9 Regressão linear simples

9.1 Modelos de regressão

Objectivo

Modelação da relação de uma variável aleatória (Y – **variável resposta**) com uma ou várias **variáveis explicativas** ($x_1, \ldots, x_k$) de forma que, tanto quanto possível, a variação observada de Y possa ser atribuída ao efeito das variáveis explicativas.

Variação observada em Y
=
Variação previsível + Variação aleatória

Forma funcional

$Y = f(x_1, \ldots, x_k; \theta) + E$ em que E é uma variável aleatória

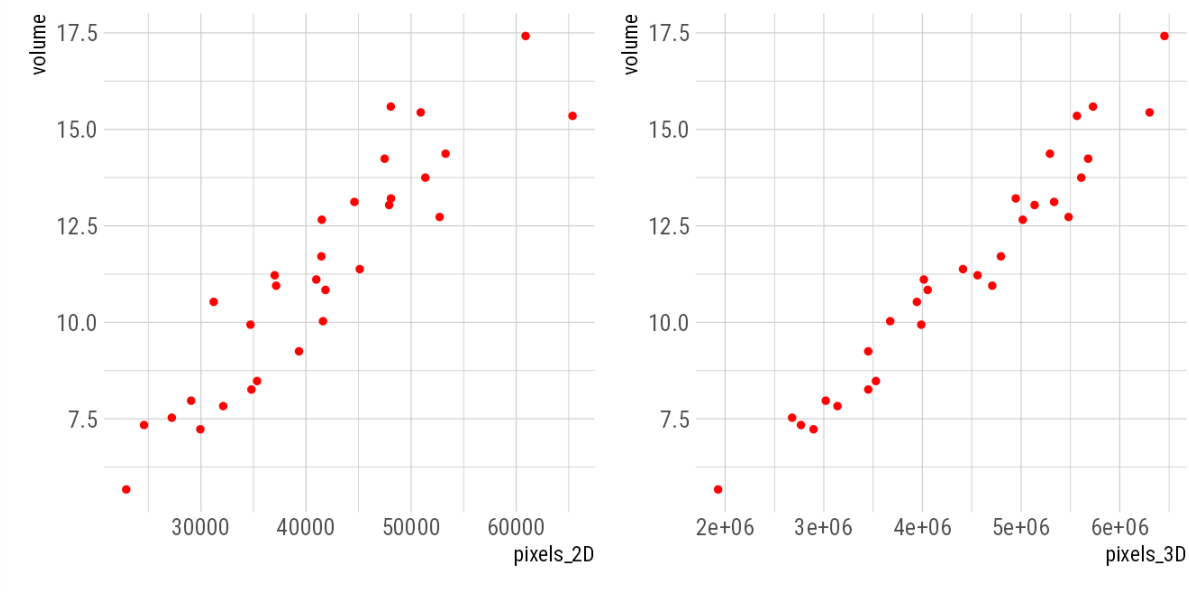
Exemplo – Produção industrial de ostras

Na venda para consumo alimentar as ostras são classificadas de acordo com o seu volume.

Uma empresa produtora planeia implementar um sistema de classificação automática das ostras e avalia 2 propostas:

- 2D – contagem de pixels numa imagem da ostra
- 3D – contagem de pixels numa representação tridimensional da ostra

Exemplo – Os dados recolhidos



Dados

Um conjunto de pontos observáveis (x_i, Y_i), $i = 1, \ldots, n$

Definição

Modelo de Regressão Linear Simples (MRLS)

$$Y_i = \beta_0 + \beta_1 x_i + E_i,\ i = 1, \ldots, n$$

$\beta_0,\ \beta_1$ – parâmetros do MRLS

E_i – erro aleatório associado a $Y_i = (Y \mid x = x_i)$

Pressupostos usuais do MRLS

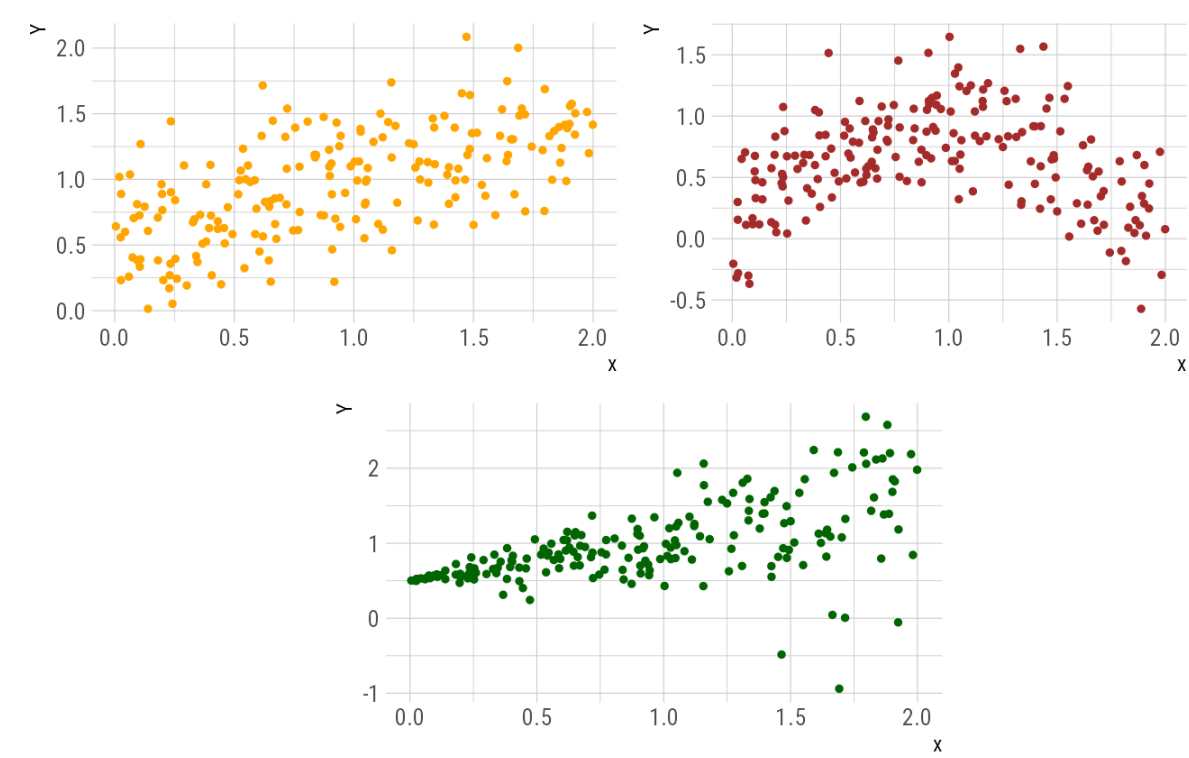
1. $E[E_i] = 0 \iff E[Y_i] = \beta_0 + \beta_1 x_i$
2. $Var[E_i] = \sigma^2 \iff Var[Y_i] = \sigma^2$
3. E_i's não correlacionados $\iff$ Y_i's não correlacionadas

Interpretação dos parâmetros do MRLS

$\beta_0 =$ ordenada na origem $= E[Y \mid x = 0]$

$\beta_1 =$ declive da recta $= E[Y \mid x = x_0 + 1] - E[Y \mid x = x_0]$

Aplicabilidade e validade do MRLS



9.2 Inferências no MRLS

$$SQ(\beta_0, \beta_1) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min SQ(\beta_0, \beta_1)$ – **estimador de mínimos quadrados**

$$\begin{cases} \frac{\partial SQ(\beta_0,\beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial SQ(\beta_0,\beta_1)}{\partial \beta_1} = 0 \end{cases} \iff \begin{cases} \hat{\beta}_1 & = & \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ \hat{\beta}_0 & = & \bar{Y} - \hat{\beta}_1 \bar{x} \end{cases}$$

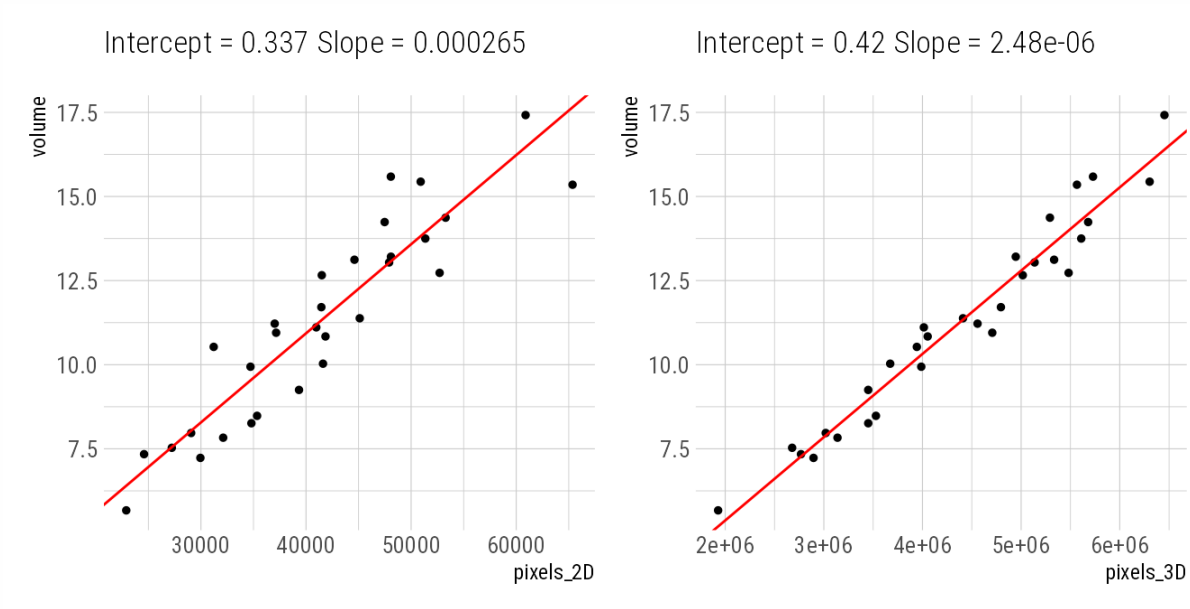
Equação de regressão estimada

$$\hat{E}[Y \mid x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

Nota

$E[\hat{\beta}_i] = \beta_i, i = 1, 2$

Exemplo – As retas de regressão estimadas



Alternativa: método da máxima verosimilhança

Pressuposto adicional:

$$E_i \sim N\left(0, \sigma^2\right) \iff Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

$$\mathcal{L}\left(\beta_0, \beta_1, \sigma^2 \mid \mathbf{y}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2}$$

Os estimadores de máxima verosimilhança de β_0 e β_1 coincidem com os anteriores e

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

Nota

$E[\hat{\sigma}^2_{MV}] = \frac{n-2}{n}\sigma^2$

Inferências sobre β_1

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)} \text{ - variável fulcral para } \beta_1$$

Hipóteses importantes

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

Exemplo – Significância do modelo

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

- 2D: $t_0 = 12.406$, valor-p = 6.77×10^{-13}
- 3D: $t_0 = 24.019$, valor-p = 3.165×10^{-20}

Inferências sobre β_0

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)} \text{ - variável fulcral para } \beta_0$$

Estimação da resposta esperada

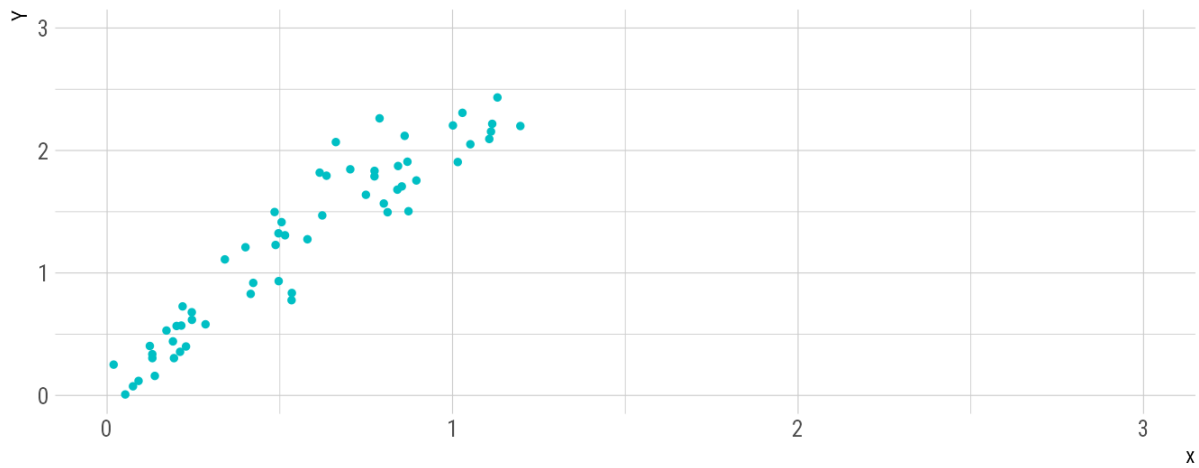
$E[Y_0] = E[Y \mid x = x_0] = \beta_0 + \beta_1 x_0$

Estimador pontual: $\hat{E}[Y_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

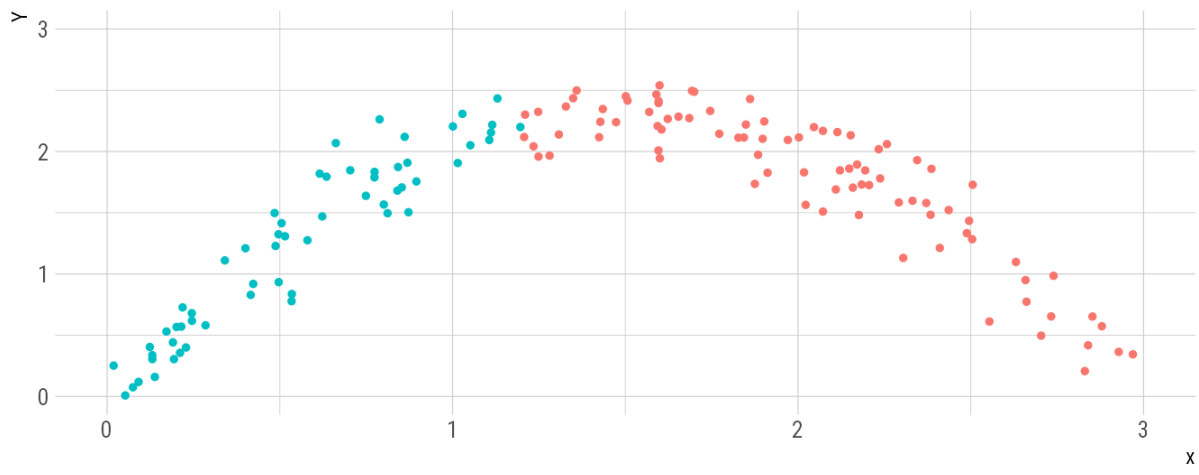
Nota

As inferências podem não ser válidas fora do intervalo de valores de x considerado – **extrapolação**.



Nota

As inferências podem não ser válidas fora do intervalo de valores de x considerado – **extrapolação**.



Exemplo – Estimação da resposta esperada

- 2D: $x_0 = 50000$ px

  $\hat{E}[Y_0] = 13.581\ cm^3$

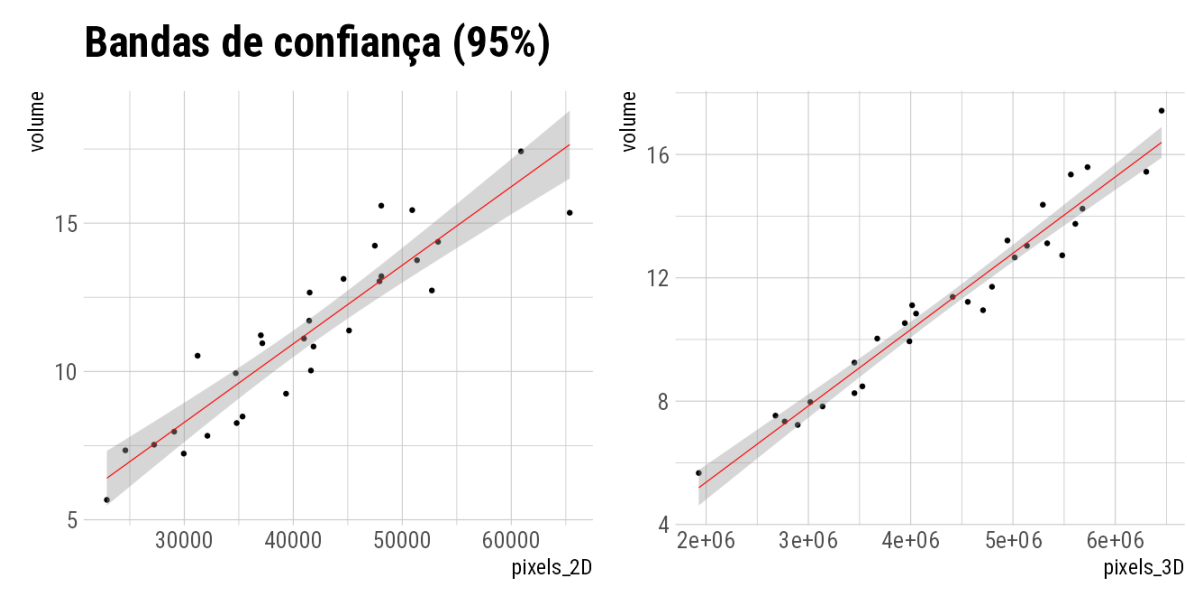
  $\text{IC}_{0.95}(E[Y_0]) = [12.997, 14.165]\ cm^3$

- 3D: $x_0 = 5000000$ px

  $\hat{E}[Y_0] = 12.796\ cm^3$

  $\text{IC}_{0.95}(E[Y_0]) = [12.52, 13.071]\ cm^3$

Exemplo – Estimação da resposta esperada



9.3 Avaliação do MRLS

Há um grande número de técnicas para avaliar a qualidade do ajustamento de um MRLS. Vejamos uma das mais simples.

Sendo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ pode mostrar-se que

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$\iff SQT = SQE + SQR$$

variação total em Y = variação devida ao erro aleatório + variação explicada pelo MRLS

Coeficiente de determinação

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \left(\hat{\beta}_1\right)^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}$$

R^2 – proporção da variação em Y explicada pelo MRLS

Por definição $0 \leq R^2 \leq 1$.

$R = +\sqrt{R^2}$ – coeficiente de correlação empírico

Exemplo – Qualidade dos modelos

Qual é o melhor método de classificação de ostras?

- 2D: $R^2 = 0.841$
- 3D: $R^2 = 0.952$

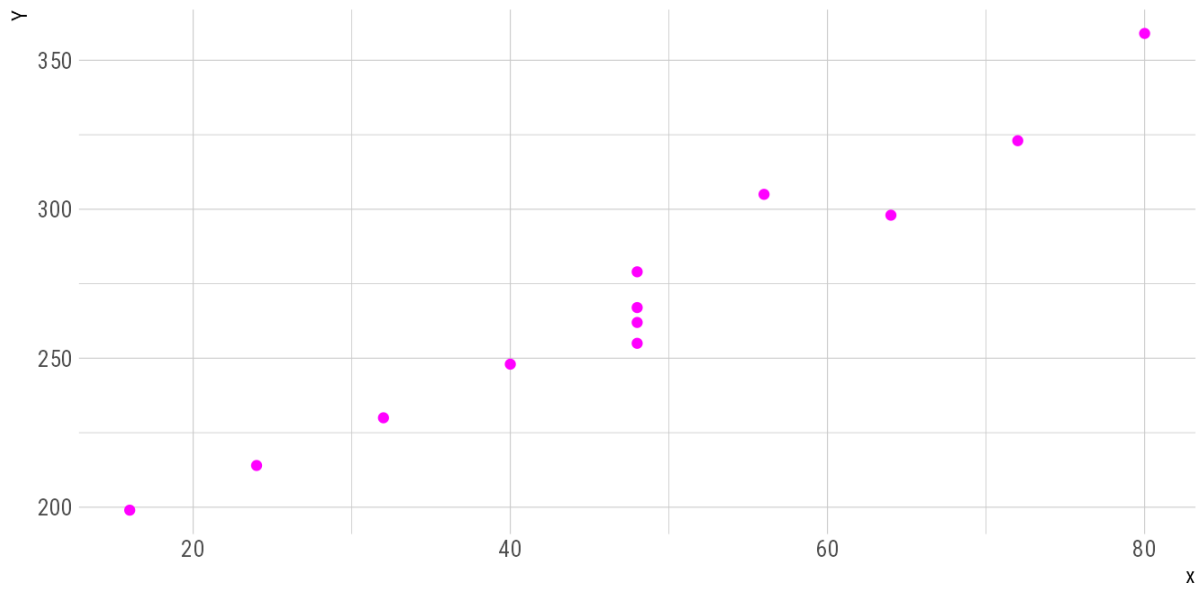
Exercício

Interessa estudar a relação entre a resistência de um determinado tipo de plástico (Y) e o tempo que decorre a partir da conclusão do processo de moldagem até ao momento de medição da resistência (x $[horas]$). As observações que se seguem foram efectuadas em 12 peças construídas com este plástico, escolhidas aleatoriamente.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
x_i	32	48	72	64	48	16	40	48	48	24	80	56
y_i	230	262	323	298	255	199	248	279	267	214	359	305

Exercício

a. Represente graficamente as observações e desenhe a recta que, no seu entender, melhor se ajusta às observações.



Exercício

b. Considere um modelo de regressão linear simples para explicar as observações. Obtenha a estimativa dos mínimos quadrados dos coeficientes da recta de regressão e desenhe-a no gráfico.

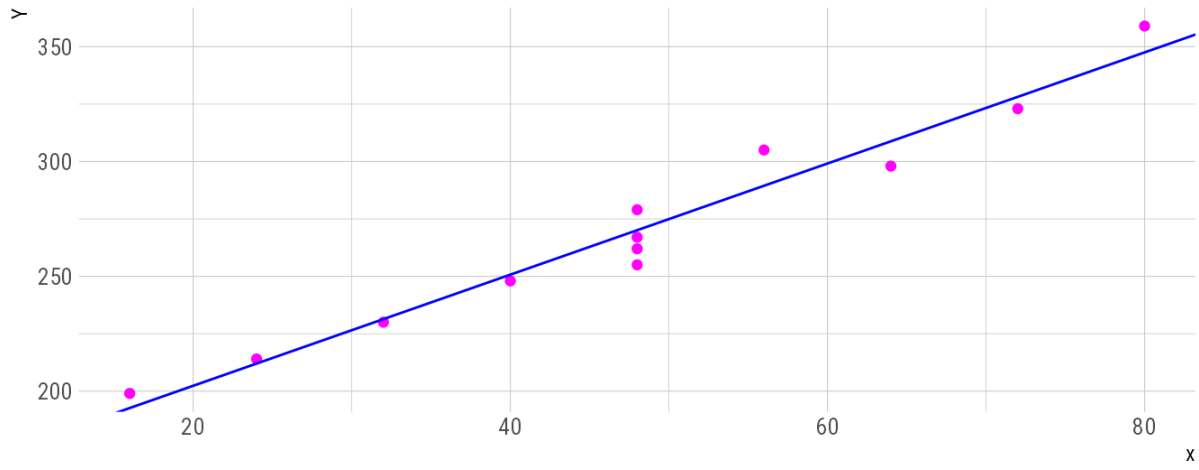
$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{12} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{12} x_i^2 - n\bar{x}^2} = \frac{164752 - 12 \times 48 \times 269.92}{31486 - 12 \times 48^2} = \\
&= 2.4167
\end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 269.92 - 2.4167 \times 48 = 153.9167$$

Exercício

Equação de regressão estimada:

$$\widehat{E}[Y \mid x] = \hat{\beta}_0 + \hat{\beta}_1 x = 153.92 + 2.42x$$



Exercício

c. Calcule o coeficiente de determinação e comente o valor obtido.

$$\begin{aligned}
R^2 &= \frac{\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} = \\
&= \frac{(9278.08)^2}{3838 \times 23378.92} = 0.9593
\end{aligned}$$

Isto é, 95.93% da variabilidade total da resistência do plástico é explicada pelo modelo de regressão com o tempo decorrido entre a moldagem e a medição da resistência.

Exercício

d. Proceda ao teste da hipótese "O coeficiente angular é nulo". Qual o interesse desta hipótese?

Hipóteses

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

Estatística de teste

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_i x_i^2 - 12\bar{x}^2}}} \overset{H_0}{\sim} t_{(10)},$$

Exercício

Valor observado da estatística de teste: $t_0 = 15.35$.

Valor-p

$p = 2 \times P(T_0 > 15.35) = 2.81 \times 10^{-8}$.

Note-se que $p < 0.001 = 2 \times 0.0005$ pois $F^{-1}_{t(10)}(0.9995) = 4.587$.

Conclusão

Rejeita-se H_0 para níveis de significância de pelo menos 2.81×10^{-8}, ou seja, há evidência contra H_0, isto é, o tempo decorrido entre a moldagem e a medição da resistência influencia significativamente a resistência do plástico.

Exercício

e. Calcule o intervalo de confiança a 95% para o valor esperado da resistência obtida 48 horas depois de concluída a moldagem. Acha legítimo usar o mesmo procedimento tratando-se de um período de 10 horas em vez de 48 horas? Justifique a sua resposta.

Variável fulcral para $E[Y|x = x_0] = \beta_0 + \beta_1 x_0$:

$$W = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_i x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(10)}$$

Exercício

Intervalo aleatório de confiança para $E[Y|x = x_0]$ a 95%:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm F^{-1}_{t(10)}(0.975)\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_i x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}$$

$P(-2.228 \leq W \leq 2.228) = 0.95$ pois $F^{-1}_{t(10)}(0.975) = 2.228$

Exercício

Estimativa pontual:

$$\hat{E}[Y|x = 48] = 153.91 + 2.4167 \times 48 = 269.91$$

Intervalo de confiança para $E[Y|x = 48]$ a 95%:

$$[263.035; 276.785]$$

Exercício

Não é aconselhável considerarmos x_0 fora do domínio dos dados observados, visto que não há informação fora desse domínio. O que acontece com $x_0 = 10 \notin [16, 80]$.