Regressão linear simples

Gilberto Pereira Sassi

Universidade Federal da Bahia
Instituto de Matemática e Estatística
Departamento de Estatística

Motivação: Regressão linear simples

Modelo determinístico

Modelo matemático não inclui incerteza ou aleatoriedade. Por exemplo, considere a posição inicial d_0 de corpo e a velocidade v, então a posição deste corpo no momento t é dado por

$$d_t = d_0 + v \cdot t.$$

Modelo estatístico

Modelo matemático inclui incerteza ou aleatoriedade. Por exemplo, considere y o consumo de energia elétrica de uma residência e x o tamanho em metros quadrados da casa. Casas maiores tendem a gastar mais energia elétrica, mas algumas podem ser econômicas e gastarem menos e outras podem gastar mais. Então, o consumo de energia elétrica pode ser descrita pelo modelo:

$$y = a + b \cdot x + \epsilon.$$

Uso pode ser inspirado através:

- Justificativa teórica;
- Diagrama de dispersão. Os pontos (x_i, y_i) estão próximos da reta $y = a + b \cdot x$.

Assumimos que $\epsilon \sim N(0, \sigma^2)$.

Regressão linear simples

Apresentação: regressão linear simples

Sejam X e Y duas variáveis quantitativas associadas com valores observados conforme ilustrado na Tabela 1.

Tabela 1: Observações de X e Y.

X	x_1	$\cdots$	x_n
Y	y_1	$\cdots$	y_n

Queremos encontrar valores a, chamado de intercepto, e b, chamado de inclinação, tal que

$$y_i = a + bx_i + \epsilon_i, \quad i = 1, \ldots, n,$$

e a e b é o valor que minimiza $S(a, b) = (y_1 - a - bx_1)^2 + \cdots + (y_n - a - bx_n)^2$.

Interpretação

Considere o método descrito pela equação $y = a + b \cdot x + \epsilon$.

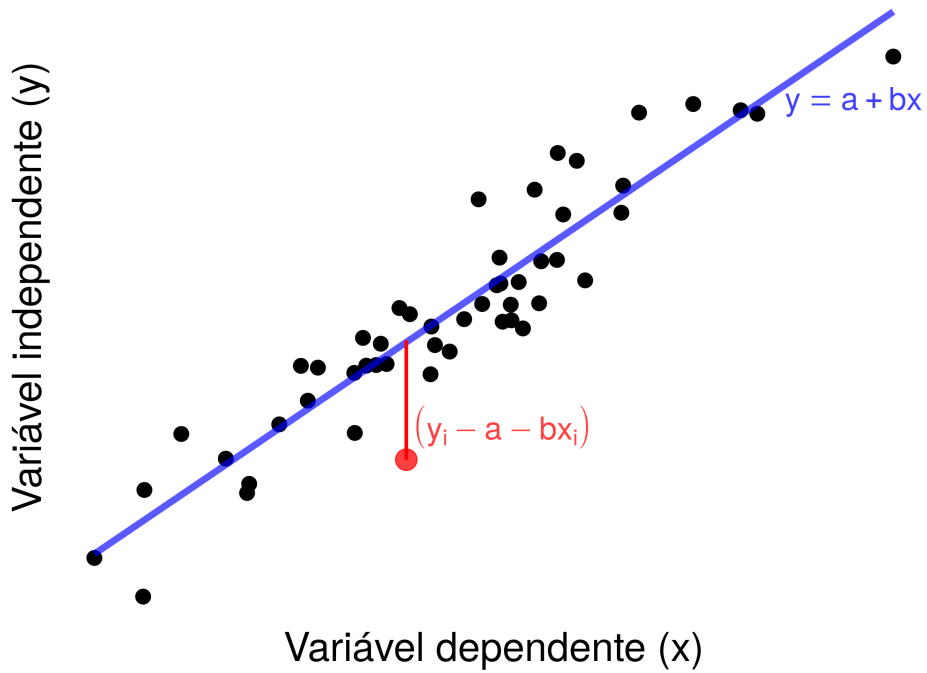
- a parte $y = a + b \cdot x$ representa os valores de y que podem ser explicados por x através da equação da reta $y = a + b \cdot x$;
- ϵ representa a parte dos valores de y que não podem ser explicados por por x através da equação da reta $y = a + b \cdot x$;
- chamamos x de variável independente, variável regressora ou variável explicativa;
- chamamos de y de variável dependente ou variável reposta.

Os valores ϵ idealmente são pequenos, com média zero e com distribuição normal e com variância σ^2. A Figura 1 ilustra a ideia da regressão linear simples.

Ilustração da regressão linear simples

Desejamos encontrar a e b de tal que forma que as distâncias $(y_i - a - bx_i)$ sejam as menores possíveis (em valores quadráticos).

Figura 1: Motivação de regressão linear simples.



Estimativas para a e b

Imagine que temos n observações da variável X e n observações da variável Y, conforme Tabela 1 e queremos encontrar a "melhor reta" como ilustrado na Figura 1. O modelo de regressão linear simples é dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

O critério para determinar a "melhor reta" é denominado de Método dos Mínimos Quadrados e consiste em encontrar $\hat{a}$ e $\hat{b}$ que minimiza a equação

$$S(a, b) = (y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdots + (y_n - a - bx_n)^2.$$

O ponto de mínimo $(\hat{a}, \hat{b})$ de $S(a, b)$ são soluções das seguintes equações:

$$\begin{aligned} \frac{\partial S(\hat{a}, \hat{b})}{\partial a} &= 0 \\ \frac{\partial S(\hat{a}, \hat{b})}{\partial b} &= 0 \end{aligned} \tag{1}$$

Regressão linear simples

As estimativas de Mínimos Quadrados para o intercepto (a) e a inclinação (b) são

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

$$\hat{b} = \frac{S_{xy} - n\bar{x}\bar{y}}{S_{x^2} - n\bar{x}^2}$$

em que

$$\begin{aligned} S_x &= x_1 + \cdots + x_n \\ S_{xy} &= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \\ S_{y^2} &= y_1^2 + y_2^2 + \cdots + y_n^2 \\ S_{(y-\bar{y})(x-\bar{x})} &= (y_1 - \bar{y})(x_1 - \bar{x}) + \cdots + (y_n - \bar{y})(x_n - \bar{x}) \end{aligned}$$

$$\begin{aligned} S_y &= y_1 + \cdots + y_n \\ S_{x^2} &= x_1^2 + x_2^2 + \cdots + x_n^2 \\ S_{(x-\bar{x})^2} &= (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \end{aligned}$$

Além disso,

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n}, \\ \bar{y} &= \frac{y_1 + y_2 + \cdots + y_n}{n}, \\ S_{(y-\bar{y})(x-\bar{x})} &= S_{xy} - n\bar{x}\bar{y} \\ S_{(x-\bar{x})^2} &= S_{x^2} - n\bar{x}^2. \end{aligned}$$

Note que a reta estimada é

$$\hat{y} = \hat{a} + \hat{b}x,$$

e para cada observação na amostra com x_i e y_i, temos que a relação

$$y_i = \hat{a} + \hat{b}x_i + e_i, \quad i = 1, \ldots, n,$$

em que $e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$. Chamamos $e_i, \ i = 1, \ldots, n$ de resíduos.

Propriedades da regressão linear simples

Esperamos que os resíduos são valores observados da variável erro $\epsilon \sim N(0, \sigma^2)$.
Uma aproximação de σ^2 pode ser aproximada

$$\hat{\sigma}^2 = \frac{SQ_E}{n-2}$$

em que

$$SQ_E = e_1^2 + \cdots + e_n^2 = (y_1 - \hat{y}_1)^2 + \cdots + (y_n - \hat{y}_n)^2 = S_{y^2} - n\bar{y}^2 - \hat{b}\,(S_{xy} - n\bar{x}\bar{y})\,.$$

SQ_E é chamado de Soma dos Quadrados dos Erros.

Note que $\hat{b} \sim N\left(b; \mathrm{Var}(\hat{b}) = \frac{\sigma^2}{S_{(x-\bar{x})^2}}\right)$ e $\hat{a} \sim N\left(a; \mathrm{Var}(\hat{a}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{s_{(x-\bar{x})^2}}\right]\right)$.

E

$$\frac{\hat{a} - a}{\sqrt{\widehat{\mathrm{Var}}(\hat{a})}} \sim t_{n-2} \qquad \frac{\hat{b} - b}{\sqrt{\widehat{\mathrm{Var}}(\hat{b})}} \sim t_{n-2}.$$

em que

$$\widehat{\mathrm{Var}}(\hat{a}) = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{s_{(x-\bar{x})^2}}\right] = \frac{S_{y^2} - n\bar{y}^2 - \hat{b}\,(S_{xy} - n\bar{x}\bar{y})}{n-2} \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{s_{(x-\bar{x})^2}}\right],$$

$$\widehat{\mathrm{Var}}(\hat{b}) = \frac{\hat{\sigma}^2}{S_{(x-\bar{x})^2}} = \frac{S_{y^2} - n\bar{y}^2 - \hat{b}\,(S_{xy} - n\bar{x}\bar{y})}{(n-2)\cdot(S_{x^2} - n\bar{x}^2)}.$$

Teste de hipóteses para a inclinação: b.

Considere n pares $(y_1, x_1), \ldots, (y_n, x_n)$, e o modelo dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Considere o nível de significância α.

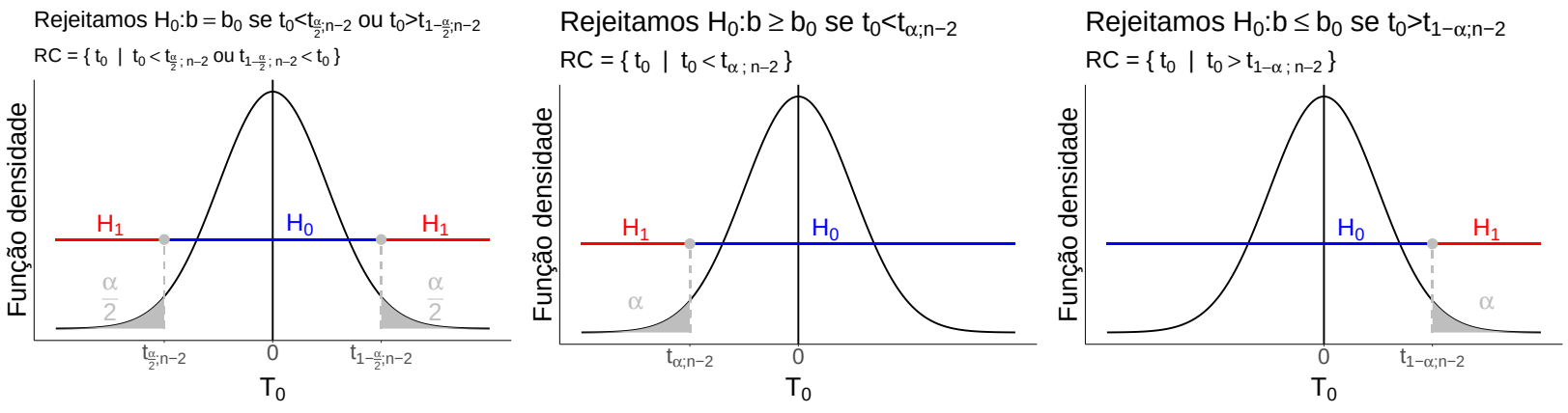
Queremos testar as seguintes hipóteses:

- **Teste bilateral:** $H_0 : b = b_0$ e $H_1 : b \neq b_0$;
- **Teste unilateral:** $H_0 : b \leq b_0$ e $H_1 : b > b_0$;
- **Teste unilateral:** $H_0 : b \geq b_0$ e $H_1 : b < b_0$.

Ideia: Primeiro calculamos a distância padronizada entre b e $\hat{b}$: $T_0 = \frac{b - \hat{b}}{\sqrt{\widehat{\text{Var}}(\hat{b})}}$. Então,

- **Teste bilateral:** Rejeitamos $H_0 : b = b_0$ se $|T_0|$ for grande;
- **Teste unilateral:** Rejeitamos $H_0 : b \leq b_0$ se T_0 for grande;
- **Teste bilateral:** Rejeitamos $H_0 : b \geq b_0$ se T_0 for pequeno.

Teste de hipóteses para a inclinação: *b*.



(a) Teste bilateral. **(b)** Teste bilateral. **(c)** Teste bilateral.

Figura 2: Regiões críticas para testes sobre a inclinação *b*.

Teste de hipóteses para a inclinação: b.

- Na Figura 2a, testamos $H_0 : b = b_0$ versus $H_1 : b \neq b_0$. Rejeitamos H_0 se $t_0 = \frac{b-\hat{b}}{\sqrt{\widehat{\text{Var}}(\hat{b})}} \in RC = \left\{ t_0 \mid t_0 < t_{\frac{\alpha}{2};n-2} \text{ ou } t_{1-\frac{\alpha}{2};n-2} < t_0 \right\}$, em que $P\left(t_{n-2} < t_{\frac{\alpha}{2},n-2}\right) = \frac{\alpha}{2}$ e $P\left(t_{n-2} < t_{1-\frac{\alpha}{2},n-2}\right) = 1 - \frac{\alpha}{2}$;

- Na Figura 2b, testamos $H_0 : b \geq b_0$ versus $H_1 : b < b_0$. Rejeitamos H_0 se $t_0 = \frac{b-\hat{b}}{\sqrt{\widehat{\text{Var}}(\hat{b})}} \in RC = \{t_0 \mid t_0 < t_{\alpha;n-2}\}$, em que $P\left(t_{n-2} < t_{\alpha,n-2}\right) = \alpha$;

- Na Figura 2c, testamos $H_0 : b \leq b_0$ versus $H_1 : b > b_0$. Rejeitamos H_0 se $t_0 = \frac{b-\hat{b}}{\sqrt{\widehat{\text{Var}}(\hat{b})}} \in RC = \{t_0 \mid t_{1-\alpha;n-2} < t_0\}$, em que $P\left(t_{n-2} < t_{1-\alpha,n-2}\right) = 1 - \alpha$;

Chamamos $t_{\alpha;n-2}$, $t_{1-\alpha;n-2}$, $t_{\frac{\alpha}{2};n-2}$ e $t_{1-\frac{\alpha}{2};n-2}$ de valores críticos.

Teste de hipóteses para o intercepto: a.

Considere n pares $(y_1, x_1), \ldots, (y_n, x_n)$, e o modelo dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Considere o nível de significância α.

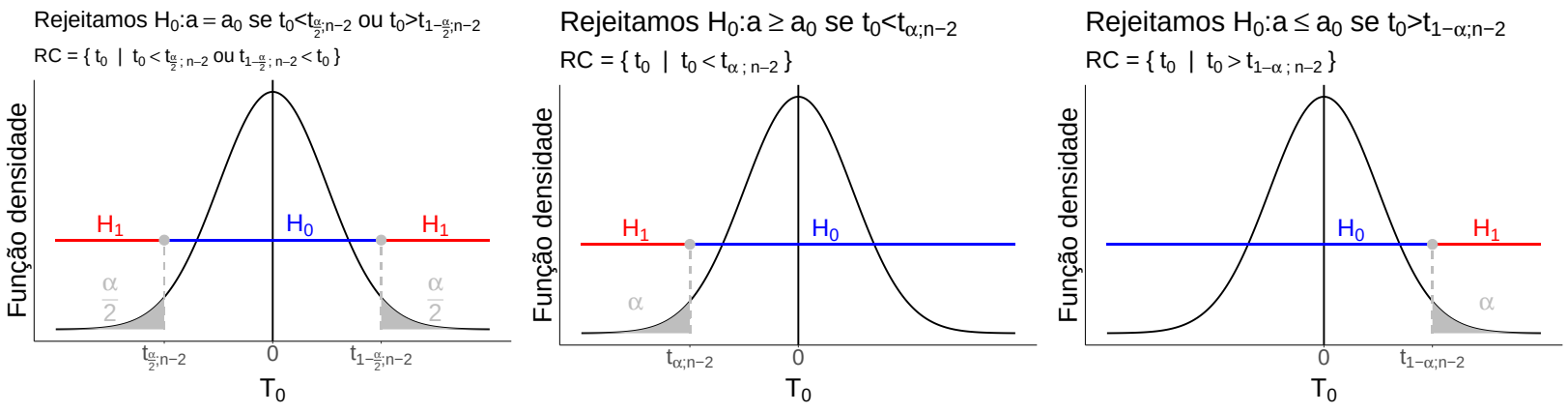
Queremos testar as seguintes hipóteses:

- **Teste bilateral:** $H_0 : a = a_0$ e $H_1 : a \neq a_0$;
- **Teste unilateral:** $H_0 : a \leq a_0$ e $H_1 : a > a_0$;
- **Teste unilateral:** $H_0 : a \geq a_0$ e $H_1 : a < a_0$.

Ideia: Primeiro calculamos a distância padronizada entre a e $\hat{a}$: $T_0 = \frac{a - \hat{a}}{\sqrt{\widehat{\text{Var}}(\hat{a})}}$. Então,

- **Teste bilateral:** Rejeitamos $H_0 : a = a_0$ se $|T_0|$ for grande;
- **Teste unilateral:** Rejeitamos $H_0 : a \leq a_0$ se T_0 for grande;
- **Teste bilateral:** Rejeitamos $H_0 : a \geq a_0$ se T_0 for pequeno.

Teste de hipóteses para o intercepto: *a*.



(a) Teste bilateral. **(b)** Teste bilateral. **(c)** Teste bilateral.

Figura 3: Regiões críticas para testes sobre a inclinação *b*.

Teste de hipóteses para o intercepto: a.

- Na Figura 3a, testamos $H_0 : a = a_0$ versus $H_1 : a \neq a_0$. Rejeitamos H_0 se $t_0 = \frac{a - \hat{a}}{\sqrt{\widehat{\text{Var}}(\hat{a})}} \in RC = \left\{ t_0 \mid t_0 < t_{\frac{\alpha}{2};n-2} \text{ ou } t_{1-\frac{\alpha}{2};n-2} < t_0 \right\}$, em que $P\left(t_{n-2} < t_{\frac{\alpha}{2},n-2}\right) = \frac{\alpha}{2}$ e $P\left(t_{n-2} < t_{1-\frac{\alpha}{2},n-2}\right) = 1 - \frac{\alpha}{2}$;

- Na Figura 3b, testamos $H_0 : a \geq a_0$ versus $H_1 : a < a_0$. Rejeitamos H_0 se $t_0 = \frac{a - \hat{a}}{\sqrt{\widehat{\text{Var}}(\hat{a})}} \in RC = \{ t_0 \mid t_0 < t_{\alpha;n-2} \}$, em que $P\left(t_{n-2} < t_{\alpha,n-2}\right) = \alpha$;

- Na Figura 3c, testamos $H_0 : a \leq a_0$ versus $H_1 : a > a_0$. Rejeitamos H_0 se $t_0 = \frac{a - \hat{a}}{\sqrt{\widehat{\text{Var}}(\hat{a})}} \in RC = \{ t_0 \mid t_{1-\alpha;n-2} < t_0 \}$, em que $P\left(t_{n-2} < t_{1-\alpha,n-2}\right) = 1 - \alpha$;

Chamamos $t_{\alpha;n-2}$, $t_{1-\alpha;n-2}$, $t_{\frac{\alpha}{2};n-2}$ e $t_{1-\frac{\alpha}{2};n-2}$ de valores críticos.

Intervalo de confiança para o intercepto e a inclinação.

Considere n pares $(y_1, x_1), \ldots, (y_n, x_n)$, e o modelo dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Intervalo de confiança para o intercepto: a.

Se o intercepto populacional é a, então $\frac{\hat{a}-a}{\sqrt{\widehat{\text{Var}(\hat{a})}}} \sim t_{n-2}$ e

$$\gamma = 1 - \alpha = P\left(t_{\frac{\alpha}{2};n-2} \leq \frac{\hat{a} - a}{\sqrt{\widehat{\text{Var}\,(\hat{a})}}} \leq t_{1-\frac{\alpha}{2}.n-2} \right),$$

e o intervalo de confiança para a com coeficiente de confiança $\gamma = 1 - \alpha$ é dado por

$$IC(a, \gamma) = \left(t_{\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{a})}} + \hat{a}; t_{1-\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{a})}} + \hat{a} \right).$$

Intervalo de confiança para a inclinação: b.

Se a inclinação populacional é b, então $\frac{\hat{b}-b}{\sqrt{\widehat{\text{Var}\left(\hat{b}\right)}}} \sim t_{n-2}$ e

$$\gamma = 1 - \alpha = P\left(t_{\frac{\alpha}{2};n-2} \leq \frac{\hat{b} - b}{\sqrt{\widehat{\text{Var}\left(\hat{b}\right)}}} \leq t_{1-\frac{\alpha}{2}.n-2} \right),$$

e o intervalo de confiança para b com coeficiente de confiança $\gamma = 1 - \alpha$ é dado por

$$IC(b, \gamma) = \left(t_{\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{b})}} + \hat{b}; t_{1-\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{b})}} + \hat{b} \right).$$

Intervalo de confiança para o intercepto e a inclinação.

Considere n pares $(y_1, x_1), \ldots, (y_n, x_n)$, e o modelo dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Considere $x_0 \notin \{x_1, \ldots, x_n\}$ e não conhecemos o valor de $Y_0 = a + b \cdot x_0 + \epsilon_0$.

Estimativa pontual para Y_0.

Uma estimativa pontual para Y_0 é $\hat{y}_0 = \hat{a} + \hat{b} \cdot x_0$.

Intervalo de confiança para Y_0.

Pode-se provar que

$$\mathrm{E}\left[Y_0 - \hat{Y}_0\right] = 0; \mathrm{Var}\left[Y_0 - \hat{Y}_0\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{(x-\bar{x})^2}}\right]; \frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{(x-\bar{x})^2}}\right]}} \sim t_{n-2},$$

em que $\hat{\sigma}^2 = \frac{SQE}{n-2} = \frac{S_{y^2} - n\bar{y}^2 - \hat{b}(S_{xy} - n\bar{x}\bar{y})}{n-2}$. Então,

$$\gamma = 1 - \alpha = P\left(t_{\frac{\alpha}{2};n-2} \leq \frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{(x-\bar{x})^2}}\right]}} \leq t_{1-\frac{\alpha}{2};n-2}\right),$$

e o intervalo de confiança para Y_0 com coeficiente de confiança $\gamma = 1 - \alpha$ é dado por

$$IC(Y_0, \gamma) = \left(t_{\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{(x-\bar{x})^2}}\right]} + \hat{Y}_0; t_{1-\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{(x-\bar{x})^2}}\right]} + \hat{Y}_0\right).$$

Análise de variância

Considere n pares $(y_1, x_1), \ldots, (y_n, x_n)$, e o modelo dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Seja $\hat{y}_i = \hat{a} + \hat{b}x_i$ e $e_i = y_i - \hat{y}_i$. Então, pode-se provar que

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SQ_T=(n-1)s_y^2} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SQ_R=\hat{b}S_{(x-\bar{x})(y-\bar{y})}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}_{SQ_E=(n-1)s_e^2},$$

em que

- $s_y^2 = \frac{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}{n-1}$;
- $s_e^2 = \frac{(e_1-\bar{e})^2+(e_2-\bar{e})^2+\cdots+(e_n-\bar{e})^2}{n-1}$, em que $\bar{e} = \frac{e_1+\cdots+e_n}{n}$.

Pode-se provar que $\mathrm{E}[SQ_R] = \sigma^2 + b^2 S_{(x-\bar{x})^2}$ e $\mathrm{E}[SQ_E] = (n-2)\sigma^2$.

Considere $QM_E = \frac{SQ_E}{n-2}$ e $QM_R = \frac{SQ_R}{1}$. Se $H_0 : b = 0$ é verdadeira, então $F_0 = \frac{QM_R}{QM_E} \sim F_{1,n-2}$.

De forma semelhante a ANOVA, chamamos: SQ_T de soma de quadrados totais; SQ_R de soma de quadrados de regressão; SQ_E de soma de quadrados dos erros; QM_R de quadrados médios de regressão; e QM_E de quadrados médios dos erros.

Análise de variância

Imagine que desejamos decidir entre duas hipóteses: $H_0 : b = 0$ e $H_1 : b \neq 0$. Rejeitamos H_0 se F_0 for grande, e a rejeição crítica é dada por $RC = \{f_0 \mid f_0 > f_{1-\alpha;1,n-2}\}$. Na Figura 4, ilustramos esta região crítica.

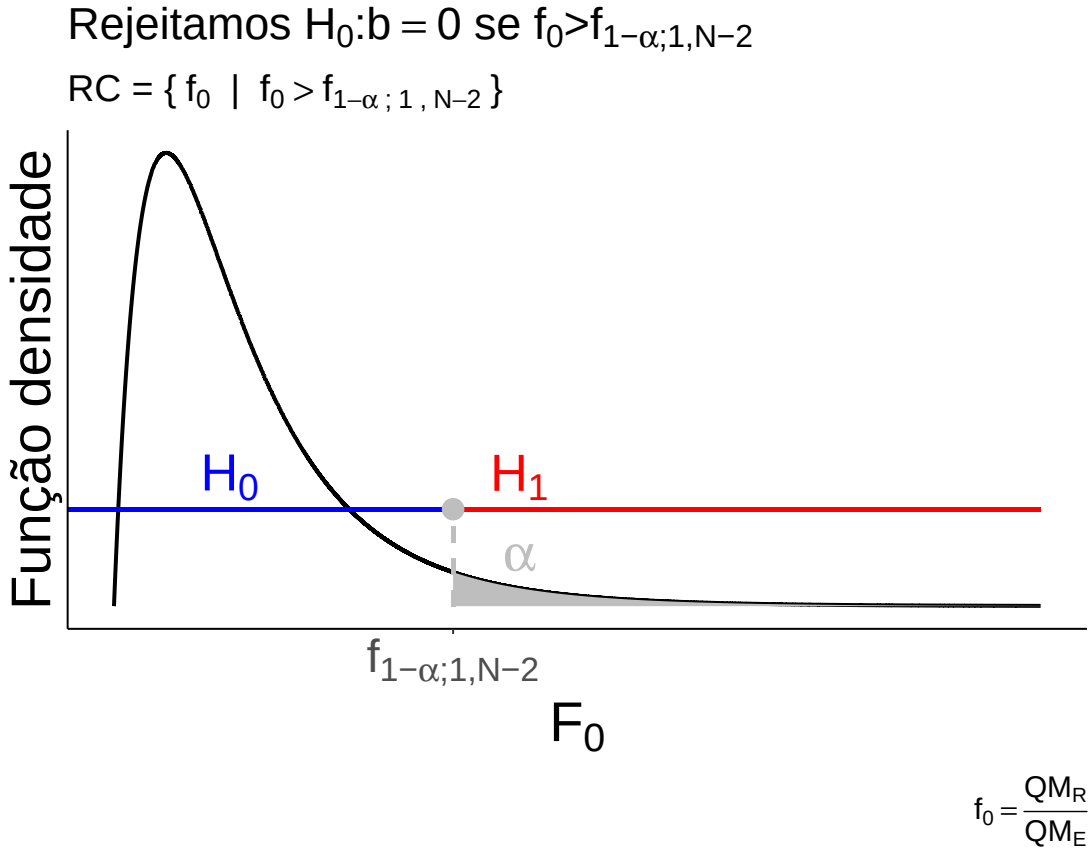


Figura 4: Região para análise de variância para regressão linear simples.

Checando as suposições do modelo matemática.

Análise de resíduos.

Considere n pares $(y_1, x_1), \ldots, (y_n, x_n)$, e o modelo dado por

$$y_i = a + b \cdot x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

em que $\epsilon_i \in N(0, \sigma^2)$, $i = 1, \ldots, n$. Chamamos $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, de resíduos. Precisamos verificar se as seguintes suposições estão satisfeitas:

(a) **Linearidade:** para cada par (x_i, e_i) desenhamos um ponto no plano cartesiano. Se não existe qualquer padrão e tendência, concluímos que a relação $y = a + b \cdot x$ captou toda influência de x sobre y;

(b) **Normalidade:** para cada par $\left(q_{(i)}; \frac{e_{(i)} - \bar{e}}{s_e}\right)$ desenhamos um ponto no plano cartesiano, em que $\Phi(q_{(i)}) = \frac{i-0,5}{n}$, $i = 1, \ldots, n$. Se os pontos estão sobre ou próximos da reta $y = x$, concluímos que as variáveis aleatórias ϵ_i, $i = 1, \ldots, n$, têm distribuição normal;

(c) **Independência:** para cada par (i, d_i) desenhamos um ponto no plano cartesiano, em que $d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}$ – chamamos d_i de resíduo padronizado. Se não existe padrão ou tendência, concluímos que as variáveis aleatórias ϵ_i são independentes;

(d) **Ponto exterior:** para cada par (i, d_i) desenhamos um ponto no plano cartesiano, em que $d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}$ – chamamos d_i de resíduo padronizado. Pontos abaixo de -3 ou acima de 3 são pontos exteriores;

(e) **Igualdade da variância (homoscedasticidade):** para cada par $(e_i, \hat{y}_i)$ desenhamos um ponto no plano cartesiano. Se não existe padrão ou tendência, concluímos que as variáveis aleatórias ϵ_i tem a mesma variância.

Exemplo

Um motor de foguete é produzido por uma liga de dois tipos de propelentes: um iniciador e um mantenedor. Imagina-se que a força da liga y é uma função linear da idade do propelente x quando o motor é lançado. A Tabela 2 fornece 20 observações. Estude a associação entre X e Y. Ajuste uma regressão linear simples e verifique se a regressão linear é significativa. Qual a força da liga para um propelente com 20 semanas (construa intervalo de confiança). Use $\alpha = 5\%$ e $\gamma = 95\%$.

Idade em semanas (x)	Força (y)
15,50	1823,01
23,75	1945,73
8,00	2446,80
17,00	2113,32
5,00	2512,34
19,00	1923,63
24,00	1676,79
2,50	2464,78
7,50	2463,42
11,00	2382,73
13,00	1999,45
3,75	2373,36
25,00	1754,28
9,75	2306,62
22,00	1735,37
18,00	1798,60
6,00	2545,93
12,50	2255,47
2,00	2459,01
21,50	1858,06

Tabela 2: Dados sobre propelentes de foguetes.

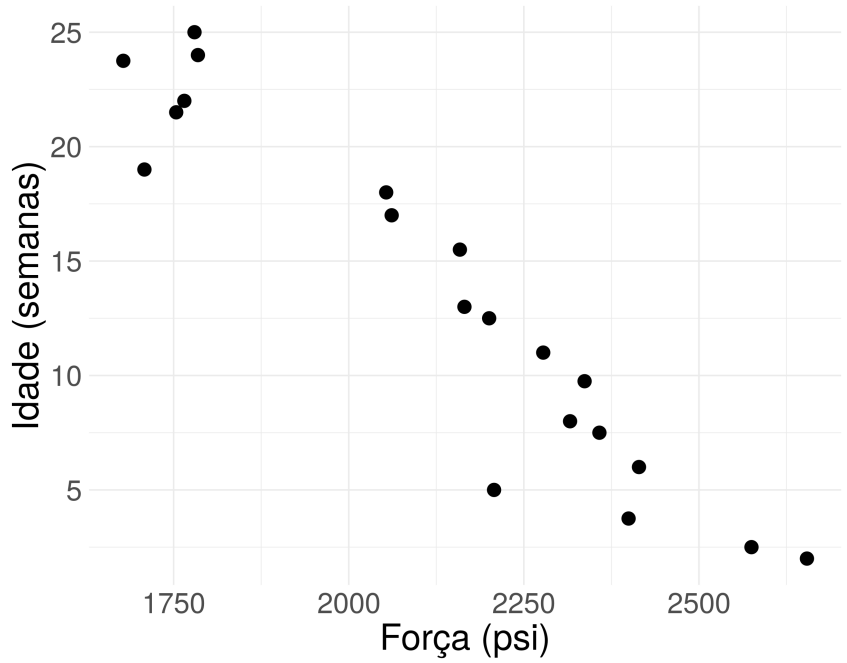
$S_x = 266,75$ | $S_y = 42838,7$
$S_{xy} = 530297,5$ | $S_{x^2} = 4672,438$
$S_{y^2} = 93550117$ | $S_{(x-\bar{x})^2} = 1114,659$
$S_{(y-\bar{y})(x-\bar{x})} = -41063,62$

Exemplo

Solução – diagrama de dispersão.

Na Figura 5, X (Idade em semanas) e Y (força) estão negativamente e fortemente associadas.

Figura 5: Gráfico de dispersão.



Exemplo

Solução – Ajuste da regressão linear

Nesse momento, queremos ajustar uma regressão linear simples, ou seja, queremos encontrar as estimativas $\hat{a}$ e $\hat{b}$ para o intercepto e para a inclinação da equação da reta.
Primeiro vamos calcular a inclinação:

$$\begin{aligned}\hat{b} &= \frac{S_{x\cdot y} - n\bar{x}\bar{y}}{S_{x^2} - n\bar{x}^2}, \\ &= \frac{530297,5 - 20 \cdot 2141,935 \cdot 13,3375}{4672,438 - 20 \cdot 13,3375^2}, \\ &= -36,84.\end{aligned}$$

Agora podemos calcular o intercepto:

$$\begin{aligned}\hat{a} &= \bar{y} - \hat{b}\bar{x} = 2141,935 + 36,84 \cdot 13,3375 \\ &= 2633,28.\end{aligned}$$

Finalmente, vamos estimar a variância do erro:

$$\begin{aligned}\hat{\sigma}^2 &= SQE = S_{y2} - n\bar{y}^2 - \hat{b}(S_{xy} - n\bar{x}\bar{y}) \\ &= 93550117 - 20 \cdot \left(\frac{42838,7}{20}\right)^2 - (-36,84) \cdot \left(530297,5 - 20 \cdot \frac{266,75}{20}\frac{42838,7}{20}\right) \\ &= 9811,212.\end{aligned}$$

Exemplo

Solução – teste de hipóteses

Vamos decidir entre as hipóteses: $H_0 : b = 0$ e $H_1 : b \neq 0$.

Passo 1) Queremos decidir entre as hipóteses: $H_0 : b = 0$ e $H_1 : b \neq 0$;

Passo 2) Nível de significância $\alpha = 5\%$;

Passo 3) Rejeitamos H_0 se $|T_0| = \left| \frac{b - \hat{b}}{\sqrt{\widehat{\text{Var}}(\hat{b})}} \right|$ for grande. Ou seja,

$RC = \left\{ t_0 \mid t_0 < t_{\frac{\alpha}{2};n-2} \text{ ou } t_0 < t_{1-\frac{\alpha}{2};n-2} \right\}$;

Passo 4) Vamos encontrar o valor crítico:

- $P\left(t_{n-2} \leq t_{\frac{\alpha}{2};n-2}\right) = P\left(t_{28} \leq t_{0,025;18}\right) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;18} = -2,101$;
- $P\left(t_{n-2} \leq t_{1-\frac{\alpha}{2};n-2}\right) = P\left(t_{28} \leq t_{0,975;18}\right) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;18} = 2,101$;

Passo 5) Note que $\hat{b} = -36,84$, $\widehat{\text{Var}}(\hat{b}) = \frac{\hat{\sigma}^2}{S_{(x-\bar{x})^2}} = \frac{9811,212}{1114,659} = 8,80$ e

$t_0 = \frac{-36,84-0}{\sqrt{8,80}} = -12,42 \in RC$, e rejeitamos H_0.

Ao nível de significância $\alpha = 5\%$, a inclinação populacional na regressão linear simples b é diferente de zero.

Exemplo

Solução – teste de hipóteses

Vamos decidir entre as hipóteses: $H_0 : a = 0$ e $H_1 : a \neq 0$.
Passo 1) Queremos decidir entre as hipóteses: $H_0 : a = 0$ e $H_1 : a \neq 0$;
Passo 2) Nível de significância $\alpha = 5\%$;
Passo 3) Rejeitamos H_0 se $|T_0| = \left| \frac{a - \hat{a}}{\sqrt{\widehat{\text{Var}}(\hat{a})}} \right|$ for grande. Ou seja,

$$RC = \left\{ t_0 \mid t_0 < t_{\frac{\alpha}{2};n-2} \text{ ou } t_0 < t_{1-\frac{\alpha}{2};n-2} \right\};$$

Passo 4) Vamos encontrar o valor crítico:

- $P\left(t_{n-2} \leq t_{\frac{\alpha}{2};n-2}\right) = P\left(t_{28} \leq t_{0,025;18}\right) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;18} = -2,101$;
- $P\left(t_{n-2} \leq t_{1-\frac{\alpha}{2};n-2}\right) = P\left(t_{28} \leq t_{0,975;18}\right) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;18} = 2,101$;

Passo 5) Note que $\hat{a} = 2633,28$, $\bar{x} = \frac{S_x}{n} = \frac{266,75}{20} = 13,34$,

$\widehat{\text{Var}}(\hat{a}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{(x-\bar{x})^2}} \right] = 9811,212 \left[\frac{1}{20} + \frac{13,34^2}{11114,659} \right] = 2056,923$ e

$t_0 = \frac{a - \hat{a}}{\sqrt{\widehat{\text{Var}}(\hat{a})}} = \frac{2633,28 - 0}{\sqrt{2056,923}} = 58,06 \in RC$, e rejeitamos H_0.

Ao nível de significância $\alpha = 5\%$, o intercepto populacional na regressão linear simples a é diferente de zero.

Exemplo

Intervalo de confiança para o intercepto: a.

Note que $\hat{a} = 2633,28$ e $\widehat{\text{Var}(\hat{a})} = 2056,923$. Vamos usar o coeficiente de confiança $\gamma = 1 - \alpha = 95\%$ e $\alpha = 5\%$.
Primeiro vamos encontrar os quantis da distribuição t-Student:

- $P(t_{n-2} \leq t_{\frac{\alpha}{2};n-2}) = P(t_{18} \leq t_{0,025;18}) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;18} = -2,101$;
- $P(t_{n-2} \leq t_{1-\frac{\alpha}{2};n-2}) = P(t_{18} \leq t_{0,975;18}) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;18} = 2,101$.

Então o intervalo de confiança para o intercepto com coeficiente de confiança $\gamma = 0,95$ é dado por

$$\begin{aligned} IC\,(a;95\%) &= \left(t_{\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{a})}} + \hat{a}; t_{1-\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{a})}} + \hat{a} \right) \\ &= \left(-2,101 \cdot \sqrt{2056,923} + 2633,28; 2,101 \cdot \sqrt{2056,923} + 2633,28 \right) \\ &= (2537,993; 2728,567) \end{aligned}$$

Exemplo

Intervalo de confiança para a inclinação: b.

Note que $\hat{b} = -36,84$ e $\widehat{\text{Var}(\hat{b})} = 8,80$. Vamos usar o coeficiente de confiança $\gamma = 1 - \alpha = 95\%$ e $\alpha = 5\%$.
Primeiro vamos encontrar os quantis da distribuição t-Student:

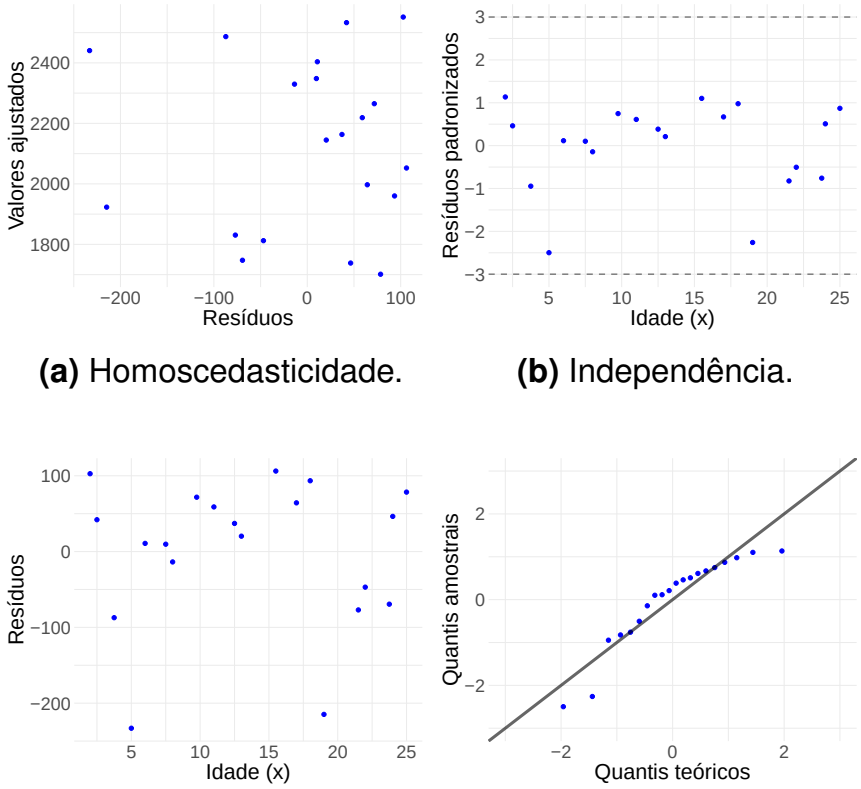
- $P(t_{n-2} \leq t_{\frac{\alpha}{2};n-2}) = P(t_{18} \leq t_{0,025;18}) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;18} = -2,101$;
- $P(t_{n-2} \leq t_{1-\frac{\alpha}{2};n-2}) = P(t_{18} \leq t_{0,975;18}) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;18} = 2,101$.

Então o intervalo de confiança para a inclinação com coeficiente de confiança $\gamma = 0,95$ é dado por

$$\begin{aligned} IC\,(b; 95\%) &= \left(t_{\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{b})}} + \hat{b}; t_{1-\frac{\alpha}{2};n-2}\sqrt{\widehat{\text{Var}(\hat{b})}} + \hat{b} \right) \\ &= \left(-2,101 \cdot \sqrt{8,80} - 36,84; 2,101 \cdot \sqrt{8,80} - 36,84 \right) \\ &= (-43,07; -30,61) \end{aligned}$$

Exemplo

Solução – análise de resíduos



(a) Homoscedasticidade.

(b) Independência.

(c) Linearidade.

(d) Gráfico normal de probabilidade.

Figura 6: Análise de resíduos.

Exemplo

Solução – análise de resíduos

(a) Na Figura 6c, não encontramos padrão e tendência. Então, a regressão linear simples, $a + b \cdot x$, é adequada para este conjunto de dados;

(b) Na Figura 6b, não encontramos padrão e tendência, e nenhum ponto está abaixo de -3 ou acima de 3. Então, as variáveis aleatórias ϵ_i, $i = 1 \ldots, n$, são independentes e não temos pontos exteriores;

(c) Na Figura 6a, não encontramos padrão ou tendência. Então, as variáveis aleatórias ϵ_i, $i = 1, \ldots, n$, tem a mesma variância;

(d) Na Figura 6d, os pontos estão perto da reta $y = x$. Então, as variáveis aleatórias ϵ_i, $i = 1, \ldots, n$, têm distribuição normal.

Exemplo

Solução – ANOVA.

Passo 1) Queremos decidir entre as hipóteses: $H_0; b = 0$ e $H_1; b \neq 0$;
Passo 2) Nível de significância $\alpha = 5\%$;
Passo 3) Rejeitamos H_0 se $F_0 = \frac{QM_R}{QM_E}$. Ou seja, $RC = \{f_0 \mid f_0 > f_{1-\alpha;1,n-2}\}$;
Passo 4) Vamos encontrar o valor crítico:

- $P\left(F_{1,n-2} \leq F_{1-\alpha;1,n-2}\right) = P\left(F_{1,18} \leq f_{0,95;1,18}\right) = 1 - \alpha = 0,95$, então $f_{0,95;1,18} = 4,4139$;

Passo 5) Vamos fazer uma Análise de Variância na Tabela 3.

Fator de variação	Graus de liberdade	Soma dos quadrados	Quadrados médios	F_0
Idade	1	$SQ_R = \hat{b}S_{(x-\bar{x})(y-\bar{y})} = 1522819,11$	$QM_R = \frac{SQ_R}{1} = 1522819,11$	$\frac{QM_R}{QM_E} = 155,21$
Erro	18	$SQ_E = 176601,82$	$QM_E = \frac{SQ_E}{n-2} = 9811,21$	
Total	19	$SQ_T = (n-1)s^2 = 1699420,93$		

Tabela 3: Tabela ANOVA para regressão linear simples.

Como $f_0 = 155,21 > f_{0,95;1,18} = 4,4139$, rejeitamos H_0.

Exemplo

Solução – predição para $Y_0 = a + b \cdot x_0 + \epsilon_0$.

- **Estimativa pontual** para Y_0: $\hat{y}_0 = \hat{a} + \hat{b}x_0 = 2633,28 - 36,84 \cdot 20 = 1896,48$;
- **Intervalo de confiança** para Y_0. Primeiro calculamos os quantis:
  - $P(t_{n-2} = t_{\frac{\alpha}{2};n-2}) = P(t_{18} = t_{0,025;18}) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;18} = -2,101$;
  - $P(t_{n-2} = t_{\frac{\alpha}{2};n-2}) = P(t_{18} = t_{0,975;18}) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;18} = 2,101$;

  Note que $\hat{\sigma} = 9811,21$, $x_0 = 20$, $\bar{x} = 13,34$ e $S_{(x-\bar{x})} = 1114,659$. Então o intervalo de confiança com coeficiente de confiança $\gamma = 95\%$ é dado por:

$$IC(Y_0, \gamma) = \left(t_{\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(x-\bar{x})^2}}\right]} + \hat{y}_0; t_{1-\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{(x-\bar{x})^2}}\right]} + \hat{y}_0 \right)$$

$$= \left(-2,101 \cdot \sqrt{9811,212\left[1 + \frac{1}{20} + \frac{(20 - 13,24)^2}{1114,659}\right]} + 1896,48; \right.$$

$$\left. 2,101 \cdot \sqrt{9811,212\left[1 + \frac{1}{20} + \frac{(20 - 13,24)^2}{1114,659}\right]} + 1896,48 \right)$$

$$= (1679,23; 2113,73)$$

A força da liga para um proponente com 20 semanas está entre $1679,23$ e $2113,73$ com coeficiente de confiança $\gamma = 95\%$.