$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$ sent $\varepsilon_i \sim N(0, \sigma)$

Y_i : valor de la variable resposta Y en el cas i-èsim

X_i : valor que pren la condició X en el cas i-èsim

ε_i : error aleatori o distància a la recta del cas i-èsim

β_0 com a constant, β_1 com a pendent i σ^2 variancia dels ε_i

Exemple:

Homes adults i sans del Bcn: Y és pes en kg; X és alçada en cm. Suposem com a model una recta amb paràmetres:

$\beta_0 = -100 kg$ $\beta_1 = +1 kg/cm$ $\sigma = 6 kg$

Quin pes correspon a un señor de 160 cm?

$Y = -100 + 1 \cdot 160 = 60 kg$

I un de 180 cm?

$Y = -100 + 1 \cdot 180 = 80 kg$

Què significa "correspon"?

'Esperat, en mitjana'

Què significa $\sigma = 6 kg$?

Separar-se uns 6 kg del pes esperat és habitual. (Rarament és més del doble)

Què opines de l'etiqueta 'pes ideal' en algunes Farmàcies?

Que ignoren la variabilitat natural.

• Estimació dels paràmetres

β_0, β_1 i σ^2 són valors poblacionals, autèntics, desconeguts a 'estimar'. L'estimació dels dos primers, dóna lloc a la recta estimada:

$$\hat{y}_i = b_0 + b_1 \cdot X_i$$

error de predicció:

$$e_i = y_i - \hat{y}_i$$

$$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S^2_x} = r \cdot \frac{S_y}{S_x}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

$$\hat{\sigma}^2 = S^2 = \frac{\Sigma(e_i^2)}{n-2} = \frac{(n-1)\, S_y^2\, (1-r^2)}{n-2} = \frac{(n-1)(S_y^2 - b_1 S_{xy})}{n-2}$$

S_{xy}: covariància mostral

$r = r_{xy}$: correlació mostral

- Estimació dels paràmetres. Exemple

x cervesa	y alcohol
5	0.100
2	0.030
9	0.190
8	0.120
3	0.040
7	0.095
3	0.070
5	0.060
3	0.020
5	0.050
4	0.070
6	0.100
5	0.085
7	0.090
1	0.010
4	0.05

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{(0.1 + 0.03 + 0.19 + 0.12 + 0.04 + 0.095 + 0.07 + 0.06 + 0.02 + 0.05 + 0.07 + 0.1 + 0.085 + 0.09 + 0.01 + 0.05)}{16} = 0.07375$$

$$\bar{x} = \frac{5+2+9+8+3+7+3+5+3+5+4+6+5+7+1+4}{16} = 4.8125$$

$$S_y^2 = \frac{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}{n-1} =$$

$$\Sigma y_i^2 = (0.1^2 + 0.03^2 + 0.19^2 + 0.12^2 + 0.04^2 + 0.095^2 + 0.07^2 + 0.06^2 + 0.02^2 + 0.05^2 + 0.07^2 + 0.1^2 + 0.085^2 + 0.09^2 + 0.01^2 + 0.05^2) = 0.11625$$

$$\Sigma y_i = 1.18$$

$$S_y^2 = \frac{0.11625 - \frac{1.18^2}{16}}{16-1} = 0.001948 \qquad S^2x = 4.829167$$

$$S_{xy} = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{n-1} \qquad S_{xy} = \frac{6.98 - \frac{77 \cdot 1.18}{16}}{16-1} = 0.08675$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{0.08675}{\sqrt{4.829167} \cdot \sqrt{0.001948}} = 0.89441$$

Resultats de la regressió:

$$b_1 = \frac{S_{xy}}{S^2x} = r_{xy} \cdot \frac{S_y}{S_x} = 0.01796 \qquad \hat{Y} = -0.0127 + 0.01796 \cdot X_i$$

$$\cdot S \approx 0.077$$

$$b_0 = \bar{Y} - b_1 \bar{X} = -0.0127 \qquad e_i = 0.100 - 0.077 = 0.023$$

$$S = \sqrt{\frac{\Sigma(e_i^2)}{n-2}} = 0.0204$$

- **Validació del model lineal. Anàlisi dels residus**

e_i vs "Fitted Values" → Linealitat i homoscedasticitat

e_i vs ordre observacions → Independència

Qqnorm dels residus (e_i) → Premissa de Normalitat

Histograma dels residus (e_i) → Premissa de Normalitat

- **e_i versus fitted values**

- **Linealitat**: El núvol de punts ha de mantenir sempre la mateixa alçada
  - que formen más o menos una línea recta

- **Homoscedasticitat**: La variabilitat dels residus ha de mantenir-se constant independentment dels valors predits.
  - la distribución de los puntos tine que ser igual, o todos juntitos o todos separados

ei vs ordre de les observacions

- Independència: Els residus no han de mostrar cap patró enfront l'ordre

- que no se alternen subidas y bajadas y no haya ningun patrón

qqnorm i histograma de residus

- Normalitat: Els residus han de ser normals: situarse sobre la recta en el qqnorm i forma de campana a l'histograma.

Generalment amb poques dades és difícil avaluar les premisses: l'opció més prudent és acceptar-les a no ser que no veiem clar que alguna d'elles s'infringeixi