Regressão linear simples

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

Destaques

Aplicações das áreas de exatas

Aplicações das áreas biológicas

Introdução

- Foi visto na aula anterior que o coeficiente de correlação de Pearson é utilizado para mensurar o grau de associação entre duas variáveis quantitativas.
- E se o nosso interesse for ir além disso, ou seja, se estivermos interessados em saber o quanto o aumento no número de horas de treinamento de um empregado irá reduzir o número de acidentes daquele empregado?
- Ou ainda, o aumento em uma hora de sono irá aumentar em quanto o tempo de reação de uma pessoa?
- E se estivéssimos interessados em prever o tempo de reação de uma pessoa para uma determinada quantidade de horas de sono. Como fazer?
- Para responder a estas perguntas utilizaremos a análise de regressão linear simples.

Objetivo

- Estudar a relação funcional entre duas variáveis quantitativas.
- Estabelecer um modelo para entender a relação funcional entre as variáveis.
- Fazer predições como o modelo ajustado principalmente para valores que não foram observados na amostra.

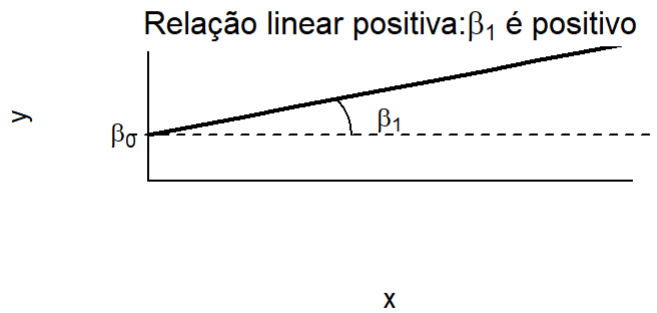
O modelo

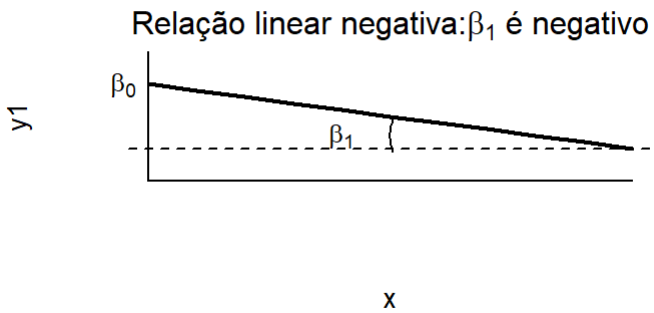
- O modelo matemático que estabele a relação funcional entre duas variáveis é definido como:

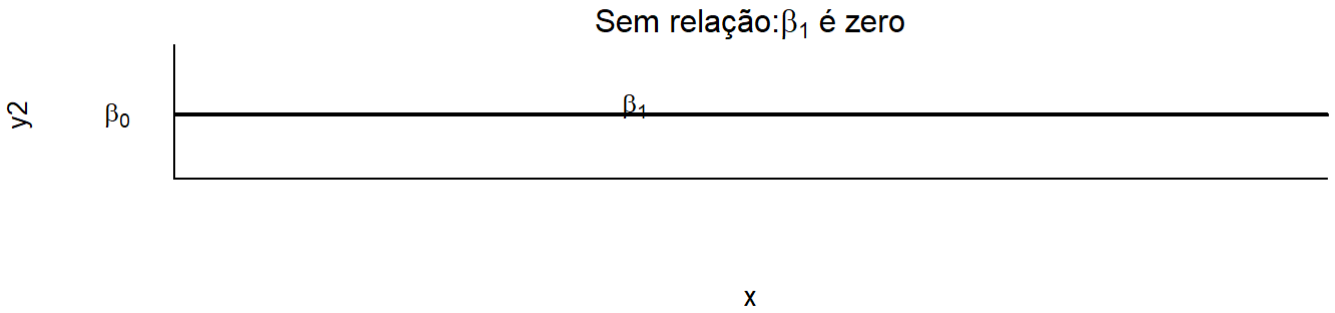
$$y = \beta_0 + \beta_1 x + \varepsilon$$

em que:
y= é a variável dependente
β_0= é o coeficiente linear ou intercepto da reta de regressão
β_1= é o coeficiente angular ou inclinação (declive) da reta de regressão
x= é a variável independente
ε= é o erro aleatório referente a variabilidade em y quem não pode ser explicada pela variável x.

Possíveis retas de regressão linear simples







Entendendo o comportamento de β_0 e β_1

Ver simulação!

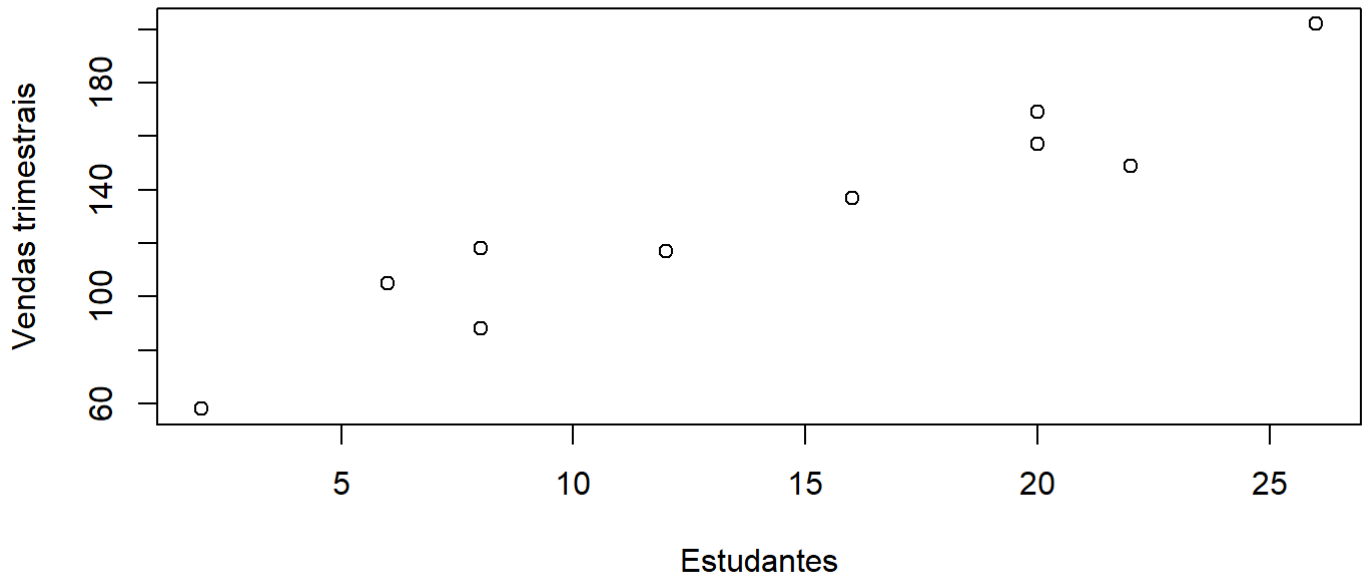
Ajuste da regressão

- Agora que já entendemos como funciona os parâmetros de uma regressão, chegou a hora de ajustarmos um modelo de regressão aos dados provenientes de uma amostra.
- Observemos a seguinte situação
  - Suponhamos que o dono de uma rede de restaurantes esteja interessado em saber a relação entre a quantidade de estudantes que almoçam em seus restaurantes com o lucro obtido trimestralmente.
  - Uma amostra de dez restaurantes foi coletado e os dados podem ser visualizados a seguir:

restaurante	estudantes	vendas_trimestrais
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

- O primeiro passo é verificarmos qual variável é *causa* e qual é *efeito*, ou seja, quem é a variável independente (x) e quem é a variável dependente (y).
- No nosso exemplo, verificamos que as vendas estão em função da quantidade de estudantes, ou seja, quem determina a venda é a quantidade de estudantes que adentram o restaurante.

- Logo, a variável *estudante* será considerada independente e *vendas* a variável dependente.
- O segundo passo, é elaborarmos um diagrama de dispersão para detectarmos o tipo de relação existente entre as variáveis.



- Agora vem a grande pergunta.
- Como podemos ajustar uma regressão que explique o máximo de variabilidade possível dos dados e com um mínimo de erro?

Método dos mínimos quadrados

- A idéia é encontrar valores de **b** e **a** que faça com que a reta de regressão passe na menor distância possível entre os pontos observados, minimizando o máximo possível o erro e fazendo com que o modelo explique o máximo possível a variabilidade dos dados.
- As letras **b** e **a** são os estimadores dos parâmetros β_0 e β_1 respectivamente.
- Primeiramente vamos demonstrar o método de maneira interativa, ou seja, vamos tentar encontrar valores de **b** e **a** que minimiza a soma de quadrados do erro.

Ver simulação!

- Matematicamente o que queremos é minimizar a seguinte quantidade:

$$\sum \varepsilon^2 = \sum (y - \beta_0 - \beta_1 x)^2$$

- Qual é o mínimo que desejamos para $\sum \varepsilon^2$?
  - Obviamente que a resposta é **zero**.
- Então basta substituirmos $\sum \varepsilon^2$ por zero, e derivarmos a expressão em relação a β_0 e β_1.

- Portanto, tem-se as seguintes equações para determinar os valores de **b** e **a**:

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

e

$$a = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Aplicação

1. Os dados a seguir foram coletados para determinar a relação entre a pressão e a escala de leitura para uma determinada proposta de calibração.

pressao	leitura
10	13
10	18
10	16
10	15
10	20
50	86
50	90
50	88
50	88
50	92

a. Encontre a equação de regressão.

Por partes temos:

$$\sum_{i=1}^{10} x_i = 300 \quad \sum_{i=1}^{10} x_i^2 = 1300$$

$$\sum_{i=1}^{10} x_i \sum_{j=1}^{10} y_j = 157800 \quad \sum_{i=1}^{10}\sum_{j=1}^{10} x_i y_j = 23020$$

$$\bar{y} = 52.6 \quad \bar{x} = 30$$

$$a = \frac{23020 - \frac{157800}{10}}{13000 - \frac{300^2}{10}} = 1.81$$

$$b = 52.6 - 1.81 * 30 = -1.7$$

Logo, a equação de regressão ajustada é:
$\hat{y} = -1.7 + 1.81 \cdot x$

1. Jackson et al. (1996) developed a novel specific assay for measuring bone alkaline phosphatase activity, an enzyme which reflects bone metabolism. They were interested to know whether this measure, the wheatgerm lectin precipitated bone alkaline phosphatase activity (wBAP), was correlated with an independent marker of bone formation, the carboxy-terminal propeptide of Type I collagen (PICP). The data of 46 adult horses are disponible here.

a. Find the equation of regression.

Interpretando a equação de regressão

- Muitos pesquisadores utilizam a regressão linear simples, mas não sabem o que fazer com ela.
- Com a nossa equação ajustada, podemos tirar as seguintes conclusões:
  - O aumento em uma unidade na variável x, provoca o aumento em 1.81 unidades a variável y. O que isso quer dizer na prática?
    - O aumento em uma libra/sq da pressão, aumenta em **média** em 1.81 a escala de leitura.
- Uma outra utilidade da equação de regressão, é podermos estimar **y** para pontos em **x** que não foram observados.

Aplicação

2. A quantidade de um composto químico y que dissolveu em 100 gramas de água em várias temperaturas foram as seguintes:

temp	composto	temp	composto
0	8	45	31
0	6	45	33
0	8	45	28
15	12	60	44
15	10	60	39
15	14	60	42
30	25	75	48
30	21	75	51
30	24	75	44

a. Encontre a equação de regressão e estime a quantidade de composto químico dissolvido em 100 gramas de água quando a temperatura for de 20°C.

   A equação de regressão estimada foi: $\hat{y} = 5.825 + 0.5676x$. Logo, a quantidade de composto químico estimada para uma temperatura de 20°C é: $\hat{y} = 5.825 + 0.5676 * 20 = 17.18g$.

2. A study was conducted to investigate the relationship between the sheep's live weight (LW, kg) and its chest girth (CG, cm). The data of 66 sheep are disponible here.

a. Find the equation of regression and estimate the live weight when the chest girth is 66 cm.

Coeficiente de determinação (r^2)

- E se perguntássemos qual foi a qualidade do ajuste pelo modelo de regressão?
- É possível determinarmos o quanto a equação ajustada explica a variabilidade dos dados?
- Sim! O coeficiente de determinação representado pela letra r^2 **minúsculo** responde as perguntas anteriores.
- Em alguns livros ainda terá a notação R^2, mas já está ultrapassada, sendo este último utilizado para outros tipos de regressão.
- Logo, o coeficiente de determinação é calculado da seguinte forma:

$$r^2 = 1 - \frac{SQerro}{SQtotal}$$

- As quantidades $SQerro$ e $SQtotal$ são a *soma de quadrados do erro* e *soma de quadrados total* respectivamente.
- A $SQerro$ é calculada como:

$$SQerro = \sum (y - \hat{y})^2$$

ou seja, é a soma dos desvios ao quadrado do que foi observado na amostra menos o que foi estimado pela equação de regressão.

- A $SQtotal$ é calculada como:

$$SQtotal = \sum y^2 - \frac{(\sum y)^2}{n}$$

Aplicação

3. Considerando a aplicação 2, calcule o coeficiente de determinação.

   Vamos calcular o $\hat{y}$ para os mesmos valores da variável independente com a equação de regressão estimada. Tem-se:

$$\hat{y}_0 = 5.825 + 0.5676 * 0 = 5.825$$
$$\hat{y}_{15} = 5.825 + 0.5676 * 15 = 14.339$$
$$\hat{y}_{30} = 5.825 + 0.5676 * 30 = 22.853$$
$$\hat{y}_{45} = 5.825 + 0.5676 * 45 = 31.367$$
$$\hat{y}_{60} = 5.825 + 0.5676 * 60 = 39.881$$
$$\hat{y}_{75} = 5.825 + 0.5676 * 75 = 48.395$$

   Logo, a soma de quadrados do erro (SQerro) é:

	y	$\hat{y}$	$y-\hat{y}$	$y-\bar{y}$	$(y-\hat{y})^2$	$(y-\bar{y})^2$
1	8	5.825	2.175	-19.111	4.731	365.235
2	6	5.825	0.175	-21.111	0.031	445.679
3	8	5.825	2.175	-19.111	4.731	365.235
4	12	14.339	-2.339	-15.111	5.471	228.346
5	10	14.339	-4.339	-17.111	18.827	292.79
6	14	14.339	-0.339	-13.111	0.115	171.901
7	25	22.853	2.147	-2.111	4.61	4.457
8	21	22.853	-1.853	-6.111	3.434	37.346
9	24	22.853	1.147	-3.111	1.316	9.679
10	31	31.367	-0.367	3.889	0.135	15.123
11	33	31.367	1.633	5.889	2.667	34.679
12	28	31.367	-3.367	0.889	11.337	0.79
13	44	39.881	4.119	16.889	16.966	285.235
14	39	39.881	-0.881	11.889	0.776	141.346
15	42	39.881	2.119	14.889	4.49	221.679
16	48	48.395	-0.395	20.889	0.156	436.346
17	51	48.395	2.605	23.889	6.786	570.679
18	44	48.395	-4.395	16.889	19.316	285.235
SQerro					105.895	
SQtotal						3911.78

Portanto, o coeficiente de determinação é:

$$r^2 = 1 - \frac{SQerro}{SQtotal} = 1 - \frac{105.895}{3911.78} = 0.9729$$

Logo, podemos afirmar que a equação estimada explica em 97,29% a variabilidade dos dados.

3. Considering the application 2, calculate the coefficient of determination.

Inferência para β_1

Pressupostos do modelo

- Para realizarmos inferência a cerca dos parâmetros da regressão, algumas suposições se faz necessário sobre o termo ε do modelo de regressão.
- Estas suposições são:
  - Os erros devem ser normalmente distribuídos com média 0 e variância igual para todo x;
  - Os erros são independentes, ou seja, o valor de ε para um valor particular de x, não está relacionado ao valor de ε para qualquer outro valor de x;
- Veremos adiante como checar estes pressupostos.

Teste de hipótese

- Iremos abordar a parte inferencial apenas para β_1, pois é ele que determina a relação funcional entre as variáveis x e y.
- Uma vez que a equação de regressão foi ajustada a partir de uma amostra, poderíamos nos perguntar se de fato a relação entre x e y não ocorreu por mero acaso.
- Repare que se β_1 for igual a zero, o valor médio de y não dependente de x e, portanto, concluiríamos que x e y não estão linearmente relacionados.
- Logo, o seguinte teste de hipótese se faz necessário:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

- Precisamos agora estimar a variância do erro.

$$s^2_{erro} = \frac{SQerro}{n-2}$$

- E portanto, o erro padrão residual (do erro) é $s_{erro} = \sqrt{s^2_{erro}}$
- O desvio padrão de a é calculado como:

$$s_a = \frac{s_{erro}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

- Logo, tem-se a seguinte estatística de teste:

$$t = \frac{a}{s_a}$$

em que t tem distribuição t de student com n-2 graus de liberdade.

Aplicação

4. Considerando ainda a aplicação 2, teste a hipótese de que $\beta_1 \neq 0$ considerando um $\alpha = 0.01$.

   Segue os cálculos.

$$s^2_{erro} = \frac{SQerro}{n-2} = \frac{105.895}{18-2} = 6.6184$$

$$s_{erro} = \sqrt{s^2_{erro}} = \sqrt{6.6184} = 2.5726$$

$$s_b = \frac{s_{erro}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} = \frac{2.5726}{\sqrt{37125 - \frac{455625}{18}}} = 0.02367$$

$$t = \frac{0.5676}{0.02367} = 23.98 \quad p - valor = 0.0000$$

   Logo, como o p-valor é menor que α podemos afirmar com 99% de confiabilidade que o $\beta 1$ é diferente de zero.

4. Considering yet the application 2, test the hypothesis of $\beta_1 \neq 0$ with $\alpha = 0.01$.

Análise de resíduos

- Na verdade, após ajustarmos um modelo de regressão o primeiro passo é analisarmos se o modelo é válido ou não, ou seja, se os pressupostos citados anteriormente estão sendo satisfeitos.
- Isto é necessário para que as inferências feitas a cerca dos parâmetros sejam válidas.
- Primeiro vamos avaliar se os resíduos são normalmente distribuídos. Iremos utilizar o gráfico quantil-quantil que é mais fácil de verificarmos quando uma reta se ajusta aos pontos do que se uma curva se ajusta a um histograma, principalmente com pequenas amostras.

- Para elaborarmos o gráfico devemos seguir os seguintes passos:
  - Ordenar a variável
  - Determinar as probabilidades teóricas de acordo com a seguinte expressão:

$$\frac{i - 0.5}{n}$$

em que i é a posição da variável e

n é o tamanho da amostra

  - Determinar os quantis teóricos. No nosso caso iremos usar a função `qnorm`.
  - Fazer um gráfico de dispersão do par ordenado (variável ordenada,quantil teórico).

- ...
  - Fazer uma reta de referência para compararmos e julgarmos se a variável tem distribuição normal ou não.
    - Já sabemos que uma reta é composta por uma função do primeiro grau e portanto precisamos encontrar a e b.
    - Como é uma reta empírica, o coeficiente a (inclinação da reta) é determinado como:

$$a = \frac{3^{o}quartil_{dados} - 1^{o}quartil_{dados}}{3^{o}quartil_{teórico} - 1^{o}quartil_{teórico}}$$

$$b = 1^{o}quartil_{dados} - a * (1^{o}quartil_{teórico})$$

    - Então a linha é elaborada como: $y = b + a \cdot \text{quantil teórico}$

- Para avaliar se a variância é constante, ou seja, a mesma para cada x usaremos um gráfico no qual no eixo y será colocado os erros e no eixo x a variável independente (estudantes no caso).

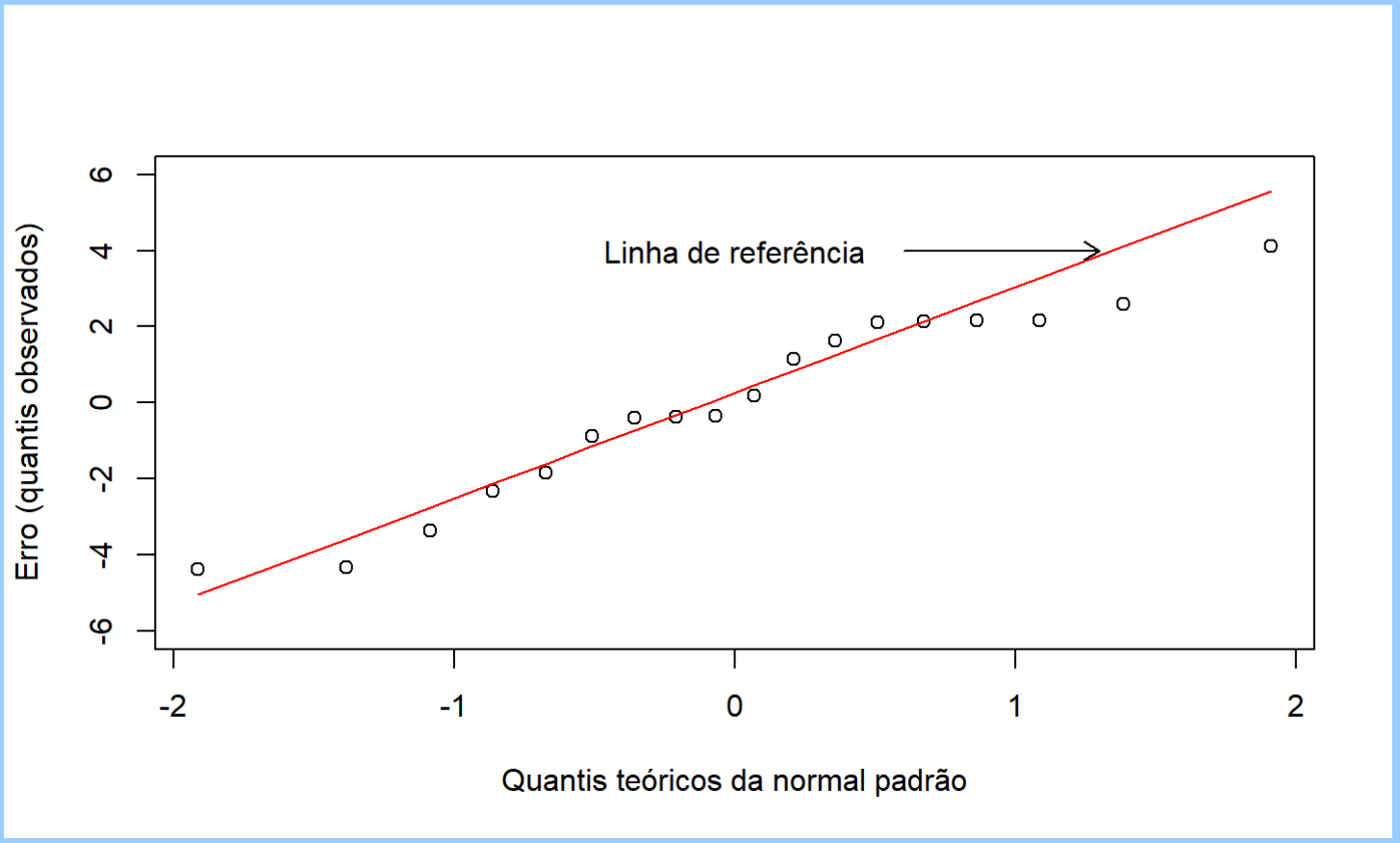
Aplicação

5. Avalie a validade da regressão ajustada na aplicação 2.

   Precisamos verificar se os erros são normalmente distribuídos e se a variância é constante. Vamos avaliar primeiro a normalidade dos erros. Os erros já foram calculados na aplicação 3 e portanto vamos pular esta etapa. Determinando as probabilidades, os quantils teóricos e a linha de referência temos os seguintes valores:

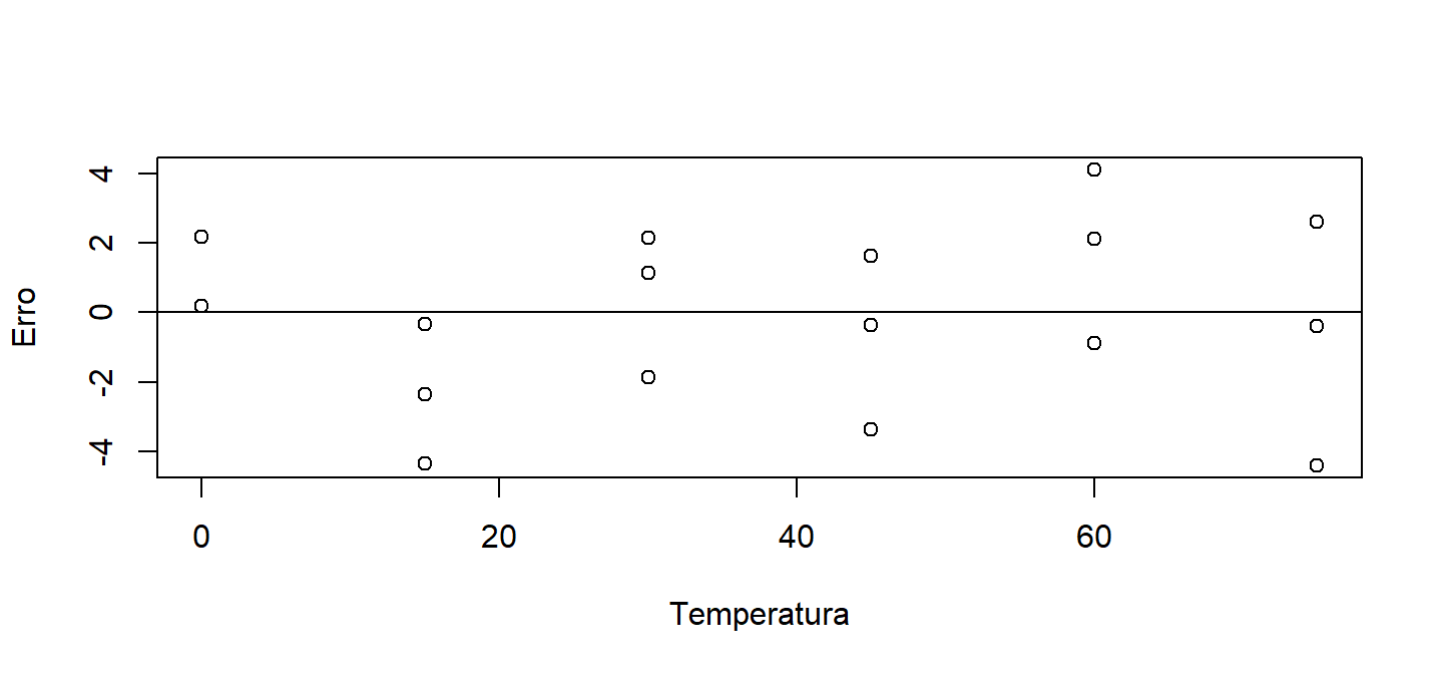
erro	prob. teórica	quantil teórico	linha referência
-4.395	0.028	-1.911	-5.0473491
-4.339	0.083	-1.385	-3.5851326
-3.367	0.139	-1.085	-2.7511689
-2.339	0.194	-0.863	-2.1340357
-1.853	0.250	-0.674	-1.6086386
-0.881	0.306	-0.507	-1.1443987
-0.395	0.361	-0.356	-0.7246370
-0.367	0.417	-0.210	-0.3187746
-0.339	0.472	-0.070	0.0704085
0.175	0.528	0.070	0.4595915
1.147	0.583	0.210	0.8487746
1.633	0.639	0.356	1.2546370
2.119	0.694	0.507	1.6743987
2.147	0.750	0.674	2.1386386
2.175	0.806	0.863	2.6640357
2.175	0.861	1.085	3.2811689
2.605	0.917	1.385	4.1151326
4.119	0.972	1.911	5.5773491

Segue então o gráfico de dispersão do erro (quantis amostrais) em função do quantis teóricos.



Podemos observar que os pontos se ajustam bem a reta e portanto podemos inferir que os erros são normalmente distribuídos.

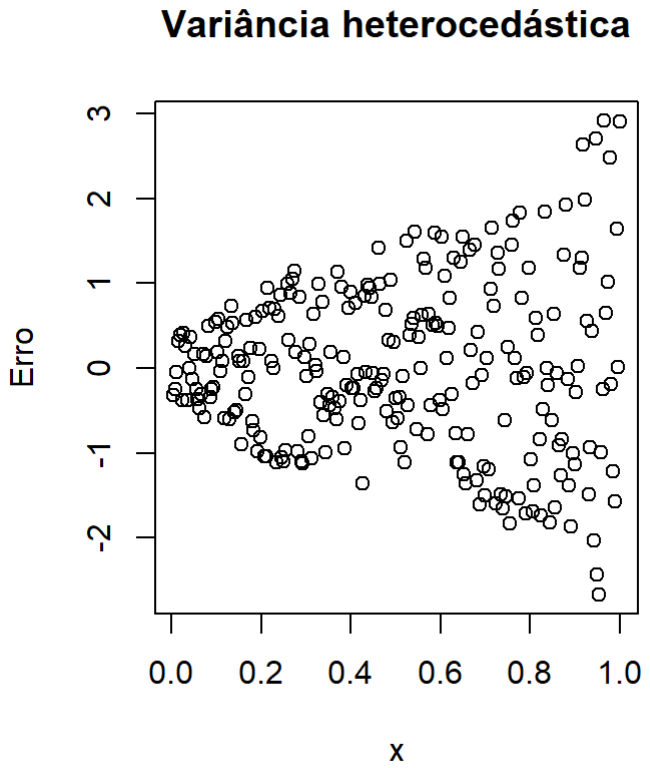
Para avaliar se a variância é constante, ou seja, a mesma para cada x usaremos um gráfico no qual no eixo y será colocado os erros e no eixo x a variável independente (temperatura).

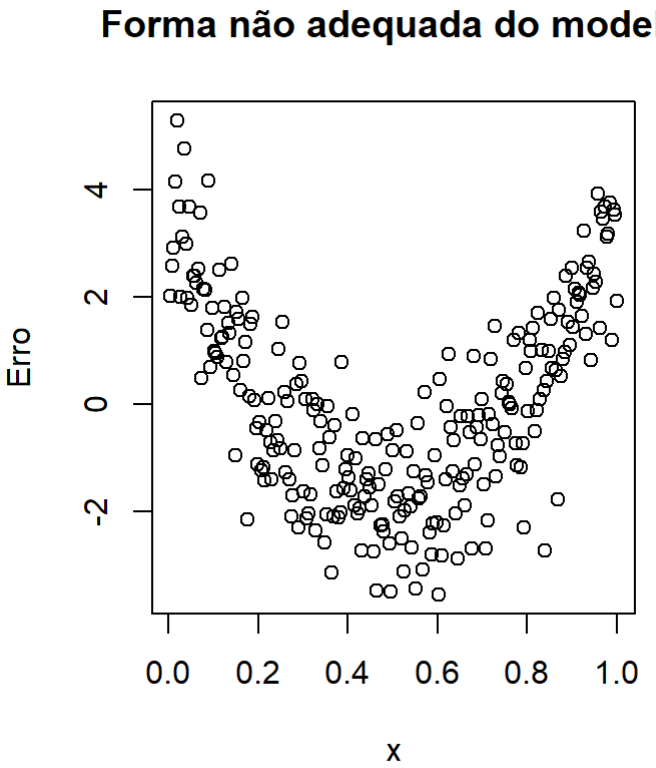


Segundo o gráfico não há nenhuma tendência nos resíduos, o que indica variância constante.

5. Evaluate regression validation in application 2.

- Nos gráficos a seguir, seguem exemplos de resíduos que violam o pressuposto de variância constante.





Intervalo de Confiança do Valor Médio de y

- Quando utilizamos a equação de regressão para estimarmos y em função de x, estamos fazendo uma estimação pontual. No entanto, qual seria a margem de erro associado a esta estimação pontual?
- Para repondermos a esta pergunta, precisamos determinar a variância associado ao y estimado no ponto x desejado.

$$s^2_{\hat{y}} = s^2_{erro}\left[\frac{1}{n} + \frac{(x_e - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

em que:

- s^2_{erro} é a variância do erro, ou seja, é a soma de quadrados do erro dividido por n - 2.
- x_e é o valor no qual se quer predizer y.
- $\bar{x}$ é a média da variável independente.
- x_i são os valores observados da variável independente.

- Portanto, a margem de erro associada a um valor estimado é determinada como:

$$ME = t_{(\alpha/2, n-2)} \sqrt{s^2_{\hat{y}}}$$

Aplicação

6. Aproveitando os dados da aplicação 2, determine um intervalo de confiança de 95% para $\hat{y}$ quando a temperatura for de 30°C.

   Na aplicação 3 já foi determinado o valor estimado de y em todas as temperaturas. Para a temperatura igual a 30, o valor estimado é de 22.853. Vamos determinar a variância do valor estimado.

$$s^2_{\hat{y}} = 6.229 \cdot \left[\frac{1}{18} + \frac{(30 - 37.5)^2}{11812.5} \right]$$

$$s^2_{\hat{y}} = 0.3757$$

Logo, a margem de erro é:

$$ME = |2.1199| \cdot \sqrt{0.3757}$$
$$ME = 1.299$$

Portanto o intervalo de confiança de 95% para o valor médio quando a temperatura for de 30°C é $(22.853 - 1.299; 22.853 + 1.299) = (21.554; 24.152)$.

6. Taking advantage of the data from application 2, determine a 95% confidence interval for $\hat{y}$ when the chest girth is 83 cm.