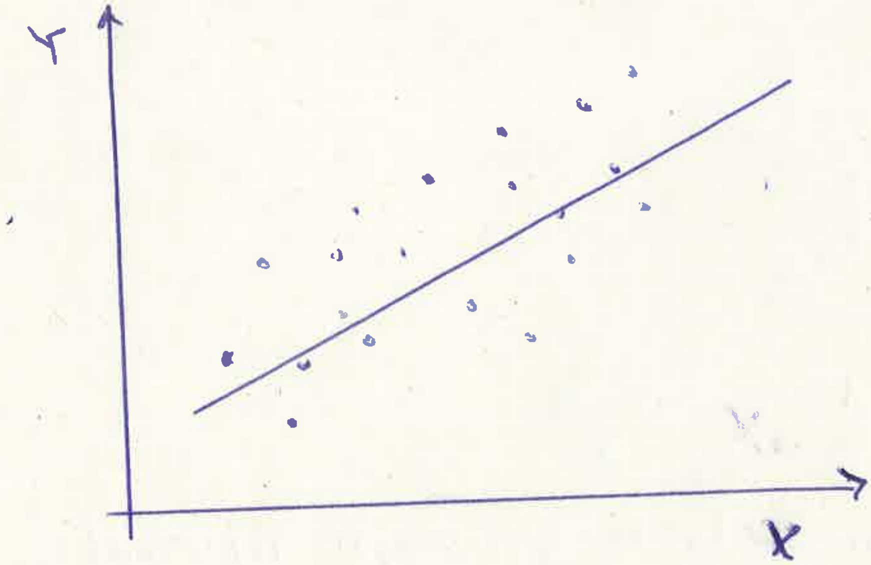
Regresión Lineal Simple



Preguntas de interés:

- ¿Existe relación entre X e Y?
- ¿Cómo cambia Y en función de X?
- ¿Qué valor de Y se puede esperar dado un valor de X?

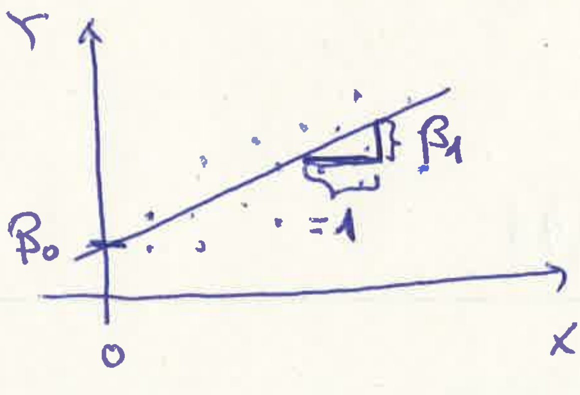
Modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , i = 1, \dots, n$$

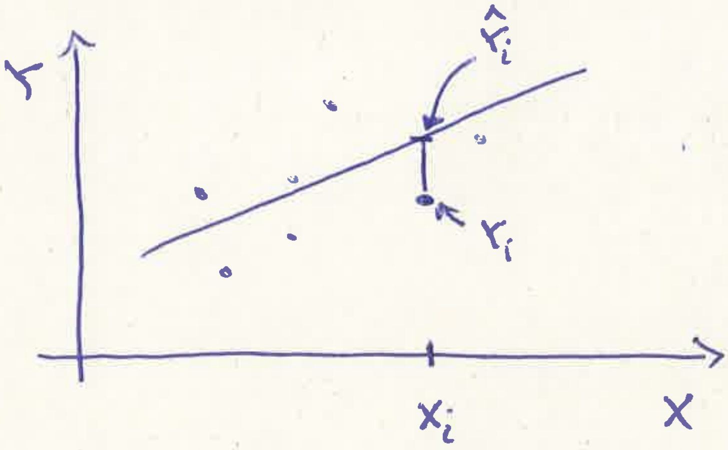
$$\varepsilon_i \sim N(0, \sigma^2)$$

Interpretación de los parametros

- β_0 : $E(Y_i \mid X_i = 0) = \beta_0$
- β_1 : $E(Y_i \mid X_i + 1) - E(Y_i \mid X_i) = \beta_1$



Estimación de parámetros: ¿$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$?



¿Cuál es la mejor recta, la que mejor representa la relación entre X e Y?

Idea: Minimizar $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot X_i))^2$

→ Metódo de mínimos cuadrados

→ Solución ⇝ Transpa 12

Ejemplo Bundesliga

- Y: Puntos $\qquad \bar{Y} = 46{,}67, \quad s_Y = 13{,}2$

  X: Goles a favor $\qquad \bar{X} = 48{,}8, \quad s_X = 12{,}97$

  $S_{XY} = 145{,}42 \rightsquigarrow r = 0{,}85$

- $b_1 = \hat{\beta}_1 = \dfrac{S_{XY}}{S_X^2} = r \cdot \dfrac{S_Y}{S_X}$

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

- $\hat{\beta}_1 = 0{,}865, \quad \hat{\beta}_0 = 4{,}46$

$$\hat{\beta}_1 = b_1 \sim N(\beta_1, \sigma^2_{b_1})$$

$$= \frac{\sigma^2}{(n-1)\cdot s_x^2}$$

$$\Leftrightarrow \quad \frac{b_1 - \beta_1}{\sigma_{b_1}} \sim N(0,1)$$

$$\Rightarrow \quad \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

$$= \sqrt{\frac{s^2}{(n-1)\cdot s_x^2}}$$

$$\rightsquigarrow \quad IC(\beta_1; 1-\alpha) = b_1 \mp t_{n-2;\,1-\alpha/2} \cdot s_{b_1}$$

$$= 0{,}865 \mp t_{376;\,0.975} \cdot 0{,}028 = [0{,}81;\ 0{,}92]$$

Ejemplo Bundesliga

→ $n = 378$

→ $\alpha = 0{,}05$

Pruebas de hipótesis sobre β_1

- $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

  $\rightarrow$ Implica que X e Y son independientes

- $H_0: \beta_1 \leq 0{,}8$ vs. $H_1: \beta_1 > 0{,}8$

- Estadístico de contraste:

  - $b_1 = \hat{\beta}_1 \sim N(\beta_1, \sigma_{b_1})$

  - $H_0: \beta_1 = \beta_1^*$

  $$\Longrightarrow \frac{b_1 - \beta_1^*}{s_{b_1}} \overset{H_0}{\sim} t_{n-2}$$

- Ejemplo (Bundesliga)

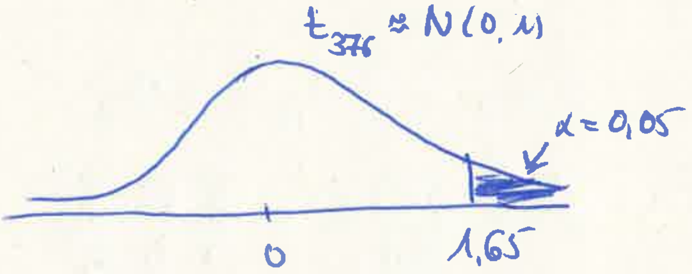
  - $H_0: \beta_1 \leq 0{,}8$ vs. $H_1: \beta_1 > 0{,}8$

  - $\alpha = 0{,}05$; $n = 378$

  - $\hat{\beta}_1 = 0{,}865$, $s_{b_1} = 0{,}028$

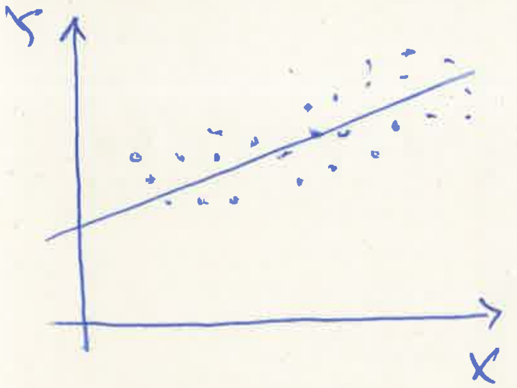
  $\rightsquigarrow t = 2{,}321$

  - Valor crítico



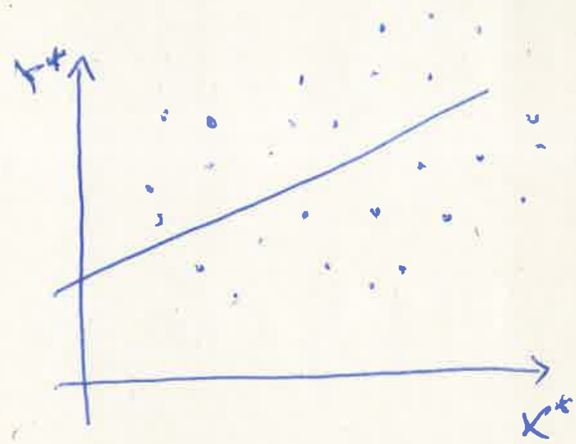
$t > 1{,}65 \Rightarrow$ Se rechaza H_0.

Coeficiente de determinación (R^2)



$b_1 \approx b_1^*$

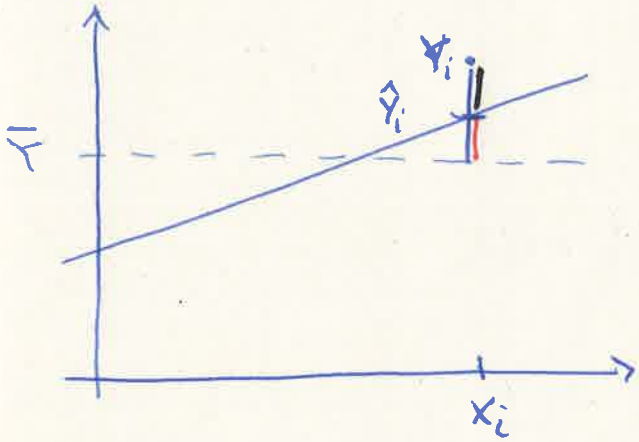
$b_0 \approx b_0^*$



→ X "predice" Y bastante bien

→ X^* es una variable entre varias que predicen Y^*

¿Qué medida nos cuantifica el grado de asociación entre la variable respuesta y la(s) explicativa(s)?



- Variabilidad total de Y:

$$SQ_T = \sum_i (y_i - \bar{y}_i)^2$$

- Variabilidad explicada por X

$$SQ_E = \sum_i (\hat{y}_i - \bar{y})^2$$

- Variabilidad residual

$$SQ_R = \sum_i (y_i - \hat{y}_i)^2$$

- $SQ_T = SQ_E + SQ_R$

- $R^2 = \dfrac{SQ_E}{SQ_T} = \dfrac{\sum (\hat{y}_i - \bar{y})^2}{(n-1)\cdot s_Y^2}$

"Variabilidad de Y explicada por X."

- $0 \leq R^2 \leq 1$
- RLS: $R^2 = r_{XY}^2$
- $R^2 = 1 \iff y_i = \hat{y}_i \ , \forall i$
- $R^2 = 0 \iff \hat{y}_i = \bar{y} \ , \forall i$ (= recta con pendiente 0)