

# CE071 - Análise de Regressão Linear

Cesar Augusto Taconeli

21 de março, 2018

## Aula 2 - Regressão linear simples

# Definição e propriedades

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

$$y = \beta_0 + \beta_1 x + \epsilon, \quad (1)$$

em que  $\beta_0$  é o intercepto e  $\beta_1$  a inclinação da reta, e  $\epsilon$  representa o erro aleatório.

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

# Definição e propriedades

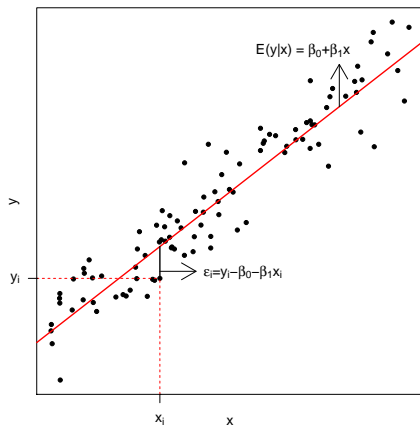
- Condicional a um valor observado  $x$ , a média da distribuição de  $y$  fica dada por:

$$E(y|x) = \beta_0 + \beta_1 x. \quad (2)$$

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

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

# Definição e propriedades

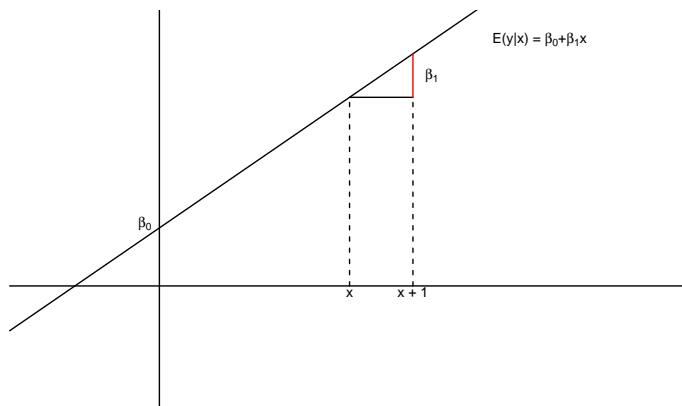


**Figura 1:** Regressão linear simples.

# Definição e propriedades

- Interpretação dos parâmetros do modelo:
  - $\beta_1$  expressa a alteração no valor esperado de  $y$  associada ao acréscimo de uma unidade em  $x$ ;
  - $\beta_0$  é o valor esperado de  $y$  quando  $x = 0$  (caso  $x = 0$  faça parte do suporte do problema).

# Definição e propriedades



**Figura 2:** Interpretação dos parâmetros.

# Estimação por mínimos quadrados

- A estimação de  $\beta_0$  e  $\beta_1$  por mínimos quadrados baseia-se em  $n$  observações para as quais se dispõe dos valores de  $x$  e  $y$ , ou seja,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n. \quad (4)$$

- O método de mínimos quadrados baseia-se na determinação de  $\beta_0$  e  $\beta_1$  tal que a soma de quadrados dos erros, definida na sequência, seja mínima:

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

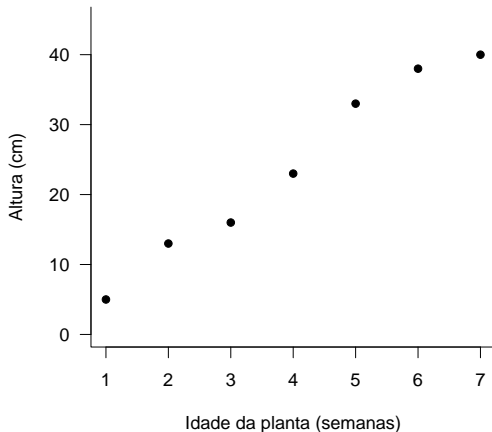


## Exemplo 1

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

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

# Estimação por mínimos quadrados



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

# Estimação por mínimos quadrados

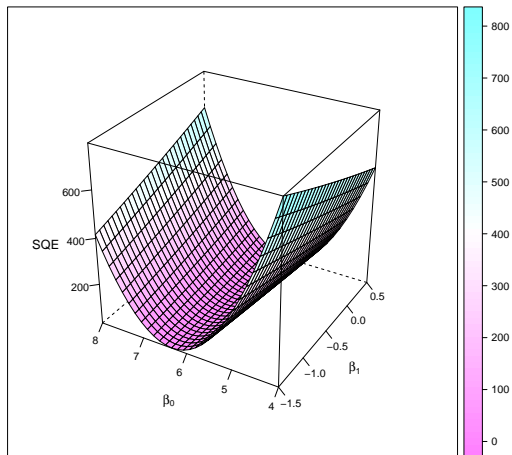


Figura 4: Ilustração da estimação por mínimos quadrados.

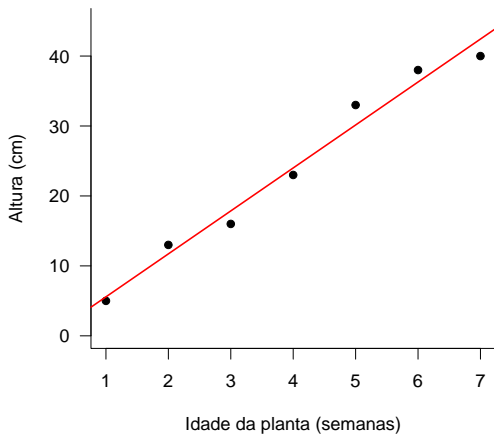
# Estimação por mínimos quadrados

- Observando a figura 4, as estimativas de mínimos quadrados para  $\beta_0$  e  $\beta_1$  (denotadas por  $\hat{\beta}_0$  e  $\hat{\beta}_1$ ) correspondem aos valores de  $\beta_0$  e  $\beta_1$  tais que  $SQE$  seja mínimo.
- Para o presente problema, as estimativas de mínimos quadrados são dadas por  $\hat{\beta}_0 = -0.57$  e  $\hat{\beta}_1 = 6.14$ .
- O modelo ajustado é usualmente expresso da seguinte forma:

$$\hat{y} = -0.57 + 6.14x, \quad (6)$$

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

# Estimação por mínimos quadrados



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

# Estimação por mínimos quadrados

- Os estimadores de mínimos quadrados devem satisfazer:

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

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

# Estimação por mínimos quadrados

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (9)$$

e

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

# Estimação por mínimos quadrados

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \quad (11)$$

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

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n. \quad (12)$$

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



# Propriedades dos estimadores de mínimos quadrados

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

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

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

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

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

## Teorema de Gauss Markov

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

## Estimação de $\sigma^2$

- A estimação de  $\sigma^2$  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  $\sigma^2$  é baseado na soma de quadrados de resíduos:

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

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

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

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

# Regressão com dados centrados

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

$$y_i = \beta'_0 + \beta'_1(x_i - \bar{x}) + \epsilon_i. \quad (18)$$

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

# Regressão com dados centrados

- Como resultado, apenas o intercepto do modelo fica alterado para  $\beta'_0 = \beta_0 + \beta_1\bar{x}$ , em que  $\beta_0$  e  $\beta_1$  são os parâmetros do modelo com a variável  $x$  não centrada.
- O estimador de mínimos quadrados de  $\beta'_0$  fica dado por  $\bar{y}$ , e o estimador de  $\beta_1$  não é afetado pela transformação. Portanto, o modelo ajustado fica dado por:

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

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

- Neste ponto teremos que assumir, adicionalmente, que os erros são normalmente distribuídos (isto é, os erros são independentes com  $\epsilon \sim Normal(0, \sigma^2)$ ).
- A suposição de que os erros têm distribuição Normal implica  $y|x \stackrel{ind}{\sim} Normal(\beta_0 + \beta_1 x, \sigma^2)$ .

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

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

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

- De maneira semelhante:

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

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

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

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

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



- Vamos considerar o teste de que  $\beta_1$  é igual a um particular valor postulado constante  $\beta_{10}$ :

$$H_0 : \beta_1 = \beta_{10} \text{ vs } H_1 : \beta_1 \neq \beta_{10}. \quad (23)$$

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

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

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

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

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

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

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

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

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

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

- Com base no resultado (30) pode-se conduzir o teste da hipótese  $H_0 : \beta_1 = \beta_{10}$ .
- Fixando o nível de significância em  $\alpha$ ,  $H_0$  será rejeitada se  $|t| > |t_{n-2; \alpha/2}|$ , em que  $t_{n-2; \alpha/2}$  é o quantil  $\alpha/2$  da distribuição  $t_{n-2}$ .

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

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

- Um intervalo de confiança  $100(1 - \alpha)\%$  para  $\beta_1$  é definido pelo par de limites:

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

# Teste da significância da regressão

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

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

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

# Teste da significância da regressão

- É importante ressaltar que a não rejeição de  $H_0 : \beta_1 = 0$  permite concluir que não há relação linear entre  $y$  e  $x$ , mas não que não se tenha relação entre as variáveis.
- Além disso, ainda que  $H_0$  seja rejeitada, isso não implica que um modelo não linear (como um polinômio, por exemplo), seja mais adequado para explicar a relação entre as variáveis.

# Testes de hipóteses e intervalos de confiança para $\beta_0$

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

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

sob a suposição de que a hipótese nula é verdadeira.

# Testes de hipóteses e intervalos de confiança para $\beta_0$

- Fixando o nível de significância em  $\alpha$ , 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  $\beta_0$  é definido pelo par de limites:

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



# Intervalo de confiança para $\sigma^2$

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

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

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

# Intervalo de confiança para a resposta média

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

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

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

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

# Intervalo de confiança para a resposta média

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

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

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

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

# Intervalo de confiança para a resposta média

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

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

## Predição de uma nova observação

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

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad (39)$$

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

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

# Predição de uma nova observação

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

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

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

# Estimação por máxima verossimilhança

- A estimação de  $\beta_0$  e  $\beta_1$  por máxima verossimilhança baseia-se, novamente, em  $n$  observações para as quais se dispõe dos valores de  $x$  e  $y$ , ou seja,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ :
- 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:

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

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

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



- Os estimadores de máxima verossimilhança devem satisfazer a:

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

# Estimação por máxima verossimilhança

- Observe que maximizar  $\ln L(\beta_0, \beta_1, \sigma^2; \mathbf{y}, \mathbf{x})$  com relação a  $\beta_0$  e  $\beta_1$  equivale a maximizar  $-\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = -SQE$  em função desses parâmetros;
- Lembre que na estimação por mínimos quadrados a obtenção dos estimadores dos parâmetros do modelo era obtida pela minimização de  $SQE = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ ;
- Uma vez que minimizar  $SQE$  é equivalente a maximizar  $-SQE$ , os estimadores de máxima verossimilhança para  $\beta_0$  e  $\beta_1$  são idênticos aos de mínimos quadrados.

# Estimação por máxima verossimilhança

- O estimador de máxima verossimilhança de  $\sigma^2$ , por sua vez, é dado por:

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

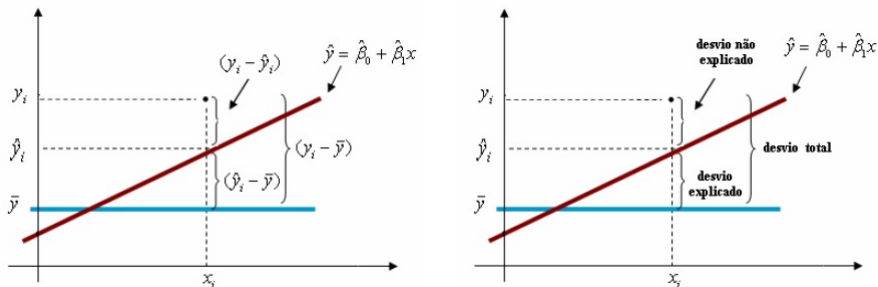
que, diferentemente do estimador estudado anteriormente, é viciado para  $\sigma^2$  (mas **assintoticamente** não viciado).

# Análise de variância aplicada à regressão linear simples

- A análise de variância é uma técnica que permite particionar a variação total dos dados em parcelas atribuíveis a diferentes fontes.
- No contexto de regressão, a análise de variância baseia-se na seguinte identidade:

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

# Análise de variância aplicada à regressão linear simples



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

# Análise de variância aplicada à regressão linear simples

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

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (47)$$

$SQ_{Total}$                        $SQ_{Reg}$                        $SQ_{Res}$

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  $SQ_{Reg}$  em detrimento a  $SQ_{Res}$ , maior a parcela da variação total dos dados explicada pela regressão.
- Associado a cada componente dessa decomposição temos:
  - $n - 1$  graus de liberdade para  $SQ_{Total}$  (perda de um grau devido à estimação da média);
  - $n - 2$  graus de liberdade para  $SQ_{Res}$  (perda de dois graus devido à estimação de  $\beta_0$  e  $\beta_1$ );
  - $(n - 1) - (n - 2) = 1$  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	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$QM_{Reg} = \frac{SQ_{Reg}}{1}$	$F = \frac{QM_{Reg}}{QM_{Res}}$
Resíduos	n-2	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$QM_{Res} = \frac{SQ_{Res}}{n-2}$	
Total	n-1	$\sum_{i=1}^n (y_i - \bar{y})^2$		

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



# Análise de variância aplicada à regressão linear simples

- O teste da significância do modelo via ANOVA baseia-se em:
  - $\frac{(n-2)QM_{Res}}{\sigma^2} \sim \chi_{n-2}$ ;
  - Sob a hipótese nula (isso é, se  $\beta_1 = 0$ ), então  $\frac{SQ_{Reg}}{\sigma^2}$  tem distribuição  $\chi_1$ ;
  - $SQ_{Reg}$  e  $SQ_{Res}$  são independentes.
- Então:

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

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

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

# Análise de variância aplicada à regressão linear simples

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

$$R^2 = \frac{SQ_{Reg}}{SQ_{Total}}, \quad (49)$$

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

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

# Análise de variância aplicada à regressão linear simples

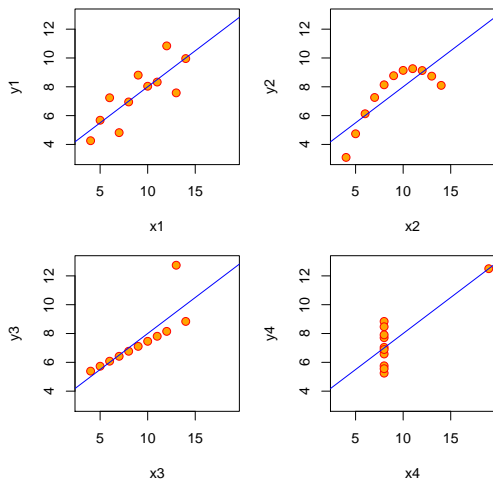


Figura 7: Quatro conjuntos de dados que produzem mesmo valor de  $R^2$

# Caso em que $x$ também é aleatório - análise de correlação

- Em algumas situações, pode não ser razoável admitir que a variável explicativa  $x$  seja fixa.
- Como exemplo, num experimento na agronomia em que está se estudando produção vegetal, pode ser pouco realista assumir a altura das plantas ou o número de folhas como não sendo aleatórios;
- Vamos estudar agora o caso em que  $x$  e  $y$  são variáveis aleatórias e o estudo da distribuição conjunta.

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

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

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

em que  $\mu_x$  e  $\sigma_x^2$  são a média e a variância de  $x$ ;  $\mu_y$  e  $\sigma_y^2$  são a média e a variância de  $y$  e

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

é o coeficiente de correlação entre  $x$  e  $y$ .

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

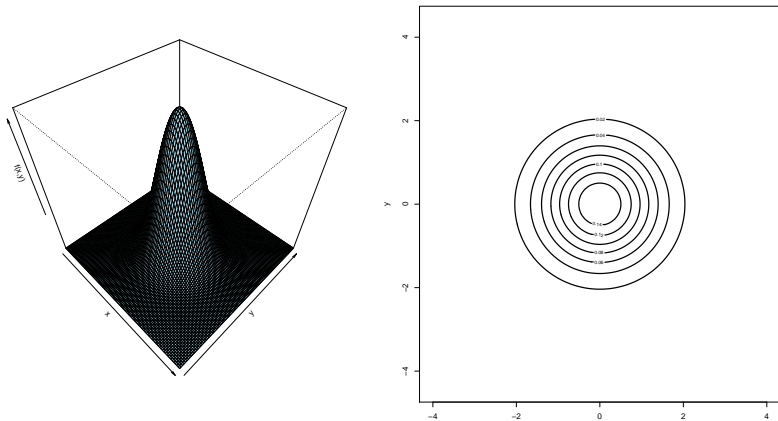
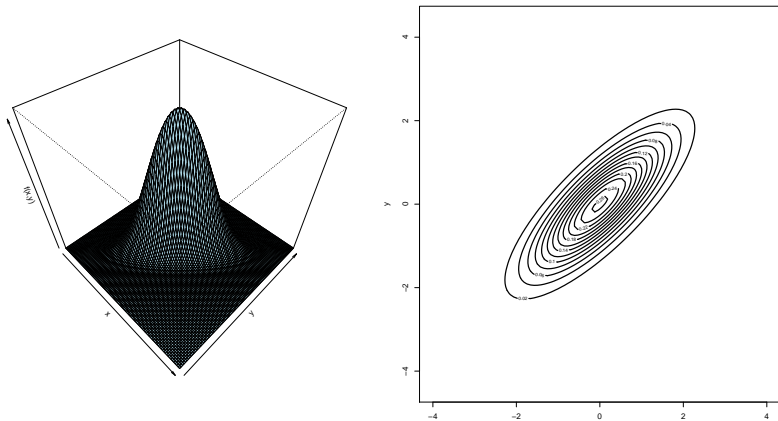


Figura 8: Distribuição normal bivariada:  $\rho^x = 0$ .

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



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

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

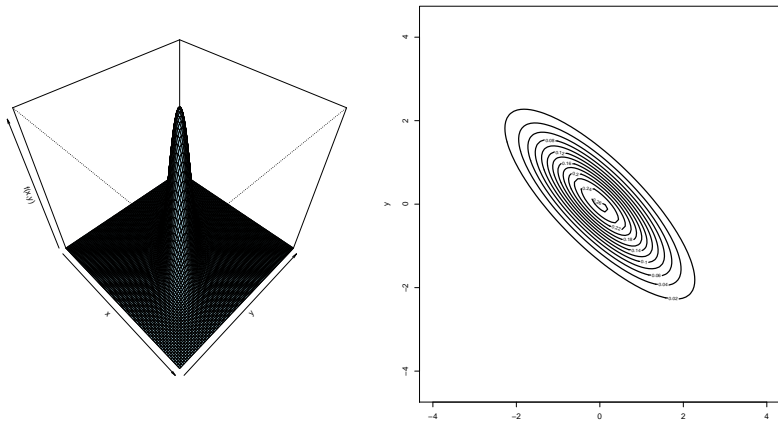


Figura 10: Distribuição normal bivariada:  $\rho = -0.8$ .



# Análise de correlação

- O estimador de  $\rho$  é o coeficiente de correlação amostral, dados por:

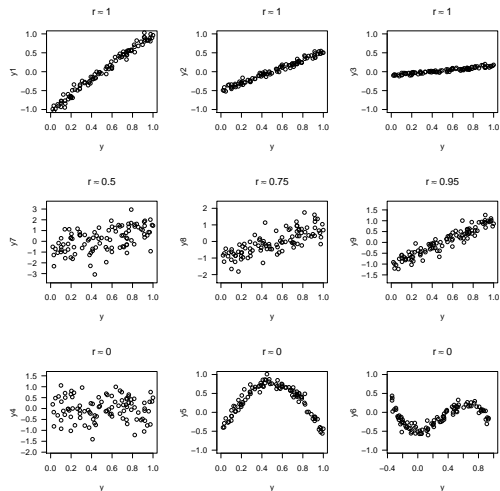
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2]^{1/2}}. \quad (52)$$

- Verifica-se facilmente que:

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

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

# Análise de correlação



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

# Análise de correlação

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

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$$

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

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

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

# Análise de correlação

- Assim, a hipótese de correlação nula deverá ser rejeitada, ao nível de significância de  $\alpha$ , se  $|t| > |t_{n-2;\alpha/2}|$ .
  
- O nível descritivo do teste pode ser calculado por  $p = 2 \times P(X > |t|)$ , sendo  $X \sim t_{n-2}$ .

# Análise de correlação

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

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

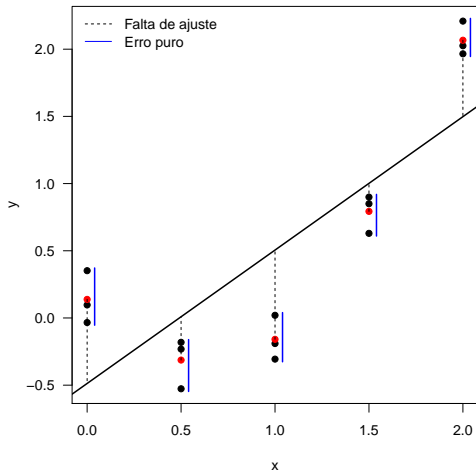
em que:

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

# Teste da falta de ajuste da regressão linear

- O teste da falta de ajuste permite testar formalmente a adequação do ajuste do modelo de regressão.
- Neste ponto assumimos 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 e o segundo ao erro puro.

# Teste da falta de ajuste da regressão linear



**Figura 12:** 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  $x$ .
  
- Dispondo de replicações de  $y$  em diferentes valores de  $x$ , temos condições de obter uma estimativa para a variância do erro ( $\sigma^2$ ) 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  $x_i$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, n_i$ ,  $n = \sum_{i=1}^m n_i$ . Então:

$$r_i = y_{ij} - \hat{y}_i = \underbrace{(y_{ij} - \bar{y}_i)}_{\text{Resíduo}} + \underbrace{(\bar{y}_i - \hat{y}_i)}_{\text{Erro puro}} + \underbrace{(\bar{y}_i - \hat{y}_i)}_{\text{Falta de ajuste}}, \quad (57)$$

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:

$$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \underbrace{\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{SQRes} + \underbrace{\sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2}_{SQFA}. \quad (58)$$

# Teste da falta de ajuste da regressão linear

- Assim, sob a suposição de variância constante SQEP 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 3:** 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	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$QM_{Reg} = \frac{SQ_{Reg}}{1}$	$F = \frac{QM_{Reg}}{QM_{Res}}$
Resíduos	n-2	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$QM_{Res} = \frac{SQ_{Res}}{n-2}$	
Falta de ajuste	m-2	$\sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2$	$QM_{FA} = \frac{SQ_{FA}}{m-2}$	$F = \frac{QM_{FA}}{QM_{EP}}$
Erro puro	n-m	$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$QM_{EP} = \frac{SQ_{EP}}{n-m}$	
Total	n-1	$\sum_{i=1}^n (y_i - \bar{y})^2$		

# Teste da falta de ajuste da regressão linear

- Se a função de regressão verdadeira de fato é linear, então tanto  $QM_{EP}$  quanto  $QM_{FA}$  são estimadores não viciados de  $\sigma^2$ .
- Caso a real função de regressão seja não linear, então  $E(QM_{FA}) > \sigma^2$ .
- Sob a hipótese nula de que não há falta de ajuste (a função de regressão verdadeira é linear), então:

$$F_0 = \frac{SQ_{FA}/(m-2)}{SQ_{EP}/(n-m)} = \frac{QM_{FA}}{QM_{EP}} \quad (59)$$

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 (a regressão de fato é linear) deverá ser rejeitada, ao nível de significância  $\alpha$ , 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).

# Modelos intrinsecamente lineares

- Em alguns casos em a relação entre as variáveis é não linear mas pode ser linearizada aplicando alguma transformação adequada.
- Os modelos de regressão resultantes são denominados *modelos intrinsecamente lineares*.
- Usar transformações pode remediar o não atendimento de diferentes pressupostos do modelo (como variância não constante ou ausência de normalidade).
- Neste ponto vamos nos ater à aplicação de transformações com o objetivo de linearizar a relação entre as variáveis.

**Tabela 4:** Exemplos de modelos intrinsecamente lineares

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

# Modelos intrinsecamente lineares

- Qualquer uma dessas transformações requer que os erros **na escala transformada** sejam independentes, normalmente distribuídos com média zero e variância  $\sigma^2$ .
  
- Quando o método de mínimos quadrados é aplicado após transformação as propriedades dos estimadores, que estudamos anteriormente, valem para os dados transformados e não necessariamente para os dados originais.