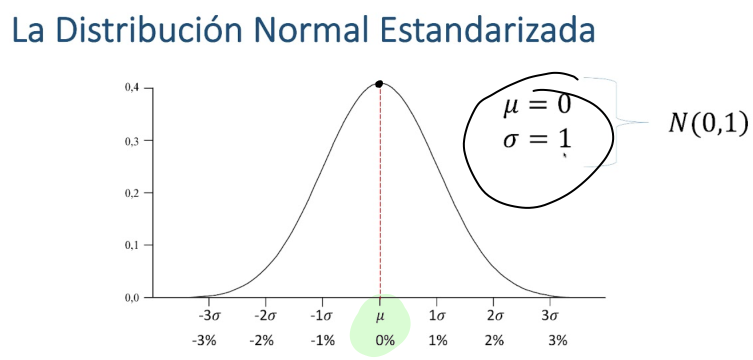
Repaso Estadística - Inferencial

Distribuciones Probabilidad - Dan la probabilidad de ocurrencia de eventos para un experimento

Normal - Continua



Poisson - Discreta - Número de observaciones en un periodo de tiempo

Bernoulli - Lanzar una moneda - Ejemplo Típico

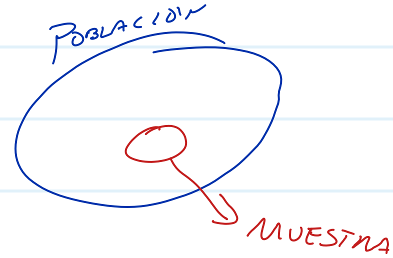
$X \in \{0, 1\}$

$$f(x) = \mathbb{P}(X = w) = p^{w} (1-p)^{1-w}$$

Binomial - En un experimento Bernoulli que repetimos n veces cuántas veces obtengo un éxito

Exponencial - Tiempo que pasa antes de que ocurra un evento en un proceso Poisson.

Nos interesa saber la edad promedio de alumnos de maestrías matemáticas en México



Población - Todos esos alumnos

Muestra - Subconjunto de esa población

(a) $x_1, x_2, \dots, x_n$ variables aleatorias $\quad n$: # datos muestra

(b) Independencia // x_i no afecta al valor de x_j

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad //$$

(c) Provienen de una distribución idéntica

$$x_i \sim N(\mu, \sigma^2)$$

Distribuyen como

μ : Media Poblacional

σ^2 : Varianza Poblacional

3 maneras de inferirlos:

→ Estimación Puntual: Con una función de la muestra quiero estimar esos parámetros

(a) Media Muestral

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Esperanza $\quad \mathbb{E}[\bar{x}] = \mu \quad$ Propiedad de insesgamiento

- Varianza $\quad Var[\bar{x}] = \sigma^2/n$

Distribución $\quad \bar{x} \sim N(\mu, \sigma^2/n)$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \Longrightarrow \underbrace{n\bar{x}}_{n\text{-componentes}} - \underbrace{\sum_{i=1}^{n} x_i}_{n\text{-componentes}} = 0$$

$$\Rightarrow n\bar{x} - [x_1 + x_2 + \dots x_n] = 0$$

$$\Rightarrow (\bar{x} - x_1) + (\bar{x} - x_2) + \dots + (\bar{x} - x_n) = 0$$

Si yo conozco $\bar{x}$ y $(n-1)$ datos ya puedo conocer el último valor faltante

$n-1$: Grados de Libertad

(b) Varianza Muestral

$$S^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

$$\mathbb{E}[S^2] = \sigma^2 \quad \text{insesgada}$$

(c) Desviación Estándar Muestral

$$S = \sqrt{S^2}$$

→ Estimación por Intervalos

Depende de la muestra. Para construirlo se utilizan cantidades pivotales

↳ En su fórmula incluyen a los parámetros poblacionales. Pero su distribución de probabilidad no los incluye

Estandarización: $X \sim N(\mu, \sigma^2)$

Paso 1: Restar la media $\quad X-\mu$

Paso 2: Dividirlo entre desv $\quad \boxed{\dfrac{X-\mu}{\sigma}} \sim N(0,1)$

Intervalos de confianza

→ Encontrar μ conociendo σ^2

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Estandarizar

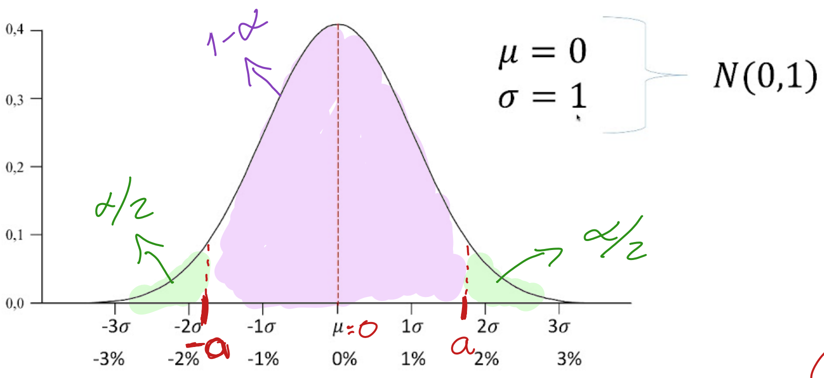
$$\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\left(\frac{\bar{x}-\mu}{\sigma^2}\right) \sim N(0,1)$$

$$\boxed{\mathbb{P}\left(-a \le \sqrt{n}\left(\frac{\bar{x}-\mu}{\sigma^2}\right) \le a\right)} = \boxed{1-\alpha}$$

→ Nivel de confianza
Ejemplo: 95%

→ Nivel de significancia
Ejemplo: 5%

La Distribución Normal Estandarizada



Cuantil de una normal estándar en $\alpha/2$

La probabilidad de ser mayor es de $\alpha/2$

Notación $\quad a = z^{(\alpha/2)}$

Intervalo de confianza

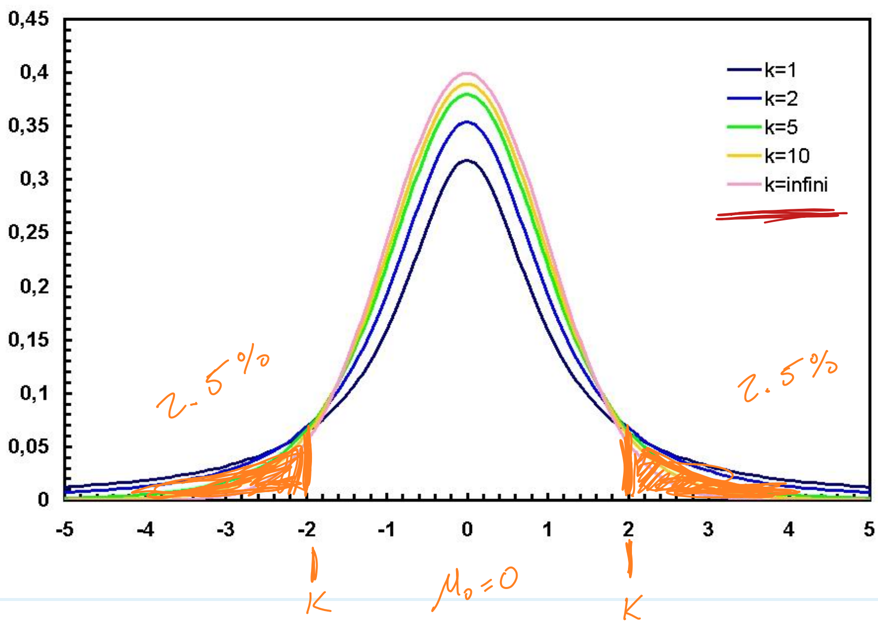
$$\mu \in \left[\bar{x} - \frac{\sigma}{\sqrt{n}} z^{(\alpha/2)} , \bar{x} + \frac{\sigma}{\sqrt{n}} z^{(\alpha/2)} \right]$$

// Si realizamos el experimento 100 veces y $1-\alpha$ es 95% vamos a encontrar a μ 95 de esas 100 veces en el intervalo//

// Si $1-\alpha$ aumenta → El intervalo aumenta//

Caso 2: Con σ^2 desconocida

$$\sqrt{n}\left(\frac{\bar{x}-\mu}{S}\right) \sim t_{(n-1)}$$ → T-student con $n-1$ grados de libertad



Conforme n tienda a infinito $t_{(n-1)} \sim N(0,1)$
$n \to \infty$

Intervalo:

$$\left[\bar{x} - \frac{S}{\sqrt{n}} t_{(n-1)}^{(\alpha/2)} , \bar{x} + \frac{S}{\sqrt{n}} t_{(n-1)}^{(\alpha/2)} \right]$$

→ Pruebas de Hipótesis

Conjetura respecto a los parámetros poblacionales

Conjetura Inicial	Conjetura Alternativa
Hipótesis Nula (H_0)	Hipótesis Alternativa (H_1)

Ya es específico

Dos Colas	$H_0: \mu = \mu_0$	$H_a: \mu \neq \mu_0$
Cola Izquierda	$H_0: \mu > \mu_0$	$H_a: \mu \leq \mu_0$
Cola Derecha	$H_0: \mu < \mu_0$	$H_a: \mu \geq \mu_0$

Debemos crear una regla de decisión para saber si rechazamos nuestra H_0

	H_0 Verdadera	H_0 Falsa
Rechazar H_0	Error tipo I (α)	Sin Error ✓
No Rechazar H_0	Sin Error ✓	Error Tipo II (β)

Se controla el error tipo I (α)

$$\alpha = \mathbb{P}(\text{Rechazar } H_0 \mid H_0 \text{ cierta})$$

¿Si rechazamos o no? Construir región de rechazo

Utilizamos un estadístico de prueba

$$T = \sqrt{n}\left(\frac{\bar{x} - \mu_0}{s}\right) \sim t_{(n-1)}$$

Suponiendo H_0 verdadera

Si T es grande (positivo o negativo) nos da evidencia para rechazar H_0

$$\begin{matrix} 5\% \\ 2\% \\ 1\% \end{matrix} \leftarrow \alpha = \mathbb{P}\Big(\underbrace{|T| > k}_{\text{Rechazar } H_0} \,\Big|\, H_0 \text{ Verdadera}\Big)$$

$$k = t_{(n-1)}^{\alpha/2}$$

Región de Rechazo: $H_0: \mu = \mu_0$

Si $|T| > t_{(n-1)}^{\alpha/2}$ hay evidencia para rechazar H_0 a un nivel de significancia α

Regresión Lineal Simple

$x \rightarrow$ Predictor (Variable Independiente) //Estatura// $x_1, x_2, x_3 \ldots$

$y \rightarrow$ Respuesta (Variable Dependiente) //Peso// $y_1, y_2, y_3 \ldots$

Análisis Individual

Media Muestral $\quad \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

Varianza Muestral $\quad S_x^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad S_y^2 = \dfrac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$

Desviación Estándar $\quad S_x = \sqrt{S_x^2} \qquad S_y = \sqrt{S_y^2}$

Covarianza Muestral

$$\widehat{Cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{S_{xy}}{n-1} = \widehat{Cov}(y,x)$$

Si $\widehat{Cov}(x,y) > 0 \implies$ Positiva Proporcional

Si $\widehat{Cov}(x,y) < 0 \implies$ Negativa Proporcional

Coeficiente de Correlación

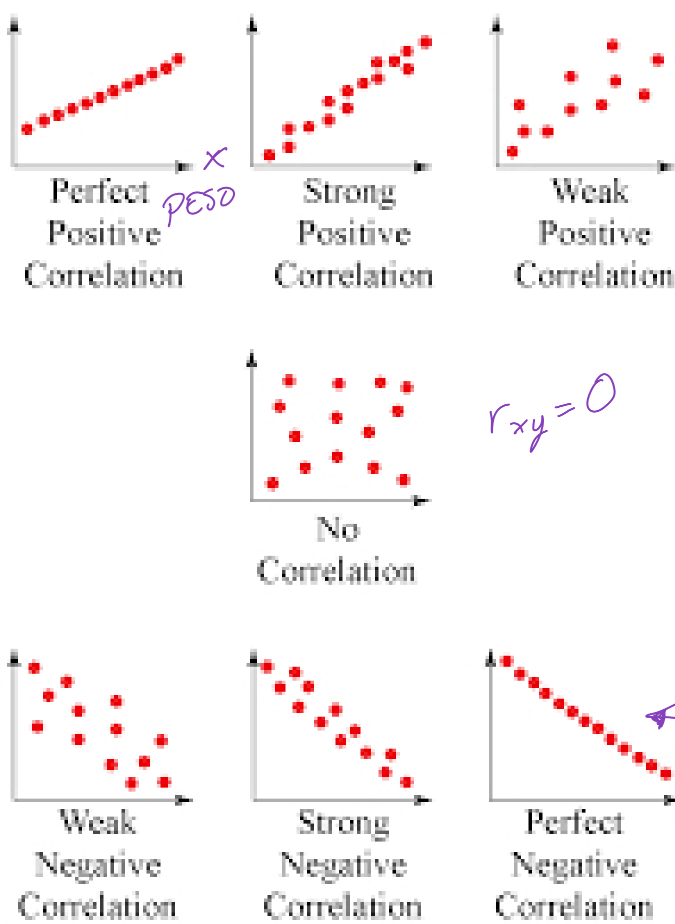
$$r_{xy} = \frac{\widehat{Cov}(x,y)}{S_x S_y}$$

$$-1 \leq r_{xy} \leq 1$$

ESTATURA y

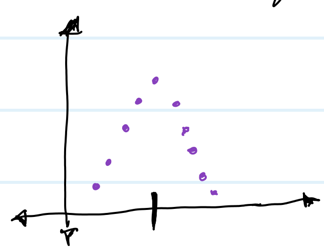
$r_{xy} = 1$

x PESO



$r_{xy} = 0$

$r_{xy} = 0$



$r_{xy} = -1$

PODEMOS PROPONER UNA RELACIÓN PARAMÉTRICA

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\varepsilon \to$ ERROR ESTOCÁSTICO // ALEATORIO //

$\beta_1 \to$ COEFICIENTE

CONSIDERA 5 SUPUESTOS:

(1) LA RELACIÓN VERDADERA ES $Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{CONOCIDO}} + \varepsilon$ $\to$ ALEATORIEDAD

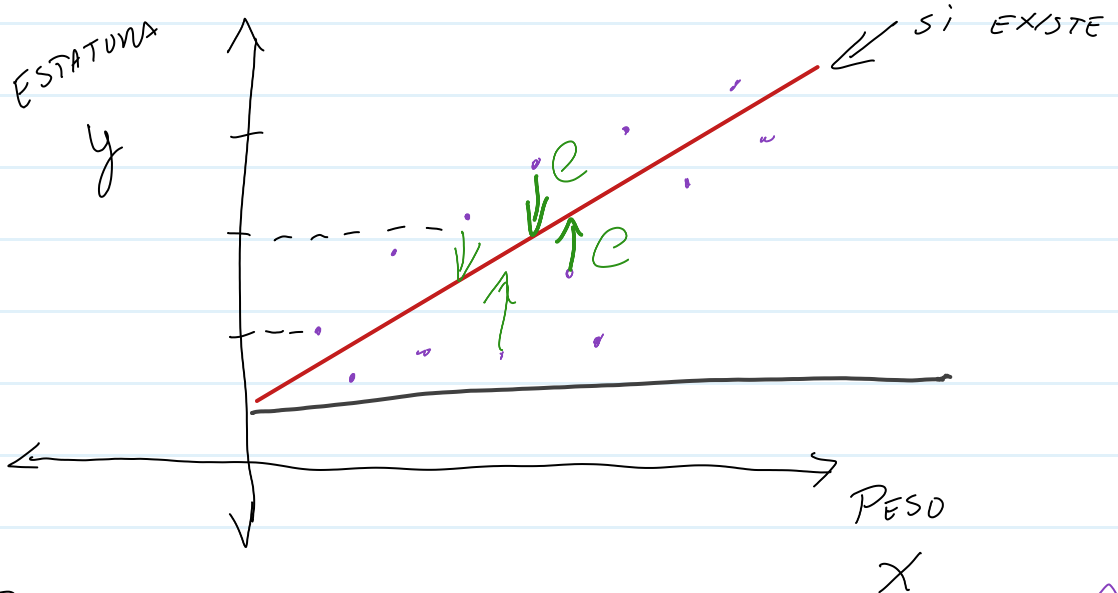
(2) X NO ES ESTOCÁSTICA // NO TIENE UN COMPORTAMIENTO ALEATORIO //

(3) ε_i SON VARIABLES ALEATORIAS : $\mathbb{E}[\varepsilon_i] = 0$

$Var[\varepsilon_i] = \sigma^2$

(4) ε_i SON INDEPENDIENTES E IDÉNTICAMENTE DISTRIBUIDOS (i.i.d)

(5) ε_i SON NORMALES $\varepsilon_i \sim N(0, \sigma^2)$



ESTIMACIÓN POR MÍNIMOS CUADRADOS:

$\Rightarrow$ PROPONEMOS $\hat{\beta}_0$, $\hat{\beta}_1$ Y CALCULAR

$$\min_{\hat{\beta}_0, \hat{\beta}_1 \in \mathbb{R}} \sum_{i=1}^{n} \Big(\underbrace{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\text{RESIDUALES } e_i} \Big)^2$$

$\hat{\beta}_0 + \hat{\beta}_1 x_i \to \hat{y}_i$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{r_{xy} \cdot S_y}{S_x}$$

ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

YA CONOCEMOS LA DISTRIBUCIÓN Y VAMOS A BUSCAR PARÁMETROS QUE LLEVEN A LA DISTRIBUCIÓN MÁS PARECIDA A NUESTROS DATOS

VEROSIMILITUD:

$$\ell(\beta_0, \beta_1, \sigma^2) = f(y_1, y_2, \dots, y_n) = \underbrace{f(y_1) f(y_2) \dots f(y_n)}_{\text{INDEPENDENCIA}}$$

FUNCIÓN DE DENSIDAD

$$= \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{y_i - \mu_i}{2\sigma^2}\right)^2}$$

NO SON IDÉNTICAMENTE DISTRIBUIDAS

LOG VEROSIMILITUD

$$L = \ln(\ell(\beta_0, \beta_1, \sigma^2))$$

¿QUÉ PARÁMETROS ARROJA?

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

RESIDUALES

σ^2_{MV} NO ES INSESGADA $\mathbb{E}[\sigma^2_{MV}] \neq \sigma^2$

$\Rightarrow$ RESTRICCIONES DE LOS RESIDUALES

$$\left.\begin{array}{l} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{array}\right\}$$

2 RESTRICCIONES

SI CONOZCO $n-2$ RESIDUALES PUEDO CONOCER LOS 2 QUE FALTAN

$n-2$: GRADOS DE LIBERTAD DE RESIDUALES

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = S^2} \longrightarrow$$

SI ES INSESGADO

$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

VARIANZAS

ESTIMADAS

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \longrightarrow \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \longrightarrow \widehat{Var}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$Var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-2} \longrightarrow \widehat{Var}(\hat{\sigma}^2) = \frac{2\hat{\sigma}^4}{n-2}$$

INTERVALOS

$$\beta_j \in \left[\hat{\beta}_j - S_e(\hat{\beta}_j) \cdot t_{(n-2)}^{(\alpha/2)}, \hat{\beta}_j + S_e(\hat{\beta}_j)\, t_{(n-2)}^{(\alpha/2)}\right]$$

PRUEBA DE HIPÓTESIS

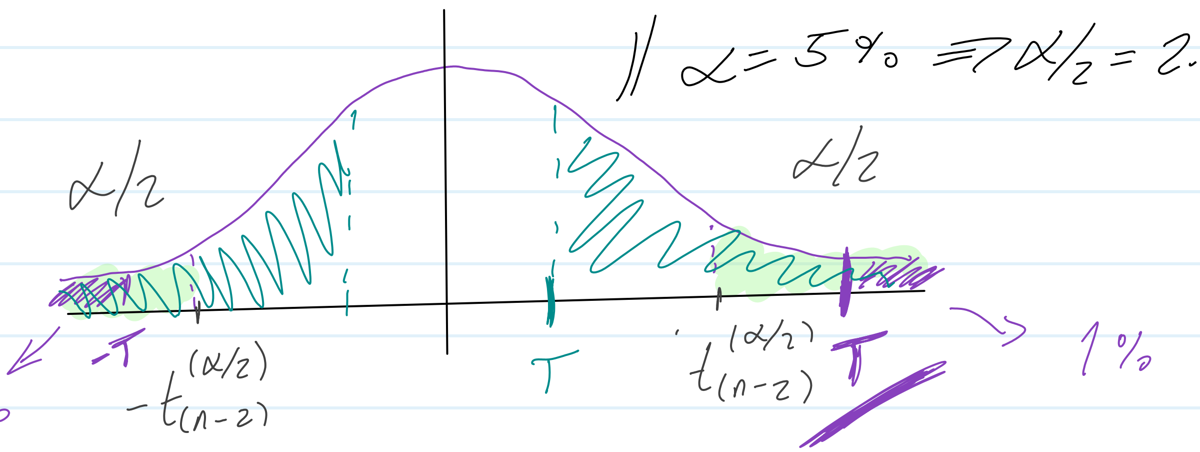
$$Y_i = \beta_0 + \overset{\overset{0}{/\!/}}{\beta_1} x_i + \varepsilon$$

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

ESTADÍSTICO DE PRUEBA $T = \dfrac{\hat{\beta}_i}{Se(\hat{\beta}_i)} \sim t_{(n-2)}$

ERROR ESTÁNDAR

$$Se(\hat{\beta}_1) = \sqrt{\hat{Var}(\hat{\beta}_1)}$$
$$Se(\hat{\beta}_0) = \sqrt{\hat{Var}(\hat{\beta}_0)}$$
$$Se(\hat{\sigma}^2) = \sqrt{Var(\hat{\sigma}^2)}$$



// $\alpha = 5\% \Rightarrow \alpha/2 = 2.5\%$ //

$$\alpha = \mathbb{P}(|T| > t^{(\alpha/2)}_{(n-2)} \mid H_0 \text{ ES VERDADERA})$$

P-VALUE

PROBABILIDAD DE OBSERVAR A T AL MENOS TAN GRANDE DADO H_0 CIERTA

SI p-VALUE $\leq \alpha$ → RECHAZO H_0

SI p-VALUE $> \alpha$ → NO RECHAZO H_0

COEFICIENTE DE DETERMINACIÓN

$R^2 = r^2$

↓

REGRESIÓN LINEAL SIMPLE

R^2: % DE VARIABILIDAD QUE NOS EXPLICA LA REGRESIÓN

$R^2 \in [0, 1]$

Regresión Lineal Múltiple

$$Y_i = \beta_0 + \beta_1 x_i + \dots + \beta_n x_n + \varepsilon_i$$

→ MINIMIZAR

PESO

ESTATURA

EDAD

Subajuste → MENOS VARIABLES

- Flexible
- No logre explicar Y

Sobreajuste → MÁS VARIABLES

- No añadir predictores insignificantes
- Suficientemente sencillo

Más variables → R^2 aumenta o se queda igual

R^2 ajustada → ENTRE MÁS GRANDE MEJOR

$$R^2_a = 1 - \frac{(n-1)}{n-k-1}(1-R^2)$$

R^2: Coef. determinación
n: # datos
k: # predictores

Criterios información

$$AIC \text{ (Akaike)} = 2(k+1) - 2\ln(L)$$

$$BIC \text{ (Bayesiano)} = \ln(L)(k+1) - 2\ln(L)$$

ENTRE MÁS PEQUEÑO MEJOR

MÁS RESTRICTIVO

MENOS RESTRICTIVO
POCOS DATOS