Regressão Linear Simples - Método: Mínimos Quadrados

Desejamos estudar o efeito de duas drogas (A e B) em 10 pacientes.

Desejamos estabelecer a relação entre os efeitos causados por essas drogas em um mesmo paciente.

- a droga A é mais econômica que a droga B
- a droga A é mais conhecida (seus efeitos são mais estudados)
- a droga B é mais avançada que a droga A
- a droga B é boa para controlar mais distúrbios que a droga A

Regressão Linear Simples - Método: Mínimos Quadrados

- Acredita-se que o efeito da droga A tem uma certa relação com o efeito da droga B
- Notação:
  - seja X_i o efeito da droga A no paciente i
  - seja Y_i o efeito da droga B no paciente i
  - $i = 1, ..., 10$

X_i	1.9	0.8	1.1	0.1	-0.1	4.4	4.6	1.6	5.5	3.4
Y_i	0.7	-1	-0.2	-1.2	-0.1	3.4	0	0.8	3.7	2

Regressão Linear Simples - Método: Mínimos Quadrados

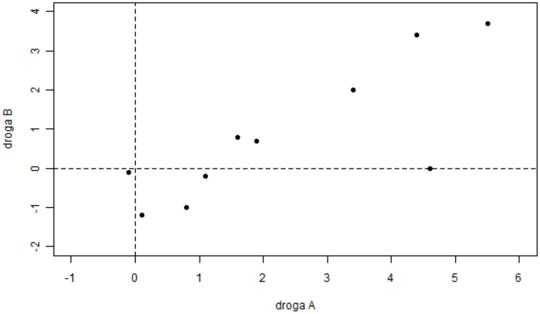
- Afirmar que existe uma relação entre o efeito das duas drogas pode ser expresso como:

$$Y = g\left(X\right)$$

  lembrando que X é o efeito da droga A, e Y o efeito da droga B
- X (variável independente) é denominada *preditor*
- Y (variável dependente) é denominada *resposta*

Regressão Linear Simples - Método: Mínimos Quadrados

- Explorando os dados:



Regressão Linear Simples - Método: Mínimos Quadrados

- Não podemos afirmar que existe alguma relacão *clara ou evidente* a partir do gráfico (relação entre X e Y), já que as medidas são afetadas por fatores que desconhecemos.
- Tentativa: ajustar o modelo $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, ..., n$
  - y: resposta
  - x: preditor
  - ε: efeito desconhecido
  - β_1, β_2: parâmetros
- Modelo ajustado: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Regressão Linear Simples - Método: Mínimos Quadrados

- Ajustar o modelo linear significa assumir um critério para achar β_1 e β_2 ótimos (segundo esse critério), tal que o erro (também segundo esse critério) seja pequeno.
- Mínimos Quadrados:
  - seja o erro do modelo: $\varepsilon_i = y_i - \hat{y}_i = y_i - (\beta_1 + \beta_2 x_i)$, $i = 1, ..., n$
  - seja $Q(\beta_1, \beta_2) = \sum_{i=1}^{n} \varepsilon_i^2$, a soma dos quadrados dos erros
  - $\hat{\beta}_1$ e $\hat{\beta}_2$ ótimos são aqueles que minimizam $Q(\beta_1, \beta_2)$

Regressão Linear Simples - Método: Mínimos Quadrados

- $Q(\beta_1, \beta_2) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$
  - $\frac{\partial}{\partial \beta_1} Q(\beta_1, \beta_2) = -2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)$
  - $\frac{\partial}{\partial \beta_2} Q(\beta_1, \beta_2) = -2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i) x_i$
  - $\frac{\partial}{\partial \beta_1} Q(\beta_1, \beta_2) = \frac{\partial}{\partial \beta_2} Q(\beta_1, \beta_2) = 0$
- Resolvendo as equações acima:
  - $\hat{\beta}_1 = \bar{y}_n - \hat{\beta}_2 \bar{x}_n$
  - $\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x}_n \bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n \bar{x}_n^2}$

Regressão Linear Simples - Método: Mínimos Quadrados

visto que $\beta_1 = \overline{y}_n - \beta_2 \overline{x}_n$, pois $-\sum_{i=1}^n y_i + n\beta_1 + \beta_2 \sum_{i=1}^n x_i = 0$.

A partir da derivada parcial em β_2, temos

$$-\sum_{i=1}^{n} x_i y_i + (\overline{y}_n - \beta_2 \overline{x}_n) n \overline{x}_n + \beta_2 \sum_{i=1}^{n} x_i^2 = 0$$

ou equivalentemente

$$-\sum_{i=1}^{n} x_i y_i + n \overline{x}_n \overline{y}_n + \beta_2 (\sum_{i=1}^{n} x_i^2 - n(\overline{x}_n)^2) = 0$$

Regressão Linear Simples - Método: Mínimos Quadrados

- Retomando o exemplo das drogas, temos então:
  - $\hat{\beta}_1 = -0.7869$
  - $\hat{\beta}_2 = 0.685$
  - Assim, o modelo é definido pela equação:

$$y_i = -0.786 + 0.685x_i$$

Propriedades das estimativas de $\hat{\beta}_1$ e $\hat{\beta}_2$

- modelo: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
- suposições teóricas:
  - $y_1, \ldots, y_n$ são funções lineares de $x_1, \ldots, x_n$ respectivamente
  - $\varepsilon_1, \ldots, \varepsilon_n$ são considerados independentes e identicamente distribuídos

$$E(\varepsilon_i) = 0 \qquad Var(\varepsilon_i) = \sigma^2 < \infty$$

- suposição extra: como o verdadeiro interesse é o estudo da variável y, podemos supor que os valores de x são fixos, já que neste processo (relativo a y) interessa a variabilidade de y e desprezamos a variabilidade de x.

Propriedades das estimativas de $\hat{\beta}_1$ e $\hat{\beta}_2$

- Proposição (a partir das suposições anteriores):
  - $E\left(\hat{\beta}_1\right) = \beta_1$
  - $Var\left(\hat{\beta}_1\right) = \sigma^2 \frac{\sum_{i=1}^n x_i^2 / n}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$
  - $E\left(\hat{\beta}_2\right) = \beta_2$
  - $Var\left(\hat{\beta}_2\right) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$

Coeficiente de Correlação Amostral

- Dada uma amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$, definimos a seguinte medida de dependência linear entre X e Y:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

- Esta medida amostral é introduzida por um outro conceito: *Correlação entre as variáveis aleatórias,* $\rho(X, Y)$
- Digamos que $r = \hat{\rho}(X, Y)$ (r estima o coeficiente de correlação)

Coeficiente de Correlação Amostral - Propriedades

- $\rho(X, Y)$ é uma medida de dependência
- $-1 \leq \rho(X, Y) \leq 1$
- $\rho(X, Y) = \pm 1 \Rightarrow Y = aX + b$
- X e Y são independente $\Rightarrow \rho(X, Y) = 0$
- $\rho(X, Y) = 0$ não significa que X e Y são independentes

Coeficiente de Correlação Amostral - Propriedades

- quando r estiver próximo de 1: Y é próximo de ser combinação linear de X ($a > 0$).

$$Y = aX + b$$

- quando r estiver próximo de -1: Y é próximo de ser combinação linear de X ($a < 0$).

$$Y = aX + b$$

- r mede o graus de associação linear entre X e Y.

Como Avaliar e Usar um Modelo Linear

- No exemplo das drogas A e B
  - fazer um gráfico X vs. Y para colocar em evidência o tipo de relação que pode ser modelada
  - calcular o coeficiente de correlação amostral ($r = 0.8022$): existe uma alta correlação linear entre as drogas A e B
  - Procedemos ajustando: $(drogaB) = -0.786 + 0.685(\text{droga A})$