Modelos lineales y aditivos en ecología

Facundo X. Palacio

facundo_palacio@fcnym.unlp.edu.ar

CONICET

Facultad de Ciencias Naturales y Museo

SECCIÓN ORNITOLOGÍA · MUSEO DE LA PLATA · FCNYM - UNLP

2 al 6 de mayo de 2022 – Universidad Nacional de Tucumán

Temas del curso

- Modelos lineales: regresión, GLMs, GLMMs
- Modelos aditivos: GAMs, GAMMs
- Otras yerbas: selección de modelos, modelos no lineales

Dinámica

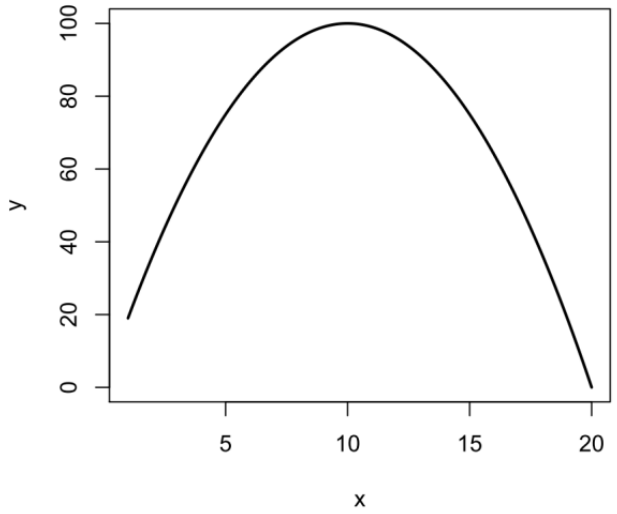
- Mañana (9-12): teórico-práctico
- Tarde (13-17): práctico

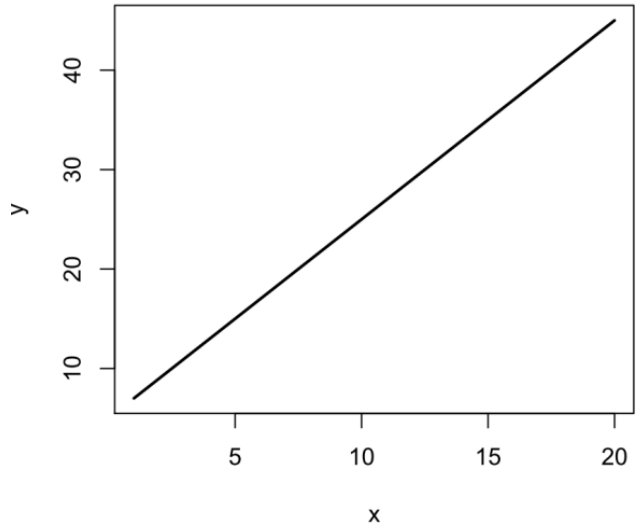
1. Lenguaje de programación.
2. Software más comprensivo (01/22 18728 paquetes).
3. Amplias posibilidades gráficas.
4. Amplia comunidad de usuarios.
5. Funciona bajo Linux, Windows y MacOS
6. Es libre, puede ser copiado, distribuido y modificado a voluntad.
7. Es gratis.

¿Qué es un modelo lineal?

- Depende de qué estemos hablando…
- Matemático: ecuación que describe la relación entre cantidades que cambian de forma proporcional = regresión lineal

$$\hat{y} = b_0 + b_1 x$$





$$\hat{y} = b_0 + b_1 x + b_2 x^2 = c + bx + ax^2$$

Modelos lineales

Objetivos:

-Describir una relación **lineal**

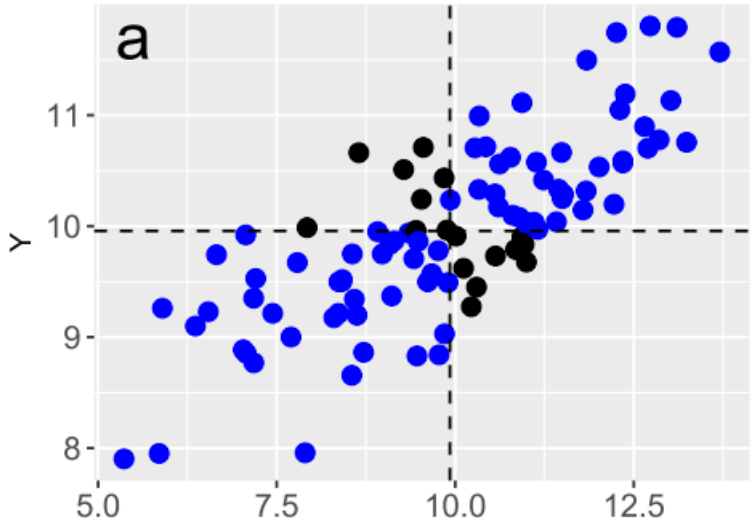
-Determinar qué variables son importantes en una relación lineal

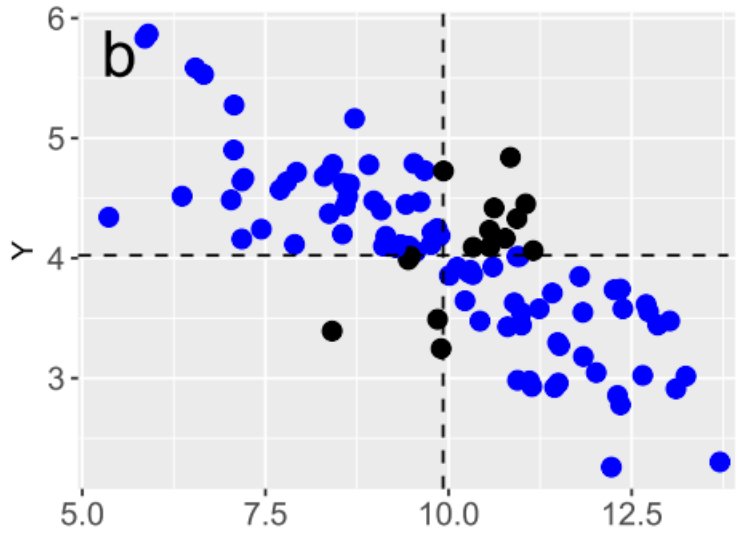
-Predecir nuevos valores de y

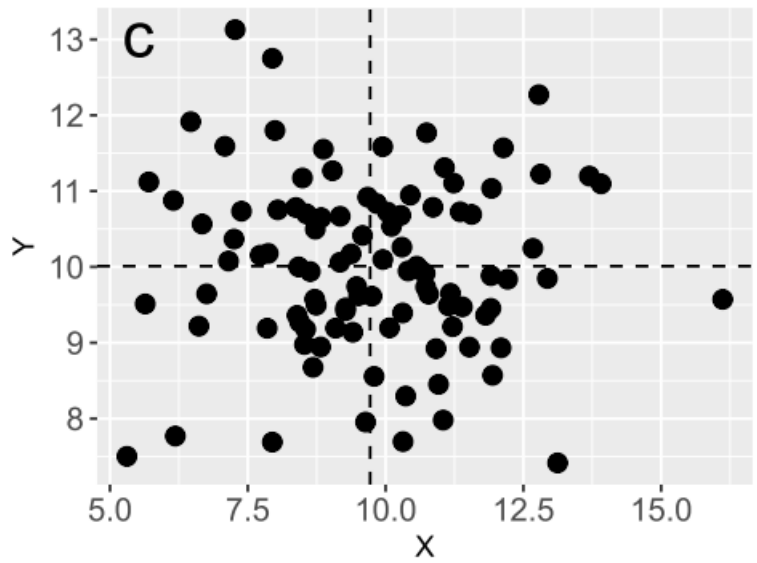
Un paréntesis… asociación entre variables

Covarianza: medida de asociación lineal entre 2 variables

$$S_{xy} = \mathrm{cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum x' y'}{\sum (x')^2}$$





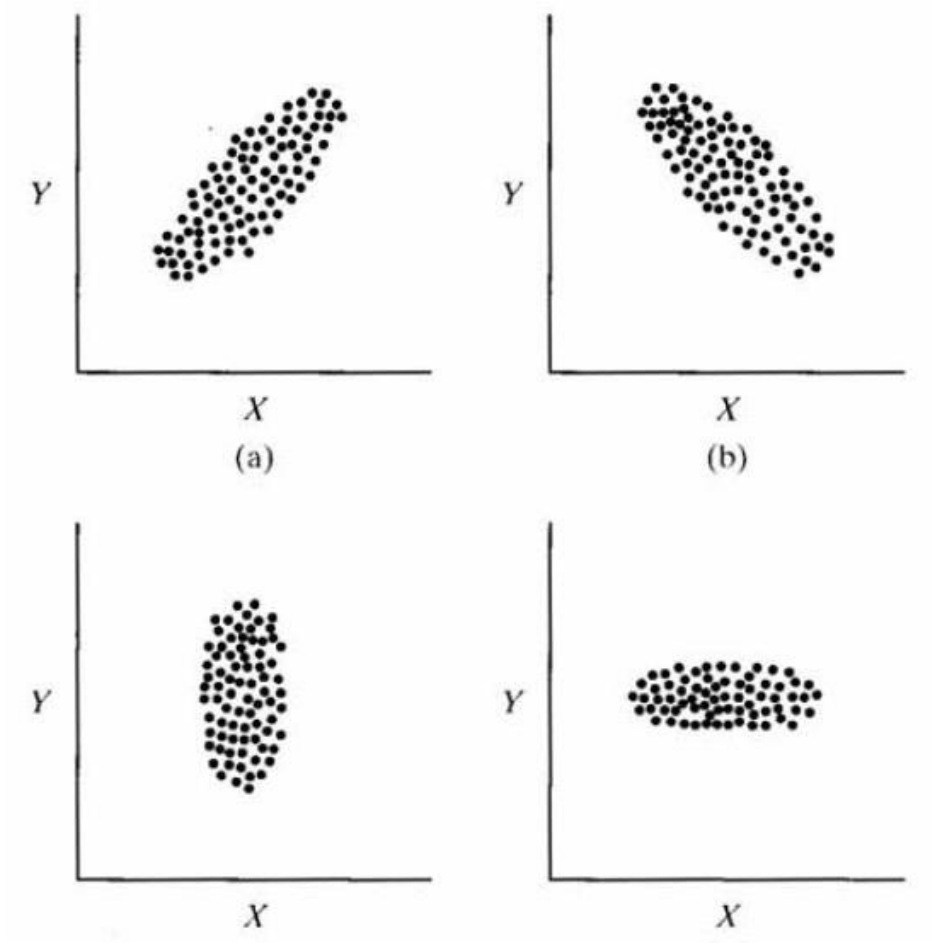


Coeficiente de correlación momento-producto de Pearson

Medida de asociación lineal entre 2 variables

$$r = \frac{\mathrm{cov}(x, y)}{S_x S_y}$$

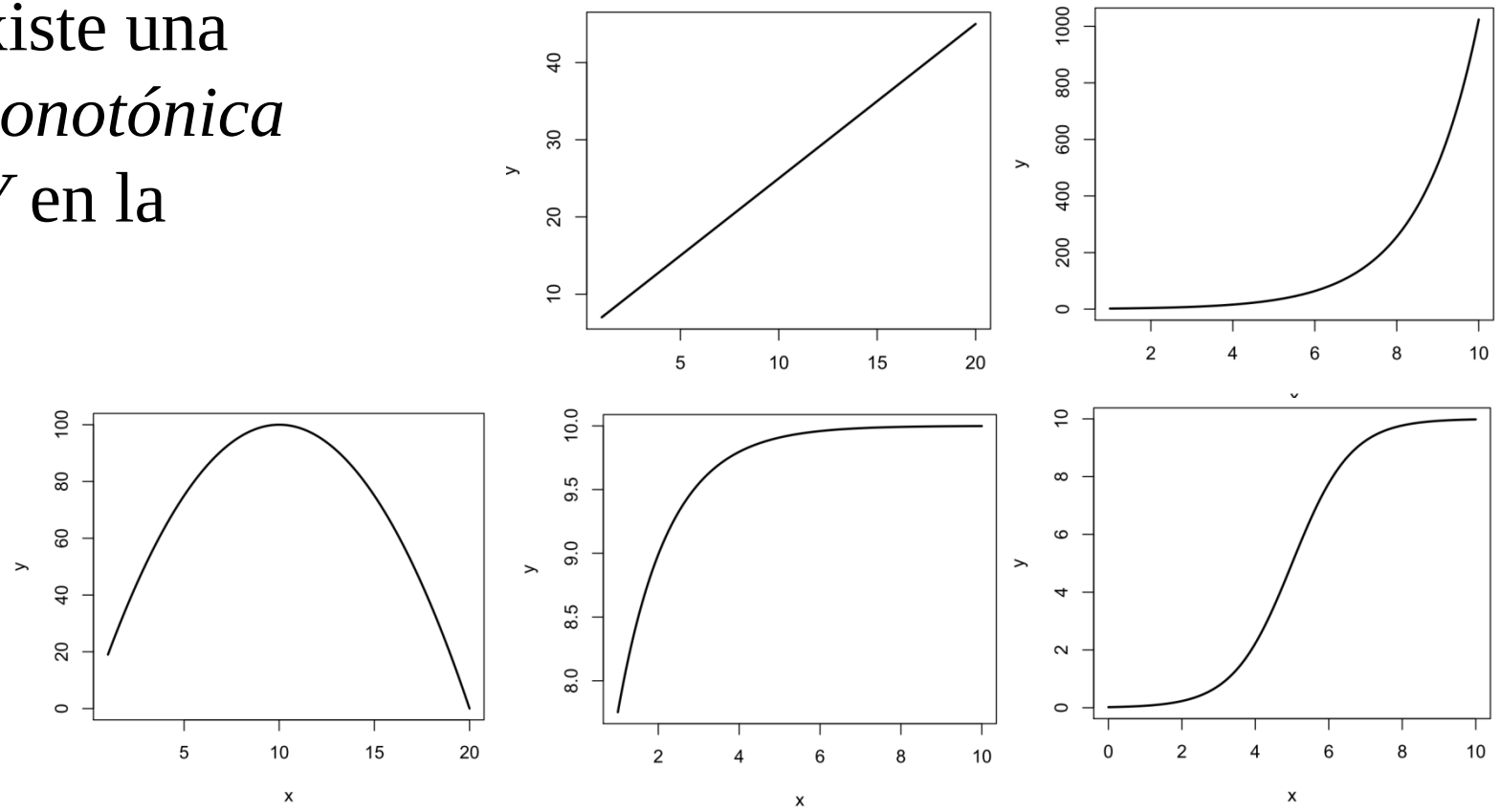
$$-1 < r < +1$$



Coeficiente de correlación de Spearman

- Se transforman las variables a rangos y se calcula el coeficiente de correlación de Pearson.

H_0 = no existe una relación *monotónica* entre X e Y en la población.



Matriz de correlación

	Altura	DAP	Densidad	Cobertura
Altura	1	0.6	-0.3	0.7
DAP	0.6	1	0.1	0.4
Densidad	-0.3	0.1	1	-0.3
Cobertura	0.7	0.4	-0.3	1

Regresión lineal simple

Objetivos:

- Describir la relación lineal entre x e y.
- Determinar cuánto de la variación total en y puede ser explicada por la relación lineal con x, y cuánto de esta variación permanece sin explicar.
- Predecir nuevos valores de y a partir de nuevos valores de x.

Regresión lineal simple

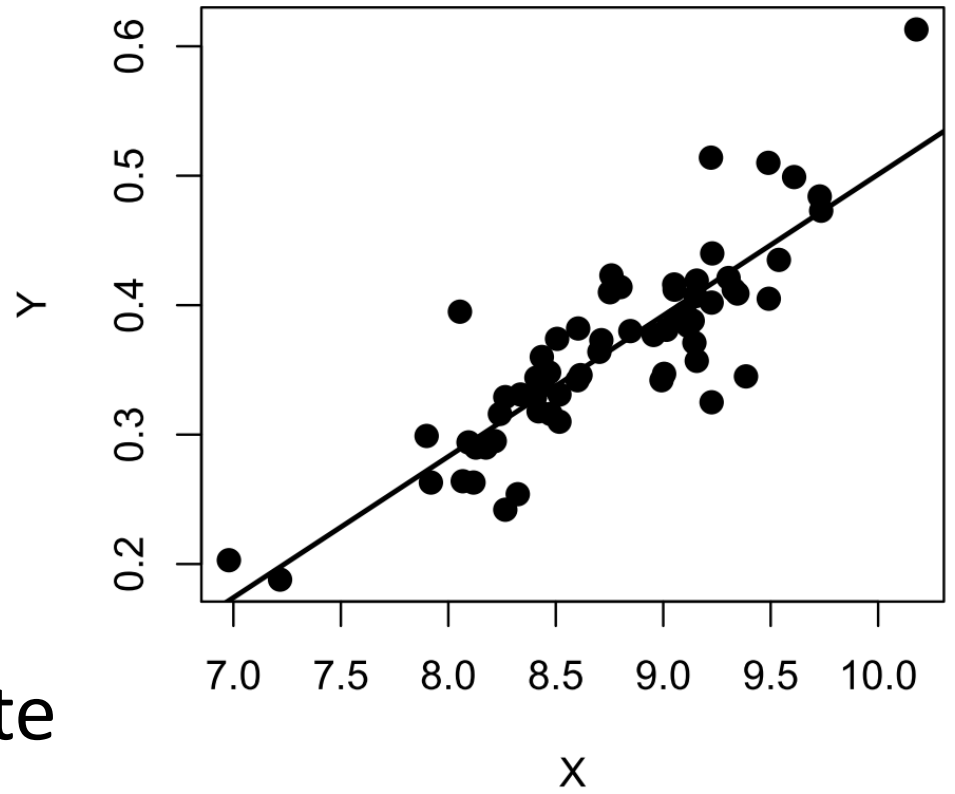
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Modelo Error

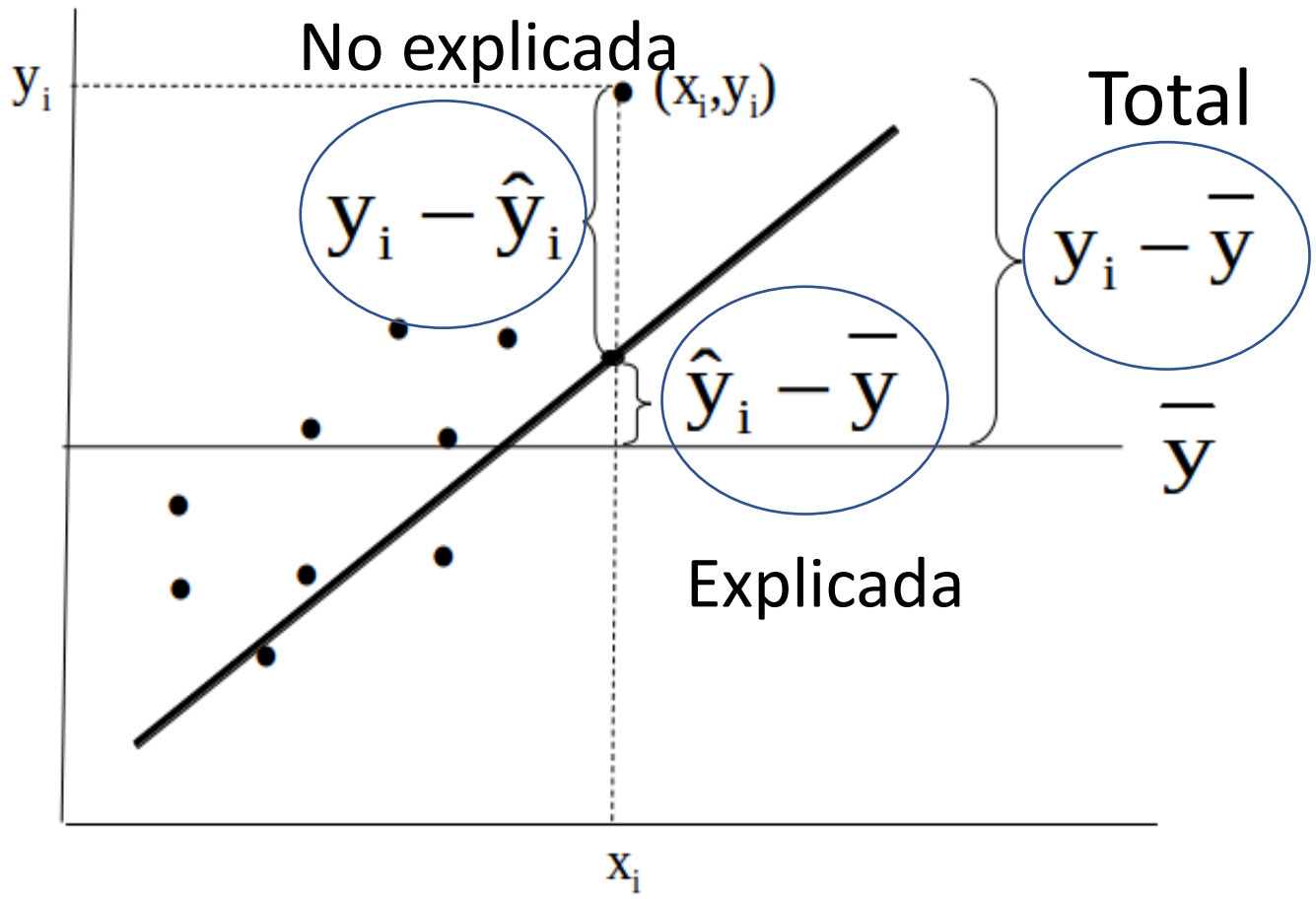
$$\hat{y} = b_0 + b_1 x$$

Ordenada al origen

Pendiente



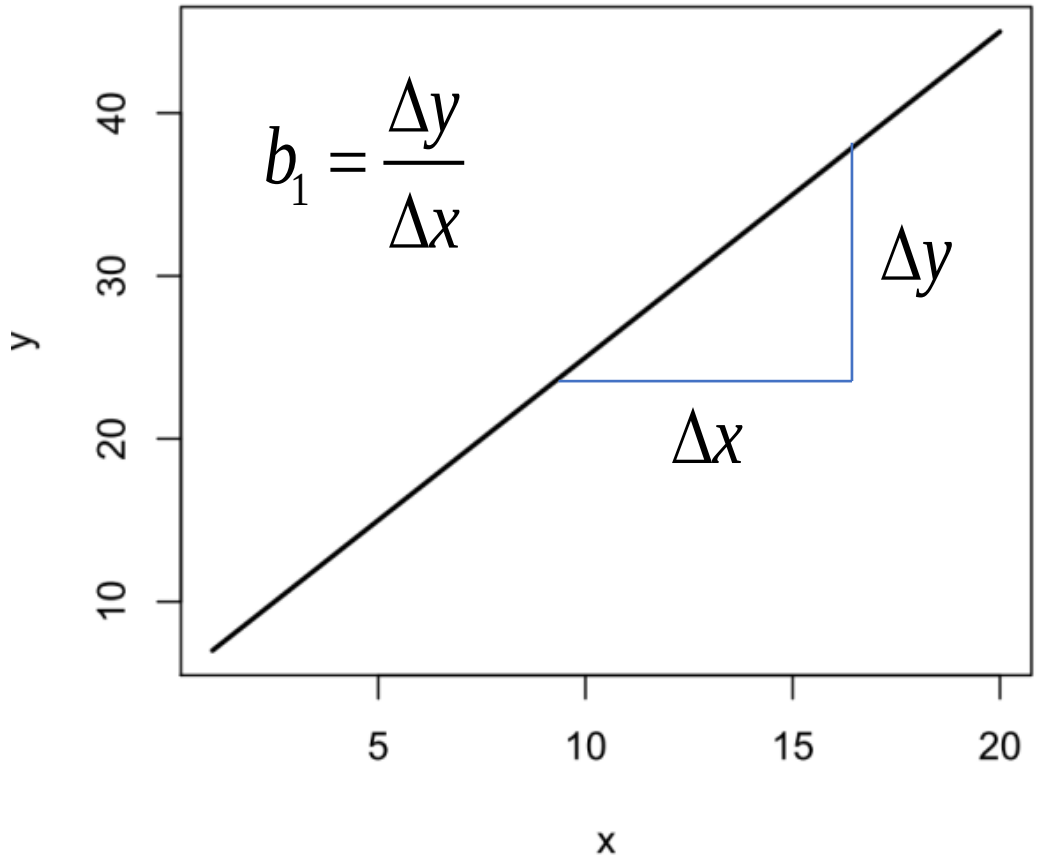
Mínimos cuadrados



$$\hat{y} = b_0 + b_1 x \quad \Rightarrow \quad \min = \sum (y_i - \hat{y}_i)^2$$

Cálculo de la pendiente

$$b_1 = \frac{\mathrm{cov}(x, y)}{S_x^2}$$



Regresión lineal simple

- **Forma 1**

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

$$t = \frac{b_1 - \beta_1}{S_b}$$

- **Forma 2**

$$H_0 : \sigma_E^2 \leq \sigma_{NE}^2$$
$$H_1 : \sigma_E^2 > \sigma_{NE}^2$$

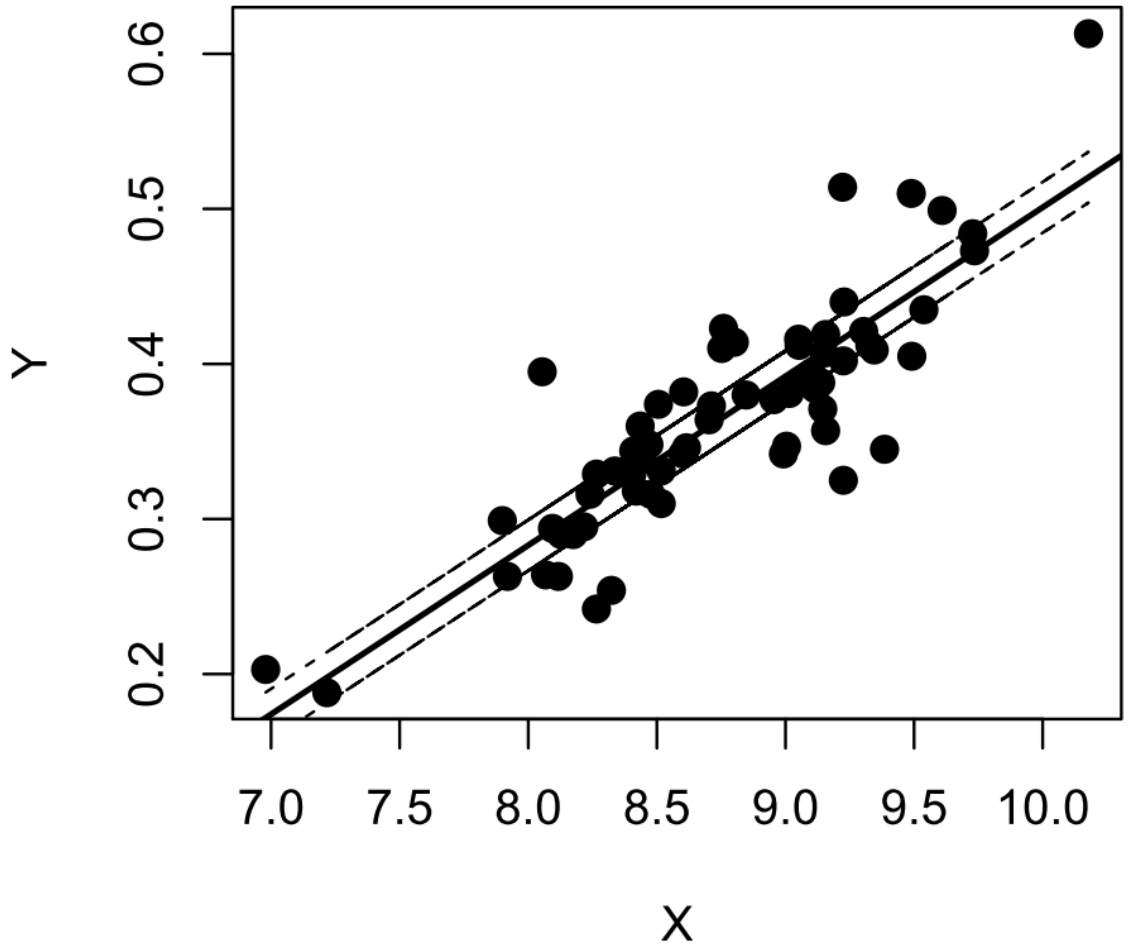
$$F = \frac{S_E^2}{S_{NE}^2}$$

Medidas de bondad de ajuste

- **Error estándar de la estimación**

$$LC = \hat{y}_i \pm 1.96 S_e$$

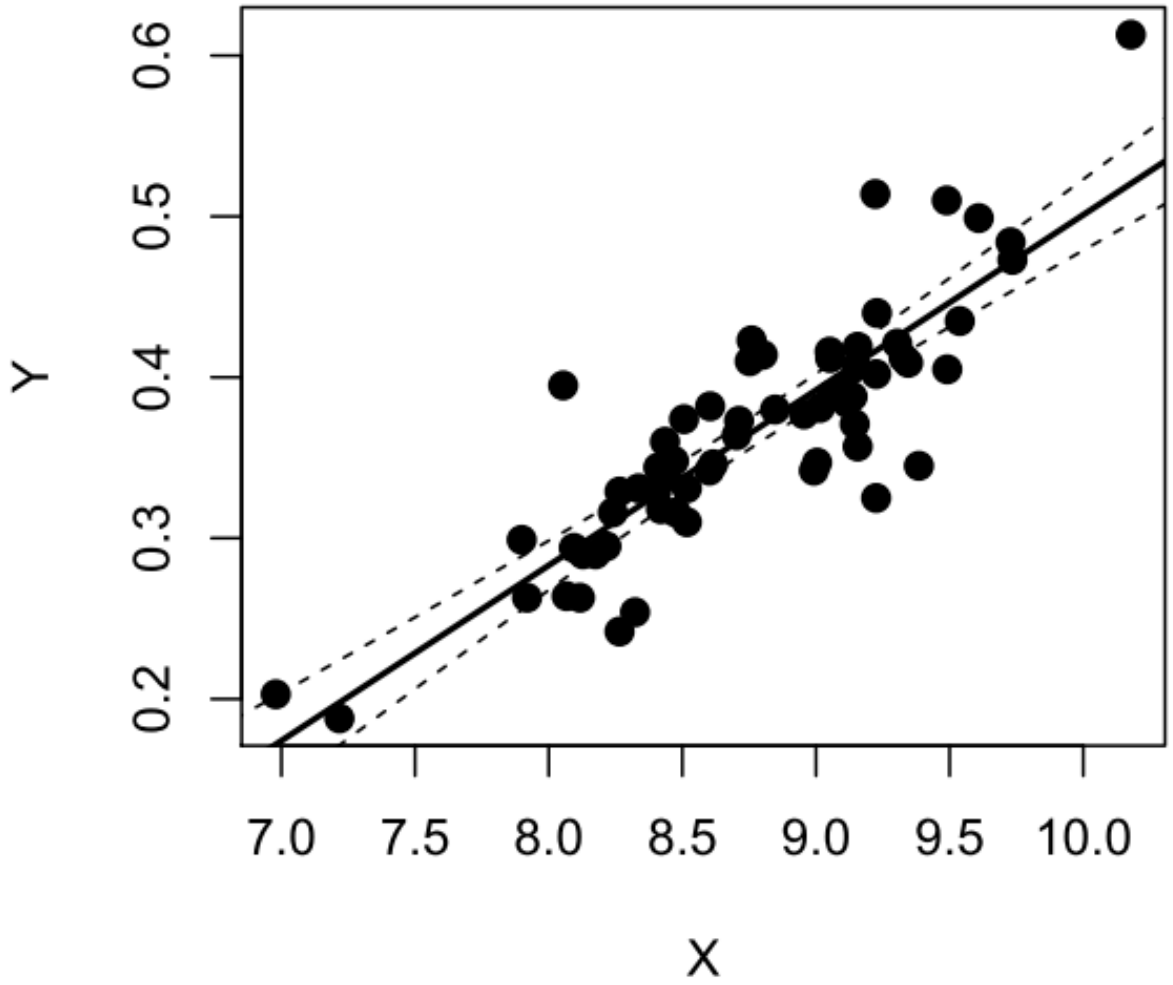
Regiones de confianza determinadas por rectas



No confundir. Error estándar de la predicción

$$LC = \hat{y}_{ii} \pm 1.96 S_p$$

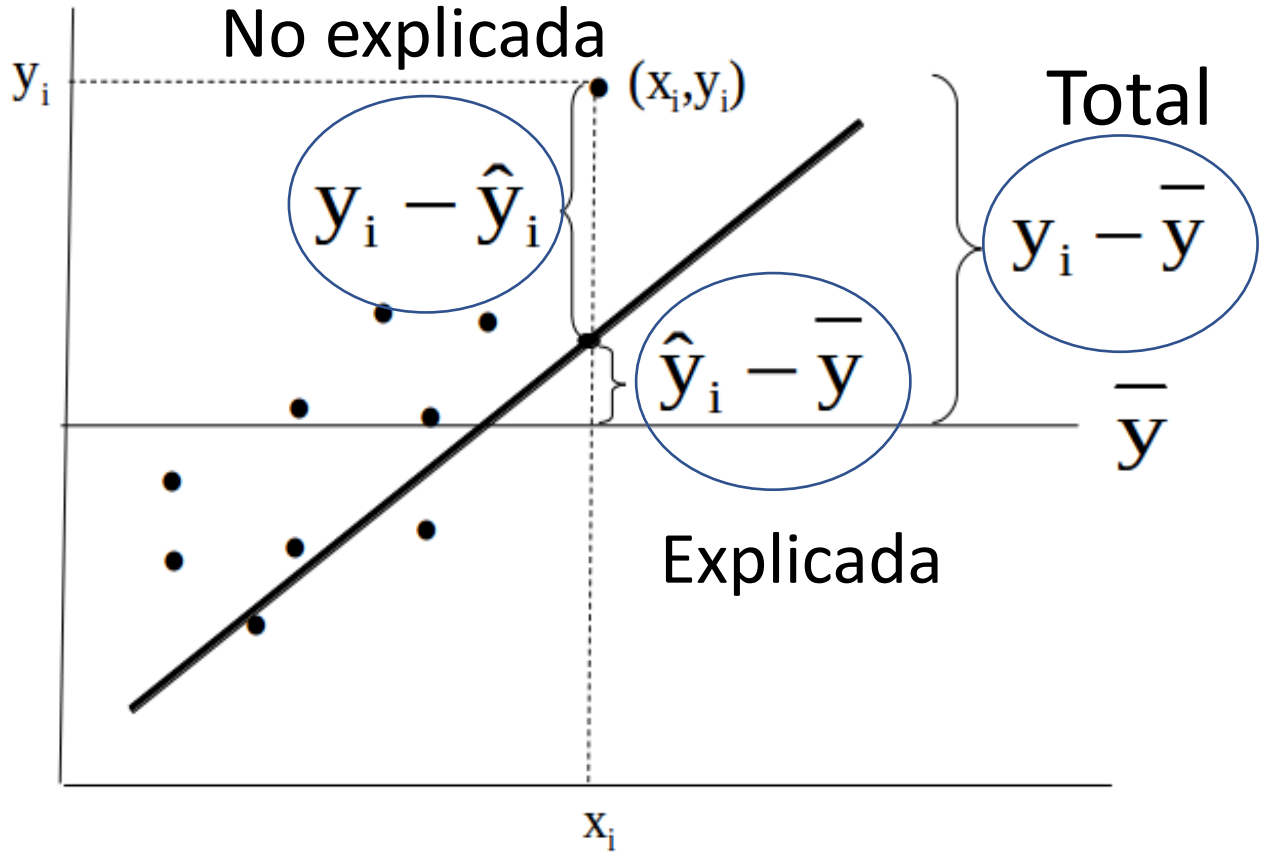
Regiones de confianza determinadas por curvas



Bondad de ajuste

- **Coeficiente de determinación**

$$r^2 = \frac{SC_E}{SC_T} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$



Relación entre regresión y correlación

$$r = \frac{\mathrm{cov}(x, y)}{S_x S_y} = b \frac{S_x}{S_y}$$

$$|r| = \sqrt{r^2}$$

Para variables estandarizadas, $r = b$

Regresión lineal múltiple

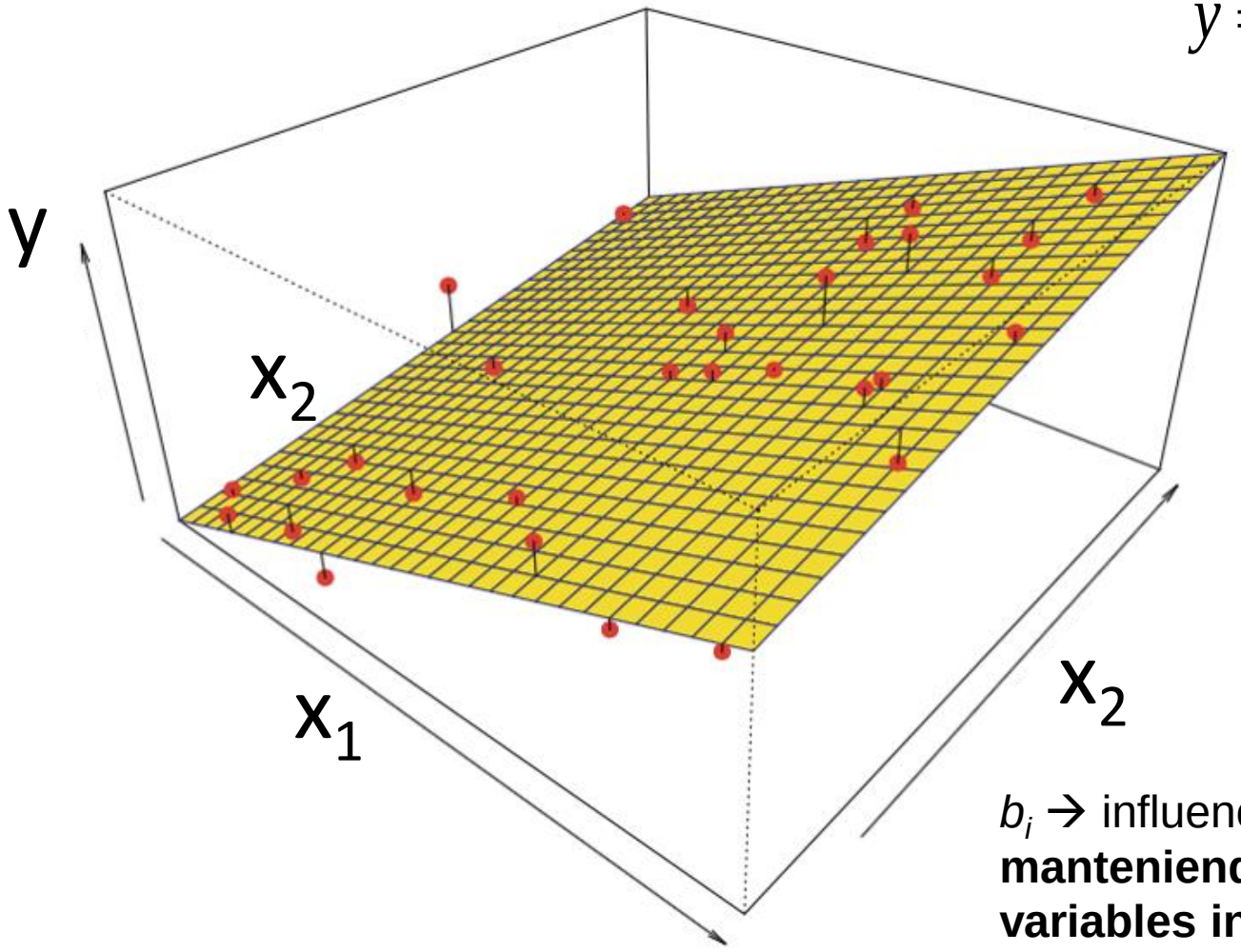
$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}_{\text{Modelo}} + \underbrace{\varepsilon}_{\text{Error}}$$

β_0 → ordenada al origen.

β_i → coeficientes de regresión parcial.

Regresión lineal múltiple

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$



Precio = b0 + b1*marca
Precio = b0 + b1*marca - b2*años

b_i → influencia de cada variable independiente sobre y, **manteniendo constante la influencia de las otras variables independientes**.

Medidas de bondad de ajuste

- **Coeficiente de determinación múltiple**

$$R^2 = \frac{SC_E}{SC_T}$$

Problemas:

- Permanece igual o aumenta con la inclusión de una nueva variable.
- No apropiado para comparar modelos con distinto nº de variables

- **Coeficiente de determinación múltiple ajustado**

$$R_a^2 = 1 - \frac{n-1}{n-k-1}\left(1-R^2\right)$$

Regresión lineal múltiple

- Test sobre el modelo global

$$H_0 : \rho^2 = 0$$

$$H_1 : \rho^2 \neq 0$$

Ó

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

H_1: al menos un par de β_i difiere

- Test sobre los coeficientes

$$t_i = \frac{b_i - \beta_i}{S_{b_i}}$$

Variables categóricas

- Variables indicadoras o “dummies”: variables categóricas dicotómicas.

$$Longitud = b_0 + b_1 Ancho + b_2 Especie$$

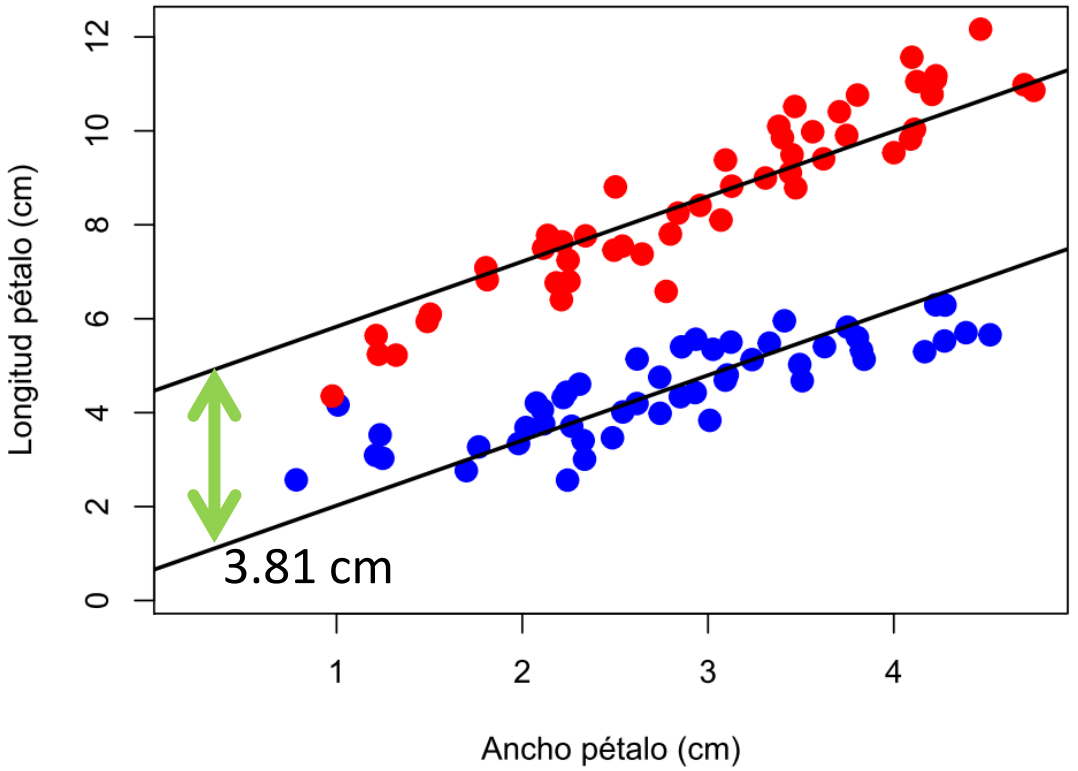
$$Longitud = 0.63 + 1.39 Ancho + 3.81 Especie$$

$$Longitud_{SpA} = 0.63 + 1.39 Ancho$$

$$Longitud_{SpB} = 0.63 + 1.39 Ancho + 3.81$$

$$= 4.44 + 1.39 Ancho$$

Sp A (azul) = 0
Sp B (rojo) = 1



Variables categóricas

$$Longitud = b_0 + b_1 Ancho + b_2 Especie + b_3 Ancho \times Especie$$

$$Longitud = 1.96 + 0.91 Ancho + 1.23 Especie + 0.89 Ancho \times Especie$$

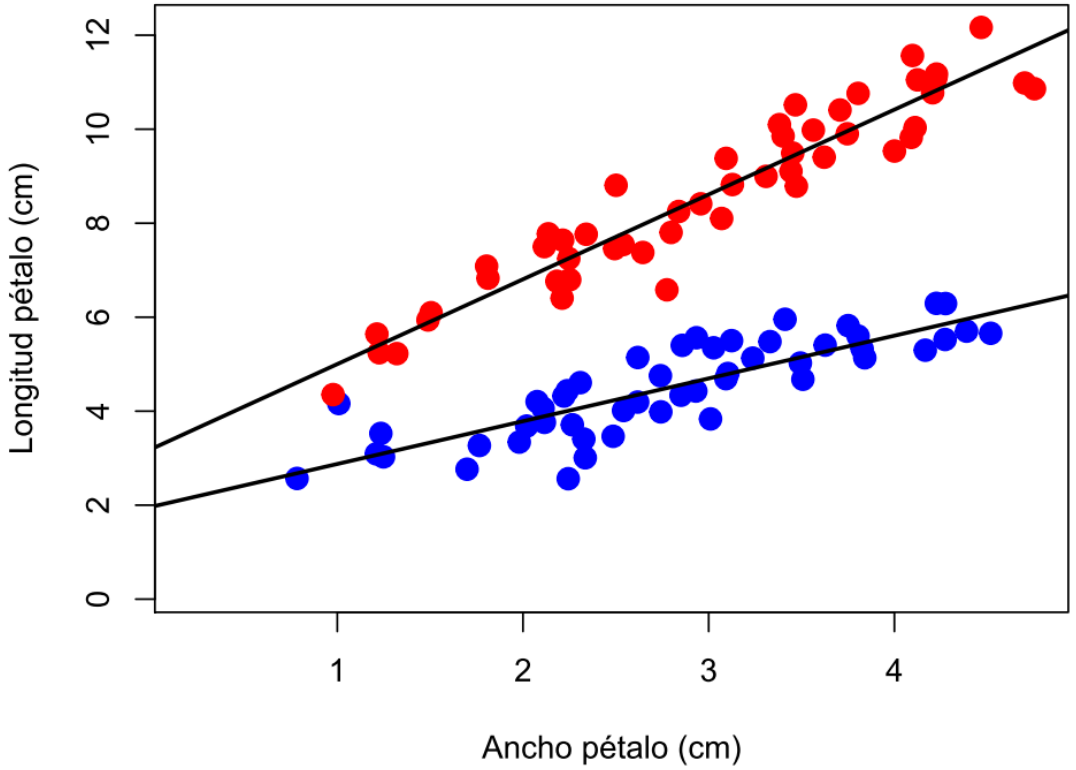
Sp A (azul) = 0
Sp B (rojo) = 1

$$Longitud_{SpA} = 1.96 + 0.91 Ancho$$

$$Longitud_{SpB} = 1.96 + 0.91 Ancho + 1.23 + 0.89 Ancho$$

$$= 3.19 + (0.91 + 0.89) Ancho$$

$$= 3.19 + 1.8 Ancho$$



Variables indicadoras

$$Longitud = b_0 + b_1 Ancho + b_2 EspecieA + b_3 EspecieB$$

Individuo	*Especie*
1	A
2	B
3	C

EspecieA	*EspecieB*	*Especie C*
1	0	0
0	1	0
0	0	1

EspecieA	*EspecieB*	Modelo
1	0	$Longitud = b_0 + b_1 Ancho + b_2 EspecieA$
0	1	$Longitud = b_0 + b_1 Ancho + b_3 EspecieB$
0	0	$Longitud = b_0 + b_1 Ancho$ → *EspecieC*

Test de *t*

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

$$t = \frac{\left(\bar{X}_A - \bar{X}_B\right) - \left(\mu_A - \mu_B\right)}{S_{\bar{X}_A - \bar{X}_B}}$$

$$t = \frac{8.75 - 9.74}{0.40}$$

$$t = -2.48$$

Droga A	Droga B
8.8	9.9
8.4	9.0
7.9	11.1
8.7	9.6
9.1	8.7
9.6	10.4
Media = 8.75	**Media** = 9.74
S = 0.58	**S** = 0.82

Test de *t* pareado

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

$$t = \frac{\bar{d} - \mu_d}{S_{\bar{d}}}$$

$$t = \frac{0.033}{0.007}$$

$$t = 3.41$$

Nutriente 1	Nutriente 2	Diferencia
1.42	1.38	+0.04
1.40	1.36	+0.04
1.44	1.47	-0.03
1.44	1.39	0.05
1.42	1.43	-0.01
1.46	1.41	0.05
1.49	1.43	0.06
1.50	1.45	0.05
1.42	1.36	0.06
1.48	1.46	0.02
Media = 1.45	**Media** = 1.47	**Media** = 0.03
S = 0.3	**S** = 0.4	**S** = 0.03

ANOVA de 1 factor

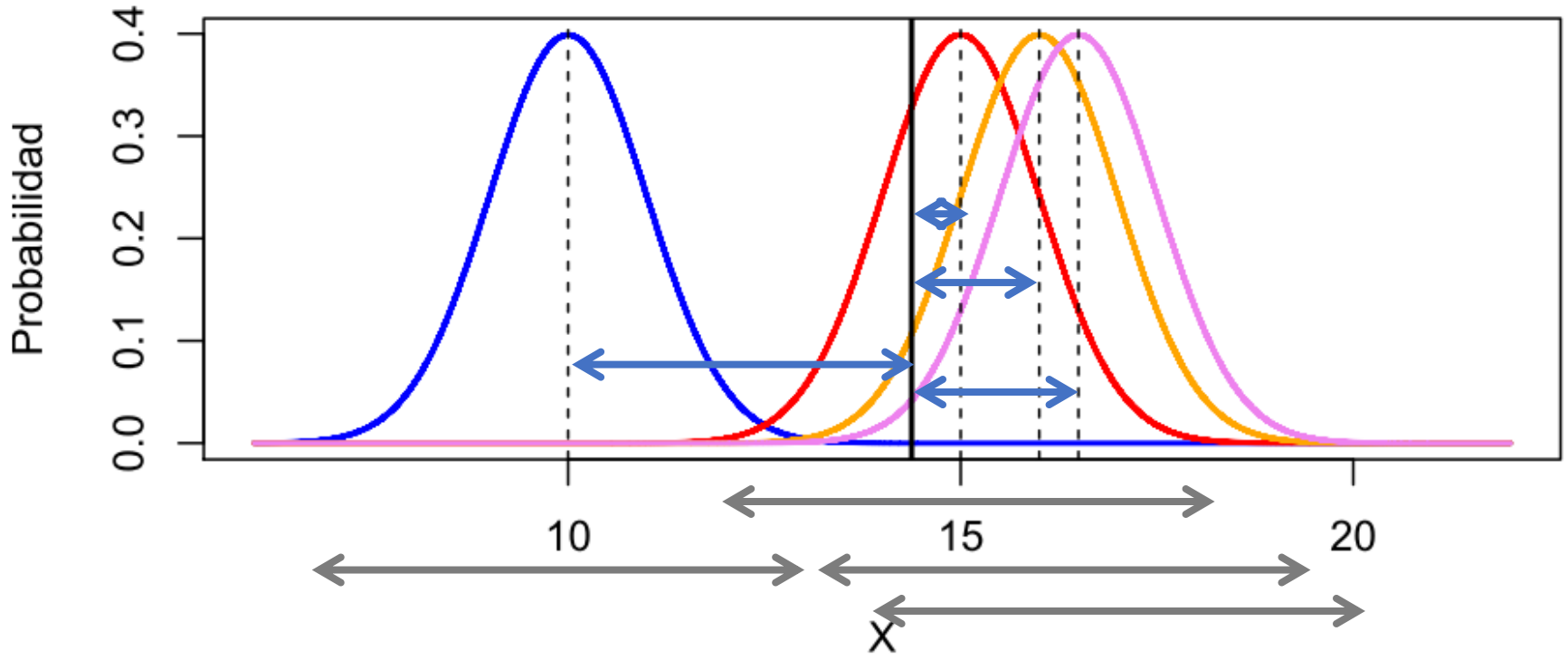
- 3 drogas + 1 control → efecto sobre el crecimiento
- Factor → droga
- Niveles del factor (= tratamientos)

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

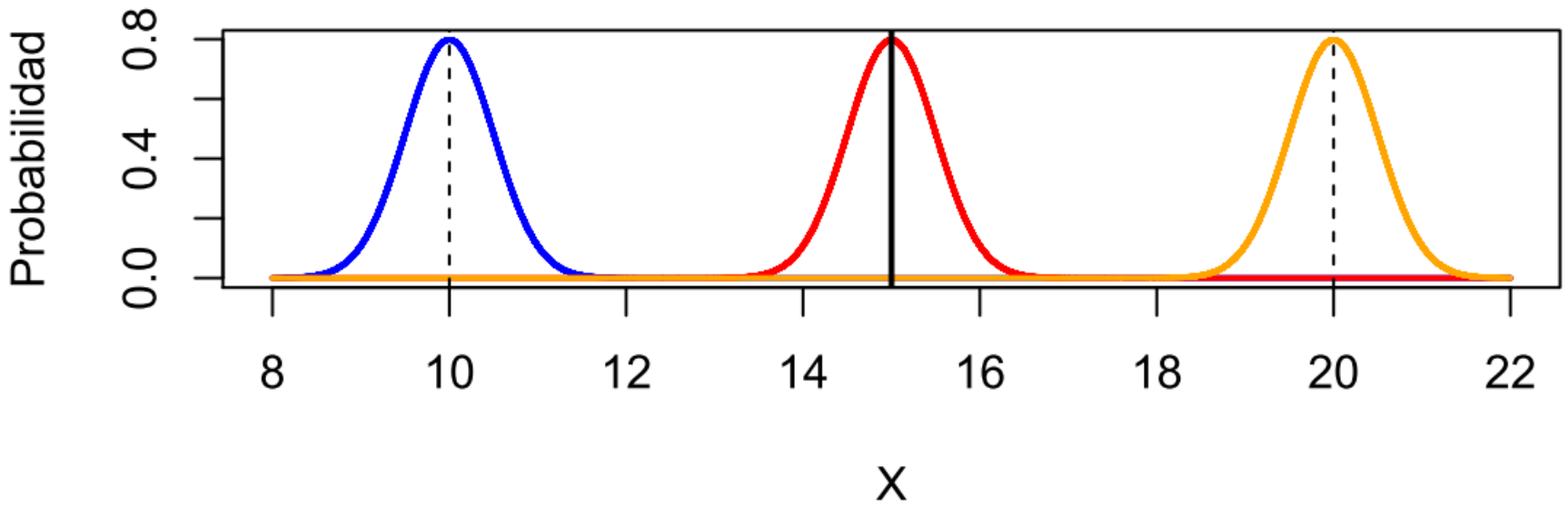
H_1: al menos un par de μ's difiere.

ANOVA de 1 factor

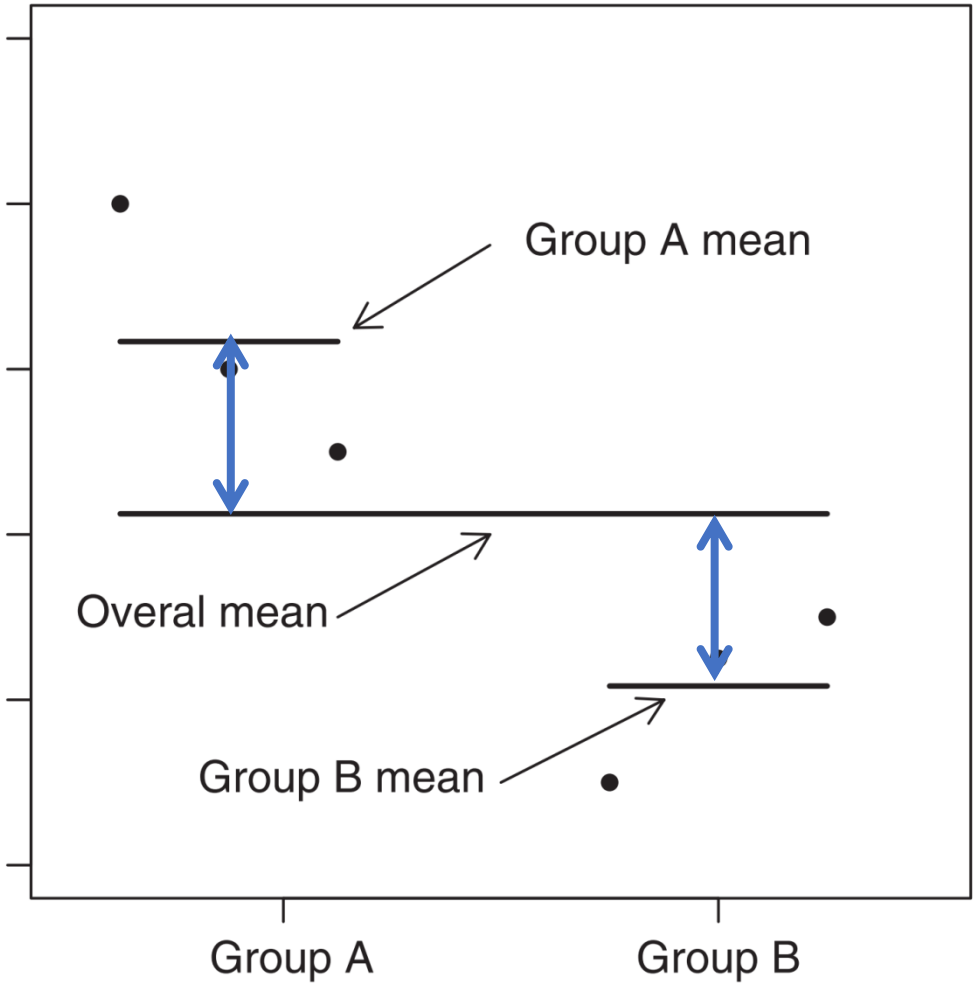
- 3 drogas + 1 control → efecto sobre el crecimiento



ANOVA de 1 factor



ANOVA de 1 factor. Varianza entre grupos



ANOVA de 1 factor. Varianza dentro de grupos

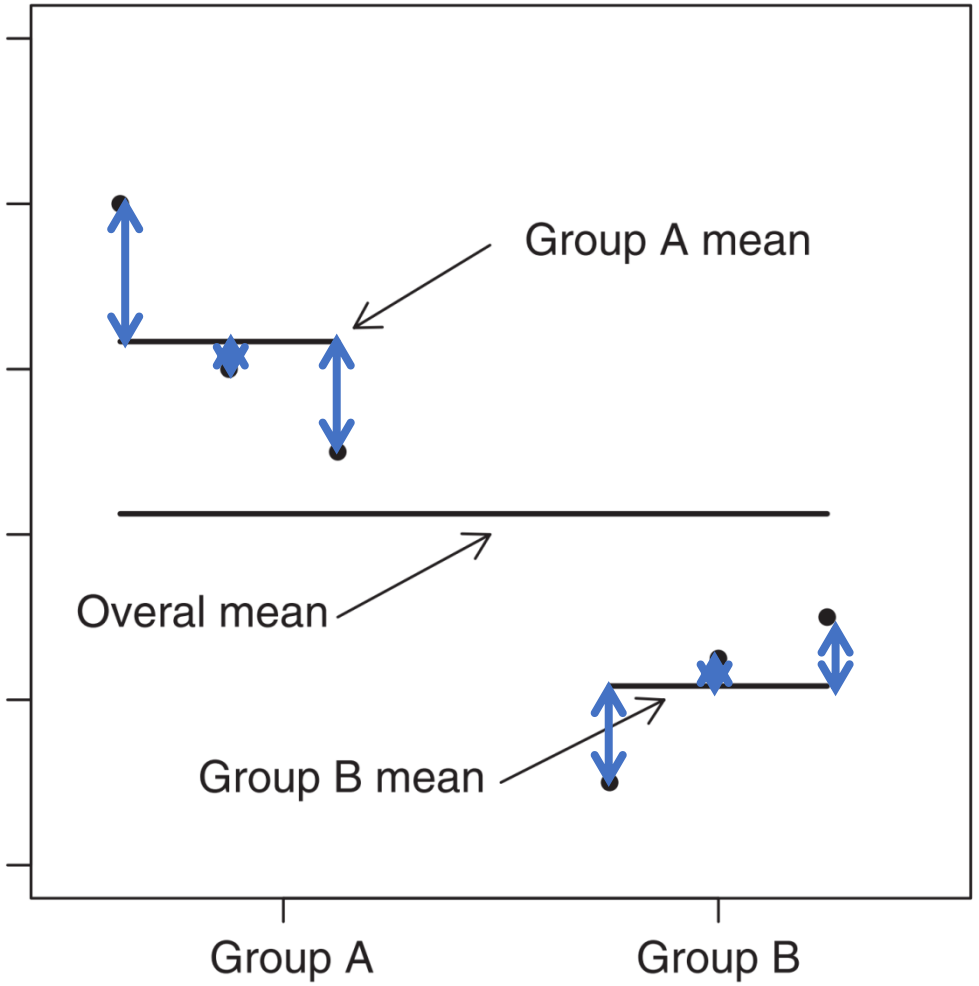


Tabla resumen de ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	*F*	*P*
Entre grupos (factor)	SCE	$gl_E = a - 1$	$S^2_E = SCE / gl_E$	S^2_E / S^2_D	
Dentro de grupos (error)	SCD	$gl_D = a\,(n - 1)$	$S^2_D = SCD / gl_D$		
Total		N - 1			

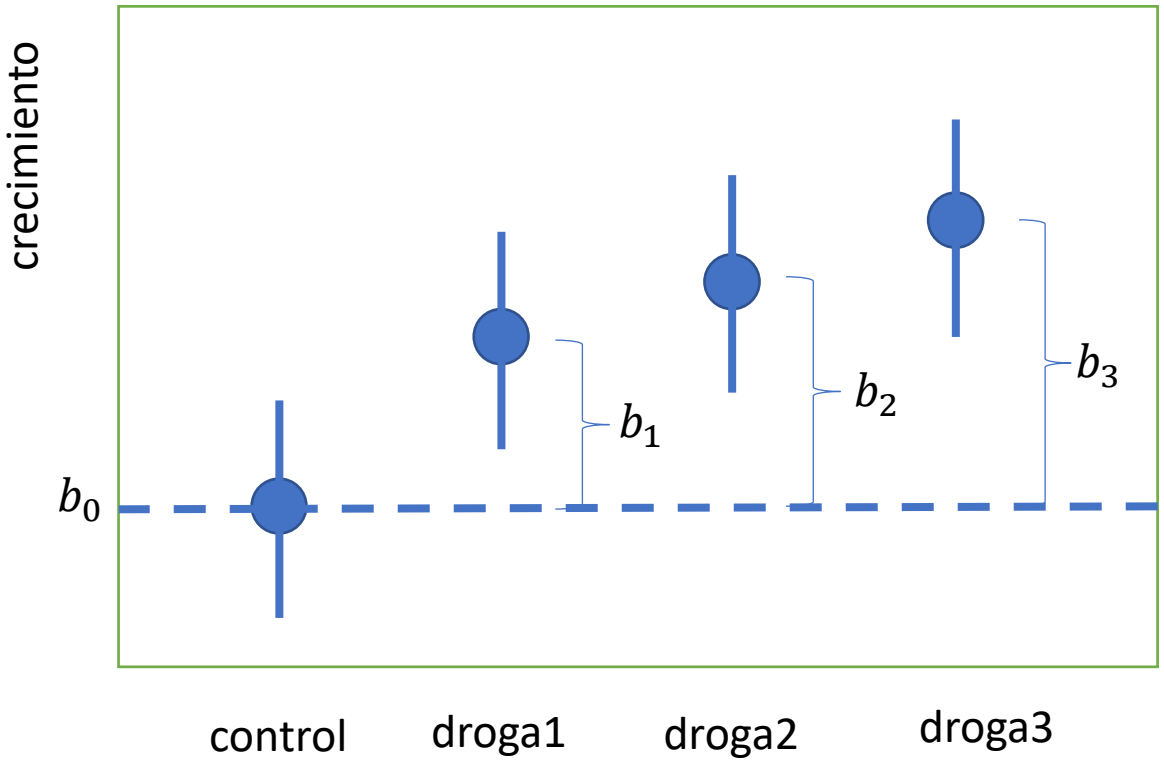
$$H_0 : \sigma_E^2 \leq \sigma_D^2$$

$$H_1 : \sigma_E^2 > \sigma_D^2$$

SCT = SCE + SCD

ANOVA como modelo lineal

$$\text{crecimiento} = b_0 + b_1\text{droga1} + b_2\text{droga2} + b_3\text{droga3}$$



Supuestos

- **ANOVA**

- Distribución normal de los residuos dentro de cada tratamiento.
- Homogeneidad de varianzas dentro de cada tratamiento.
- Observaciones independientes.
- x se mide sin error.

Supuestos
Colinealidad

- Ninguna de las variables independientes es combinación lineal de otras.

$x_1 = \text{radio}$

$x_2 = \text{diámetro} = 2\text{radio}$

$x_2 = 2x_1$

$x_1 = \text{altura}$

$x_2 = \text{dap}$

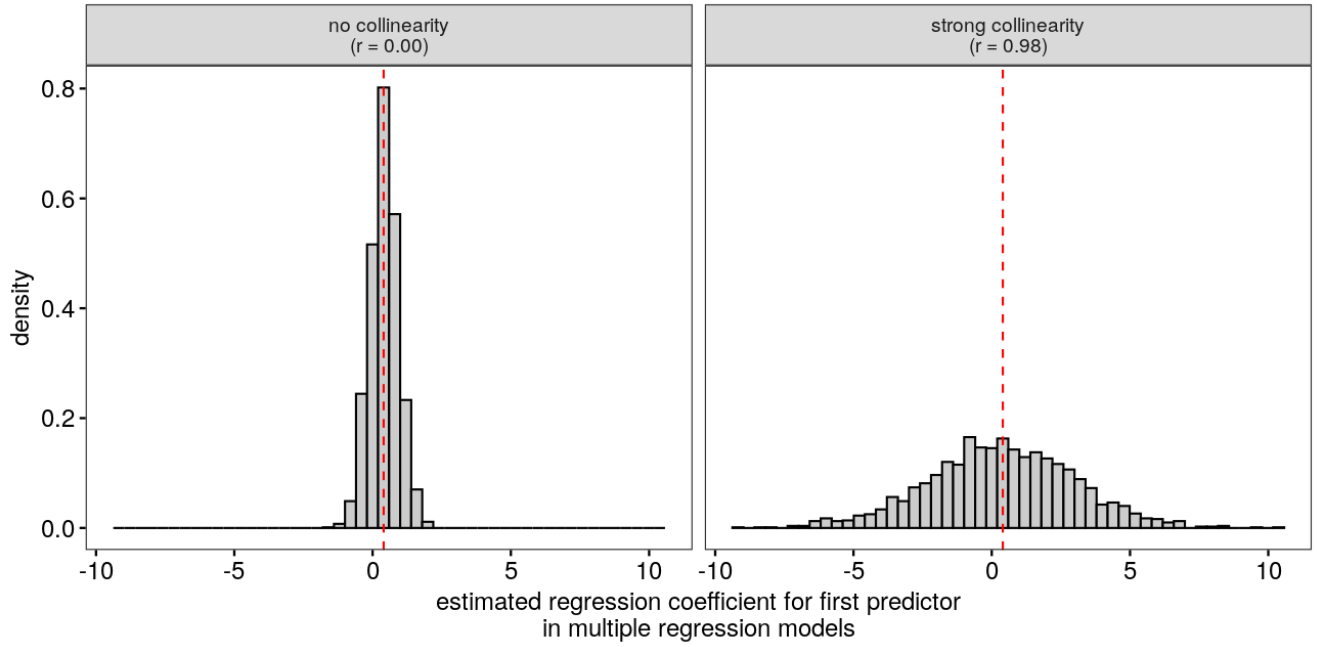
$x_3 = \text{altura} + \text{dap}$

$r = +1 \text{ ó } -1$

Colinealidad

"Problemas":

- Error estándar e intervalos de confianza inflados → se reduce la potencia estadística
- El test del modelo global puede ser significativo, sin que ninguno de los predictores sea significativo.



"La colinealidad no es una enfermedad que necesite cura"

https://janhove.github.io/analysis/2019/09/11/collinearity

Colinealidad

Criterios de identificación:

- Exploración de la matriz de correlación ($r > |0.7|$)
- Factores de inflación

de la varianza (VIF > 3)

$$T = 1 - R^2$$

$$VIF = \frac{1}{T} = \frac{1}{1 - R^2}$$

Supuestos

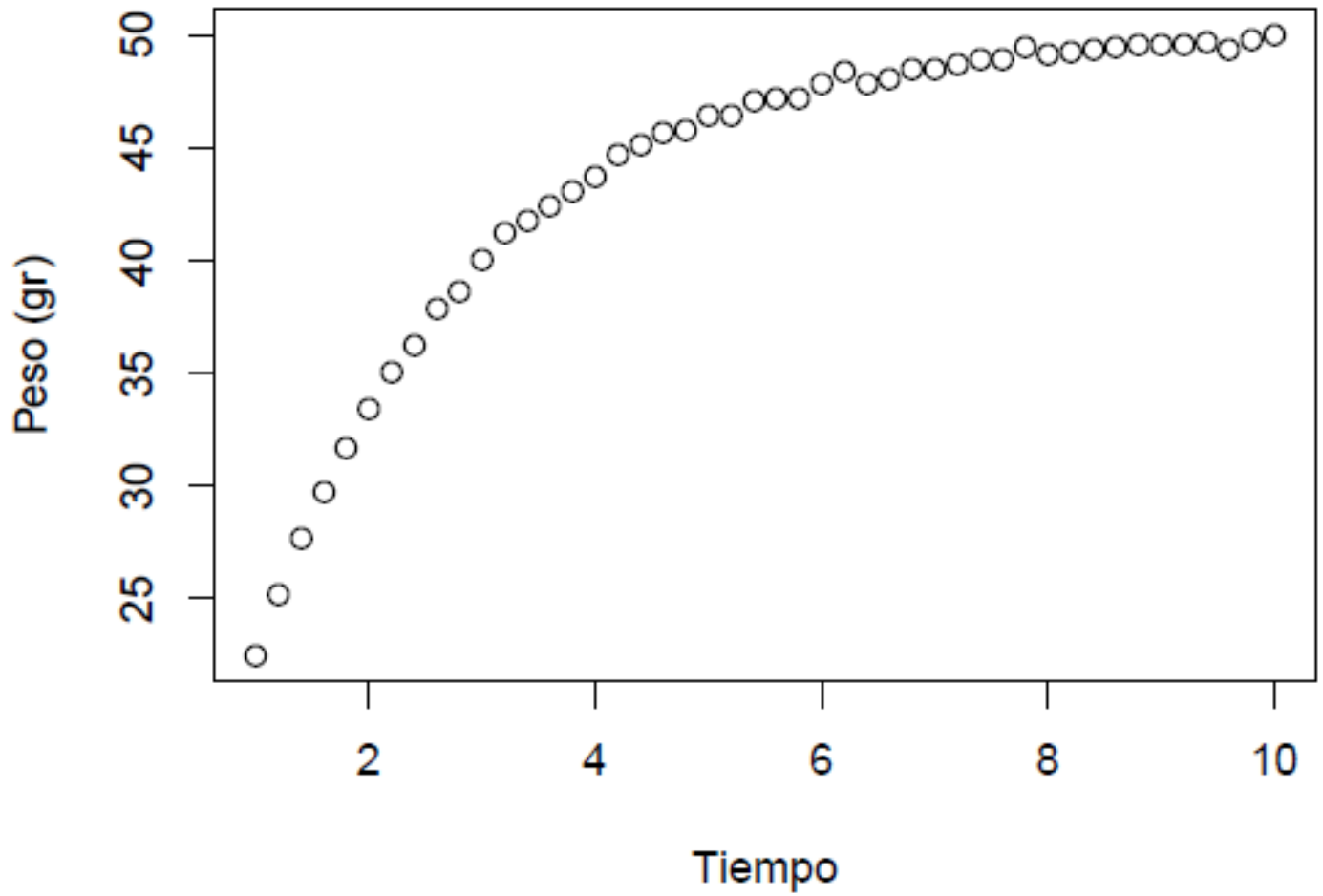
$$y - \hat{y} = \varepsilon$$ → Equivalente a la variable dependiente

Normalidad

$$\varepsilon \sim N\left(0, \sigma^2\right)$$ → Homogeneidad de varianzas

$$\rho_{\varepsilon_j \varepsilon_k} = 0$$ → Independencia de errores

Independencia de las observaciones



Diagnósticos

Test de t/ANOVA

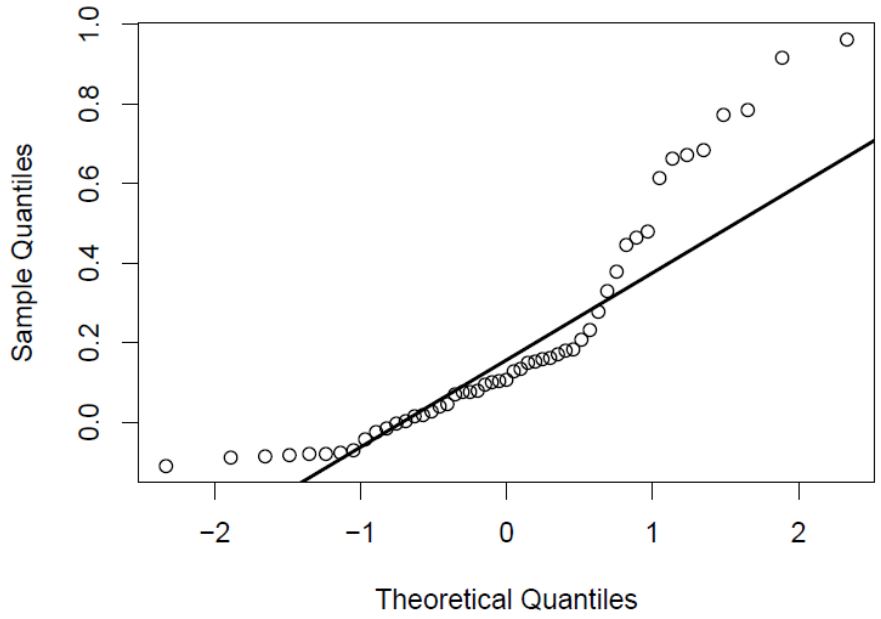
- **Normalidad:**
  - Boxplots → simetría
  - Gráfico de valores predichos (medias) vs residuos
  - QQplot
  - Tests de normalidad
- **Homoscedasticidad:**
  - Boxplots → dispersión.
  - Tests de homogeneidad de varianzas

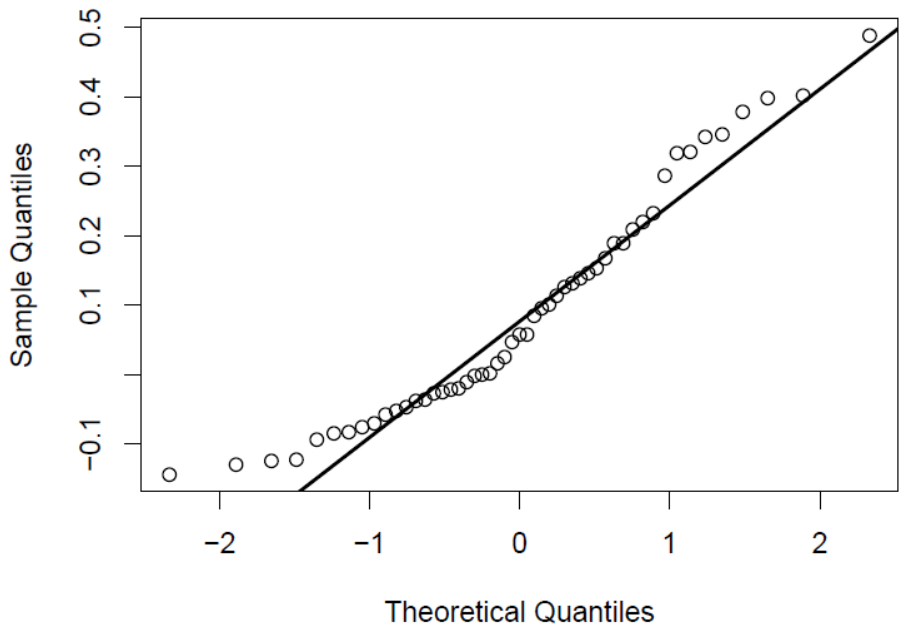
$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

Regresión

- **Normalidad:**
  - Boxplot de residuos → simetría
- **Homoscedasticidad:**
  - Gráfico de esperados (*y*) vs residuos
- **Independencia**
  - Gráfico de observados (*x*) vs residuos
- **Linealidad**
  - Gráfico de dispersión
  - Gráfico de observados (*x*) vs residuos

QQplot



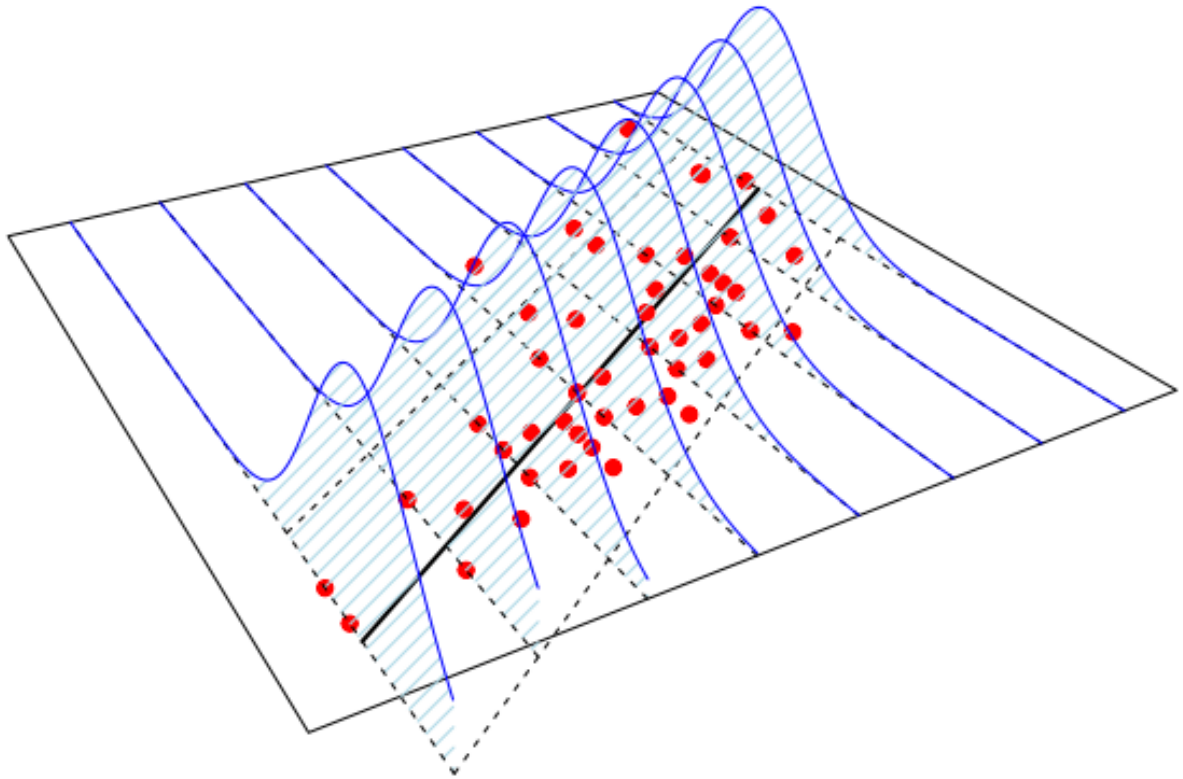


X → cuantil teórico según una distribución normal

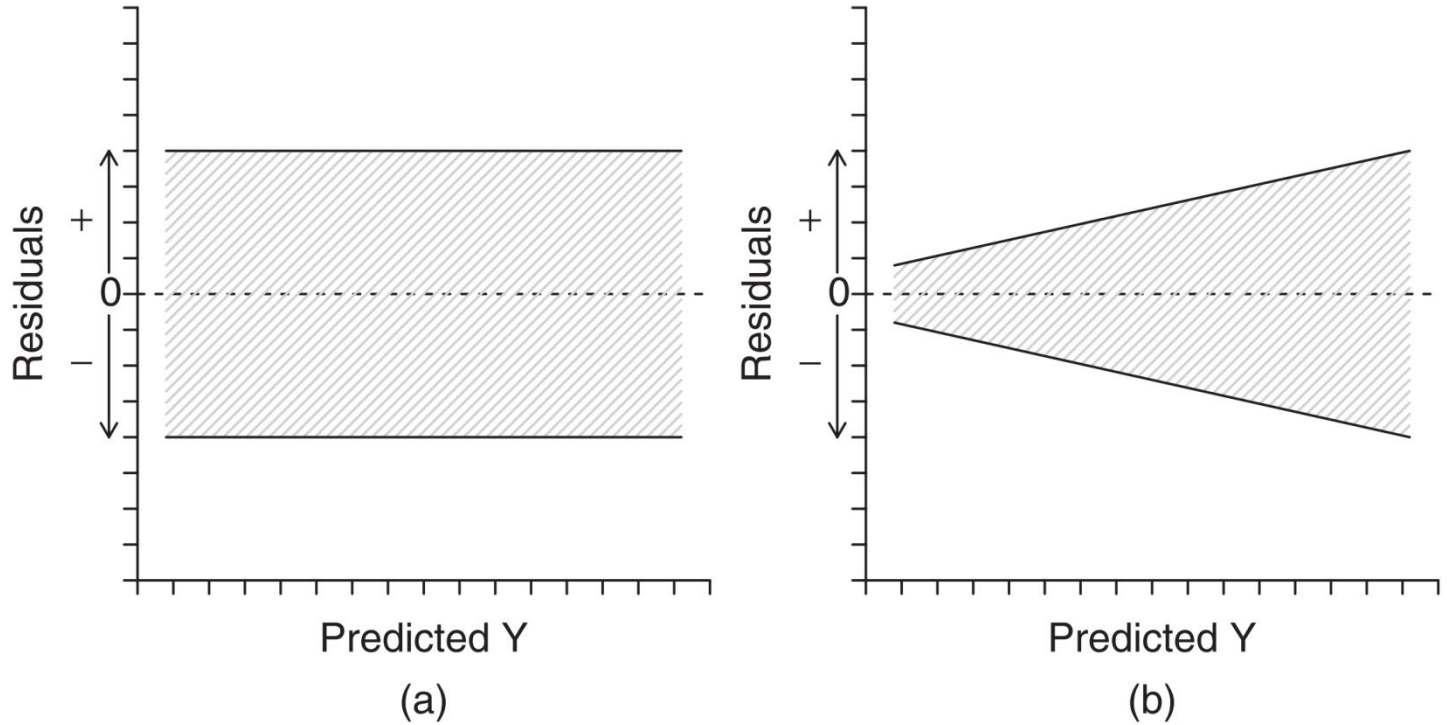
Y → cuantil observado

Supuestos

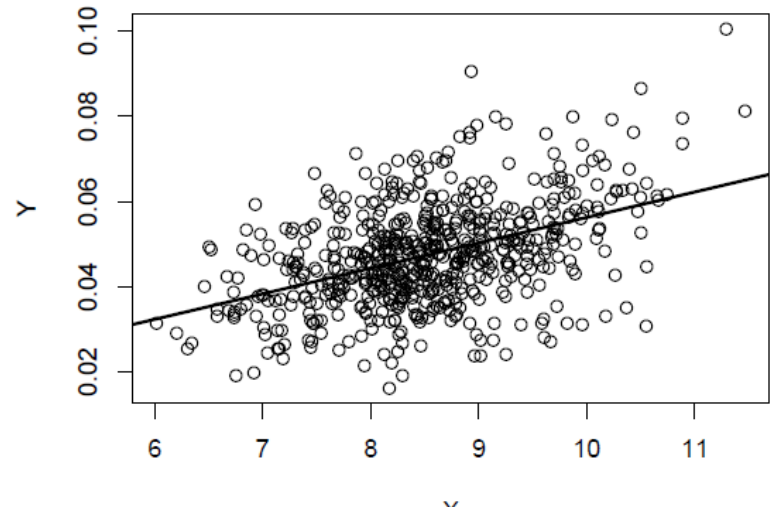
- **Regresión**

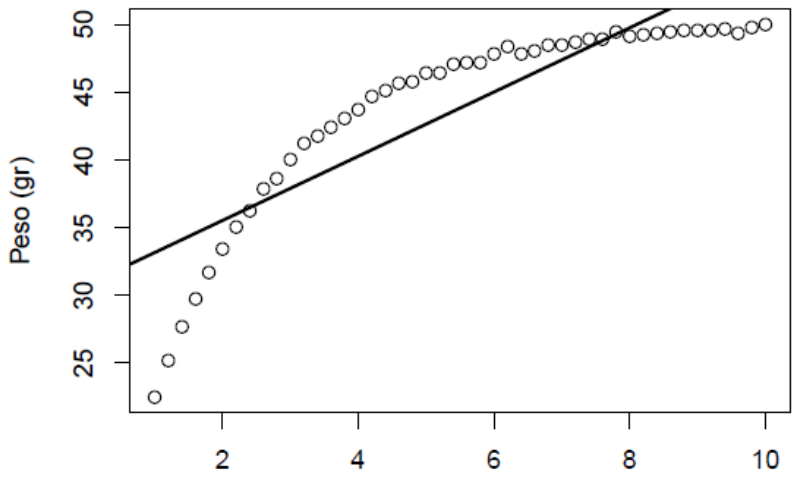


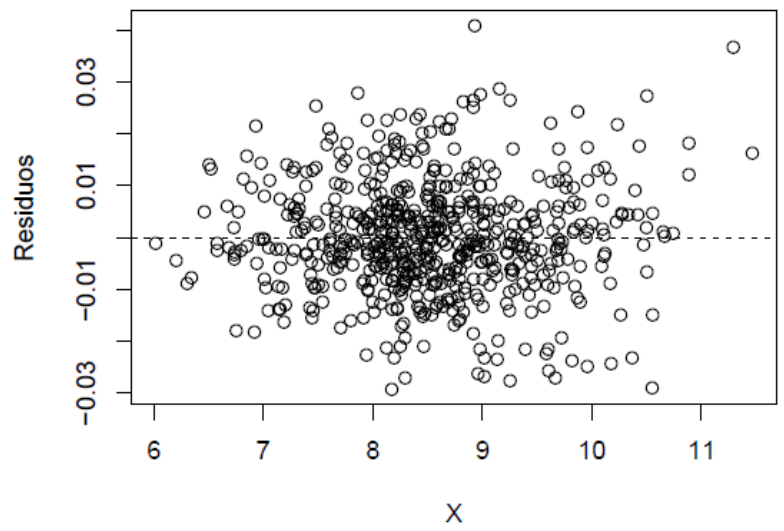
Homogeneidad de varianzas.
Gráfico de predichos vs residuos

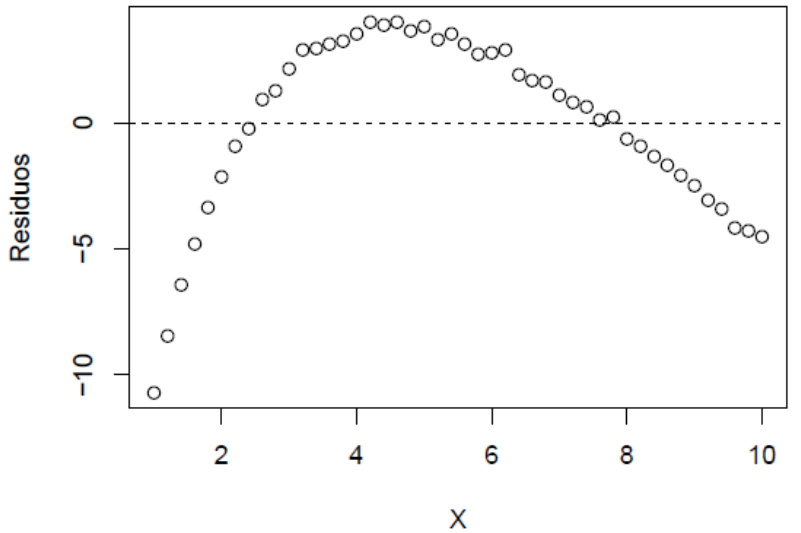


Independencia y linealidad.
Gráfico de observados vs residuos









Transformaciones

- Justificación: las escalas de medidas son arbitrarias.
- Injustificación: cambio de la variable de respuesta y de H_0

- Objetivos:

1) Normalidad.
2) Homoscedasticidad.
3) Linealidad.

Transformaciones

- Logaritmo → variables continuas sesgadas hacia la derecha, conteos
- Raíz cuadrada → densidades
- Transformación logit → proporciones
- Tranformación cuadrática → variables continuas sesgadas hacia la izquierda
- Transformación a rangos