

Introducción a la estadística

Bases indispensables y uso de 

Olivier Devineau

`olivier.devineau@fcdarwin.org.ec`

Fundación Charles Darwin

Taller interno, 27–30 abril 2010

1 / 42

Correlación y regresión

2 / 42

Dos categorías de tests estadísticos

Tests de comparación : 1 variable, ≥ 2 poblaciones

Tests de relación : ≥ 2 variables, 1 población

3 / 42

≥ 2 variables es común en biología

2 variables para el mismo individuo

- Presión sanguínea X_1 , peso X_2
- Abundancia de una especie de planta X_1 , nivel del pH en el suelo X_2 , temperatura X_3

- Datos bivariados o multivariados

\Rightarrow ¿Cuál es la relación entre las variables?

4 / 42

Relación entre ≥ 2 variables

La estadística correlacional

Varios tipos de relación

- No conexión
- Relación |*handout* : $1 > 0$ / < 0 , causal / no
- Conexión funcional \rightarrow predicción

Objetivo de la estadística correlacional

- Determinar validez y fuerza de la relación entre las variables
- Determinar la dirección de la relación

5 / 42

Estadística correlacional

Correlación: ¿Cómo 2 variables varían juntas?

Regresión: Relación entre 1 variable dependiente y ≥ 1 variable independiente

Análisis multivariados: Relación entre ≥ 2 variables independientes / dependientes / ambos

6 / 42

Noción de correlación

Ejemplo

- 1 población: 2 variables continuas
- Presión sanguínea X_1 , peso X_2
- Cada muestra i : 1 valor por cada variable: x_{i1} y x_{i2}
- ¿La presión sanguínea y el peso son correlativas?

7 / 42

Noción de correlación (2)

Definición

Correlación se define en terminos de:

- Varianza de X_1 : $var(X_1)$
- Varianza de X_2 : $var(X_2)$
- ¿Como X_1 y X_2 varían juntas? Covarianza: $cov(X_1, X_2)$

\Rightarrow Coeficiente de correlación

$$r = \frac{cov(X_1, X_2)}{\sqrt{var(X_1) \cdot var(X_2)}}$$

8 / 42

El coeficiente de correlación r

Correlación de Pearson (paramétrica)

- No unidad
- $r \in [-1, 1]$
- Magnitud: fuerza de la relación
- Signo: dirección de la relación
- Muestra: r , Población: ρ

9 / 42

¿Qué test para chequear la correlación?

X_1 : Presión sanguínea y X_2 : peso

- ¿Hipótesis nula?
- No hay una relación lineal entre la presión sanguínea y el peso
- $H_0 : \rho = 0$
- Cuando H_0 es verdadera, $r \sim \mathcal{N}(\mu, \sigma)$
 \Rightarrow uso de test t de Student

10 / 42

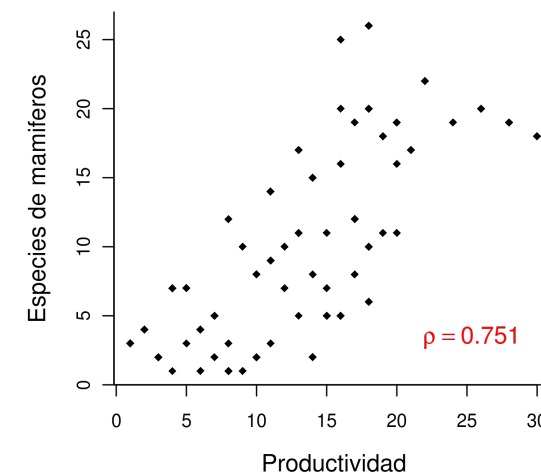
Correlación no paramétrica

- ¿Qué hacer cuando los requisitos no se cumplen?
- \Rightarrow Coeficiente de correlación de rango
- de Spearman: ρ
 - de Kendall: τ
- ¡Más conservadores!

11 / 42

La correlación depende de la escala

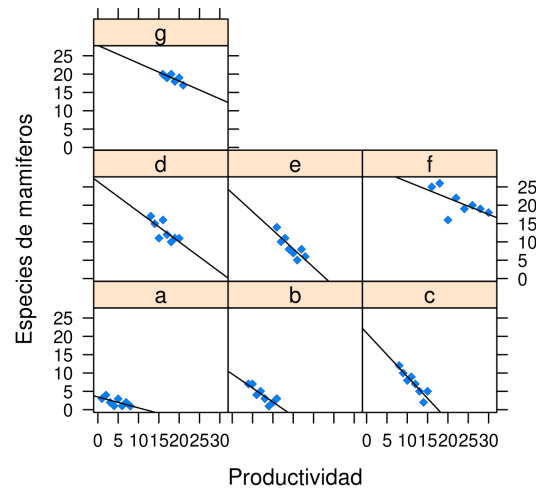
¡Las cosas no son siempre como parecen!



12 / 42

La correlación depende de la escala

¡Las cosas no son siempre como parecen!



13 / 42

Modelo lineal: concepto general

- Se puede identificar:
 - 1 variable respuesta / dependiente Y
 - ≥ 1 variable explicativa / predictiva / independiente / covariable X_1, X_2, \dots
- Cada unidad de muestra: $y_i, x_{1i}, x_{2i} \dots$
- Explicar el patrón de Y con X

14 / 42

Modelo lineal

Forma general de los modelos estadísticos

- *Variable dependiente = modelo + error*
- Modelo: covariables y parámetros
- Covariables: continuas / categóricas / ambos
- Error: parte de la variable dependiente que no está explicada por el modelo
- Se supone una distribución para el componente del error, y de ahí para la variable dependiente Y

15 / 42

¿Qué significa lineal?

- Relación de línea recta entre 2 variables
- Combinación lineal de parámetros
- No exponente, no multiplicación por otro parámetro
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

16 / 42

Análisis de regresión lineal

Contexto

- Usar datos de una muestra para estimar valores de parámetros y sus errores estándar
- ¿Cuándo se usa?
- Variables explicativa y dependiente son continuas
- Altura, peso, volumen, temperatura ...
- Nube de puntos → regresión lineal

17 / 42

Análisis de regresión lineal

Objetivos

- Describir la relación lineal entre Y y X
- Determinar cuánto de la variación en Y se explica por la relación lineal con X y cuánto de esta variación no se puede explicar
- Predecir nuevos valores de Y a partir de valores de X

18 / 42

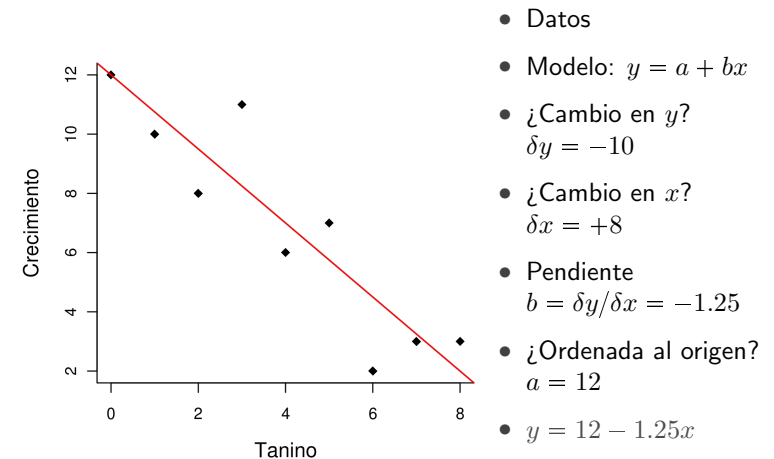
Análisis de regresión lineal

Varios tipos de regresión

- Regresión lineal: lo más simple y frecuente
- Regresión polinomial: chequear si una relación es no lineal
- Regresión no lineal
- Regresión no paramétrica: si no hay forma funcional

19 / 42

Principio de la regresión lineal



20 / 42

Principio de la regresión lineal (2)

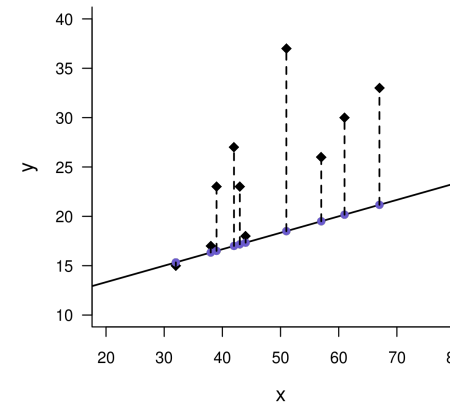
- Ajustar un modelo a los datos
- Estimar los parámetros del modelo
- Probar varios valores de parámetros hasta encontrar el mejor modelo
- Máxima verosimilitud (Maximum Likelihood ML)
- Mínimos cuadrados (Ordinary Least Square OLS)

21 / 42

Cuadrados mínimos: principio

OLS: Ordinary Least Squares

- Datos
- Modelo $y = 10 + 1/6x$
- Residual $e_i = y_i - \hat{y}_i$
- $SS = \sum (y_i - \hat{y}_i)^2 = 79.85$

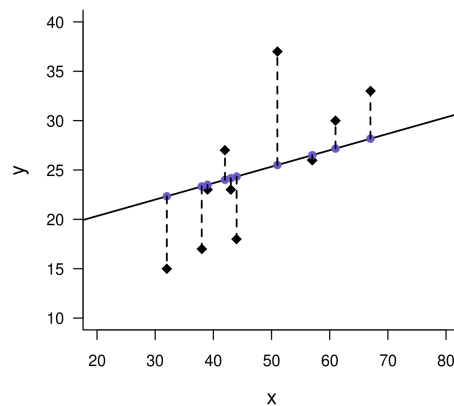


22 / 42

Cuadrados mínimos: principio

OLS: Ordinary Least Squares

- Datos
- Modelo $y = 10 + 1/6x$
- Residual $e_i = y_i - \hat{y}_i$
- $SS = \sum (y_i - \hat{y}_i)^2 = 79.85$
- $SS = 30.85$

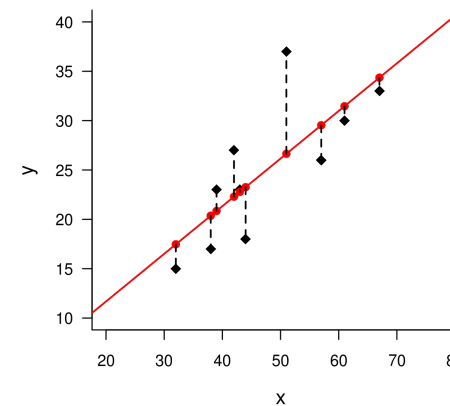


23 / 42

Cuadrados mínimos: principio

OLS: Ordinary Least Squares

- Datos
- Modelo $y = 10 + 1/6x$
- Residual $e_i = y_i - \hat{y}_i$
- $SS = \sum (y_i - \hat{y}_i)^2 = 79.85$
- $SS = 30.85$
- Modelo seleccionado:
 $SS = 19.58$
 $y = 2.03 + 0.48x$



24 / 42

Hipótesis nula en regresión

- ¿Cuál sería H_0 ?
- No hay una relación lineal entre las variables
- Pendiente $b = 0$
 - Test de Fisher: F
 - Test de Student: t

25 / 42

Varianza explicada

r^2 : coeficiente de determinación

- Variación de Y explicada por la relación con X
- (coeficiente de correlación)²
- $r^2 \in [0, 1]$
- ¿Como se mejora el ajuste del modelo con pendiente comparado a un modelo sin pendiente?
- r^2 inadecuado para comparar modelos con números de parámetros diferentes

26 / 42

Comparar varios modelos

- Evaluar varias hipótesis → varios modelos
- H_0 : modelo simple, H_1 : modelo más complejo
- Hay que comparar los modelos

27 / 42

Comparar modelos de regresión

Minimos cuadrados (OLS)

- Ajuste: proporción de varianza explicada
 - No-ajuste: proporción de varianza residual
- ⇒ Análisis de varianza

Máxima verosimilitud (ML)

- Ajuste: tamaño de la verosimilitud
- ⇒ Prueba de la razón de verosimilitud (Likelihood Ratio Test o AIC)

28 / 42

Comparar modelos de regresión (2)

Siempre la misma lógica

- Medir el ajuste de cada modelo
- Comparar los ajustes de diferentes modelos para examinar hipótesis sobre los parámetros

Ejemplo: presión sanguínea y peso

- Modelo 1: $P = \beta_0 + \varepsilon$
- Modelo 2: $P = \beta_0 + \beta_1 * peso + \varepsilon$
- Comparar M_1 y M_2 es equivalente a evaluar $H_0 : \beta_1 = 0$

29 / 42

Condiciones del análisis de regresión (1)

- Involucran de los términos de errores (ε_i)
- De la variable dependiente Y
- Importantes para intervalos de confianza
- Importantes para tests de hipótesis con distribución t o F
- Residuales importantes para chequear condiciones

30 / 42

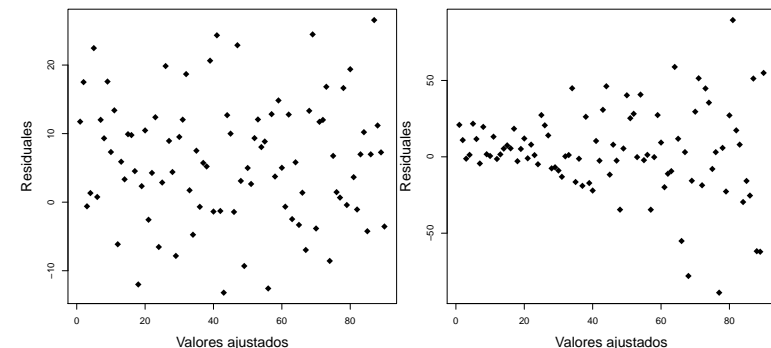
Condiciones del análisis de regresión (2)

- Normalidad: ε tiene una distribución normal
- Homogeneidad de la varianza: ε tiene la misma varianza por cada x_i : $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_i^2 = \dots = \sigma_\varepsilon^2$
- Independencia: ε son independientes: Los valores de Y para cualquier x_i no influyen los valores de Y para otra x_i

31 / 42

Homogeneidad de la varianza

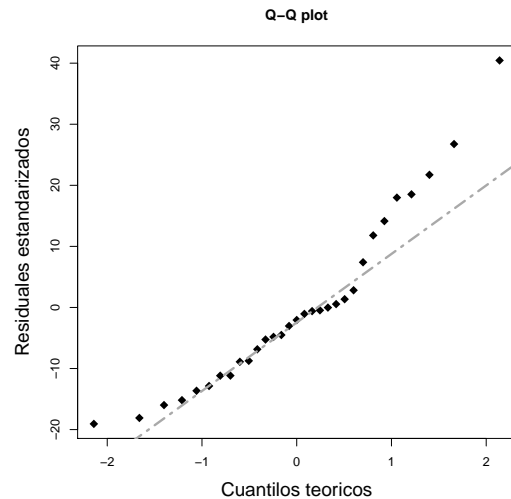
- No tendencia
- Heteroscedasticidad



- Test de Levene, test de Bartlett

32 / 42

Normalidad de los residuales



- Test de Shapiro-Wilk

33 / 42

¿Qué hacer si las condiciones no cumplen?

- Residuales no son independientes:
 - Modelos con efectos aleatorios (random effect models)
- Residuales no son normales:
 - Alternativa no paramétrica
 - Transformación de los datos *log*, *sqrt*, *exp* ...
 - Modelo lineal generalizado (Generalized Linear Model GLM)
- Heterogeneidad de la varianza:
 - GLM

34 / 42

Si el modelo es inadecuado, se puede...

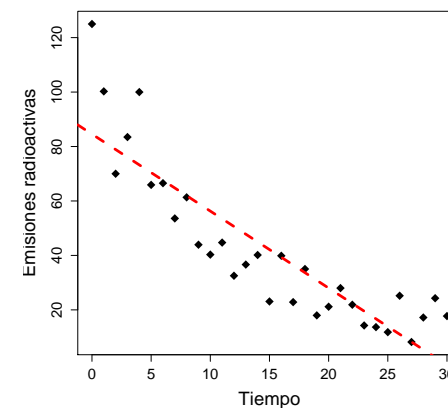
- Transformar variable dependiente
- Transformar ≥ 1 variable explicativa
- Probar otras variables explicativas
- Usar una estructura de error diferente (GLM)
- Usar alternativa no paramétrica (smoothing)
- Usar pesos diferentes por diferentes valores de y

35 / 42

Regresión polinomial

Ejemplo: Desintegración radioactiva

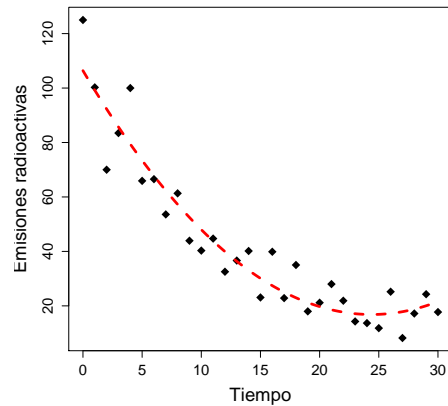
- Regresión lineal:
 $y = ax + b$



36 / 42

Regresión polinomial

Ejemplo: Desintegración radioactiva

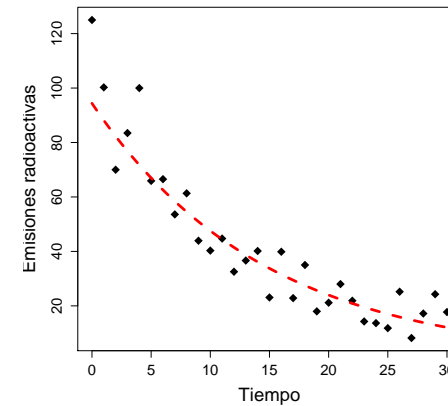


- Regresión polinómica
- $X_2 = X^2$
- $y = ax^2 + bx + c$

37 / 42

Regresión polinomial

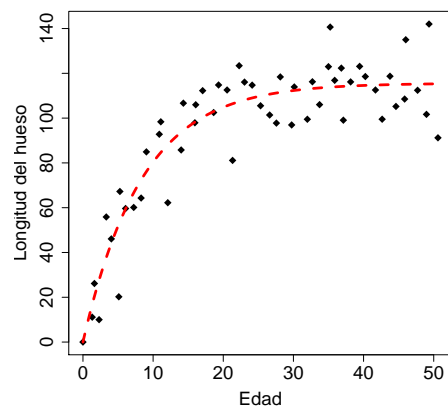
Ejemplo: Desintegración radioactiva




- $y = ae^{-bx}$
- ¡Descripción, no explicación!

38 / 42

Regresión no lineal y GAM



- : `nls()`
- Teoría:
 $y = a - be^{-cx}$
- No información:
Modelos Aditivos Generalizados
(Generalized Additive Models GAM)

39 / 42

Recordatorio de vocabulario

- Normalidad de los errores:
 - Modelos lineales
- Normalidad + var. descriptivas continuas/categorías:
 - Modelos lineales generales
- Errores no normales y/o varianza no homogénea:
 - Modelos lineales generalizados (GLM)

40 / 42

Modelos lineales generalizados (2)

Varianza no constante / residuales no normales

⇒ Se puede especificar la distribución de los errores

- Proporciones (regresión logística) → Binomial
- Conteos (modelo log-lineal) → Poisson
- Variable dependiente binaria (vivo/muerto) → Binomial
- Tiempo hasta muerte (varianza aumenta) → Exponencial

(No) enamorarse de su modelo ...

- Todos los modelos son incorrectos
- Algunos modelos son mejores que otros
- El modelo correcto nunca se puede conocer con certeza
- Cuanto mas simple el modelo mejor