativa pontual pode ser calculada por:

$$\hat{\mu}_{y|x_0} = \widehat{E(y|x = x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{34}$$

- Como $\hat{\beta}_0$ e $\hat{\beta}_1$ têm distribuição Normal, $\hat{\mu}_{y|x_0}$ também é normalmente distribuído (pois é uma combinação linear de $\hat{\beta}_0$ e $\hat{\beta}_1$).
- A variância de $\hat{\mu}_{y|x_0}$ é dada por:

$$Var(\hat{\mu}_{y|x_0}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right). \tag{35}$$

Intervalo de confiança para a resposta média

- O intervalo de confiança para $\mu_{y|x_0}$ baseia-se na seguinte distribuição amostral:

$$\hat{\mu}_{y|x_0} \sim Normal\left(\mu_{y|x_0}, \sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}\right) \qquad (36)$$

- Substituindo σ^2 por $\hat{\sigma}^2 = QMRes$:

$$\frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim t_{n-2} \qquad (37)$$

Intervalo de confiança para a resposta média

- Dessa forma, o intervalo de confiança $100(1-\alpha)\%$ para a média de y quando $x = x_0$ tem limites:

$$\hat{\mu}_{y|x_0} \mp t_{n-2;\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} \tag{38}$$

Predição de uma nova observação

- Seja $\hat{y}_0$ a predição de uma nova observação para um particular valor $x = x_0$. A estimativa pontual é a mesma de $\hat{\mu}_{y|x_0}$:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{39}$$

- A variância de $\hat{y}_0$, no entanto, é dada por:

$$\begin{gathered} var(\hat{y}_0) = Var(\hat{\mu}_{y|x_0}) + var(y_0|\mu_{y|x_0} = \hat{\mu}_{y|x_0}) = \\ \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) + \sigma^2 = \\ \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right). \end{gathered} \tag{40}$$

Predição de uma nova observação

- Um intervalo de predição $100(1-\alpha)\%$ para uma observação futura em x_0 tem os seguintes limites:

$$\hat{y}_0 \mp t_{n-2;\alpha/2}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} \tag{41}$$

- Em problemas de regressão linear com apenas uma variável explicativa, é comum representar graficamente o modelo de regressão ajustado acompanhado das **bandas de confiança** para a média e **bandas de predição** para observações futuras.

Estimação por máxima verossimilhança

- A estimação de β_0 e β_1 por máxima verossimilhança baseia-se, novamente, em n observações para as quais se dispõe dos valores de x e y, ou seja, (x_1, y_1), (x_2, y_2),..., (x_n, y_n):
- Vamos assumir $\epsilon \sim N(0, \sigma^2)$, tal que $y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$.
- Assumindo que os erros sejam independentes, a função de verossimilhança fica dada pelo produto da f.d.p. normal avaliada nas n observações:

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2; \boldsymbol{y}, \boldsymbol{x}) &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\
&= \left(2\pi\sigma^2\right)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]
\end{aligned} \tag{42}
$$

Estimação por máxima verossimilhança

- Dessa forma, a função de log-verossimilhança fica dada por:

$$\ln L(\beta_0, \beta_1, \sigma^2; \boldsymbol{y}, \boldsymbol{x}) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad (43)$$

Estimação por máxima verossimilhança

- Os estimadores de máxima verossimilhança devem satisfazer a:

$$\begin{aligned} \frac{\partial S}{\partial \beta_0}\Big|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma^2}} lnL(\beta_0,\beta_1,\sigma^2;\boldsymbol{y},\boldsymbol{x}) = 0; \\ \frac{\partial S}{\partial \beta_1}\Big|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma^2}} lnL(\beta_0,\beta_1,\sigma^2;\boldsymbol{y},\boldsymbol{x}) = 0; \\ \frac{\partial S}{\partial \sigma^2}\Big|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma^2}} lnL(\beta_0,\beta_1,\sigma^2;\boldsymbol{y},\boldsymbol{x}) = 0. \end{aligned} \tag{44}$$

Estimação por máxima verossimilhança

- Observe que maximizar $\ln L(\beta_0, \beta_1, \sigma^2; \boldsymbol{y}, \boldsymbol{x})$ com relação a β_0 e β_1 equivale a maximizar $-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = -SQE$ em função desses parâmetros;

- Lembre que na estimação por mínimos quadrados a obtenção dos estimadores dos parâmetros do modelo era obtida pela minimização de $SQE = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$;

- Uma vez que minimizar SQE é equivalente a maximizar $-SQE$, os estimadores de máxima verossimilhança para β_0 e β_1 são idênticos aos de mínimos quadrados.

Estimação por máxima verossimilhança

- O estimador de máxima verossimilhança de σ^2, por sua vez, é dado por:

$$\hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n}, \tag{45}$$

que, diferentemente do estimador estudado anteriormente, é viciado para σ^2 (mas **assintoticamente** não viciado).

Análise de variância aplicada à regressão linear simples

- A análise de variância é uma técnica que permite particionar a variação total dos dados em parcelas atribuíveis a diferentes fontes.

- No contexto de regressão, a análise de variância baseia-se na seguinte identidade:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i), \quad i = 1, 2, ..., n. \tag{46}$$

Análise de variância aplicada à regressão linear simples

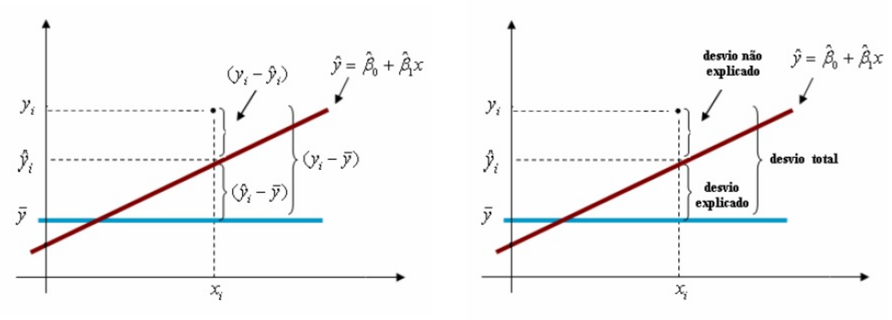

Figura 6: Decomposição da variação dos dados na regressão linear simples.

Análise de variância aplicada à regressão linear simples

- Para um conjunto de n observações, a variabilidade total dos dados (em torno da média) pode ser decomposta da seguinte forma:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SQ_{Total}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SQ_{Reg}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SQ_{Res}}, \tag{47}$$

em que:

- SQ_{Total} é a variabilidade total dos dados (corrigida pela média);
- SQ_{Reg} é a variabilidade dos dados explicada pela regressão;
- SQ_{Res} é a variabilidade dos dados não explicada pela regressão (variação residual).

Análise de variância aplicada à regressão linear simples

- Dessa forma, quanto maior SQ_{Reg} em detrimento a SQ_{Res}, maior a parcela da variação total dos dados explicada pela regressão.

- Associado a cada componente dessa decomposição temos:
  - $n-1$ graus de liberdade para SQ_{Total} (perda de um grau devido à estimação da média);
  - $n-2$ graus de liberdade para SQ_{Res} (perda de dois graus devido à estimação de β_0 e β_1);
  - $(n-1)-(n-2)=1$ graus de liberdade para SQ_{Reg}.

- O resultado da análise de variância pode ser sumarizado através do quadro da análise.

Análise de variância aplicada à regressão linear simples

Tabela 2: Quadro de análise de variância

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F
Regressão	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	$QM_{Reg} = \frac{SQ_{Reg}}{1}$	$F = \frac{QM_{Reg}}{QM_{Res}}$
Resíduos	n-2	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$QM_{Res} = \frac{SQ_{Res}}{n-2}$	
Total	n-1	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

- A significância da regressão linear pode ser testada com base na análise de variância, **com resultado idêntico** ao apresentado anteriormente no teste da hipótese $H_0 : \beta_1 = 0$.

Análise de variância aplicada à regressão linear simples

- O teste da significância do modelo via ANOVA baseia-se em:
  - $\frac{(n-2)QM_{Res}}{\sigma^2} \sim \chi_{n-2}$;
  - Sob a hipótese nula (isso é, se $\beta_1 = 0$), então $\frac{SQ_{Reg}}{\sigma^2}$ tem distribuição χ_1;
  - SQ_{Reg} e SQ_{Res} são independentes.
- Então:

$$F = \frac{SQ_{Reg}/1}{SQ_{Res}/(n-2)} = \frac{QM_{Reg}}{QM_{Res}} \tag{48}$$

tem distribuição $F - Snedecor$ com parâmetros 1 e $n-2$.

- Assim, $H_0 : \beta_1 = 0$ será rejeitada, a um nível de significância α se $F > F_{1,n-2;1-\alpha}$.

Análise de variância aplicada à regressão linear simples

- O **coeficiente de determinação** do modelo é definido por:

$$R^2 = \frac{SQ_{Reg}}{SQ_{Total}}, \tag{49}$$

tal que $0 \leq R^2 \leq 1$.

- Dessa forma, R^2 corresponde à proporção da variação dos dados explicada pela regressão.
- Para o caso da regressão linear simples, $R^2 = r^2$, em que r é o coeficiente de correlação linear.
- O valor de R^2 deve ser interpretado com cautela uma vez que um elevado valor de R^2 não implica, necessariamente, num modelo bem ajustado.

Análise de variância aplicada à regressão linear simples

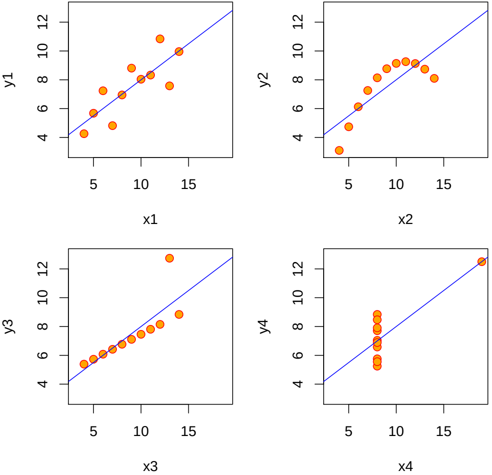

Figura 7: Quatro conjuntos de dados que produzem mesmo valor de R^2

Caso em que x também é aleatório - análise de correlação

- Em algumas situações, pode não ser razoável admitir que a variável explicativa x seja fixa.

- Como exemplo, num experimento na agronomia em que está se estudando produção vegetal, pode ser pouco realista assumir a altura das plantas ou o número de folhas como não sendo aleatórios;

- Vamos estudar agora o caso em que x e y são variáveis aleatórias e o estudo da distribuição conjunta.

O caso de x e y com distribuição normal bivariada - análise de correlação

- Considere que o par de variáveis aleatórias x e y tenha distribuição normal bivariada:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\}, \tag{50}$$

em que μ_x e σ_x^2 são a média e a variância de x; μ_y e σ_y^2 são a média e a variância de y e

$$\rho = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x\sigma_y} = \frac{Cov(x,y)}{DP(x)DP(y)} \tag{51}$$

é o coeficiente de correlação entre x e y.

O caso de x e y com distribuição normal bivariada - análise de correlação

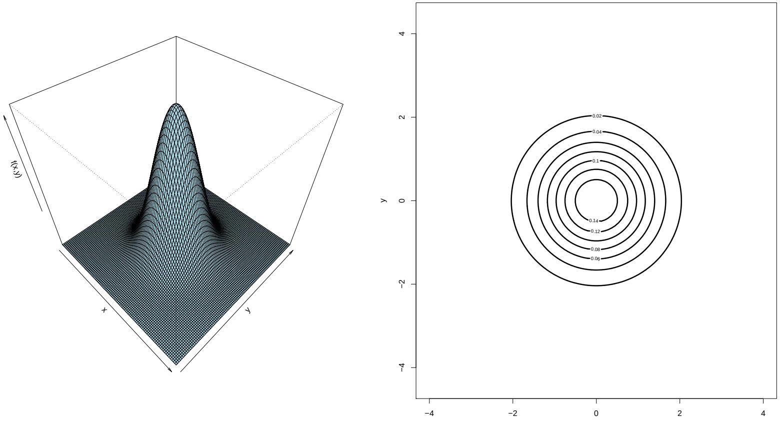

Figura 8: Distribuição normal bivariada: $\rho = 0$.

O caso de x e y com distribuição normal bivariada - análise de correlação

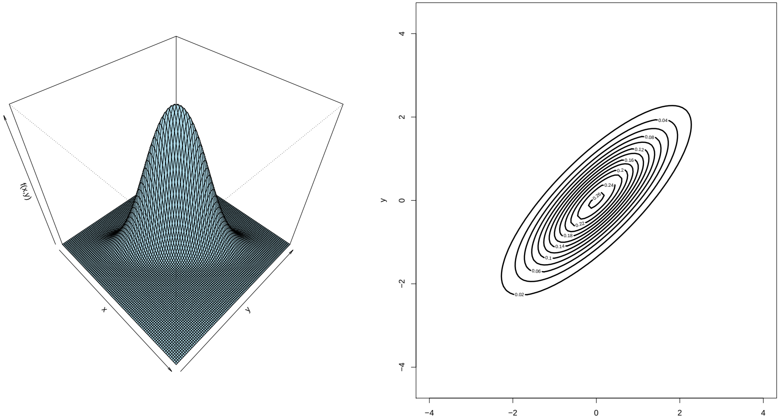

Figura 9: Distribuição normal bivariada: $\rho = 0.8$.

O caso de x e y com distribuição normal bivariada - análise de correlação

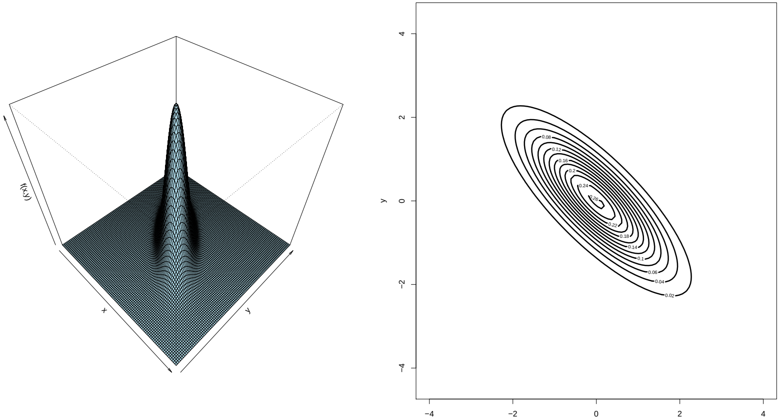

Figura 10: Distribuição normal bivariada: $\rho = -0.8$.

Análise de correlação

- O estimador de ρ é o coeficiente de correlação amostral, dados por:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}. \tag{52}$$

- Verifica-se facilmente que:

$$\hat{\beta}_1 = \left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) r, \tag{53}$$

de forma que $\hat{\beta}_1$, a inclinação da reta de mínimos quadrados, é o coeficiente de correlação amostral multiplicado por um fator de escala.

Análise de correlação

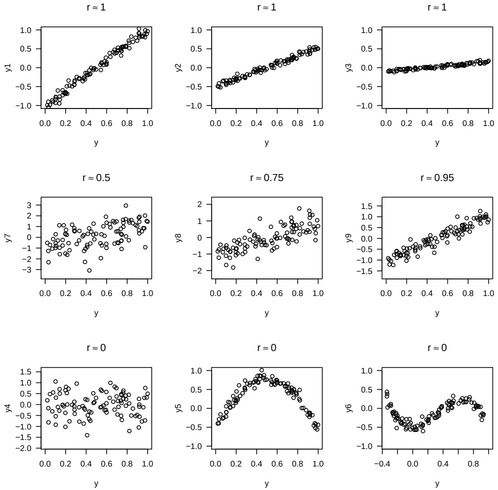

Figura 11: Ilustração de dados com diferentes níveis de correlação linear.

Análise de correlação

- Pode se testar a hipótese que a correlação linear entre um par de variáveis é igual a zero, configurando o seguinte par de hipóteses:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

.

- A estatística teste, neste caso, é dada por:

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \tag{54}$$

que, sob a hipótese nula ($\rho = 0$), tem distribuição t_{n-2}.

Análise de correlação

- Assim, a hipótese de correlação nula deverá ser rejeitada, ao nível de significância de α, se $|t| > |t_{n-2;\alpha/2}|$.

- O nível descritivo do teste pode ser calculado por $p = 2 \times P(X > |t|)$, sendo $X \sim t_{n-2}$.

- Um intervalo de confiança $100(1-\alpha)\%$ para ρ pode ser obtido da seguinte forma:

$$\tanh\left(\text{arctan}\, r - \frac{z_{\alpha/2}}{\sqrt{n-3}}; \quad \text{arctan}\, r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right), \tag{55}$$

em que:

$$\text{arctan}\, r = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}; \quad \tanh u = \frac{e^u - e^{-u}}{e^u + e^{-u}}. \tag{56}$$

Modelos intrinsicamente lineares

- Em alguns casos em a relação entre as variáveis é não linear mas pode ser linearizada aplicando alguma transformação adequada.

- Os modelos de regressão resultantes são denominados *modelos intrinsicamente lineares*.

- Usar transformações pode remediar o não atendimento de diferentes pressupostos do modelo (como variância não constante ou ausência de normalidade).

- Neste ponto vamos nos ater à aplicação de transformações com o objetivo de linearizar a relação entre as variáveis.

Modelos intrinsicamente lineares

Tabela 3: Exemplos de modelos intrinsicamente lineares

Função linearizável	Transformação	Forma linear
$y = \beta_0 x^{\beta_1}$	$y' = log(y); x' = log(x)$	$y' = log(\beta_0) + \beta_1 x$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln\beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y' = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}; x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Modelos intrinsicamente lineares

- Qualquer uma dessas transformações requer que os erros **na escala transformada** sejam independentes, normalmente distribuídos com média zero e variância σ^2.

- Quando o método de mínimos quadrados é aplicado após transformação as propriedades dos estimadores, que estudamos anteriormente, valem para os dados transformados e não necessariamente para os dados originais.