CE310 - Modelos de Regressão Linear

Regressão linear simples

Cesar Augusto Taconeli

24 de março, 2025

Introdução

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:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

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

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:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

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

- Assume-se que os erros tem média zero e variância constante, isso é, $\mathrm{E}(\epsilon) = 0$ e $\mathrm{Var}(\epsilon) = \sigma^2$.

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:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

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

- Assume-se que os erros tem média zero e variância constante, isso é, $\mathrm{E}(\epsilon) = 0$ e $\mathrm{Var}(\epsilon) = \sigma^2$.

- Supomos ainda que os erros associados a diferentes observações são não correlacionados, o que implica $Cov(\epsilon_i, \epsilon_{i'}) = 0$, para duas observações i e i' quaisquer.

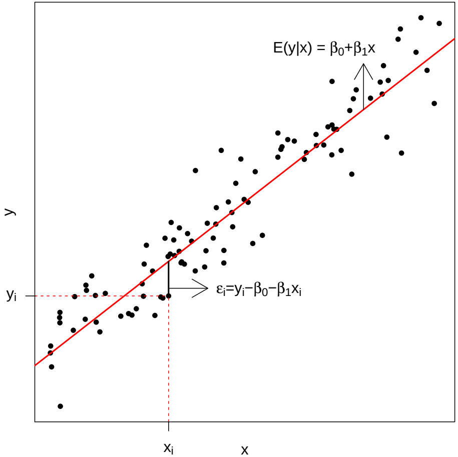


Figura 1: Regressão linear simples.

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

$$E(y|x) = \beta_0 + \beta_1 x,$$

definindo a relação linear entre as variáveis.

Definição e propriedades

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

$$E(y|x) = \beta_0 + \beta_1 x,$$

definindo a relação linear entre as variáveis.

- Já a variância de y condicional a x resulta em:

$$\mathrm{Var}(y|x) = \sigma^2,$$

que não depende do valor de x.

Definição e propriedades

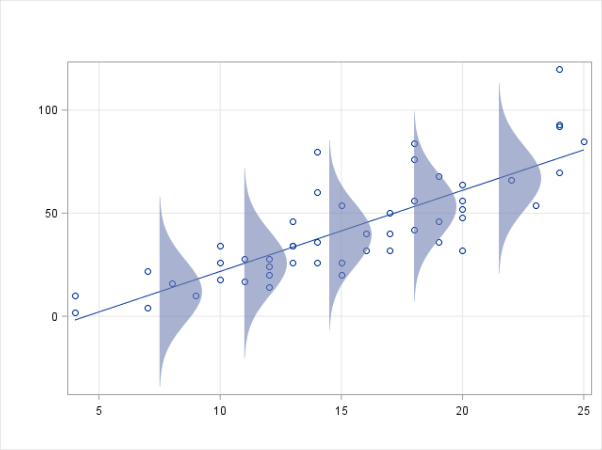


Figura 2: Regressão linear simples.

Definição e propriedades

- Interpretação dos parâmetros do modelo:

Definição e propriedades

- Interpretação dos parâmetros do modelo:
  - β_1 é a alteração no valor esperado (média) de y associada ao acréscimo de uma unidade em x;

Definição e propriedades

- Interpretação dos parâmetros do modelo:

  - β_1 é a alteração no valor esperado (média) de y associada ao acréscimo de uma unidade em x;

  - β_0 é o valor esperado de y quando $x = 0$.

Definição e propriedades

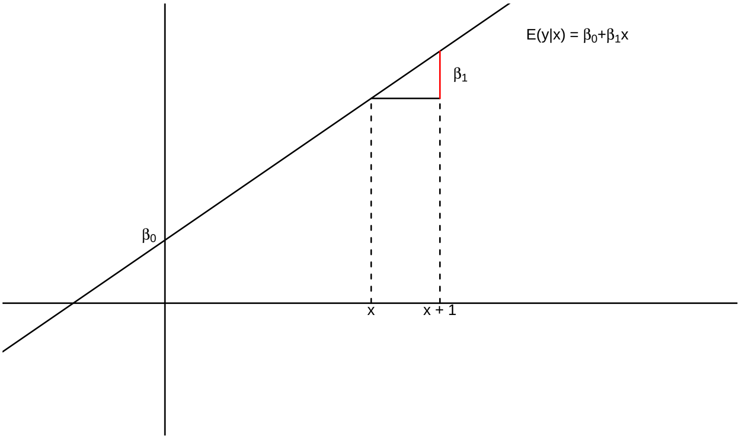


Figura 3: Interpretação dos parâmetros.

Método de mínimos quadrados

Método de mínimos quadrados

- A técnica usual para estimação dos parâmetros de um modelo de regressão linear (ajuste da regressão linear) é o **método de mínimos quadrados**.

Método de mínimos quadrados

- A técnica usual para estimação dos parâmetros de um modelo de regressão linear (ajuste da regressão linear) é o **método de mínimos quadrados**.

- Vamos motivar o problema da estimação em regressão linear por meio de um exemplo ilustrativo, usando dados de crescimento de plantas.

Exemplo- Crescimento de plantas

Exemplo- Crescimento de plantas

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

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

Exemplo - crescimento de plantas

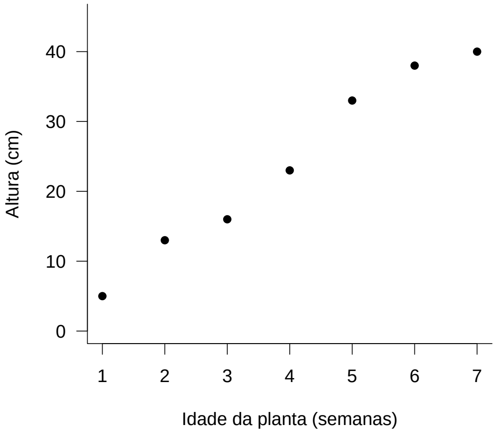


Figura 4: Gráfico de dispersão para os dados das plantas.

Exemplo- Crescimento de plantas

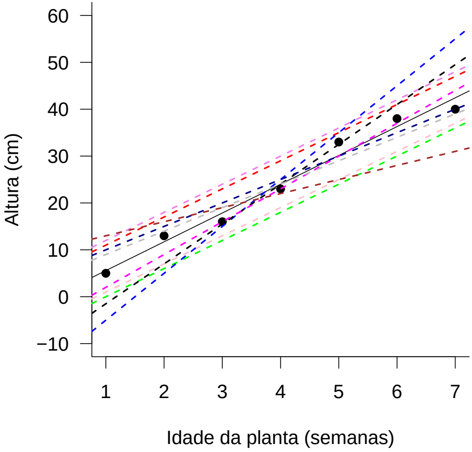


Figura 5: Gráfico de dispersão para os dados das plantas com diferentes retas ajustadas.

Estimação por mínimos quadrados

- Observe que diferentes (infinitas) retas podem ser ajustadas para explicar a altura em função da idade da planta.

Estimação por mínimos quadrados

- Observe que diferentes (infinitas) retas podem ser ajustadas para explicar a altura em função da idade da planta.

- Notadamente, algumas dessas retas proporcionam melhor ajuste aos dados.

Estimação por mínimos quadrados

- Observe que diferentes (infinitas) retas podem ser ajustadas para explicar a altura em função da idade da planta.

- Notadamente, algumas dessas retas proporcionam melhor ajuste aos dados.

- A qualidade do ajuste está relacionada à distância dos pontos à reta ajustada.

Estimação por mínimos quadrados

- Observe que diferentes (infinitas) retas podem ser ajustadas para explicar a altura em função da idade da planta.

- Notadamente, algumas dessas retas proporcionam melhor ajuste aos dados.

- A qualidade do ajuste está relacionada à distância dos pontos à reta ajustada.

- Desta forma, é desejável encontrar valores (estimativas) para os parâmetros da reta tais que as distâncias dos pontos à reta sejam mínimas.

Estimação por mínimos quadrados

- Considere 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):

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

Estimação por mínimos quadrados

- Considere 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):

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

- O método de mínimos quadrados baseia-se na determinação de β_0 e β_1 tal que a soma de quadrados dos erros (S) seja mínima:

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

Estimação por mínimos quadrados

- Os estimadores de mínimos quadrados para o modelo de regressão linear simples devem satisfazer:

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

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

Estimação por mínimos quadrados

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1}$$

e

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

Estimação por mínimos quadrados

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1}$$

e

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

Material complementar

Verificar a derivação dos estimadores de mínimos quadrados dos parâmetros do modelo de regressão linear simples.

Exemplo- Crescimento de plantas

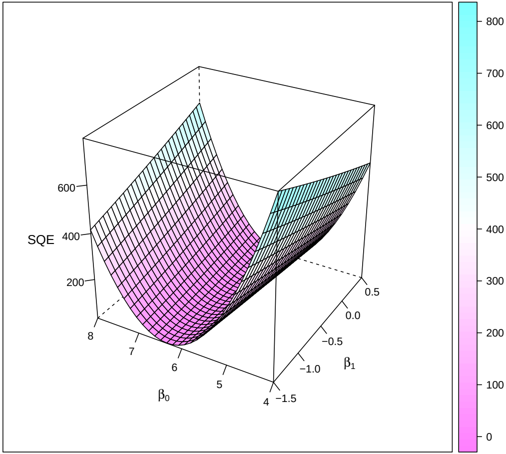


Figura 6: Ilustração da estimação por mínimos quadrados.

Exemplo- Crescimento de plantas

- Estimação de β_1:

Exemplo- Crescimento de plantas

- Estimação de β_1:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})}{\sum\left(x_i - \bar{x}\right)^2} y_i$$

Exemplo- Crescimento de plantas

- Estimação de β_1:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})}{\sum\left(x_i - \bar{x}\right)^2} y_i$$

- Como $\bar{x} = \frac{1}{n}\sum_{x_i} = \frac{1}{7} \times (1 + 2 + 3 + 4 + 5 + 6 + 7) = 4$, e:

$$\sum\left(x_i - \bar{x}\right)^2 = (1-4)^2 + (2-4)^2 + ... + (6-4)^2 + (7-4)^2 = 28$$

Exemplo- Crescimento de plantas

- Segue que:

$$\frac{1}{28} \times [(1-4) \times 5 + (2-4) \times 13 + ... + (6-4) \times 38 + (7-4) \times 40] = 6.14$$

Exemplo- Crescimento de plantas

- Segue que:

$$\frac{1}{28} \times [(1-4) \times 5 + (2-4) \times 13 + ... + (6-4) \times 38 + (7-4) \times 40] = 6.14$$

- Já para β_0, temos $\bar{y} = \frac{1}{n} \sum_{y_i} = \frac{1}{7} \times (5 + 13 + 16 + 23 + 33 + 38 + 40) = 24$, de tal forma que:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 24 - 6.14 \times 4 = -0.56$$

Exemplo- Crescimento de plantas

- Segue que:

$$\frac{1}{28} \times [(1-4) \times 5 + (2-4) \times 13 + ... + (6-4) \times 38 + (7-4) \times 40] = 6.14$$

- Já para β_0, temos $\bar{y} = \frac{1}{n} \sum_{y_i} = \frac{1}{7} \times (5 + 13 + 16 + 23 + 33 + 38 + 40) = 24$, de tal forma que:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 24 - 6.14 \times 4 = -0.56$$

- O modelo ajustado é usualmente expresso da seguinte forma:

$$\hat{y} = -0.56 + 6.14x,$$

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

Exemplo- Crescimento de plantas

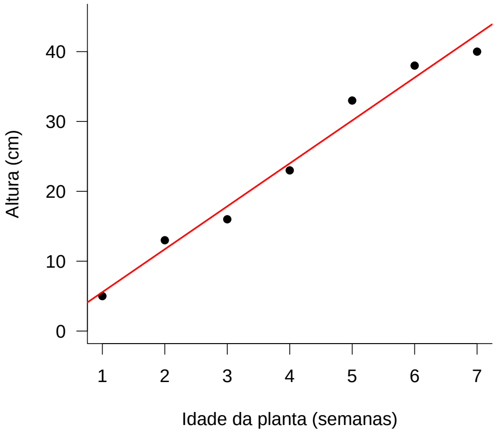


Figura 7: Gráfico de dispersão com reta de regressão ajustada por mínimos quadrados.

Estimação por mínimos quadrados

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

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

Estimação por mínimos quadrados

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

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

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

$$r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, ..., n.$$

Estimação por mínimos quadrados

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

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

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

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

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

Exemplo- Anatomia de gatos domésticos

Exemplo- Anatomia de gatos domésticos

- Nesta sessão R, vamos usar dados da anatomia de gatos domésticos. O objetivo desta aplicação é motivar a análise de regressão linear simples.

Exemplo- Anatomia de gatos domésticos

- Nesta sessão R, vamos usar dados da anatomia de gatos domésticos. O objetivo desta aplicação é motivar a análise de regressão linear simples.

- A base consiste em medidas corporais de 144 gatos domésticos (machos e fêmeas).

Exemplo- Anatomia de gatos domésticos

- Nesta sessão R, vamos usar dados da anatomia de gatos domésticos. O objetivo desta aplicação é motivar a análise de regressão linear simples.

- A base consiste em medidas corporais de 144 gatos domésticos (machos e fêmeas).

- As variáveis consideradas na análise são as seguintes:

Exemplo- Anatomia de gatos domésticos

- Nesta sessão R, vamos usar dados da anatomia de gatos domésticos. O objetivo desta aplicação é motivar a análise de regressão linear simples.

- A base consiste em medidas corporais de 144 gatos domésticos (machos e fêmeas).

- As variáveis consideradas na análise são as seguintes:
  - `Bwt-` peso corporal em kg (variável explicativa);

Exemplo- Anatomia de gatos domésticos

- Nesta sessão R, vamos usar dados da anatomia de gatos domésticos. O objetivo desta aplicação é motivar a análise de regressão linear simples.

- A base consiste em medidas corporais de 144 gatos domésticos (machos e fêmeas).

- As variáveis consideradas na análise são as seguintes:
    - `Bwt`- peso corporal em kg (variável explicativa);
    - `Hwt`- peso do coração em g (variável resposta).

Exemplo- Anatomia de gatos domésticos

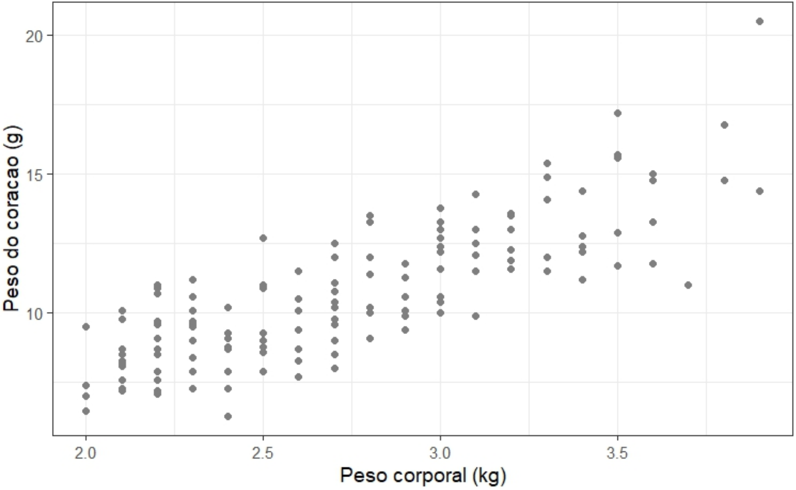


Figura 8: Dados de anatomia de gatos domésticos

Exemplo- Anatomia de gatos domésticos

- O seguinte modelo de regressão linear simples foi especificado:

$$\texttt{Hwt} = \beta_0 + \beta_1 \texttt{Bwt} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

Exemplo- Anatomia de gatos domésticos

- O seguinte modelo de regressão linear simples foi especificado:

$$\texttt{Hwt} = \beta_0 + \beta_1\texttt{Bwt} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

- O modelo apresentado pode ser escrito de maneira equivalente por:

$$\texttt{Hwt}\big|\texttt{Bwt} \sim \mathrm{N}(\mu_{\texttt{Bwt}}, \sigma^2)$$

Exemplo- Anatomia de gatos domésticos

- O seguinte modelo de regressão linear simples foi especificado:

$$\texttt{Hwt} = \beta_0 + \beta_1 \texttt{Bwt} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

- O modelo apresentado pode ser escrito de maneira equivalente por:

$$\texttt{Hwt} \big| \texttt{Bwt} \sim \mathrm{N}(\mu_{\texttt{Bwt}}, \sigma^2)$$

$$\mu_{\texttt{Bwt}} = \beta_0 + \beta_1 \texttt{Bwt}$$

Exemplo- Anatomia de gatos domésticos

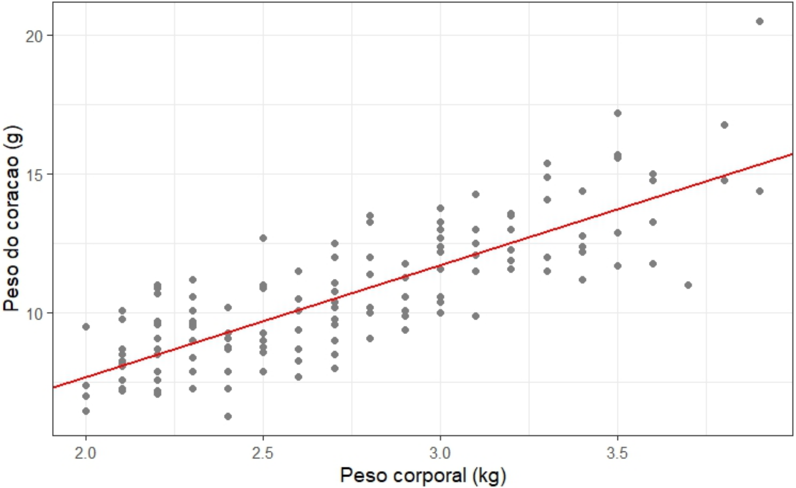


Figura 9: Dados de anatomia de gatos domésticos com reta de regressão ajustada

Exemplo- Anatomia de gatos domésticos

- Modelo ajustado por mínimos quadrados:

$$\widehat{\texttt{Hwt}} = -0.356 + 4.034\texttt{Bwt}$$

.

Exemplo- Anatomia de gatos domésticos

- Modelo ajustado por mínimos quadrados:

$$\widehat{\texttt{Hwt}} = -0.356 + 4.034\texttt{Bwt}$$

.

- Estima-se um aumento médio de 4.034 gramas no peso do coração para cada quilo corporal a mais.

Exemplo- Anatomia de gatos domésticos

- Modelo ajustado por mínimos quadrados:

$$\widehat{\texttt{Hwt}} = -0.356 + 4.034\texttt{Bwt}$$

.

- Estima-se um aumento médio de 4.034 gramas no peso do coração para cada quilo corporal a mais.

- Para 500 gramas a mais de peso corporal (meio quilo), estima-se um aumento médio de $\frac{1}{2} \times 4.034 = 2.017$ gramas no peso do coração.

Exemplo- Anatomia de gatos domésticos

- Modelo ajustado por mínimos quadrados:

$$\widehat{\texttt{Hwt}} = -0.356 + 4.034\texttt{Bwt}$$

.

- Estima-se um aumento médio de 4.034 gramas no peso do coração para cada quilo corporal a mais.

- Para 500 gramas a mais de peso corporal (meio quilo), estima-se um aumento médio de $\frac{1}{2} \times 4.034 = 2.017$ gramas no peso do coração.

- O intercepto do modelo não tem uma interpretação prática, uma vez que $\texttt{Bwt} = 0$ não faz parte do escopo dos dados.

Exemplo- Anatomia de gatos domésticos

- O primeiro gato da base tem $\texttt{Bwt} = 2$ e $\texttt{Hwt} = 7$. Para este gato, o peso do coração ajustado pelo modelo e o resíduo são dados por:

$$\widehat{\texttt{Hwt}}_1 = -0.356 + 4.034\texttt{Bwt}_1 = -0.356 + 4.034 \times 2 = 7.711\text{g}$$

$$r_1 = \texttt{Hwt}_1 - \widehat{\texttt{Hwt}}_1 = 7 - 7.711 = -0.711g$$

Exemplo- Anatomia de gatos domésticos

- O primeiro gato da base tem $\texttt{Bwt} = 2$ e $\texttt{Hwt} = 7$. Para este gato, o peso do coração ajustado pelo modelo e o resíduo são dados por:

$$\widehat{\texttt{Hwt}}_1 = -0.356 + 4.034\texttt{Bwt}_1 = -0.356 + 4.034 \times 2 = 7.711\text{g}$$

$$r_1 = \texttt{Hwt}_1 - \widehat{\texttt{Hwt}}_1 = 7 - 7.711 = -0.711g$$

- Já para o gato da linha 100, temos $\texttt{Bwt} = 3$ e $\texttt{Hwt} = 10$, produzindo:

$$\widehat{\texttt{Hwt}}_{100} = -0.356 + 4.034\texttt{Bwt}_{100} = -0.356 + 4.034 \times 3 = 11.745\text{g}$$

$$r_{100} = \texttt{Hwt}_{100} - \widehat{\texttt{Hwt}}_{100} = 10 - 11.745 = -1.745g$$

Exemplo- Anatomia de gatos domésticos

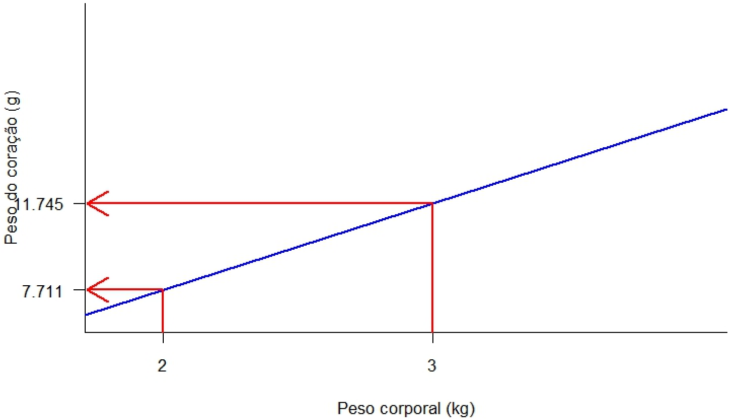


Figura 10: Predições usando o modelo de regressão linear ajustado

Propriedades dos estimadores de mínimos quadrados

- Os estimadores de mínimos quadrados são combinações lineares dos $y's$;

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:

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 \quad \mathrm{E}(\hat{\beta}_1) = \beta_1.$$

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:

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 \quad \mathrm{E}(\hat{\beta}_1) = \beta_1.$$

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

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

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

Propriedades dos estimadores de mínimos quadrados

Teorema de Gauss Markov

Satisfeitas as suposições assumidas para o modelo, os estimadores de mínimos quadrados têm menor variância que quaisquer outros estimadores não viciados que sejam combinações lineares dos $y's$.

Propriedades dos estimadores de mínimos quadrados

Teorema de Gauss Markov

Satisfeitas as suposições assumidas para o modelo, os estimadores de mínimos quadrados têm menor variância que quaisquer outros estimadores não viciados que sejam combinações lineares dos $y's$.

- Verificar a derivação das propriedades dos estimadores de mínimos quadrados dos parâmetros do modelo de regressão linear simples no material complementar.

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.

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:

$$\mathrm{SQ}_{\mathrm{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

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:

$$\mathrm{SQ}_{\mathrm{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Como o valor esperado de $\mathrm{SQ}_{\mathrm{Res}}$ é $(n-2)\sigma^2$, um estimador não viciado de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{\mathrm{SQ}_{\mathrm{Res}}}{n-2} = \mathrm{QM}_{\mathrm{Res}}.$$

Exemplo- Anatomia de gatos domésticos

- A estimativa de σ^2, para o problema dos gatos, é dada por:

$$\hat{\sigma}^2 = \frac{\text{SQ}_{\text{Res}}}{n-2} =$$

$$\frac{(\texttt{Hwt}_1 - \widehat{\texttt{Hwt}}_1)^2 + (\texttt{Hwt}_2 - \widehat{\texttt{Hwt}}_2)^2 + ... + (\texttt{Hwt}_{144} - \widehat{\texttt{Hwt}}_{144})^2}{n-2} =$$

$$\frac{(-0.711)^2 + (-0.311)^2 + ... + (5.124)^2}{144-2} = 2.109$$

Exemplo- Anatomia de gatos domésticos

- Vamos estimar as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$:

Exemplo- Anatomia de gatos domésticos

- Vamos estimar as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$\widehat{\text{Var}}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = 2.109 \left(\frac{1}{144} + \frac{7.418}{33.679} \right) = 0.479$$

Exemplo- Anatomia de gatos domésticos

- Vamos estimar as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$\widehat{\text{Var}}(\hat{\beta}_0) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = 2.109\left(\frac{1}{144} + \frac{7.418}{33.679}\right) = 0.479$$

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{2.109}{33.679} = 0.063$$

Exemplo- Anatomia de gatos domésticos

- Os erros padrões de $\hat{\beta}_0$ e $\hat{\beta}_1$ são calculados da seguinte forma:

Exemplo- Anatomia de gatos domésticos

- Os erros padrões de $\hat{\beta}_0$ e $\hat{\beta}_1$ são calculados da seguinte forma:

$$\text{EP}(\hat{\beta}_0) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_0)} = \sqrt{0.479} = 0.692$$

Exemplo- Anatomia de gatos domésticos

- Os erros padrões de $\hat{\beta}_0$ e $\hat{\beta}_1$ são calculados da seguinte forma:

$$\text{EP}(\hat{\beta}_0) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_0)} = \sqrt{0.479} = 0.692$$

$$\text{EP}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \sqrt{0.063} = 0.250$$

Exemplo- Anatomia de gatos domésticos

- Os erros padrões de $\hat{\beta}_0$ e $\hat{\beta}_1$ são calculados da seguinte forma:

$$\text{EP}(\hat{\beta}_0) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_0)} = \sqrt{0.479} = 0.692$$

$$\text{EP}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \sqrt{0.063} = 0.250$$

- Os erros padrões serão importantes posteriormente na construção de intervalos de confiança e testes de hipóteses para os parâmetros de regressão.

Exemplo computacional- parte 1

- Nesta aplicação, vamos usar simulação para ilustrar a distribuição amostral dos estimadores de mínimos quadrados.

Exemplo computacional- parte 2

- Nesta aplicação, vamos usar simulação para comparar diferentes as eficiências dos estimadores de mínimos quadrados sob diferentes delineamentos quanto aos valores fixados para a variável explicativa (x).

Exercícios

- Resolva os exercícios 1 a 10 da lista de exercícios relativa a este módulo, disponível na página da disciplina.

Testes de hipóteses e intervalos de confiança

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 $\epsilon \sim \text{Normal}(0, \sigma^2)$).

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 $\epsilon \sim \text{Normal}(0, \sigma^2)$).

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

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

- O seguinte teorma será importante para determinar a distribuição amostral de $\hat{\beta}_0$ e $\hat{\beta}_1$.

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

- O seguinte teorma será importante para determinar a distribuição amostral de $\hat{\beta}_0$ e $\hat{\beta}_1$.

- Sejam $y_1, y_2, ..., y_n$ variáveis aleatórias independentes tais que:

$$y_i \sim \mathrm{N}(\mu_i, \sigma_i^2),$$

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

- O seguinte teorma será importante para determinar a distribuição amostral de $\hat{\beta}_0$ e $\hat{\beta}_1$.

- Sejam $y_1, y_2, ..., y_n$ variáveis aleatórias independentes tais que:

$$y_i \sim \mathrm{N}(\mu_i, \sigma_i^2),$$

- Considere $c_1, c_2, ..., c_n$ um conjunto de constantes e a seguinte combinação linear dos y's:

$$z = c_1 y_1 + c_2 y_2 + ... + c_n y_n$$

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

- Segue que z também tem distribuição normal, conforme descrito na sequência:

$$z \sim \mathrm{N}\left(\mu_z = \sum_{i=1}^{n} c_i \mu_i, \;\; \sigma_z^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

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

- Segue que z também tem distribuição normal, conforme descrito na sequência:

$$z \sim \mathrm{N}\left(\mu_z = \sum_{i=1}^{n} c_i \mu_i, \;\; \sigma_z^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

- Como caso particular, se $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$, segue que:

$$z \sim \mathrm{N}\left(\mu_z = \sum_{i=1}^{n} c_i \mu_i, \;\; \sigma_z^2 = \sigma^2 \sum_{i=1}^{n} c_i^2\right)$$

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

- Segue que z também tem distribuição normal, conforme descrito na sequência:

$$z \sim \mathrm{N}\left(\mu_z = \sum_{i=1}^{n} c_i \mu_i, \;\; \sigma_z^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

- Como caso particular, se $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$, segue que:

$$z \sim \mathrm{N}\left(\mu_z = \sum_{i=1}^{n} c_i \mu_i, \;\; \sigma_z^2 = \sigma^2 \sum_{i=1}^{n} c_i^2\right)$$

- Finalmente, se $\mu_1 = \mu_2 = ... = \mu_n = \mu$,

$$z \sim \mathrm{N}\left(\mu_z = \mu \sum_{i=1}^{n} c_i, \;\; \sigma_z^2 = \sigma^2 \sum_{i=1}^{n} c_i^2\right)$$

Exercícios

- Resolva os exercícios 14 a 15 da lista de exercícios relativa a este módulo, disponível na página da disciplina.

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:

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

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:

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

- De maneira semelhante:

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

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

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

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

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

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

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

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 β_{10}:

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

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

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

Testes de hipóteses e intervalos de confiança para β_1

- Sob as especificações do modelo, é dada por:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2},$$

em que $\hat{\sigma}^2 = \text{QM}_{\text{Res}}$ e χ^2_{n-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 (2), temos:

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}, \tag{3}$$

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

Testes de hipóteses e intervalos de confiança para β_1

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

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}, \tag{3}$$

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

- Com base em (3) pode-se conduzir o teste da hipótese $H_0 : \beta_1 = \beta_{10}$.

Testes de hipóteses e intervalos de confiança para β_1

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

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}, \tag{3}$$

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

- Com base em (3) 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

- **Nota:** Ao aplicar a função `summary` a um objeto da classe `lm` no `R`, os resultados dos testes se referem aos seguintes pares de hipóteses:

- **Nota:** Ao aplicar a função `summary` a um objeto da classe `lm` no `R`, os resultados dos testes se referem aos seguintes pares de hipóteses:

$$H_0 : \beta_0 = 0 \;\; vs \;\; H_1 : \beta_0 \neq 0$$

- **Nota:** Ao aplicar a função `summary` a um objeto da classe `lm` no `R`, os resultados dos testes se referem aos seguintes pares de hipóteses:

$$H_0 : \beta_0 = 0 \ \ vs \ \ H_1 : \beta_0 \neq 0$$

$$H_0 : \beta_1 = 0 \ \ vs \ \ H_1 : \beta_1 \neq 0$$

Testes de hipóteses e intervalos de confiança para β_1

- **Nota:** Ao aplicar a função `summary` a um objeto da classe `lm` no `R`, os resultados dos testes se referem aos seguintes pares de hipóteses:

$$H_0 : \beta_0 = 0 \;\; vs \;\; H_1 : \beta_0 \neq 0$$

$$H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$$

- Vale reforçar que testar a nulidade de β_1 é de particular importância, pois $\beta_1 = 0$ implica que não há relação entre a variável explicativa e a resposta (assumindo que as suposições do modelo sejam atendidas).

Testes de hipóteses e intervalos de confiança para β_1

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

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

Teste da significância da regressão

- Uma importante hipótese a ser testada é $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$.

Teste da significância da regressão

- Uma importante hipótese a ser testada é $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$.

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

Teste da significância da regressão

- Uma importante hipótese a ser testada é $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$.
- Chamamos esse teste de **teste da significância da regressão linear simples**.
- Neste caso, a estatística do teste fica dada por:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}, \tag{4}$$

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

Exemplo- Anatomia de gatos domésticos

Um pesquisador postulou que um aumento de um quilograma no peso corporal dos gatos está associado a um aumento médio de 5 gramas no peso dos corações. Teste esta firmativa ao nível de 5% de significância.

Exemplo- Anatomia de gatos domésticos

- Hipóteses:

$$H_0 : \beta_1 = 5 \quad vs \quad H_1 : \beta_1 \neq 5$$

Exemplo- Anatomia de gatos domésticos

- Hipóteses:

$$H_0 : \beta_1 = 5 \quad vs \quad H_1 : \beta_1 \neq 5$$

- Estatística do teste:

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{4.034 - 5}{\sqrt{\frac{2.109}{33.679}}} = -3.860$$

Exemplo- Anatomia de gatos domésticos

- Regra de decisão: Devemo rejeitar H_0, ao nível de significância $\alpha = 5\%$, se:

$$|t| > |t_{144-2;0.05/2}| = 1.976$$

Exemplo- Anatomia de gatos domésticos

- Regra de decisão: Devemo rejeitar H_0, ao nível de significância $\alpha = 5\%$, se:

$$|t| > |t_{144-2;0.05/2}| = 1.976$$

- Decisão: Como $|t| = 3.860 > 1.976$, então temos evidências para rejeitar H_0 ao nível de significância de 5%.

Exemplo- Anatomia de gatos domésticos

- Regra de decisão: Devemo rejeitar H_0, ao nível de significância $\alpha = 5\%$, se:

$$|t| > |t_{144-2;0.05/2}| = 1.976$$

- Decisão: Como $|t| = 3.860 > 1.976$, então temos evidências para rejeitar H_0 ao nível de significância de 5%.

- Conclusão: A postulação do pesquisador pode ser rejeitada, de maneira que descartamos a hipótese de um aumento médio de 5 gramas no peso dos corações para cada quilograma corporal adicional.

Exemplo- Anatomia de gatos domésticos

- A conclusão anterior se manteria para um nível de significância $\alpha = 1\%$?

Exemplo- Anatomia de gatos domésticos

- A conclusão anterior se manteria para um nível de significância $\alpha = 1\%$?

- Neste caso, H_0 deveria ser rejeitada se:

$$|t| > |t_{144-2;0.01/2}| = 2.611$$

Exemplo- Anatomia de gatos domésticos

- A conclusão anterior se manteria para um nível de significância $\alpha = 1\%$?

- Neste caso, H_0 deveria ser rejeitada se:

$$|t| > |t_{144-2;0.01/2}| = 2.611$$

- Decisão: Como $|t| = 3.860 > 2.611$, então ainda temos evidências para rejeitar H_0 ao nível de significância de 1%, e a conclusão se mantém.

Exemplo- Anatomia de gatos domésticos

- Vamos calcular o p-valor do teste. Seja T uma variável aleatória com distribuição $t - Student$ com $df = 142$ graus de liberdade. O p-valor é dado por:

Exemplo- Anatomia de gatos domésticos

- Vamos calcular o p-valor do teste. Seja T uma variável aleatória com distribuição $t - Student$ com $df = 142$ graus de liberdade. O p-valor é dado por:

$$p = P(|T| > 3.860) = P(T < -3.860) + P(T > 3.860) = 2 \times 0.0001 = 0.0002$$

Exemplo- Anatomia de gatos domésticos

- Finalmente, vamos calcular um intervalo de 95% de confiança para β_1:

Exemplo- Anatomia de gatos domésticos

- Finalmente, vamos calcular um intervalo de 95% de confiança para β_1:

$$IC(\beta_1; 95\%) = \hat{\beta}_1 \pm t_{n-2;0.025}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

Exemplo- Anatomia de gatos domésticos

- Finalmente, vamos calcular um intervalo de 95% de confiança para β_1:

$$IC(\beta_1; 95\%) = \hat{\beta}_1 \pm t_{n-2;0.025}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$4.034 \pm 1.976\sqrt{\frac{2.109}{33.679}} = 4.034 \pm 0.494 =$$

Exemplo- Anatomia de gatos domésticos

- Finalmente, vamos calcular um intervalo de 95% de confiança para β_1:

$$IC(\beta_1; 95\%) = \hat{\beta}_1 \pm t_{n-2;0.025}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$4.034 \pm 1.976\sqrt{\frac{2.109}{33.679}} = 4.034 \pm 0.494 =$$

$$(3.539; 4.528)$$

Exemplo- Anatomia de gatos domésticos

- Finalmente, vamos calcular um intervalo de 95% de confiança para β_1:

$$IC(\beta_1; 95\%) = \hat{\beta}_1 \pm t_{n-2;0.025}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$4.034 \pm 1.976\sqrt{\frac{2.109}{33.679}} = 4.034 \pm 0.494 =$$

$$(3.539; 4.528)$$

- Desta forma, podemos afirmar com 95% de confiança que o intervalo (3.539;4.528) contém o valor desconhecido de β_1.

Teste da significância da regressão

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

Teste da significância da regressão

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

- Dessa forma, ainda que H_0 seja rejeitada, pode-se ter alguma relação não linear 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.

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.

- Se as suposições do modelo forem atendidas, então:

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

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}.

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 fica dado por:

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

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, sendo definido pelos seguintes limites:

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, sendo definido pelos seguintes limites:

$$IC(\sigma^2; 1-\alpha) = \left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;1-\alpha/2}} \;\; ; \;\; \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;\alpha/2}} \right),$$

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, sendo definido pelos seguintes limites:

$$IC(\sigma^2; 1-\alpha) = \left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;1-\alpha/2}} \;\; ; \;\; \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;\alpha/2}} \right),$$

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.

Exemplo de simulação

- Neste exemplo de simulação, vamos usar simulação para ilustrar os conceitos de intervalos de confiança e testes de hipóteses, aplicados na regressão linear simples.

Estimação da resposta média e predição de novas observações

Intervalo de confiança para a resposta média

- Suponha que se deseja estimar a média de y para um particular valor $x = x_0$.

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:

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

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:

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

- 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 $\hat{\beta}_1$).

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:

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

- 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:

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

Intervalo de confiança para a resposta média

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

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

Intervalo de confiança para a resposta média

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

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

- Substituindo σ^2 por $\hat{\sigma}^2 = \text{QM}_{\text{Res}}$, temos:

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

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:

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

Predição de uma nova observação

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

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

Predição de uma nova observação

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

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

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

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

Predição de uma nova observação

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

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

Predição de uma nova observação

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

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

- 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 da resposta média e predição de novas observações

Material complementar

Verificar a derivação das propriedades do estimador da resposta média e da predição de novas observações para o modelo de regressão linear simples.

Exemplo- Anatomia de gatos domésticos

- Determine um intervalo de confiança de 95% para o peso médio do coração da população de gatos com `Bwt` $= 2.5$kg.

Exemplo- Anatomia de gatos domésticos

- Determine um intervalo de confiança de 95% para o peso médio do coração da população de gatos com `Bwt` $= 2.5$kg.

- Estimativa pontual:

$$\hat{\mu}_{\texttt{Hwt}|\texttt{Bwt=2.5}} = -0.356 + 4.034 \times 2.5 = 9.729$$

Exemplo- Anatomia de gatos domésticos

- Intervalo de confiança:

Exemplo- Anatomia de gatos domésticos

- Intervalo de confiança:

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

Exemplo- Anatomia de gatos domésticos

- Intervalo de confiança:

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

$$9.729 \pm 1.976 \times \sqrt{2.109\left(\frac{1}{144} + \frac{(2.5 - 2.723)^2}{33.679}\right)} =$$

Exemplo- Anatomia de gatos domésticos

- Intervalo de confiança:

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

$$9.729 \pm 1.976 \times \sqrt{2.109\left(\frac{1}{144} + \frac{(2.5-2.723)^2}{33.679}\right)} =$$

$$9.729 \pm 0.263 = (9.466; 9.992)$$

Exemplo- Anatomia de gatos domésticos

- Intervalo de confiança:

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

$$9.729 \pm 1.976 \times \sqrt{2.109\left(\frac{1}{144} + \frac{(2.5-2.723)^2}{33.679}\right)} =$$

$$9.729 \pm 0.263 = (9.466; 9.992)$$

- Estima-se, com 95% de confiança, que o peso médio do coração para a população de gatos com peso corporal de 2.5kg esteja no intervalo $(9.466; 9.992)$.

Exemplo- Anatomia de gatos domésticos

- O gato Bob pesa 2.5kg. Baseado no modelo de regressão ajustado, apresente um intervalo de predição de 95% de confiança para o peso do coração do gato Bob.

Exemplo- Anatomia de gatos domésticos

- O gato Bob pesa 2.5kg. Baseado no modelo de regressão ajustado, apresente um intervalo de predição de 95% de confiança para o peso do coração do gato Bob.

- Estimativa pontual:

$$\widehat{\texttt{Hwt}}_{\texttt{Bwt=2.5}} = -0.356 + 4.034 \times 2.5 = 9.729$$

Exemplo- Anatomia de gatos domésticos

- Intervalo de predição:

Exemplo- Anatomia de gatos domésticos

- Intervalo de predição:

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

Exemplo- Anatomia de gatos domésticos

- Intervalo de predição:

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

$$9.729 \pm 1.976 \times \sqrt{2.109\left(1+\frac{1}{144}+\frac{(2.5-2.723)^2}{33.679}\right)} =$$

Exemplo- Anatomia de gatos domésticos

- Intervalo de predição:

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

$$9.729 \pm 1.976 \times \sqrt{2.109\left(1+\frac{1}{144}+\frac{(2.5-2.723)^2}{33.679}\right)} =$$

$$9.729 \pm 2.882 = (6.847; 12.611)$$

Exemplo- Anatomia de gatos domésticos

- Intervalo de predição:

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

$$9.729 \pm 1.976 \times \sqrt{2.109\left(1+\frac{1}{144}+\frac{(2.5-2.723)^2}{33.679}\right)} =$$

$$9.729 \pm 2.882 = (6.847; 12.611)$$

- Podemos afirmar, com 95% de confiança, que o peso do coração do gato Bob está no intervalo $(6.847; 12.611)$.

Exemplo- Anatomia de gatos domésticos

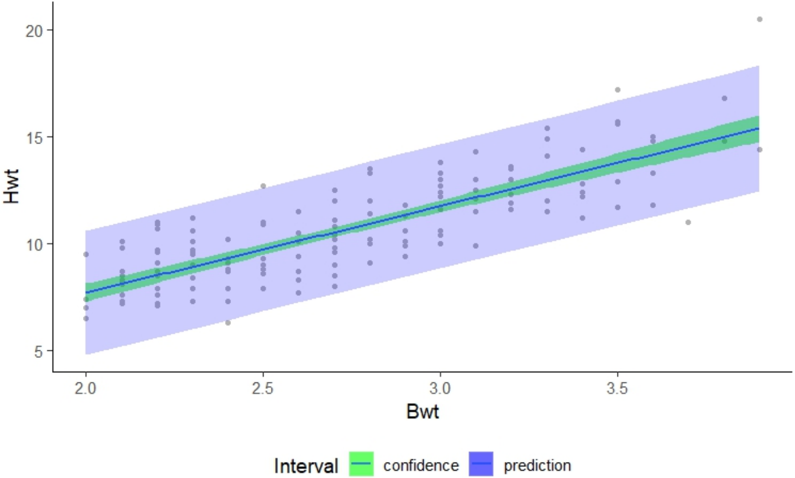


Figura 11: Bandas de confiança para a resposta média e de predição

Análise de variância para a regressão linear simples

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.

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:

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

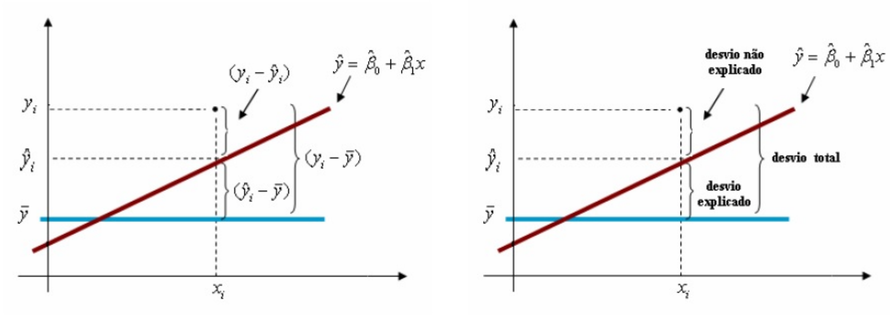


Figura 12: Partição da variação dos dados na regressão linear simples.

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:

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

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:

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

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 $\mathrm{SQ}_{\mathrm{Reg}}$ em detrimento a $\mathrm{SQ}_{\mathrm{Res}}$, maior a parcela da variação total dos dados explicada pela regressão.

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 partição 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}.

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 partição 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	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	$\text{QM}_{\text{Reg}} = \frac{\text{SQ}_{\text{Reg}}}{1}$	$F = \frac{\text{QM}_{\text{Reg}}}{\text{QM}_{\text{Res}}}$
Resíduos	n-2	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\text{QM}_{\text{Res}} = \frac{\text{SQ}_{\text{Res}}}{n-2}$	
Total	n-1	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Tabela 2: Quadro de análise de variância

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F
Regressão	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	$\text{QM}_{\text{Reg}} = \frac{\text{SQ}_{\text{Reg}}}{1}$	$F = \frac{\text{QM}_{\text{Reg}}}{\text{QM}_{\text{Res}}}$
Resíduos	n-2	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\text{QM}_{\text{Res}} = \frac{\text{SQ}_{\text{Res}}}{n-2}$	
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:
  - $\frac{(n-2)\mathrm{QM}_{\mathrm{Res}}}{\sigma^2} \sim \chi^2_{n-2}$;
  - Sob a hipótese nula (isso é, se $\beta_1 = 0$), então $\frac{\mathrm{SQ}_{\mathrm{Reg}}}{\sigma^2}$ tem distribuição χ^2_1;
  - $\mathrm{SQ}_{\mathrm{Reg}}$ e $\mathrm{SQ}_{\mathrm{Res}}$ são independentes.

Análise de variância aplicada à regressão linear simples

- O teste da significância do modelo via ANOVA baseia-se em:
  - $\frac{(n-2)\mathrm{QM}_{\mathrm{Res}}}{\sigma^2} \sim \chi^2_{n-2}$;
  - Sob a hipótese nula (isso é, se $\beta_1 = 0$), então $\frac{\mathrm{SQ}_{\mathrm{Reg}}}{\sigma^2}$ tem distribuição χ^2_1;
  - $\mathrm{SQ}_{\mathrm{Reg}}$ e $\mathrm{SQ}_{\mathrm{Res}}$ são independentes.

- Então:

$$F = \frac{\mathrm{SQ}_{\mathrm{Reg}}/1}{\mathrm{SQ}_{\mathrm{Res}}/(n-2)} = \frac{\mathrm{QM}_{\mathrm{Reg}}}{\mathrm{QM}_{\mathrm{Res}}}$$

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

Análise de variância aplicada à regressão linear simples

- O teste da significância do modelo via ANOVA baseia-se em:
  - $\frac{(n-2)\mathrm{QM}_{\mathrm{Res}}}{\sigma^2} \sim \chi^2_{n-2}$;
  - Sob a hipótese nula (isso é, se $\beta_1 = 0$), então $\frac{\mathrm{SQ}_{\mathrm{Reg}}}{\sigma^2}$ tem distribuição χ^2_1;
  - $\mathrm{SQ}_{\mathrm{Reg}}$ e $\mathrm{SQ}_{\mathrm{Res}}$ são independentes.

- Então:

$$F = \frac{\mathrm{SQ}_{\mathrm{Reg}}/1}{\mathrm{SQ}_{\mathrm{Res}}/(n-2)} = \frac{\mathrm{QM}_{\mathrm{Reg}}}{\mathrm{QM}_{\mathrm{Res}}}$$

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:

$$R^2 = \frac{\text{SQ}_{\text{Reg}}}{\text{SQ}_{\text{Total}}},$$

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

Análise de variância aplicada à regressão linear simples

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

$$R^2 = \frac{\text{SQ}_{\text{Reg}}}{\text{SQ}_{\text{Total}}},$$

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

- Dessa forma, R^2 corresponde à proporção da variação dos dados explicada pela regressão.

Análise de variância aplicada à regressão linear simples

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

$$R^2 = \frac{\text{SQ}_{\text{Reg}}}{\text{SQ}_{\text{Total}}},$$

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

- 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 entre x e y.

Análise de variância aplicada à regressão linear simples

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

$$R^2 = \frac{\mathrm{SQ}_{\mathrm{Reg}}}{\mathrm{SQ}_{\mathrm{Total}}},$$

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

- 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 entre x e y.

- 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

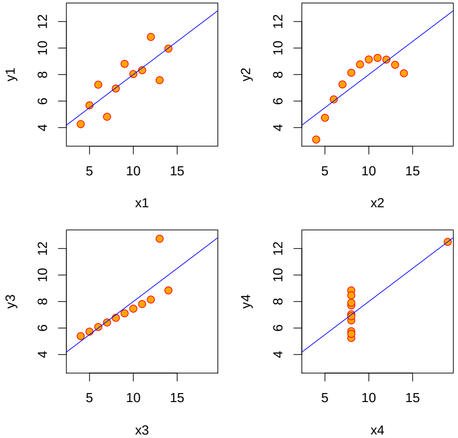


Figura 13: Quatro conjuntos de dados que produzem mesmo coeficiente de determinação ($R^2 = 0.67$).

Exemplo- Anatomia de gatos domésticos

$$\mathrm{SQ}_{\mathrm{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$$

Exemplo- Anatomia de gatos domésticos

$$\text{SQ}_{\text{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$$

$$(7.711 - 10.630)^2 + (7.711 - 10.630)^2 + ... + (15.376 - 10.630)^2 = 548.092$$

Exemplo- Anatomia de gatos domésticos

$$\text{SQ}_{\text{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$$

$$(7.711 - 10.630)^2 + (7.711 - 10.630)^2 + ... + (15.376 - 10.630)^2 = 548.092$$

$$\text{SQ}_{\text{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

Exemplo- Anatomia de gatos domésticos

$$\text{SQ}_{\text{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$$

$$(7.711 - 10.630)^2 + (7.711 - 10.630)^2 + ... + (15.376 - 10.630)^2 = 548.092$$

$$\text{SQ}_{\text{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

$$(7 - 7.711)^2 + (7.4 - 7.711)^2 + ... + (20.5 - 15.376)^2 = 299.533$$

Exemplo- Anatomia de gatos domésticos

$$\text{SQ}_{\text{Total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 =$$

Exemplo- Anatomia de gatos domésticos

$$\text{SQ}_{\text{Total}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$(7 - 10.630)^2 + (7.4 - 10.630)^2 + ... + (20.5 - 10.630)^2 = 847.625$$

$$\mathrm{SQ}_{\mathrm{Total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 =$$

$$(7 - 10.630)^2 + (7.4 - 10.630)^2 + ... + (20.5 - 10.630)^2 = 847.625$$

$$\mathrm{QM}_{\mathrm{Res}} = \frac{\mathrm{SQ}_{\mathrm{Res}}}{n-2} = \frac{299.533}{142} = 2.109$$

Exemplo- Anatomia de gatos domésticos

$$\text{SQ}_{\text{Total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 =$$

$$(7 - 10.630)^2 + (7.4 - 10.630)^2 + ... + (20.5 - 10.630)^2 = 847.625$$

$$\text{QM}_{\text{Res}} = \frac{\text{SQ}_{\text{Res}}}{n - 2} = \frac{299.533}{142} = 2.109$$

$$\text{F} = \frac{\text{QM}_{\text{Reg}}}{\text{QM}_{\text{Res}}} = \frac{548.092}{2.109} = 259.881$$

Exemplo- Anatomia de gatos domésticos

Tabela 3: Quadro de análise de variância

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F
Regressão	1	548.092	548.092	299.881
Resíduos	142	299.533	2.109	
Total	143	847.625		

Exemplo- Anatomia de gatos domésticos

- Teste da significância do modelo: a hipótese nula de que o modelo não é significativo ($H_0 : \beta_1 = 0$) será rejeitada, ao nível de 5% de significância, se:

$$\mathrm{F} > \mathrm{F}_{1,142}(0.95) = 3.908$$

Exemplo- Anatomia de gatos domésticos

- Teste da significância do modelo: a hipótese nula de que o modelo não é significativo ($H_0 : \beta_1 = 0$) será rejeitada, ao nível de 5% de significância, se:

$$\mathrm{F} > \mathrm{F}_{1,142}(0.95) = 3.908$$

- Como F = 299.881 >>> 3.908, rejeitamos a hipótese nula e a significância do modelo está comprovada.

- Cálculo do coeficiente de determinação:

$$\mathrm{R}^2 = 100 \times \frac{\mathrm{SQ}_{\mathrm{Reg}}}{\mathrm{SQ}_{\mathrm{Total}}} = 100 \times \frac{548.092}{847.625} = 64.66$$

Exemplo- Anatomia de gatos domésticos

- Cálculo do coeficiente de determinação:

$$\mathrm{R}^2 = 100 \times \frac{\mathrm{SQ}_{\mathrm{Reg}}}{\mathrm{SQ}_{\mathrm{Total}}} = 100 \times \frac{548.092}{847.625} = 64.66$$

- Isso quer dizer que o modelo ajustado é capaz de explicar aproximadamente 65% da variação dos pesos dos corações dos gatos domésticos.

Exercícios

1. Resolva os exercícios 11 a 13 da lista de exercícios relativa a este módulo, disponível na página da disciplina.

Exercícios adicionais

Exercício- Old Faithful geyser

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `faithful` do `R`.

Exercício- Old Faithful geyser

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `faithful` do `R`.

- Os dados referem-se a 272 erupções do vulcão Old Faithful geyser, Yellowstone National Park, Wyoming, USA.

Exercício- Old Faithful geyser

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `faithful` do `R`.

- Os dados referem-se a 272 erupções do vulcão Old Faithful geyser, Yellowstone National Park, Wyoming, USA.

- As variáveis a serem analisadas são as seguintes:

Exercício- Old Faithful geyser

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `faithful` do `R`.

- Os dados referem-se a 272 erupções do vulcão Old Faithful geyser, Yellowstone National Park, Wyoming, USA.

- As variáveis a serem analisadas são as seguintes:
  - `eruptions`: duração da erupção, em minutos (resposta);

Exercício- Old Faithful geyser

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `faithful` do `R`.

- Os dados referem-se a 272 erupções do vulcão Old Faithful geyser, Yellowstone National Park, Wyoming, USA.

- As variáveis a serem analisadas são as seguintes:
  - `eruptions`: duração da erupção, em minutos (resposta);
  - `waiting`: tempo decorrido desde a erupção anterior, em minutos.

Exercício- Old Faithful geyser

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `faithful` do `R`.

- Os dados referem-se a 272 erupções do vulcão Old Faithful geyser, Yellowstone National Park, Wyoming, USA.

- As variáveis a serem analisadas são as seguintes:
  - `eruptions`: duração da erupção, em minutos (resposta);
  - `waiting`: tempo decorrido desde a erupção anterior, em minutos.

- Produza uma análise de regressão linear simples.

Exercício- Old Faithful geyser

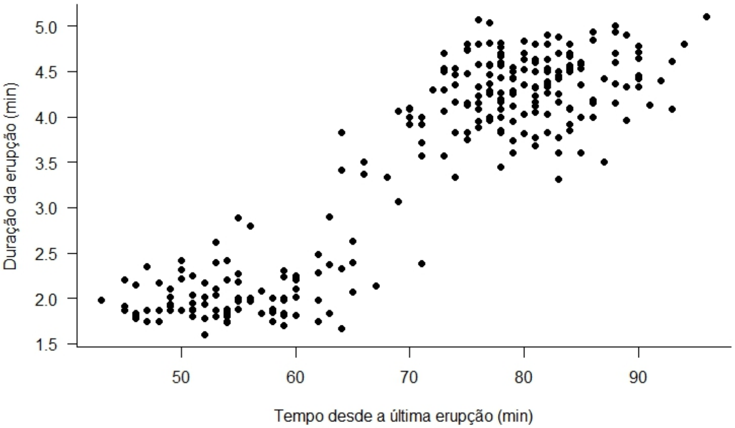


Figura 14: Dados sobre duração de erupções de um vulcão

Exercício- Demanda diária de energia de ovinos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `sheep` que pode ser acessada na biblioteca `GLMsData` do `R`.

Exercício- Demanda diária de energia de ovinos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `sheep` que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 64 ovelhas de certa espécie e rebanho.

Exercício- Demanda diária de energia de ovinos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `sheep` que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 64 ovelhas de certa espécie e rebanho.

- As variáveis a serem analisadas são as seguintes:

Exercício- Demanda diária de energia de ovinos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `sheep` que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 64 ovelhas de certa espécie e rebanho.

- As variáveis a serem analisadas são as seguintes:
  - `Energy`: demanda energética em Mcal/dia (resposta);

Exercício- Demanda diária de energia de ovinos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `sheep` que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 64 ovelhas de certa espécie e rebanho.

- As variáveis a serem analisadas são as seguintes:
  - `Energy`: demanda energética em Mcal/dia (resposta);
  - `Weight`: peso do animal em kg.

Exercício- Demanda diária de energia de ovinos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `sheep` que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 64 ovelhas de certa espécie e rebanho.

- As variáveis a serem analisadas são as seguintes:
  - `Energy`: demanda energética em Mcal/dia (resposta);
  - `Weight`: peso do animal em kg.

- Produza uma análise de regressão linear simples.

Exercício- Demanda diária de energia de ovinos

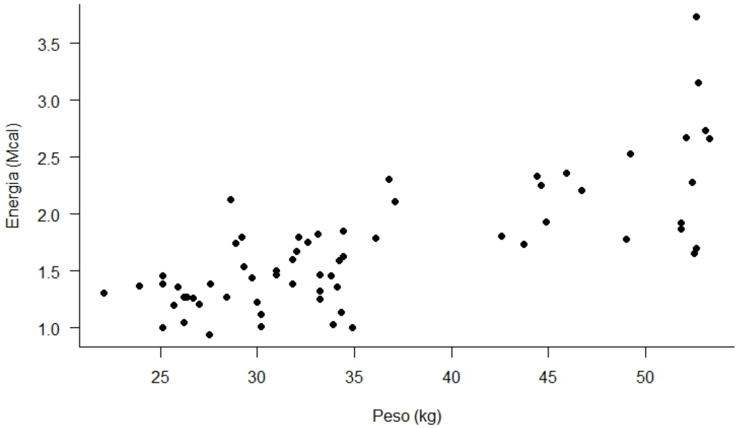


Figura 15: Dados sobre demanda de energia de ovinos

Exercício- Análise sensorial de queijos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `cheese`, que pode ser acessada na biblioteca `GLMsData` do `R`.

Exercício- Análise sensorial de queijos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `cheese`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 30 amostras de queijo submetidas a um experimento sensorial.

Exercício- Análise sensorial de queijos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `cheese`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 30 amostras de queijo submetidas a um experimento sensorial.

- As variáveis a serem analisadas são as seguintes:

Exercício- Análise sensorial de queijos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `cheese`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 30 amostras de queijo submetidas a um experimento sensorial.

- As variáveis a serem analisadas são as seguintes:
    - `Taste`: nota combinada de diversos juízes atribuida ao sabor de cada amostra de queijo (resposta). Maiores notas indicam melhor sabor;

Exercício- Análise sensorial de queijos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `cheese`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 30 amostras de queijo submetidas a um experimento sensorial.

- As variáveis a serem analisadas são as seguintes:
  - `Taste`: nota combinada de diversos juízes atribuida ao sabor de cada amostra de queijo (resposta). Maiores notas indicam melhor sabor;
  - `Lactic`: concentração de ácido lático.

Exercício- Análise sensorial de queijos

- Nesta aplicação vamos analisar os dados disponíveis na base de dados `cheese`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- Os dados referem-se a 30 amostras de queijo submetidas a um experimento sensorial.

- As variáveis a serem analisadas são as seguintes:
  - `Taste`: nota combinada de diversos juízes atribuida ao sabor de cada amostra de queijo (resposta). Maiores notas indicam melhor sabor;
  - `Lactic`: concentração de ácido lático.

- Produza uma análise de regressão linear simples.

Exercício- Análise sensorial de queijos

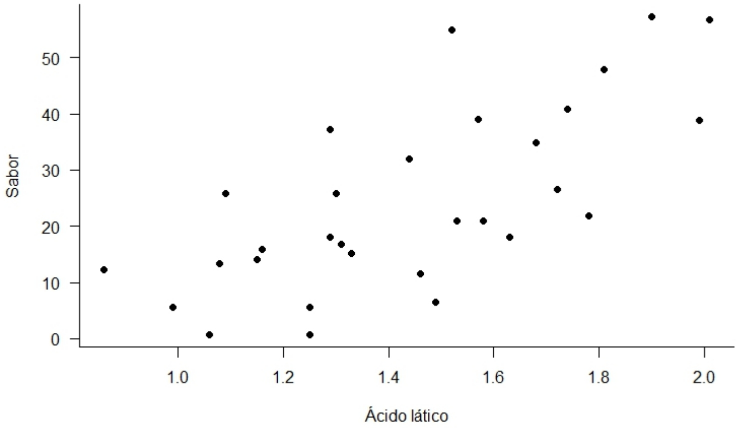


Figura 16: Dados de experimento sensorial de queijo

Análise da correlação linear

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.

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;

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 de sua 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:

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

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

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

é o coeficiente de correlação entre x e y.

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

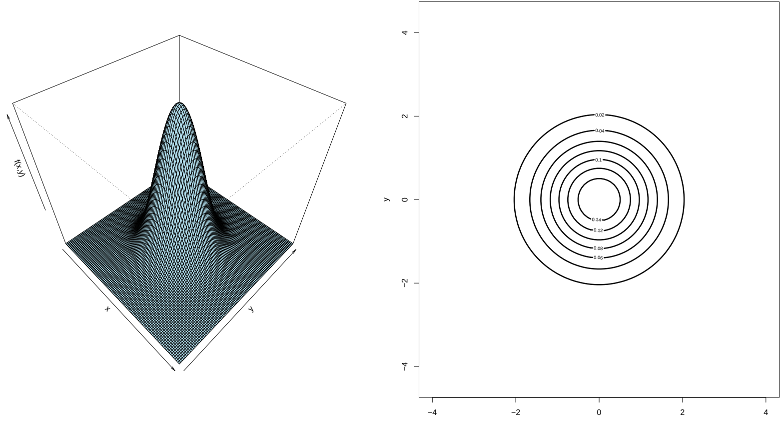


Figura 17: Distribuição normal bivariada: $\rho = 0$.

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

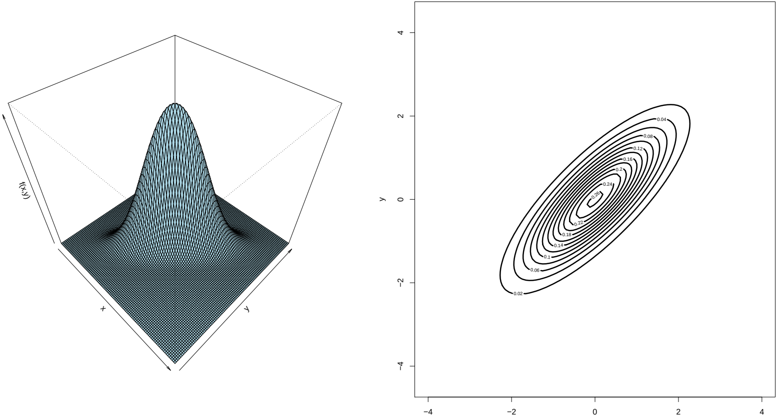


Figura 18: Distribuição normal bivariada: $\rho = 0.8$.

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

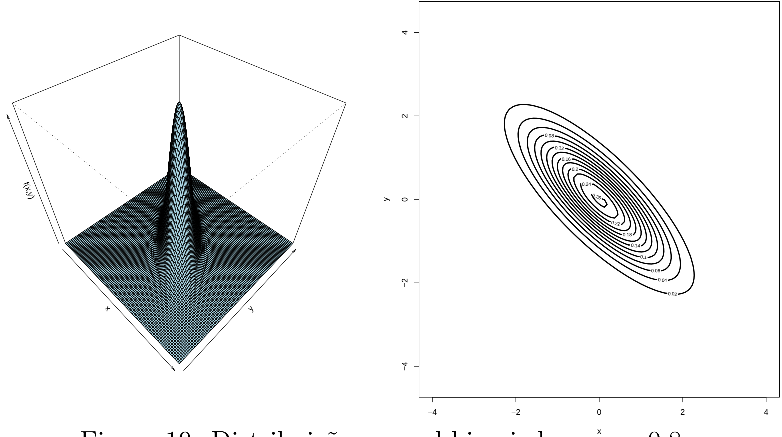


Figura 19: Distribuição normal bivariada: $\rho = -0.8$.

Análise de correlação

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

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

Análise de correlação

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

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

- Verifica-se facilmente que:

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

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.

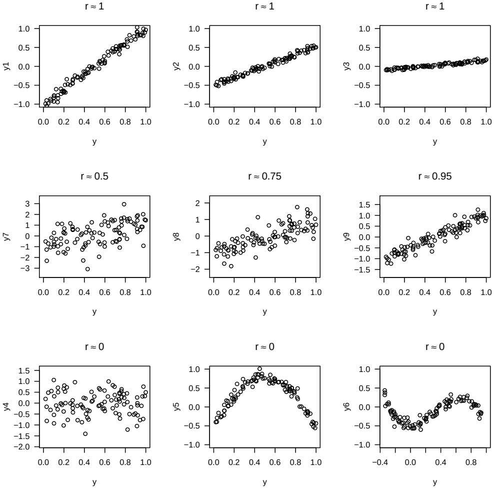
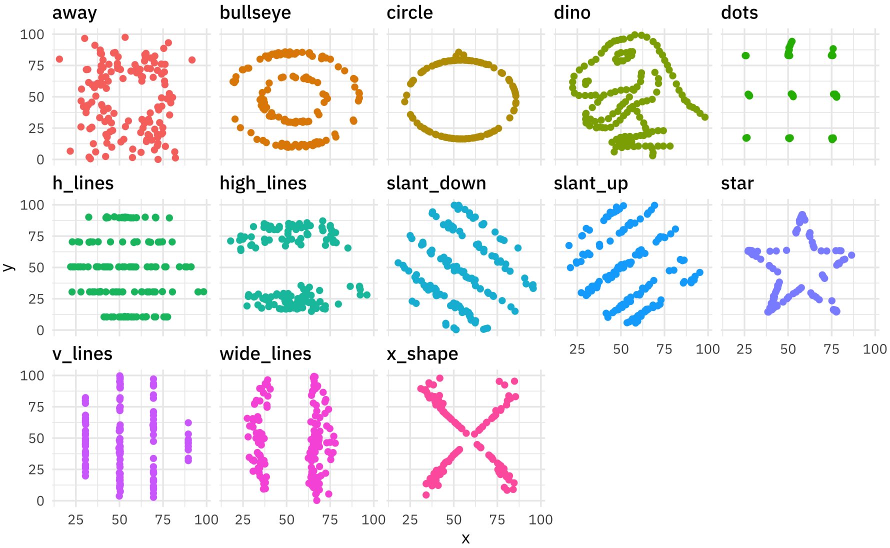


Figura 20: Ilustração de dados com diferentes níveis de correlação linear.


away
bullseye
circle
dino
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines
x_shape
y
x
0
25
50
75
100

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

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

Análise de correlação

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

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

- A estatística do teste é dada por:

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

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 α, se $|t| > |t_{n-2;\alpha/2}|$.

Análise de correlação

- Assim, a hipótese de correlação nula deverá ser rejeitada, ao nível de significância 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:

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

em que:

$$\operatorname{arctan} u = \frac{1}{2}\ln\frac{1+u}{1-u}; \quad \tanh u = \frac{e^u - e^{-u}}{e^u + e^{-u}}.$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Vamos analisar a correlação entre as variáveis:

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Vamos analisar a correlação entre as variáveis:
  - X : Taxa de analfabetismo;

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Vamos analisar a correlação entre as variáveis:
  - X : Taxa de analfabetismo;
  - Y : Probabilidade de sobrevivência aos 60 anos.

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

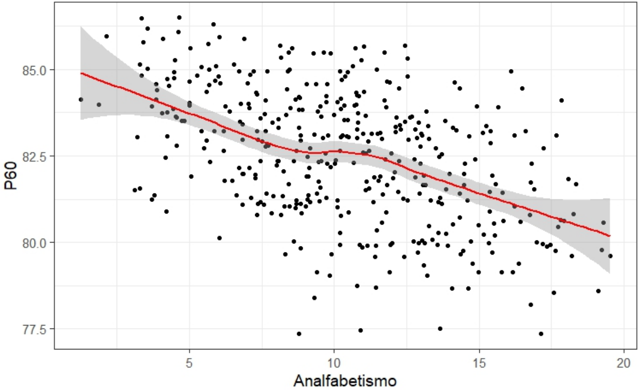


Figura 22: Probabilidade de sobrevivência aos 60 anos e taxa de analfabetismo para os municípios do Estado do Paraná

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Estimativa do coeficiente de correlação linear:

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Estimativa do coeficiente de correlação linear:

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

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Estimativa do coeficiente de correlação linear:

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

$$\bar{x} = 10.36; \quad \bar{y} = 82.49$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = (16.76-10.36)(80.80-82.49)+(16.82-10.36)(80.80-82.49)+...=-1359.63$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = (16.76-10.36)(80.80-82.49)+(16.82-10.36)(80.80-82.49)+...=-1359.63$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = (16.76-10.36)^2+(16.82-10.36)^2+...+(12.86-10.36)^2 = 5860.68$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = (16.76-10.36)(80.80-82.49)+(16.82-10.36)(80.80-82.49)+...=-1359.63$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = (16.76-10.36)^2+(16.82-10.36)^2+...+(12.86-10.36)^2=5860.68$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = (80.80-82.49)^2+(80.80-82.49)^2+...+(83.01-82.49)^2=1433.69$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

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

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

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

$$\frac{-1359.63}{\sqrt{5860.68 \times 1433.69}} = -0.47$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

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

$$\frac{-1359.63}{\sqrt{5860.68 \times 1433.69}} = -0.47$$

- Desta forma, taxa de analfabetismo e a probabilidade de sobrevivência estão negativamente correlacionadas, apresentando relação inversa.

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Vamos calcular o intervalo de confiança 95% para o coeficiente de correlação:

$$\text{arctan}(-0.47) = \frac{1}{2}\ln\left(\frac{1+(-0.47)}{1-(-0.47)}\right) = -0.51$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Vamos calcular o intervalo de confiança 95% para o coeficiente de correlação:

$$\text{arctan}(-0.47) = \frac{1}{2}\ln\left(\frac{1+(-0.47)}{1-(-0.47)}\right) = -0.51$$

$$\text{IC}(\rho, 95\%) = \tanh\left(-0.51 - \frac{1.96}{\sqrt{399-3}}; -0.51 + \frac{1.96}{\sqrt{399-3}}\right) =$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Vamos calcular o intervalo de confiança 95% para o coeficiente de correlação:

$$\text{arctan}(-0.47) = \frac{1}{2}\ln\left(\frac{1+(-0.47)}{1-(-0.47)}\right) = -0.51$$

$$\text{IC}(\rho, 95\%) = \tanh\left(-0.51 - \frac{1.96}{\sqrt{399-3}}; -0.51 + \frac{1.96}{\sqrt{399-3}}\right) =$$

$$(-0.543; -0.389)$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Finalmente, vamos testar a significância da correlação linear ao nível de significância de 5%.

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Finalmente, vamos testar a significância da correlação linear ao nível de significância de 5%.

- Hipóteses: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Finalmente, vamos testar a significância da correlação linear ao nível de significância de 5%.

- Hipóteses: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

- Estatistica teste:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0.47\sqrt{\frac{399-2}{1-(-0.47)^2}} = -10.60$$

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Regra de decisão: Devemos rejeitar H_0 ao nível de significância de 5% se $|t| > |t_{399-2}(0.025)| = 1.97$.

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Regra de decisão: Devemos rejeitar H_0 ao nível de significância de 5% se $|t| > |t_{399-2}(0.025)| = 1.97$.

- Decisão: Como $|t| = 10.60 > 1.97$, rejeitamos H_0 ao nível de significância de 5%.

Exemplo- Indicadores sócio-econômicos dos municípios do Estado do Paraná

- Regra de decisão: Devemos rejeitar H_0 ao nível de significância de 5% se $|t| > |t_{399-2}(0.025)| = 1.97$.

- Decisão: Como $|t| = 10.60 > 1.97$, rejeitamos H_0 ao nível de significância de 5%.

- Conclusão: A taxa de analfabetismo e a probabilidade de sobrevivência estão (negativamente) correlacionadas.

Exercícios adicionais

Exercício- Dados sócio econômicos dos municípios do Paraná

- Nesta aplicação vamos analisar dados sócio econômicos dos municípios do Estado do Paraná,disponíveis na página da disciplina.

Exercício- Dados sócio econômicos dos municípios do Paraná

- Nesta aplicação vamos analisar dados sócio econômicos dos municípios do Estado do Paraná,disponíveis na página da disciplina.

- As variáveis a serem analisadas são as seguintes:

Exercício- Dados sócio econômicos dos municípios do Paraná

- Nesta aplicação vamos analisar dados sócio econômicos dos municípios do Estado do Paraná,disponíveis na página da disciplina.

- As variáveis a serem analisadas são as seguintes:
  - `Analfabetismo`: taxa de analfabetismo;

Exercício- Dados sócio econômicos dos municípios do Paraná

- Nesta aplicação vamos analisar dados sócio econômicos dos municípios do Estado do Paraná,disponíveis na página da disciplina.

- As variáveis a serem analisadas são as seguintes:
    - `Analfabetismo`: taxa de analfabetismo;
    - `Renda`: renda média domiciliar.

Exercício- Dados sócio econômicos dos municípios do Paraná

- Nesta aplicação vamos analisar dados sócio econômicos dos municípios do Estado do Paraná,disponíveis na página da disciplina.

- As variáveis a serem analisadas são as seguintes:
    - `Analfabetismo`: taxa de analfabetismo;
    - `Renda`: renda média domiciliar.

- Produza uma análise de correlação linear para este par de variáveis.

Exercício- Dados sócio econômicos dos municípios do Paraná

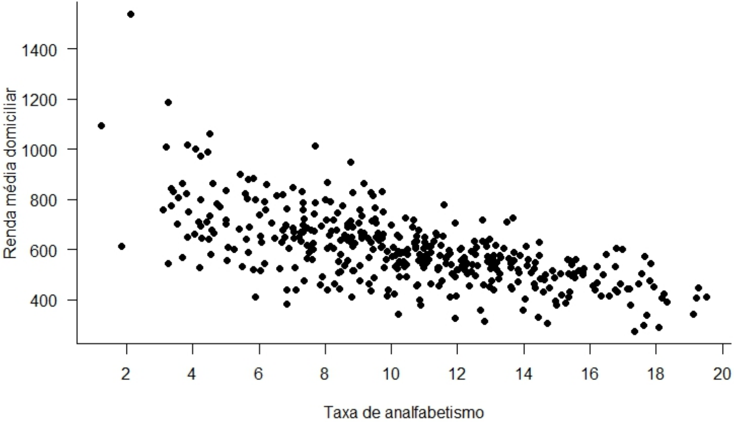


Figura 23: Dados sobre renda e analfabetismo dos municípios do Estado do Paraná

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

- Produza uma análise de correlações para o conjunto de variáveis físico-químicas (exclua da base a primeira coluna, que é um fator). Recomenda-se:

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

- Produza uma análise de correlações para o conjunto de variáveis físico-químicas (exclua da base a primeira coluna, que é um fator). Recomenda-se:
    - Consulte a documentação da base e quais são as características físico-químicas disponíveis;

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

- Produza uma análise de correlações para o conjunto de variáveis físico-químicas (exclua da base a primeira coluna, que é um fator). Recomenda-se:
  - Consulte a documentação da base e quais são as características físico-químicas disponíveis;
  - Obtenha a matriz de correlações;

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

- Produza uma análise de correlações para o conjunto de variáveis físico-químicas (exclua da base a primeira coluna, que é um fator). Recomenda-se:
  - Consulte a documentação da base e quais são as características físico-químicas disponíveis;
  - Obtenha a matriz de correlações;
  - Construa uma matriz de gráficos de dispersão (use, por exemplo, a função `ggpairs` da biblioteca `GGally`);

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

- Produza uma análise de correlações para o conjunto de variáveis físico-químicas (exclua da base a primeira coluna, que é um fator). Recomenda-se:
  - Consulte a documentação da base e quais são as características físico-químicas disponíveis;
  - Obtenha a matriz de correlações;
  - Construa uma matriz de gráficos de dispersão (use, por exemplo, a função `ggpairs` da biblioteca `GGally`);
  - Construa um correlograma (use, por exemplo, a função `corrplot` da biblioteca `corrplot`);

Exercício- Análise química de vinhos

- Nesta aplicação vamos analisar dados de características físico-químicas de 178 variedades de vinho, disponibilizadas na base de dados `wine`, que pode ser acessada na biblioteca `rattle` do `R`.

- Produza uma análise de correlações para o conjunto de variáveis físico-químicas (exclua da base a primeira coluna, que é um fator). Recomenda-se:
  - Consulte a documentação da base e quais são as características físico-químicas disponíveis;
  - Obtenha a matriz de correlações;
  - Construa uma matriz de gráficos de dispersão (use, por exemplo, a função `ggpairs` da biblioteca `GGally`);
  - Construa um correlograma (use, por exemplo, a função `corrplot` da biblioteca `corrplot`);
  - Verifique quais são as correlações estatisticamente significativas.

Alguns alertas na análise de regressão

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

- Os seguintes pontos serão aborados:

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

- Os seguintes pontos serão aborados:
    - O problema da extrapolação;

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

- Os seguintes pontos serão aborados:
  - O problema da extrapolação;
  - A disposição dos valores das covariáveis;

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

- Os seguintes pontos serão aborados:
  - O problema da extrapolação;
  - A disposição dos valores das covariáveis;
  - Presença de outliers e observações atípicas;

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

- Os seguintes pontos serão aborados:
  - O problema da extrapolação;
  - A disposição dos valores das covariáveis;
  - Presença de outliers e observações atípicas;
  - Causalidade e associação;

Alguns alertas na análise de regressão

- Nesta seção vamos abordar alguns alertas importantes que devem ser considerados na prática da análise de regressão;

- Importante destacar que esses alertas **não se limitam à regressão linear simples, nem mesmo à regressão linear**, pois se estendem para problemas gerais de análise de regressão.

- Os seguintes pontos serão aborados:
    - O problema da extrapolação;
    - A disposição dos valores das covariáveis;
    - Presença de outliers e observações atípicas;
    - Causalidade e associação;
    - Erros e perdas nas covariáveis.

O problema da extrapolação

- Numa análise de regressão, a não ser em casos específicos, limitamos os resultados aos intervalos de valores delimitados pelas covariáveis;

O problema da extrapolação

- Numa análise de regressão, a não ser em casos específicos, limitamos os resultados aos intervalos de valores delimitados pelas covariáveis;

- Extrapolar os resultados da análise para regiões de valores não observados das covariáveis pode gerar resultados pouco consistentes.

O problema da extrapolação

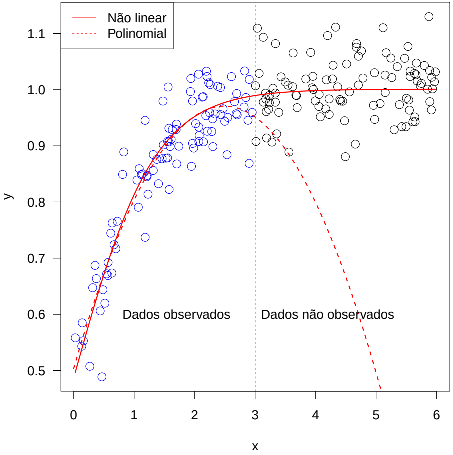


Figura 24: O problema da extrapolação.

Disposição dos valores das covariáveis

- Observações com valores extremos (atípicos) para as covariáveis têm maior peso na estimação dos parâmetros e ajuste do modelo;

Disposição dos valores das covariáveis

- Observações com valores extremos (atípicos) para as covariáveis têm maior peso na estimação dos parâmetros e ajuste do modelo;

- Nessas situações, deve-se ter cautela quanto ao efeito dessas observações nos resultados da análise;

Disposição dos valores das covariáveis

- Observações com valores extremos (atípicos) para as covariáveis têm maior peso na estimação dos parâmetros e ajuste do modelo;

- Nessas situações, deve-se ter cautela quanto ao efeito dessas observações nos resultados da análise;

- Adicionalmente, em estudos experimentais é possível controlar (fixar) os valores das covariáveis;

Disposição dos valores das covariáveis

- Observações com valores extremos (atípicos) para as covariáveis têm maior peso na estimação dos parâmetros e ajuste do modelo;

- Nessas situações, deve-se ter cautela quanto ao efeito dessas observações nos resultados da análise;

- Adicionalmente, em estudos experimentais é possível controlar (fixar) os valores das covariáveis;

- Nos casos em que a alocação das observações é definida pelo pesquisador, escolhas ótimas, quanto à investigação da relação entre as variáveis e precisão dos estimadores, podem ser buscadas.

Presença de outliers

- Outliers são observações que produzem valores para a resposta que são extremos (pouco compatíveis) em relação aos respectivos valores das covariáveis;

Presença de outliers

- Outliers são observações que produzem valores para a resposta que são extremos (pouco compatíveis) em relação aos respectivos valores das covariáveis;

- Observações que produzem elevados valores para os resíduos são potenciais outliers;

Presença de outliers

- Outliers são observações que produzem valores para a resposta que são extremos (pouco compatíveis) em relação aos respectivos valores das covariáveis;

- Observações que produzem elevados valores para os resíduos são potenciais outliers;

- Também aqui, investigar possíveis causas para o outlier e o impacto que eles produzem nos resultados da análise é fundamental.

Causalidade e associação

- A menos de situações específicas, como em experimentos planejados, modelos de regressão não permitem extrair relações de causa e efeito;

Causalidade e associação

- A menos de situações específicas, como em experimentos planejados, modelos de regressão não permitem extrair relações de causa e efeito;

- Ao identificar um resultado significativo, podemos atestar a associação entre as variáveis, mas não que a covariável está produzindo efeito na resposta;

Causalidade e associação

- A menos de situações específicas, como em experimentos planejados, modelos de regressão não permitem extrair relações de causa e efeito;

- Ao identificar um resultado significativo, podemos atestar a associação entre as variáveis, mas não que a covariável está produzindo efeito na resposta;

- Na sequência apresentamos dados de um estudo fictício, em que foram levantados os tamanhos dos pés (numeração dos calçados) e escore de habilidade verbal de crianças e adolescentes.

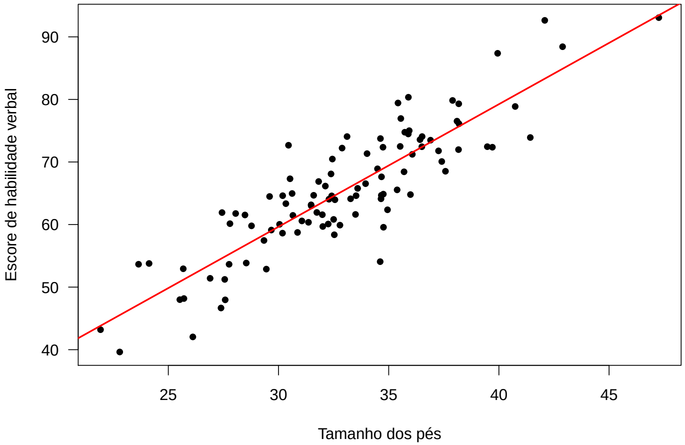


Figura 25: Relação entre tamanho dos pés e habilidade verbal.

Causalidade e associação

- Habilidade verbal e tamanho dos pés estão claramente relacionados;

Causalidade e associação

- Habilidade verbal e tamanho dos pés estão claramente relacionados;
- No entanto, seria absurdo imaginar que o tamanho dos pés esteja causando maior habilidade verbal nas crianças;

Causalidade e associação

- Habilidade verbal e tamanho dos pés estão claramente relacionados;
- No entanto, seria absurdo imaginar que o tamanho dos pés esteja causando maior habilidade verbal nas crianças;
- A princípio, o tamanho dos pés está associado à idade da criança, por consequência à sua escolaridade, que está relacionada à habilidade verbal...

Causalidade e associação

- Habilidade verbal e tamanho dos pés estão claramente relacionados;
- No entanto, seria absurdo imaginar que o tamanho dos pés esteja causando maior habilidade verbal nas crianças;
- A princípio, o tamanho dos pés está associado à idade da criança, por consequência à sua escolaridade, que está relacionada à habilidade verbal...
- Além disso, diversos outros fatores não considerados na análise podem estar associados à resposta, como escolaridade dos pais, renda, origem da criança...

Causalidade e associação

- Habilidade verbal e tamanho dos pés estão claramente relacionados;
- No entanto, seria absurdo imaginar que o tamanho dos pés esteja causando maior habilidade verbal nas crianças;
- A princípio, o tamanho dos pés está associado à idade da criança, por consequência à sua escolaridade, que está relacionada à habilidade verbal...
- Além disso, diversos outros fatores não considerados na análise podem estar associados à resposta, como escolaridade dos pais, renda, origem da criança...
- Em estudos experimentais podemos controlar fatores que possam afetar a resposta de maneira a estabelecer possível relação de causa-efeito com alguma variável de interesse.

Erros e perdas nas covariáveis

- Em alguns casos podemos ter incerteza (erro) também em relação aos valores das covariáveis;

Erros e perdas nas covariáveis

- Em alguns casos podemos ter incerteza (erro) também em relação aos valores das covariáveis;

- Em experimentos na Química, por exemplo, os valores de algumas covariáveis podem ser aferidos com nível de precisão tal que os valores obtidos estejam sujeitos a erros;

Erros e perdas nas covariáveis

- Em alguns casos podemos ter incerteza (erro) também em relação aos valores das covariáveis;

- Em experimentos na Química, por exemplo, os valores de algumas covariáveis podem ser aferidos com nível de precisão tal que os valores obtidos estejam sujeitos a erros;

- Além disso, alguns valores das covariáveis podem não ter sido registrados (dados missing), o que produzirá perda na precisão dos resultados da análise;

Erros e perdas nas covariáveis

- Em alguns casos podemos ter incerteza (erro) também em relação aos valores das covariáveis;

- Em experimentos na Química, por exemplo, os valores de algumas covariáveis podem ser aferidos com nível de precisão tal que os valores obtidos estejam sujeitos a erros;

- Além disso, alguns valores das covariáveis podem não ter sido registrados (dados missing), o que produzirá perda na precisão dos resultados da análise;

- Modelos de regressão com erro nas covariáveis e técnicas de imputação de dados devem ser considerados nos casos de erros e perdas nas covariáveis, respectivamente.

Tópicos adicionais

Análise da falta de ajuste da regressão linear

Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste permite avaliar formalmente a adequação do ajuste do modelo de regressão.

Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste permite avaliar formalmente a adequação do ajuste do modelo de regressão.

- Assumimos novamente que os pressupostos de normalidade, variância constante e independência são satisfeitos.

Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste permite avaliar formalmente a adequação do ajuste do modelo de regressão.

- Assumimos novamente que os pressupostos de normalidade, variância constante e independência são satisfeitos.

- A suposição sob teste é a de relação linear entre as variáveis.

Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste permite avaliar formalmente a adequação do ajuste do modelo de regressão.

- Assumimos novamente que os pressupostos de normalidade, variância constante e independência são satisfeitos.

- A suposição sob teste é a de relação linear entre as variáveis.

- O teste da falta de ajuste baseia-se na decomposição da variação residual em dois componentes: o primeiro atribuído à própria falta de ajuste; o segundo, ao erro puro.

Teste da falta de ajuste da regressão linear

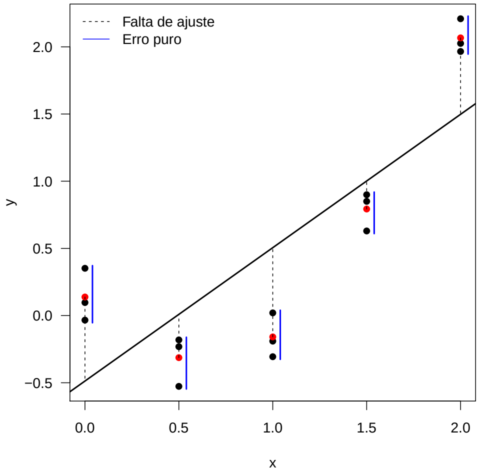


Figura 26: Ilustração da análise da falta de ajuste da regressão linear.

Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste requer que se disponha de replicações independentes de y para ao menos um valor de x.

Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste requer que se disponha de replicações independentes de y para ao menos um valor de x.

- Dispondo de replicações de y em diferentes valores de x, temos condições de obter uma estimativa para a variância (σ^2) que é independente do modelo de regressão ajustado.

Teste da falta de ajuste da regressão linear

- Seja y_{ij} a j-ésima observação da variável resposta para um particular valor x_i, $i = 1, 2, ..., m$; $j = 1, 2, ..., n_i$, $n = \sum_{i=1}^{m} n_i$. Então:

$$r_i = \underbrace{y_{ij} - \hat{y}_i}_{\text{Resíduo}} = \underbrace{(y_{ij} - \bar{y}_i)}_{\text{Erro puro}} + \underbrace{(\bar{y}_i - \hat{y}_i)}_{\text{Falta de ajuste}},$$

em que $\bar{y}_i$ é a média das n_i observações tomadas em x_i.

Teste da falta de ajuste da regressão linear

- Seja y_{ij} a j-ésima observação da variável resposta para um particular valor x_i, $i = 1, 2, ..., m$; $j = 1, 2, ..., n_i$, $n = \sum_{i=1}^{m} n_i$. Então:

$$r_i = \underbrace{y_{ij} - \hat{y}_i}_{\text{Resíduo}} = \underbrace{(y_{ij} - \bar{y}_i)}_{\text{Erro puro}} + \underbrace{(\bar{y}_i - \hat{y}_i)}_{\text{Falta de ajuste}},$$

em que $\bar{y}_i$ é a média das n_i observações tomadas em x_i.

- Tomando o quadrado de cada componente e somando-os, obtemos:

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2}_{\text{SQ}_{\text{Res}}} = \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}_{\text{SQ}_{\text{EP}}} + \underbrace{\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2}_{\text{SQ}_{\text{FA}}}.$$

Teste da falta de ajuste da regressão linear

- Assim, sob a suposição de variância constante SQ_{EP} é uma medida de dispersão dos erros independente do modelo, uma vez que é calculada com base nas variações dos $y's$ para cada valor de x_i.

Teste da falta de ajuste da regressão linear

- Assim, sob a suposição de variância constante SQ_{EP} é uma medida de dispersão dos erros independente do modelo, uma vez que é calculada com base nas variações dos $y's$ para cada valor de x_i.

- Cada valor x_i contribui com $n_i - 1$ graus de liberdade para o erro puro;

Teste da falta de ajuste da regressão linear

- Assim, sob a suposição de variância constante $\mathrm{SQ}_{\mathrm{EP}}$ é uma medida de dispersão dos erros independente do modelo, uma vez que é calculada com base nas variações dos $y's$ para cada valor de x_i.

- Cada valor x_i contribui com $n_i - 1$ graus de liberdade para o erro puro;

- Dessa forma, temos $\sum_{i=1}^{m}(n_i - 1) = n - m$ graus de liberdade para o erro puro e $(n-2) - (n-m) = m - 2$ graus de liberdade para a falta de ajuste.

Teste da falta de ajuste da regressão linear

- Assim, sob a suposição de variância constante $\mathrm{SQ}_{\mathrm{EP}}$ é uma medida de dispersão dos erros independente do modelo, uma vez que é calculada com base nas variações dos $y's$ para cada valor de x_i.

- Cada valor x_i contribui com $n_i - 1$ graus de liberdade para o erro puro;

- Dessa forma, temos $\sum_{i=1}^{m}(n_i - 1) = n - m$ graus de liberdade para o erro puro e $(n-2)-(n-m) = m-2$ graus de liberdade para a falta de ajuste.

- Os resultados da análise da falta de ajuste podem ser apresentados na forma de um quadro de análise de variância.

Teste da falta de ajuste da regressão linear

Tabela 4: Quadro de análise de variância para o teste da falta de ajuste

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F
Regressão	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	$\text{QM}_{\text{Reg}} = \frac{\text{SQ}_{\text{Reg}}}{1}$	$F = \frac{\text{QM}_{\text{Reg}}}{\text{QM}_{\text{Res}}}$
Resíduos	n-2	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\text{QM}_{\text{Res}} = \frac{\text{SQ}_{\text{Res}}}{n-2}$	
Falta de ajuste	m-2	$\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$	$\text{QM}_{\text{FA}} = \frac{\text{SQ}_{\text{FA}}}{m-2}$	$F = \frac{\text{QM}_{\text{FA}}}{\text{QM}_{\text{EP}}}$
Erro puro	n-m	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$\text{QM}_{\text{EP}} = \frac{\text{SQ}_{\text{EP}}}{n-m}$	
Total	n-1	$\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Teste da falta de ajuste da regressão linear

- Se o modelo se ajustar aos dados, então tanto QM_{EP} quanto QM_{FA} são estimadores não viciados de σ^2.

Teste da falta de ajuste da regressão linear

- Se o modelo se ajustar aos dados, então tanto QM_{EP} quanto QM_{FA} são estimadores não viciados de σ^2.

- Caso contrário, se o modelo não se ajustar aos dados, então $\text{E}(\text{QM}_{\text{FA}}) > \sigma^2$.

Teste da falta de ajuste da regressão linear

- Se o modelo se ajustar aos dados, então tanto $\mathrm{QM}_{\mathrm{EP}}$ quanto $\mathrm{QM}_{\mathrm{FA}}$ são estimadores não viciados de σ^2.

- Caso contrário, se o modelo não se ajustar aos dados, então $\mathrm{E}(\mathrm{QM}_{\mathrm{FA}}) > \sigma^2$.

- Sob a hipótese nula de que não há falta de ajuste, então:

$$F_0 = \frac{\mathrm{SQ}_{\mathrm{FA}}/(m-2)}{\mathrm{SQ}_{\mathrm{EP}}/(n-m)} = \frac{\mathrm{QM}_{\mathrm{FA}}}{\mathrm{QM}_{\mathrm{EP}}}$$

tem distribuição F-Snedecor com graus de liberdade $m-2$ e $n-m$.

Teste da falta de ajuste da regressão linear

- Assim, a hipótese nula de que não há falta de ajuste deverá ser rejeitada, ao nível de significância α, se $F_0 > F_{m-2,n-m;1-\alpha}$.

Teste da falta de ajuste da regressão linear

- Assim, a hipótese nula de que não há falta de ajuste deverá ser rejeitada, ao nível de significância α, se $F_0 > F_{m-2,n-m;1-\alpha}$.

- O nível descritivo (p-valor) do teste pode ser calculado por $P(X > F_0)$, sendo $X \sim F_{m-2,n-m}$.

Teste da falta de ajuste da regressão linear

- Assim, a hipótese nula de que não há falta de ajuste deverá ser rejeitada, ao nível de significância α, se $F_0 > F_{m-2,n-m;1-\alpha}$.

- O nível descritivo (p-valor) do teste pode ser calculado por $P(X > F_0)$, sendo $X \sim F_{m-2,n-m}$.

- No caso em que não se dispõe de réplicas de y para testar a falta de ajuste, uma estratégia consiste em agrupar indivíduos com valores próximos de x e proceder a análise (para mais informações consultar Montgomery, Peck e Vinning, 2006).

Exemplo- Corrosão de ligas metálicas

Exemplo- Corrosão de ligas metálicas

- Nesta aplicação, as seguintes variáveis são consideradas:

Exemplo- Corrosão de ligas metálicas

- Nesta aplicação, as seguintes variáveis são consideradas:
  - `loss`: corrosão em corpos de liga metálica (variável resposta);

Exemplo- Corrosão de ligas metálicas

- Nesta aplicação, as seguintes variáveis são consideradas:
  - `loss:` corrosão em corpos de liga metálica (variável resposta);
  - `Fe:` teor de ferro (variável explicativa).

Exemplo- Corrosão de ligas metálicas

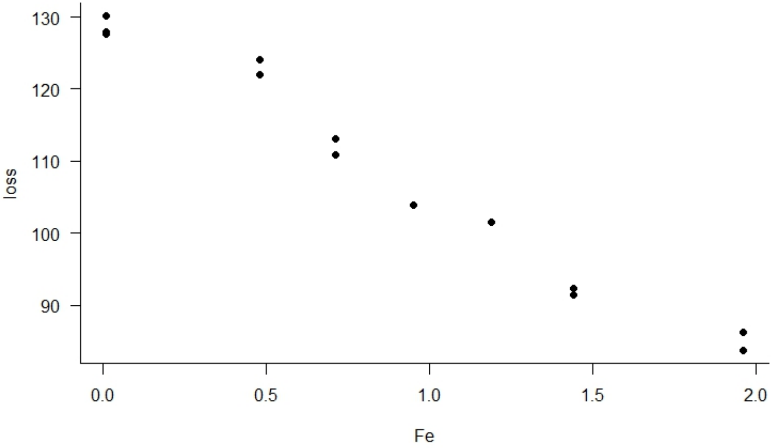


Figura 27: Dados de corrosão de ligas metálicas

Exemplo- Corrosão de ligas metálicas

- Modelo ajustado:

$$\widehat{\texttt{loss}} = 129.79 - 24.02 \times \texttt{Fe}$$

- Modelo ajustado:

$$\widehat{\texttt{loss}} = 129.79 - 24.02 \times \texttt{Fe}$$

$$\text{SQ}_{\text{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$$

Exemplo- Corrosão de ligas metálicas

- Modelo ajustado:

$$\widehat{\texttt{loss}} = 129.79 - 24.02 \times \texttt{Fe}$$

$$\text{SQ}_{\text{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$$

$$(129.54 - 108.81)^2 + (118.26 - 108.81)^2 + ... + (82.71 - 108.81)^2 = 3293.77$$

Exemplo- Corrosão de ligas metálicas

$$\text{SQ}_{\text{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

Exemplo- Corrosão de ligas metálicas

$$\mathrm{SQ}_{\mathrm{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

$$(127.6 - 129.54)^2 + (124 - 118.26)^2 + ... + (86.2 - 82.71)^2 = 102.85$$

Exemplo- Corrosão de ligas metálicas

$$\mathrm{SQ}_{\mathrm{Total}} = \mathrm{SQ}_{\mathrm{Reg}} + \mathrm{SQ}_{\mathrm{Res}} = 3293.77 + 102.85 = 3396.62$$

Exemplo- Corrosão de ligas metálicas

$$\text{SQ}_{\text{Total}} = \text{SQ}_{\text{Reg}} + \text{SQ}_{\text{Res}} = 3293.77 + 102.85 = 3396.62$$

$$\text{QM}_{\text{Res}} = \frac{\text{SQ}_{\text{Res}}}{n-2} = \frac{102.85}{11} = 9.35$$

Exemplo- Corrosão de ligas metálicas

$$\mathrm{SQ}_{\mathrm{Total}} = \mathrm{SQ}_{\mathrm{Reg}} + \mathrm{SQ}_{\mathrm{Res}} = 3293.77 + 102.85 = 3396.62$$

$$\mathrm{QM}_{\mathrm{Res}} = \frac{\mathrm{SQ}_{\mathrm{Res}}}{n-2} = \frac{102.85}{11} = 9.35$$

$$\mathrm{F} = \frac{\mathrm{QM}_{\mathrm{Reg}}}{\mathrm{QM}_{\mathrm{Res}}} = \frac{3293.77}{9.35} = 352.27$$

Exemplo- Corrosão de ligas metálicas

Tabela 5: Tabela auxiliar para o teste da falta de ajuste

Fe	n_i	$\bar{y}_i$	$\hat{y}_i$
0.01	3	128.57	129.55
0.48	2	123.00	118.26
0.71	2	111.95	112.73
0.95	1	103.90	106.97
1.19	1	101.50	101.20
1.44	2	91.85	95.20
1.96	2	84.95	82.71

Exemplo- Corrosão de ligas metálicas

$$\mathrm{SQ}_{\mathrm{FA}} = \sum_{i=1}^{m} n_i(\hat{y}_i - \bar{y}_i)^2 =$$

Exemplo- Corrosão de ligas metálicas

$$\text{SQ}_{\text{FA}} = \sum_{i=1}^{m} n_i(\hat{y}_i - \bar{y}_i)^2 =$$

$$3 \times (129.55 - 128.57)^2 + 2 \times (118.26 - 123)^2 + ... + 2 \times (82.71 - 84.95)^2 = 91.03$$

Exemplo- Corrosão de ligas metálicas

$$\mathrm{SQ}_{\mathrm{EP}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 =$$

Exemplo- Corrosão de ligas metálicas

$$\mathrm{SQ}_{\mathrm{EP}} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 =$$

$$(127.6 - 128.57)^2 + (124.0 - 128.57)^2 + ... + (86.2 - 84.95)^2 = 11.78$$

Exemplo- Corrosão de ligas metálicas

$$\text{SQ}_{\text{EP}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 =$$

$$(127.6 - 128.57)^2 + (124.0 - 128.57)^2 + ... + (86.2 - 84.95)^2 = 11.78$$

$$\text{QM}_{\text{FA}} = \frac{\text{SQ}_{\text{FA}}}{m-2} = \frac{91.03}{7-2} = 18.21$$

Exemplo- Corrosão de ligas metálicas

$$\text{SQ}_{\text{EP}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 =$$

$$(127.6 - 128.57)^2 + (124.0 - 128.57)^2 + ... + (86.2 - 84.95)^2 = 11.78$$

$$\text{QM}_{\text{FA}} = \frac{\text{SQ}_{\text{FA}}}{m-2} = \frac{91.03}{7-2} = 18.21$$

$$\text{QM}_{\text{EP}} = \frac{\text{SQ}_{\text{EP}}}{n-m} = \frac{11.78}{13-7} = 1.96$$

Exemplo- Corrosão de ligas metálicas

$$\text{SQ}_{\text{EP}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 =$$

$$(127.6 - 128.57)^2 + (124.0 - 128.57)^2 + ... + (86.2 - 84.95)^2 = 11.78$$

$$\text{QM}_{\text{FA}} = \frac{\text{SQ}_{\text{FA}}}{m-2} = \frac{91.03}{7-2} = 18.21$$

$$\text{QM}_{\text{EP}} = \frac{\text{SQ}_{\text{EP}}}{n-m} = \frac{11.78}{13-7} = 1.96$$

$$\text{F} = \frac{\text{QM}_{\text{FA}}}{\text{QM}_{\text{EP}}} = \frac{18.21}{1.96} = 9.29$$

Exemplo- Corrosão de ligas metálicas

Tabela 6: Quadro de análise de variância para o teste da falta de ajuste

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F
Regressão	1	3293.77	3293.77	352.27
Resíduos	11	102.85	9.35	
Falta de ajuste	5	91.03	18.21	9.29
Erro puro	6	11.78	1.96	
Total	12	3396.62		

Exemplo- Corrosão de ligas metálicas

- A hipótese nula de que o modelo não sofre de falta de ajuste deve ser rejeitada, ao nível de 5% de significância, se:

$$\mathrm{F} > \mathrm{F}_{m-2,n-m}(0.95) = 4.39$$

Exemplo- Corrosão de ligas metálicas

- A hipótese nula de que o modelo não sofre de falta de ajuste deve ser rejeitada, ao nível de 5% de significância, se:

$$\mathrm{F} > \mathrm{F}_{m-2,n-m}(0.95) = 4.39$$

- Como $\mathrm{F} = 9.29 > 4.39$, rejeitamos a hipótese nula, e concluímos que o modelo sofre de falta de ajuste.

Exercício adicional

Exercício- Efeito de Nitrofen na eclosão de ovos

- Nesta aplicação vamos analisar os dados de um experimento sobre eclosão de ovos de uma espécie marinha, disponíveis em script `R` na página da disciplina (ver as informações adicionais no próprio arquivo).

Exercício- Efeito de Nitrofen na eclosão de ovos

- Nesta aplicação vamos analisar os dados de um experimento sobre eclosão de ovos de uma espécie marinha, disponíveis em script `R` na página da disciplina (ver as informações adicionais no próprio arquivo).

- Os dados referem-se a uma amostra de 50 C.dubia (pequeno animal invertebrado aquatico de agua doce), que foram submetidos a dosagens diferentes do herbicida Nitrofen.

Exercício- Efeito de Nitrofen na eclosão de ovos

- Nesta aplicação vamos analisar os dados de um experimento sobre eclosão de ovos de uma espécie marinha, disponíveis em script `R` na página da disciplina (ver as informações adicionais no próprio arquivo).

- Os dados referem-se a uma amostra de 50 C.dubia (pequeno animal invertebrado aquatico de agua doce), que foram submetidos a dosagens diferentes do herbicida Nitrofen.

- As variáveis a serem analisadas são as seguintes:

Exercício- Efeito de Nitrofen na eclosão de ovos

- Nesta aplicação vamos analisar os dados de um experimento sobre eclosão de ovos de uma espécie marinha, disponíveis em script `R` na página da disciplina (ver as informações adicionais no próprio arquivo).

- Os dados referem-se a uma amostra de 50 C.dubia (pequeno animal invertebrado aquatico de agua doce), que foram submetidos a dosagens diferentes do herbicida Nitrofen.

- As variáveis a serem analisadas são as seguintes:
  - `tovos`: número de ovos eclodidos (resposta);

Exercício- Efeito de Nitrofen na eclosão de ovos

- Nesta aplicação vamos analisar os dados de um experimento sobre eclosão de ovos de uma espécie marinha, disponíveis em script `R` na página da disciplina (ver as informações adicionais no próprio arquivo).

- Os dados referem-se a uma amostra de 50 C.dubia (pequeno animal invertebrado aquatico de agua doce), que foram submetidos a dosagens diferentes do herbicida Nitrofen.

- As variáveis a serem analisadas são as seguintes:
  - `tovos`: número de ovos eclodidos (resposta);
  - `dose`: dose aplicada do herbicida Nitrofen: 0, 80, 160, 235 e 310 mg/l.

Exercício- Efeito de Nitrofen na eclosão de ovos

- Nesta aplicação vamos analisar os dados de um experimento sobre eclosão de ovos de uma espécie marinha, disponíveis em script `R` na página da disciplina (ver as informações adicionais no próprio arquivo).

- Os dados referem-se a uma amostra de 50 C.dubia (pequeno animal invertebrado aquatico de agua doce), que foram submetidos a dosagens diferentes do herbicida Nitrofen.

- As variáveis a serem analisadas são as seguintes:
  - `tovos`: número de ovos eclodidos (resposta);
  - `dose`: dose aplicada do herbicida Nitrofen: 0, 80, 160, 235 e 310 mg/l.

- Teste a falta de ajuste da regressão linear simples.

Exercício- Efeito de Nitrofen na eclosão de ovos

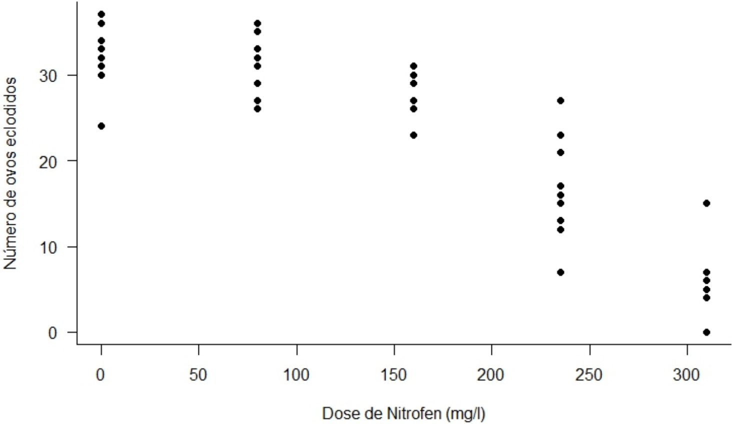


Figura 28: Dados de experimento sobre eclosão de ovos sob diferentes doses de Nitrofen

Mudança de escala em regressão

Mudança de escala em regressão

- Na análise de regressão é comum recorrer a alguma mudança de escala das variáveis.

Mudança de escala em regressão

- Na análise de regressão é comum recorrer a alguma mudança de escala das variáveis.

- Como exemplos, podemos converter dados referentes a pesos de gramas para quilogramas; medidas de extensão de quilômetros para metros; dados financeiros de reais para milhares de reais…

Mudança de escala em regressão

- Na análise de regressão é comum recorrer a alguma mudança de escala das variáveis.

- Como exemplos, podemos converter dados referentes a pesos de gramas para quilogramas; medidas de extensão de quilômetros para metros; dados financeiros de reais para milhares de reais...

- Mudanças de escala têm diferentes propósitos na análise de regressão, como veremos ao longo da disciplina.

Mudança de escala em regressão

- Duas mudanças de escala bastante usadas consistem em *centrar* e *escalonar* os valores de uma ou mais covariáveis.

Mudança de escala em regressão

- Duas mudanças de escala bastante usadas consistem em *centrar* e *escalonar* os valores de uma ou mais covariáveis.

- Sejam $x_1^* = (x_1 - \bar{x}), x_2^* = (x_2 - \bar{x}), ..., x_n^* = (x_n - \bar{x})$ os valores centrados de uma covariável x;

Mudança de escala em regressão

- Duas mudanças de escala bastante usadas consistem em *centrar* e *escalonar* os valores de uma ou mais covariáveis.
- Sejam $x_1^* = (x_1 - \bar{x}), x_2^* = (x_2 - \bar{x}), ..., x_n^* = (x_n - \bar{x})$ os valores centrados de uma covariável x;
- Neste caso:

$$y = \beta_0 + \beta_1(x - \bar{x} + \bar{x}) + \epsilon' = \underbrace{\beta_0 + \beta_1\bar{x}}_{\beta_0'} + \beta_1 \underbrace{(x - \bar{x})}_{x*} + \epsilon',$$

de maneira que o intercepto da regressão fica alterado para $\beta_0' = \beta_0 + \beta_1\bar{x}$, mas a inclinação fica inalterada em relação à regressão original (dados não centrados).

Mudança de escala em regressão

- Sejam $x_1^* = \frac{x_1 - \bar{x}}{s}, x_2^* = \frac{x_2 - \bar{x}}{s}, ..., x_n^* = \frac{x_n - \bar{x}}{s}$ os valores centrados e escalonados de uma covariável x;

Mudança de escala em regressão

- Sejam $x_1^* = \frac{x_1 - \bar{x}}{s}, x_2^* = \frac{x_2 - \bar{x}}{s}, ..., x_n^* = \frac{x_n - \bar{x}}{s}$ os valores centrados e escalonados de uma covariável x;

- Neste caso:

$$y = \beta_0 + \beta_1 \left[s \left(\frac{x - \bar{x}}{s} \right) + \bar{x} \right] + \epsilon' = \underbrace{\beta_0 + \beta_1 \bar{x}}_{\beta_0'} + \underbrace{\beta_1 s}_{\beta_1'} \underbrace{\left(\frac{x - \bar{x}}{s} \right)}_{x^*} + \epsilon',$$

de maneira que o intercepto da regressão fica alterado conforme no modelo apenas centrado, enquanto o parâmetro de inclinação fica multiplicado pelo desvio padrão da covariável.

Mudança de escala em regressão

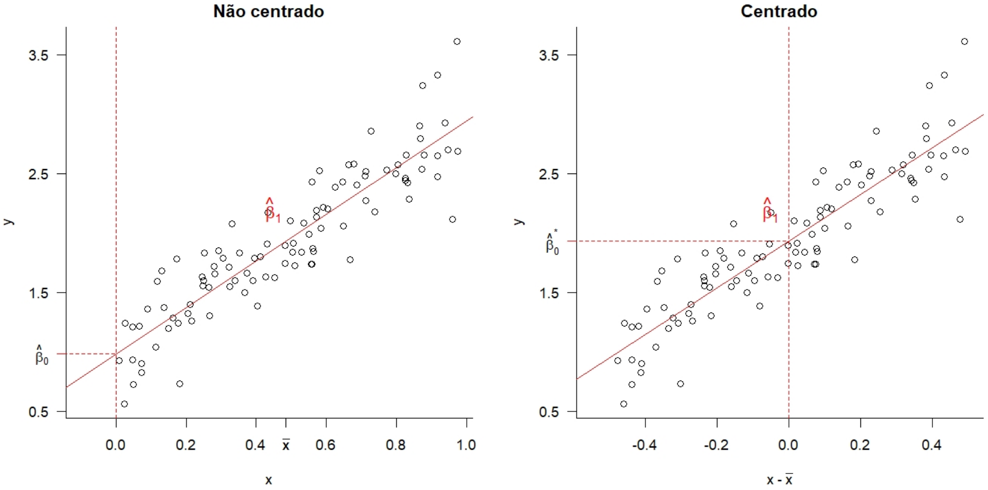


Figura 29: Efeito de centrar os valores da variável explicativa na regressão linear simples.

Mudança de escala em regressão

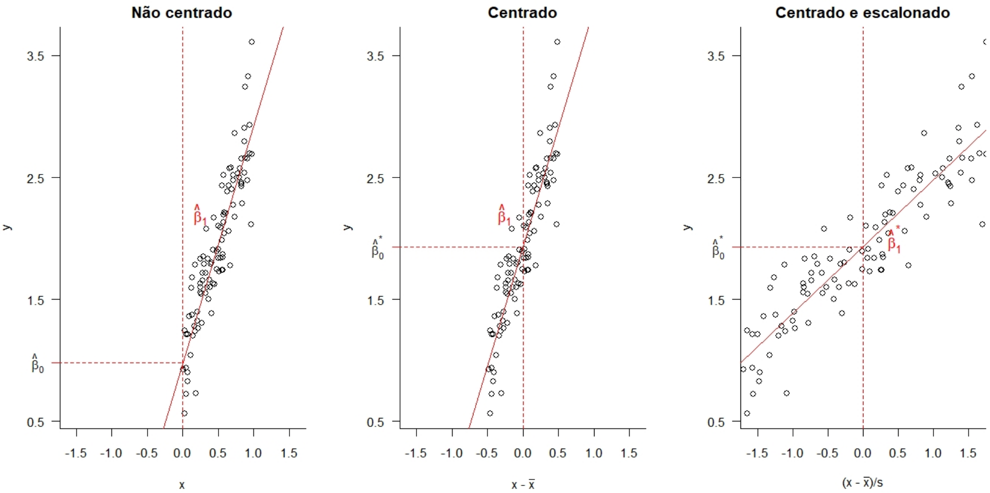


Figura 30: Efeito de centrar e escalonar os valores da variável explicativa na regressão linear simples.

Mudança de escala em regressão

- De maneira geral, se considerarmos uma mudança de escala do tipo $x^* = a + bx$, então teremos:

$$y = \beta_0 + \beta_1 \left[(1/b)(a + bx) - a\right] + \epsilon' = \underbrace{\beta_0 - \beta_1 a}_{\beta_0'} + \underbrace{\frac{\beta_1}{b}}_{\beta_1'} \underbrace{(a + bx)}_{x^*} + \epsilon'.$$

Exemplo- Anatomia de gatos domésticos

- A média amostral e o desvio padrão para os valores dos pesos corporais dos felinos são respectivamente $\overline{\texttt{Bwt}} = 2.723$kg e $s_{\texttt{Bwt}} = 0.485$kg.

Exemplo- Anatomia de gatos domésticos

- A média amostral e o desvio padrão para os valores dos pesos corporais dos felinos são respectivamente $\overline{\texttt{Bwt}} = 2.723$kg e $s_{\texttt{Bwt}} = 0.485$kg.

- Primeiramente, o modelo de regressão linear ajustado com a variável explicativa centrada na média é:

$$\widehat{\texttt{Hwt}} = 10.631 + 4.034(\texttt{Bwt} - \overline{\texttt{Bwt}})$$

Exemplo- Anatomia de gatos domésticos

- Interpretação dos parâmetros:

Exemplo- Anatomia de gatos domésticos

- Interpretação dos parâmetros:
  - **Intercepto:** Estima-se em 10.631g o peso médio do coração dos gatos com peso corporal igual à média amostral ($\texttt{Bwt} = \overline{\texttt{Bwt}} = 2.723$kg);

Exemplo- Anatomia de gatos domésticos

- Interpretação dos parâmetros:
  - **Intercepto:** Estima-se em 10.631g o peso médio do coração dos gatos com peso corporal igual à média amostral ($\texttt{Bwt} = \overline{\texttt{Bwt}} = 2.723$kg);
  - **Inclinação:** Estima-se um aumento médio de 4.034g no peso do coração a cada um quilograma a mais no peso corporal dos gatos (mesma interpretação do modelo anterior).

Exemplo- Anatomia de gatos domésticos

- Agora, o modelo com a variável explicativa centrada na média e escalonada:

$$\widehat{\texttt{Hwt}} = 10.631 + 1.958 \frac{(\texttt{Bwt} - \overline{\texttt{Bwt}})}{s_{\text{Bwt}}}$$

Exemplo- Anatomia de gatos domésticos

- Agora, o modelo com a variável explicativa centrada na média e escalonada:

$$\widehat{\texttt{Hwt}} = 10.631 + 1.958\frac{(\texttt{Bwt} - \overline{\texttt{Bwt}})}{s_{\text{Bwt}}}$$

- Interpretação dos parâmetros:

Exemplo- Anatomia de gatos domésticos

- Agora, o modelo com a variável explicativa centrada na média e escalonada:

$$\widehat{\texttt{Hwt}} = 10.631 + 1.958\frac{(\texttt{Bwt} - \overline{\texttt{Bwt}})}{s_{\text{Bwt}}}$$

- Interpretação dos parâmetros:
  - **Intercepto:** Estima-se em 10.631g o peso médio do coração dos gatos com peso corporal igual à média amostral ($\texttt{Bwt} = \overline{\texttt{Bwt}} = 2.723$kg);

Exemplo- Anatomia de gatos domésticos

- Agora, o modelo com a variável explicativa centrada na média e escalonada:

$$\widehat{\texttt{Hwt}} = 10.631 + 1.958\frac{(\texttt{Bwt} - \overline{\texttt{Bwt}})}{s_{\text{Bwt}}}$$

- Interpretação dos parâmetros:
  - **Intercepto:** Estima-se em 10.631g o peso médio do coração dos gatos com peso corporal igual à média amostral ($\texttt{Bwt} = \overline{\texttt{Bwt}} = 2.723$kg);
  - **Inclinação:** Estima-se um aumento médio de 1.958g no peso do coração a cada $s_{\text{Bwt}} = 0.485$kg a mais no peso corporal dos gatos.

Estimação por máxima verossimilhança

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).

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 $y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$.

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:

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

Estimação por máxima verossimilhança

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

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

Estimação por máxima verossimilhança

- Os estimadores de máxima verossimilhança são resultantes do seguinte sistema de derivadas parciais:

$$
\begin{aligned}
\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma^2}} \ln\left[L(\beta_0,\beta_1,\sigma^2;\boldsymbol{y},\boldsymbol{x})\right] &= 0;\\
\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma^2}} \ln\left[L(\beta_0,\beta_1,\sigma^2;\boldsymbol{y},\boldsymbol{x})\right] &= 0;\\
\left.\frac{\partial S}{\partial \sigma^2}\right|_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma^2}} \ln\left[L(\beta_0,\beta_1,\sigma^2;\boldsymbol{y},\boldsymbol{x})\right] &= 0.
\end{aligned}
$$

Estimação por máxima verossimilhança

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

Estimação por máxima verossimilhança

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

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

Estimação por máxima verossimilhança

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

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

- 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:

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

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