Biometría

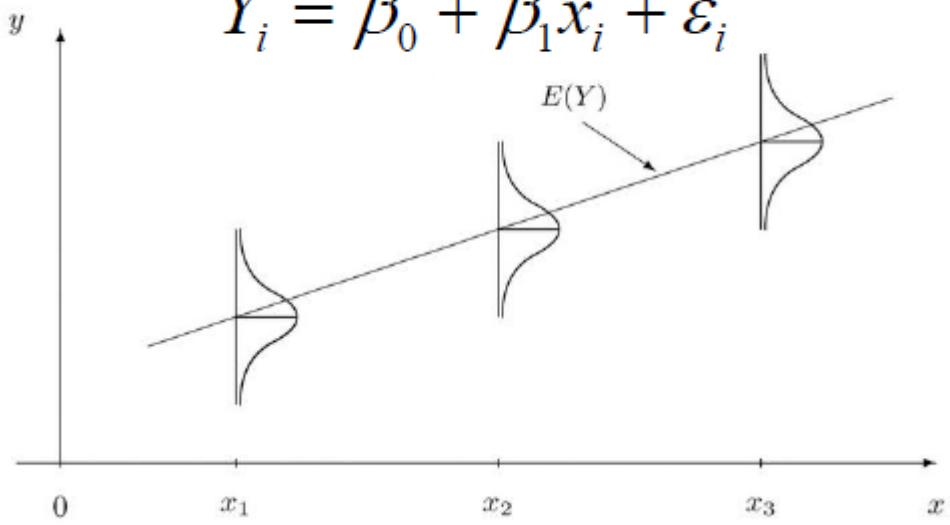
Modelos Lineales Generales

Regresión y correlación

Modelo estadístico: Simplificación de la realidad. Es una expresión matemática que indica cómo una variable aleatoria (VR, Y), con una distribución de probabilidades dada, se relaciona con una o más variables explicatorias VE, aleatorias o no, consideradas en el diseño experimental

$$E(Y) = \beta_0 + \beta_1 x_i$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



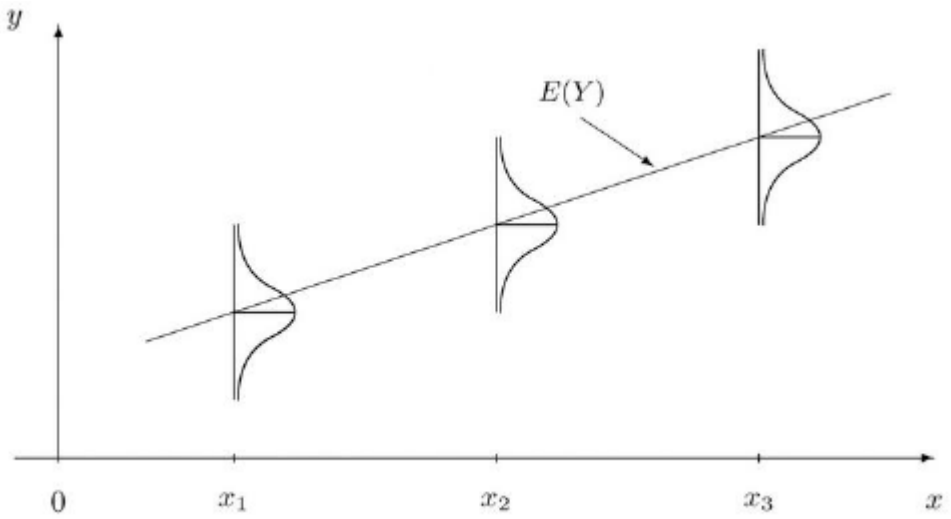
Modelo estadístico: Dado un valor de X, la esperanza de Y queda determinada unívocamente. Existe variación aleatoria (error)

Modelo estadístico

VR — parámetros — VE

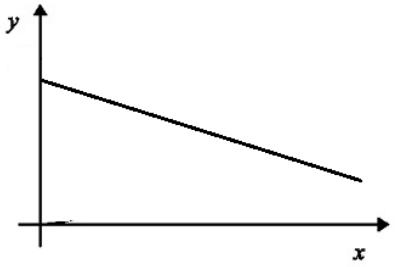
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

determinístico — estocástico



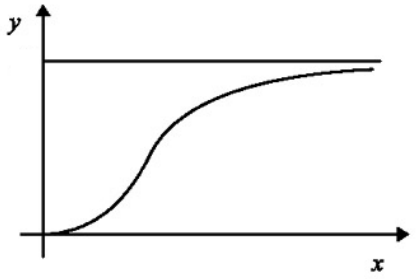
Modelos lineales en los parámetros (los parámetros aparecen sumando; ningún parámetro aparece como exponente o multiplicado o dividido por otro parámetro). La VR es una combinación lineal de las VE

$$Y_i = \beta_0 + \beta_1 x_i + ... + \varepsilon_i$$



Modelos no lineales

$$y = \beta_0 e^{\beta_1 x} + \varepsilon_i$$



CLASIFICACIÓN DE LOS MODELOS SEGÚN LA VARIABLE RESPUESTA:

Modelo lineal general

- ✓ VR cuantitativa con distribución normal
  - Anova, regresión lineal

Modelos lineales generalizados (GLM)

- ✓ VR cualitativa o cuantitativa discreta o cuantitativa continua con cualquier distribución de probabilidades
  - Regresión logística: VR dicotómica (Si/No), distribución Bernoulli)
  - Regresión binomial: VR discreta (cantidad de éxitos en muestra n
  - Regresión de Poisson : VR discreta (conteos)
  - Otras: distribución normal (el modelo lineal general es un caso particular de los GLM); distribución gamma, etc

Modelos estadísticos

- Se elige un modelo que representa el comportamiento de la variable respuesta como función de las variables explicatorias

Respuesta observada = modelo matemático + error

F (VARIABLES EXPLICATORIAS)

- V.respuesta cuantitativa y V.explicatoria cualitativa → **ANÁLISIS DE LA VARIANZA**

- V.respuesta cuantitativa y V.explicatoria cuantitativa → **ANÁLISIS DE REGRESIÓN**

Regresión

- Es una técnica que permite determinar cómo los cambios en variable/s explicatoria/s afectan a una variable respuesta.
- Sirve para predecir una variable en función de otra (u otras)
  - La variable **respuesta o dependiente** es la variable de interés; es la que deseamos predecir, es aleatoria y se denomina Y
  - La/s variable/s **explicatorias o independientes o predictoras** intentan explicar la variable respuesta, no son aleatorias y se denominan X
  - Todas las variables involucradas deben ser **cuantitativas**

- Los modelos de regresión pueden ser:
  - **Simples:** Una sola variable explicatoria
  - **Múltiples:** Más de una variable explicatoria

Expresión de la α-actinina durante el desarrollo del músculo cardíaco

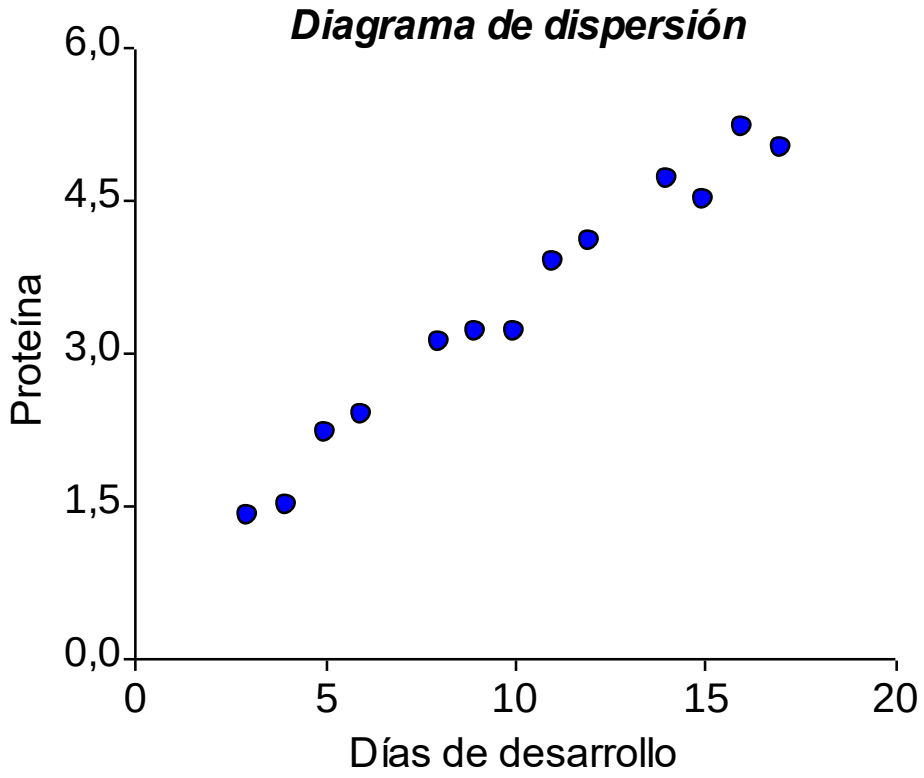
Se ha visto que la modulación de la expresión de ciertas proteínas constitutivas del citoesqueleto de la célula muscular está directamente relacionada con el grado de maduración celular.
Dado que la α -actinina juega un papel decisivo en el proceso de contracción de la célula muscular cardíaca, interesa estudiar el patrón de expresión durante el desarrollo del músculo cardíaco.

Para ello 50 µg de proteínas totales de embriones de pollos de 3 a 17 días fueron sometidas a electroforesis en gel de poliacrilamida e inmunoblotting. Los resultados corresponden a la cantidad de proteína/mg de tejido cuantificada por densitometría:

3 17 días

Expresión de la α-actinina durante el desarrollo del músculo cardíaco

Días de desarrollo	Proteína (µg)
3	1.40
4	1.50
5	2.20
6	2.40
8	3.10
9	3.20
10	3.20
11	3.90
12	4.10
14	4.70
15	4.50
16	5.20
17	5.00



n =

El análisis de regresión se propone:

- Estimar la relación funcional entre X e Y
- Formular hipótesis con respecto a los parámetros del modelo
- Predecir valores de la variable respuesta para valores específicos de X (*dentro* del rango analizado)

Modelo de regresión lineal

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- β_0 es la ordenada al origen poblacional
- β_1 es la pendiente o coeficiente de regresión poblacional
  - β_0 y β_1 son parámetros
- ε_i es el error aleatorio, no explicado por el modelo y que se debe a todas las variables no contempladas en el análisis más la variabilidad natural entre individuos
- siendo

$$\mu_{(y/x)} = E_{(y/x)} = \beta_0 + \beta_1 x_i$$

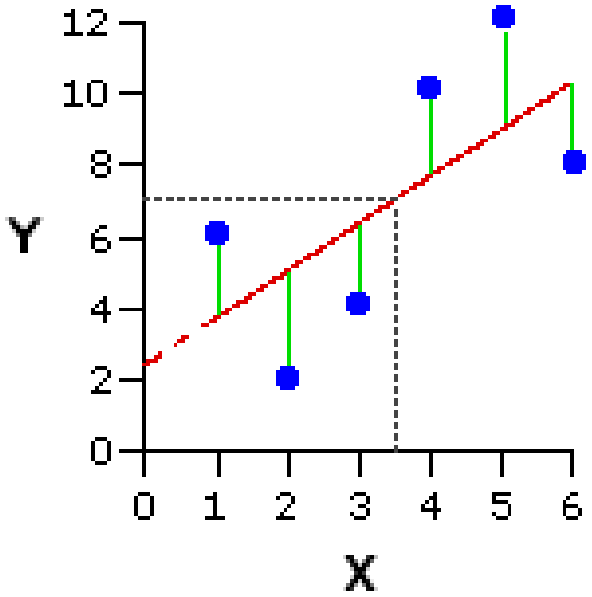
Recta de regresión estimada

- La función anterior no es observable directamente, sino que debe ser estimada a través de una muestra:

$$\hat{y}_i = b_0 + b_1 x_i$$

- b_0 es la ordenada al origen muestral
- b_1 es la pendiente o coeficiente de regresión muestral
  - b_0 y b_1 son estimadores
- Los errores aleatorios no aparecen en la ecuación que describe el comportamiento de la recta.
- La recta + ε_i = valores de la var repuesta y_i

Ecuación de la recta: Método de Cuadrados mínimos



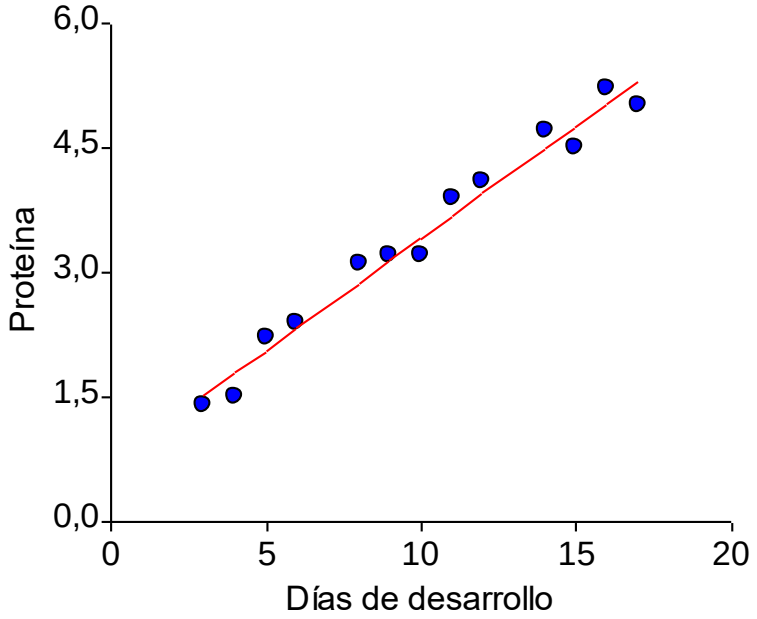
- Se busca la recta que, pasando por el centro $\bar{x}, \bar{y}$ haga mínima

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^{\,2}$$

$$b_1 = \frac{\Sigma(xy_i - \bar{x}\bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{Sxy}{SCx} = \frac{Sxy}{Sxx}$$

la recta pasa por $\bar{x}, \bar{y}$ ➡ $\bar{y} = b_0 + b_1\bar{x}$ ➡ $b_0 = \bar{y} - b_1\bar{x}$

Ecuación de la recta

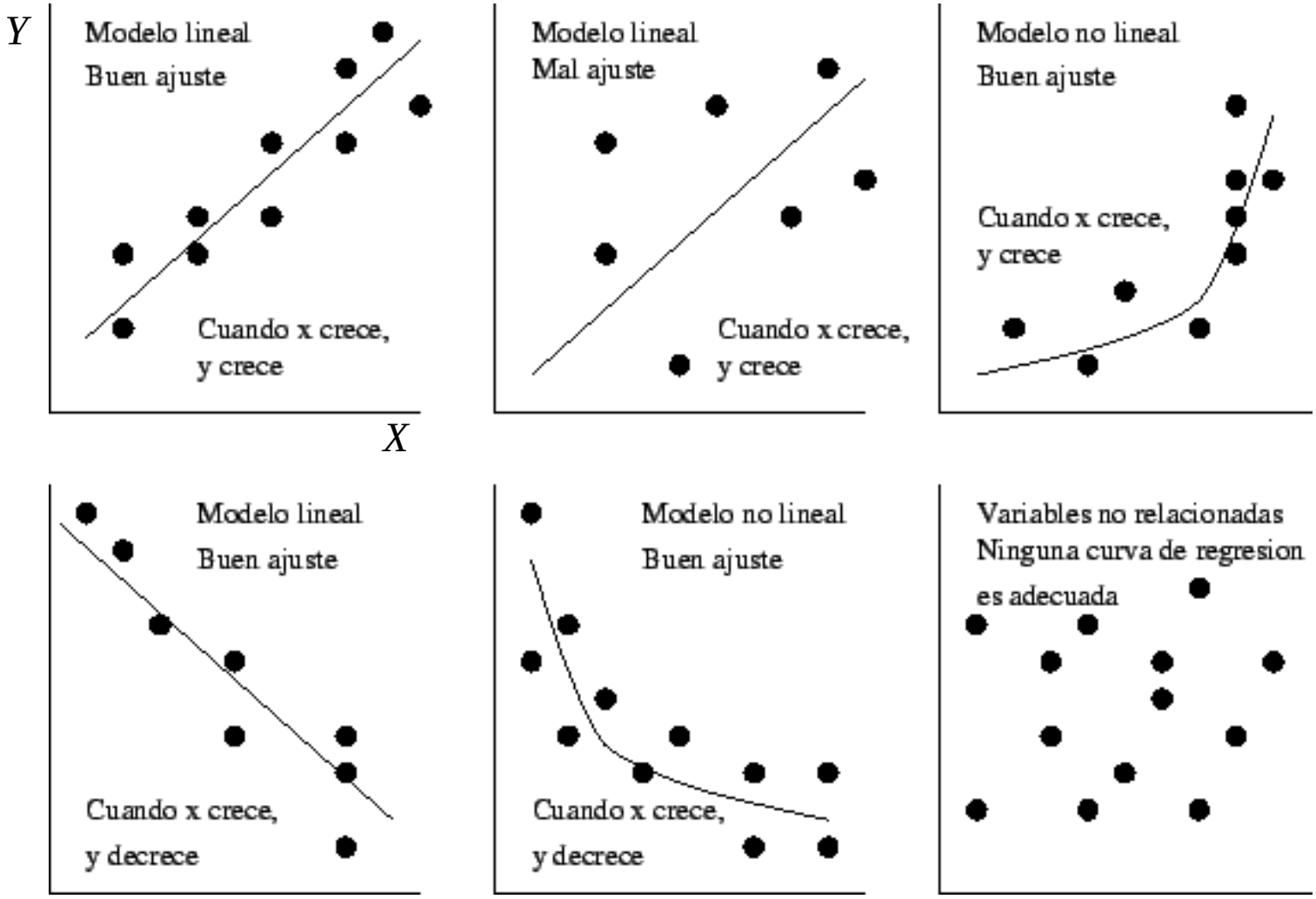


Efectos fijos

	Value	Std.Error	t-value	p-value
(Intercept)	0.713	0.148	4.821	0.0005
Dias.de.desarrollo	0.270	0.013	20.027	<0.0001

$$\hat{y} = 0.713 + 0.27x$$

Diferentes nubes de puntos



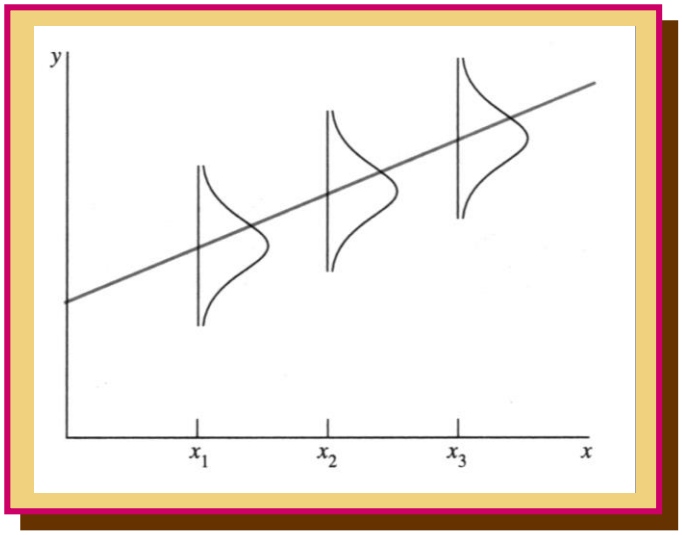
Supuestos del modelo de regresión lineal

- X medida sin error
- las observaciones Y son independientes
- los valores esperados de Y, para cada valor de X, están alineados, es decir $E(Y) = \beta_0 + \beta_1 x$
- para cada valor de X, la correspondiente subpoblación de Y sigue una distribución normal
- las varianzas de las subpoblaciones de Y son iguales

Estos supuestos se pueden resumir en:

ε_i deben ser independientes y tener distribución normal

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



Supuestos: análisis de residuos

- La diferencia entre el valor **observado** y el **pronosticado** por el modelo se llama residuo

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

- Su promedio es **cero** (se compensan los residuos positivos con los negativos)
- El análisis de residuos permite:
  - Detectar **outliers o datos atípicos** (datos con residuos grandes)
  - Determinar si el modelo está **bien especificado** (los residuos deberían distribuirse aleatoriamente)
  - Determinar si la variabilidad es **constante**
  - Determinar si ajustan a la distribución **normal**

Análisis de residuos

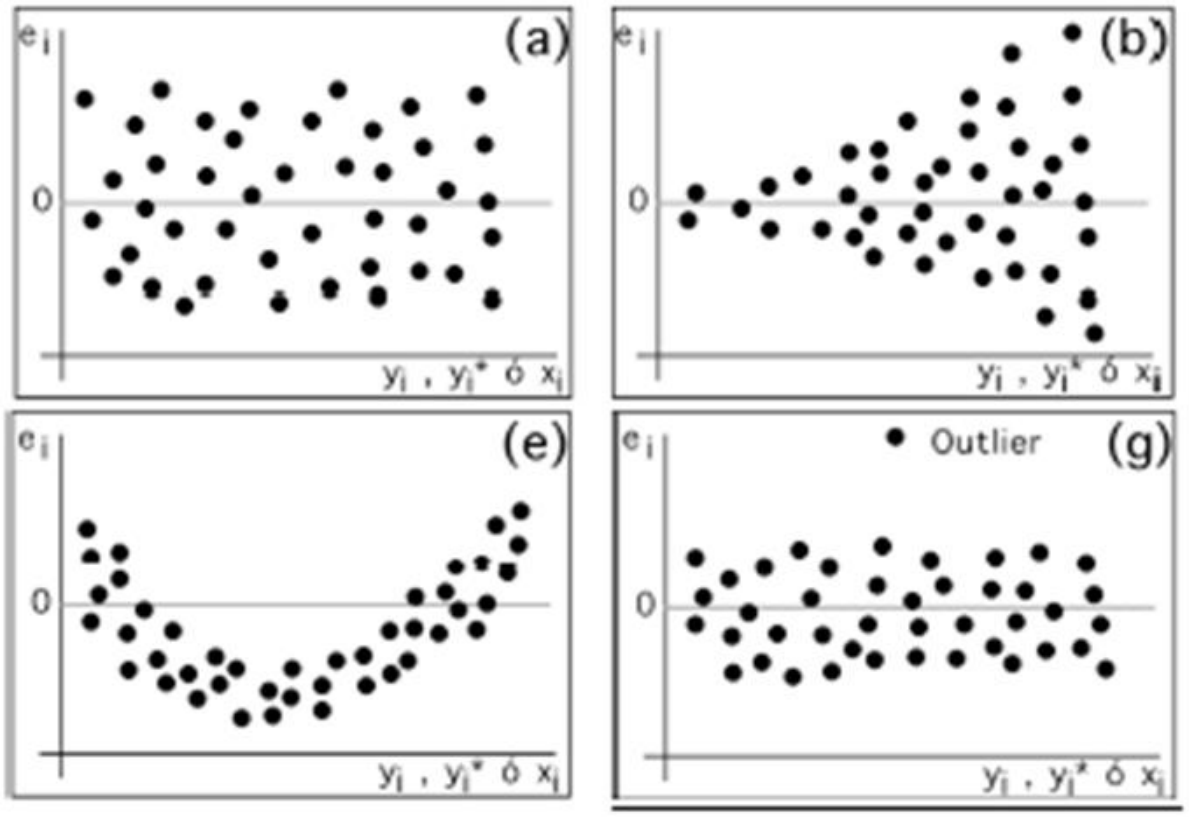
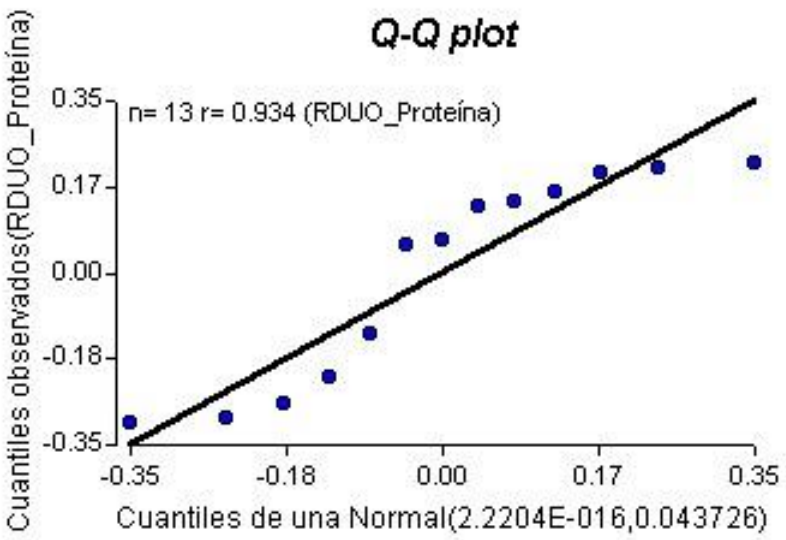
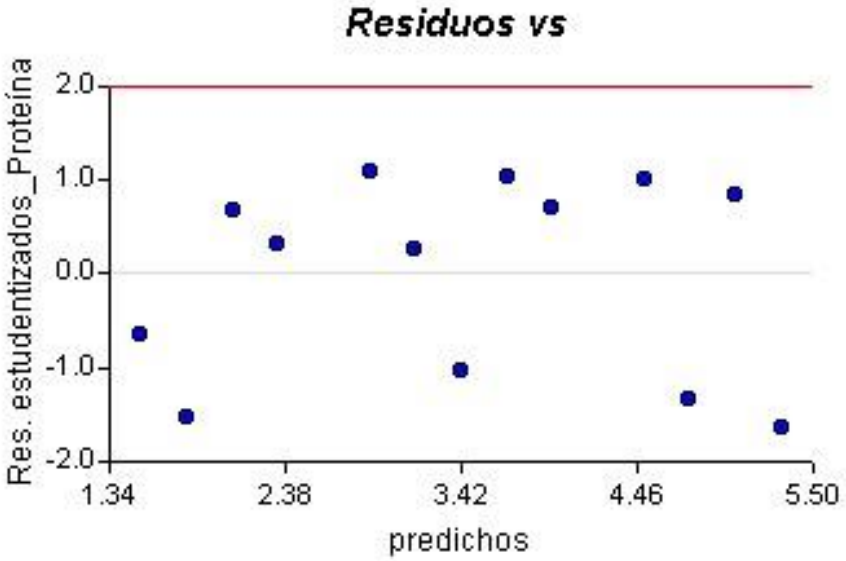


Gráfico de dispersión de residuos vs predichos

Análisis de residuos en el ejemplo





Normalidad

- ✓ Homocedasticidad
- ✓ Correcta especificación del modelo lineal
- ✓ Inexistencia de outliers

Evaluación del modelo de regresión: Prueba de hipótesis para la pendiente

Ho: $\beta_1 = 0$

Y no depende linealmente de X; el modelo lineal no es válido

H_1: $\beta_1 \neq 0$

Y sí depende linealmente de X; el modelo lineal es válido: el contenido de proteína depende linealmente de los días de desarrollo

Hay dos formas equivalentes de poner a prueba estas hipótesis:

- **Prueba t**
- **Anova**

Prueba t

- Parámetro: β_1
- Estimador: b_1
- Distribución muestral del estimador:
  - Esperanza: β_1
  - Error estándar $S_{b_1} = \sqrt{\frac{CM_{res}}{S_{xx}}}$
  - Distribución de probabilidades: t de Student con $n-2$ GL

$$t_m = \frac{b_1 - \beta_1}{S_{b_1}}$$

Efectos fijos

	Value	Std.Error	t-value	p-value
(Intercept)	0.713	0.148	4.821	0.0005
Dias.de.desarrollo	0.270	0.013	20.027	<0.0001

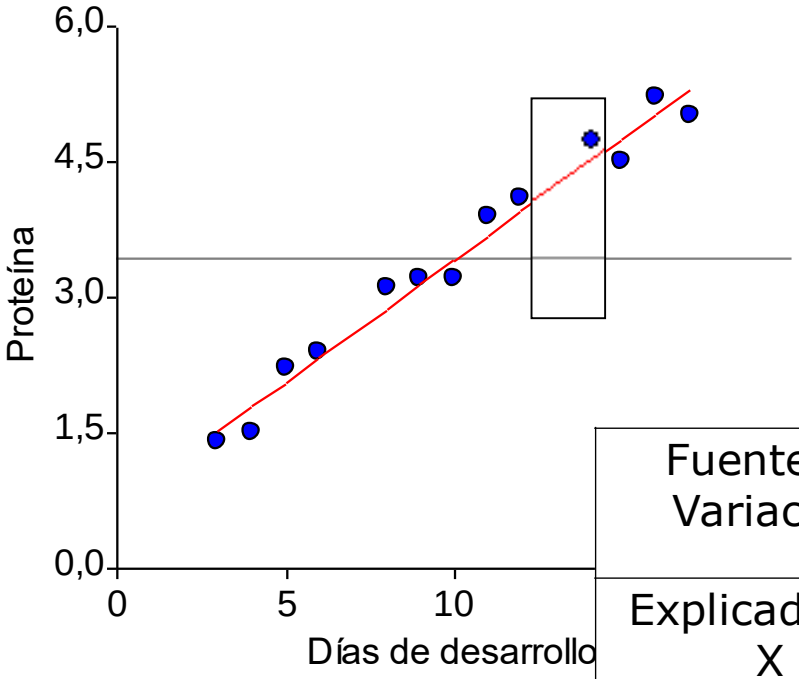
Anova

Descompone la variabilidad de Y en:
explicada por X y no explicada o error

Fuente de Variación	SC	GL	CM	F
Explicada por X	$\Sigma(\hat{y}_i - \bar{y})^2$	1	$\frac{SC_{EX}}{GL_{EX}}$	$\frac{CM_{EX}}{CMr}$
No explicada (residual o error)	$\Sigma(y_i - \hat{y}_i)^2$	n-2	$\frac{SCr}{GLr}$	
Total	$\Sigma(y_i - \bar{y})^2$	n-1		

Anova

Descompone la variabilidad de Y en: explicada por X y no explicada o error



Fuente de Variación	SC	GL	CM	F
Explicada por X	$\Sigma(\hat{y}_i - \bar{y})^2$	1	$\frac{SC_{EX}}{GL_{EX}}$	$\frac{CM_{EX}}{CMr}$
No explicada (residual o error)	$\Sigma(y_i - \hat{y}_i)^2$	n-2	$\frac{SCr}{GLr}$	
Total	$\Sigma(y_i - \bar{y})^2$	n-1		

Anova

Descompone la variabilidad de Y en: explicada por X y no explicada o error

X	Y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$(\hat{y}-\bar{y})$	$(\hat{y}-\bar{y})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
3	1,40	1,524	-0,124	0,015	-1,89	3,58	-2,02	4,06
4	1,50	1,794	-0,294	0,086	-1,62	2,63	-1,92	3,67
5	2,20	2,064	0,136	0,018	-1,35	1,83	-1,22	1,48
6	2,40	2,334	0,066	0,004	-1,08	1,17	-1,02	1,03
8	3,10	2,875	0,225	0,051	-0,54	0,29	-0,32	0,10
9	3,20	3,145	0,055	0,003	-0,27	0,07	-0,22	0,05
10	3,20	3,415	-0,215	0,046	0,00	0,00	-0,22	0,05
11	3,90	3,686	0,214	0,046	0,27	0,07	0,48	0,23
12	4,10	3,956	0,144	0,021	0,54	0,29	0,68	0,47
14	4,70	4,496	0,204	0,041	1,08	1,17	1,28	1,65
15	4,50	4,767	-0,267	0,071	1,35	1,83	1,08	1,18
16	5,20	5,037	0,163	0,027	1,62	2,63	1,78	3,18
17	5,00	5,307	-0,307	0,094	1,89	3,58	1,58	2,51
	3,42		0,000	0,525		19,13		19,66

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	19.13	1	19.13	401.09	<0.0001
Días de maduración	19.13	1	19.13	401.09	<0.0001
Error	0.52	11	0.05		
Total	19.66	12			

Evaluación del modelo de regresión: Coeficiente de determinación

- El buen ajuste de un modelo de regresión se mide usando el coeficiente de determinación R^2
- Mide la proporción de variabilidad de la variable respuesta explicada por el modelo de regresión.

$$R^2 = \frac{SC\,\mathrm{exp}\,lic}{SCtotal}$$

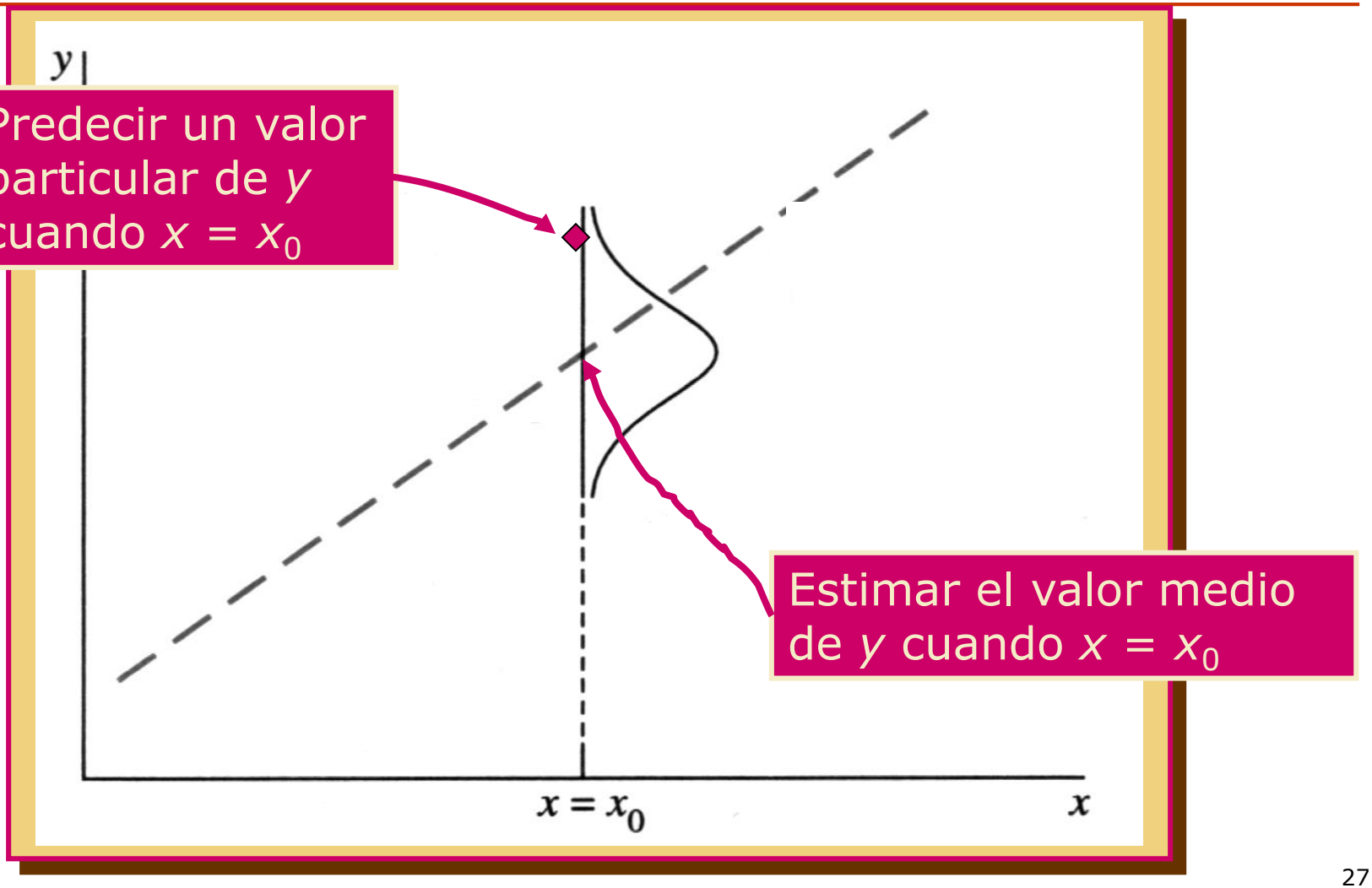
- R^2 es una cantidad adimensional que sólo puede tomar valores entre 0 y 1
- Cuando un ajuste es bueno, R^2 será cercano a uno.
- Cuando un ajuste es malo R^2 será cercano a cero.
- Es una medida de la capacidad predictiva del modelo *dentro* del rango considerado
- En el ejemplo **$R^2 = 0.97$**

Estimación y predicción

- Una vez estimado y validado el modelo, una de sus aplicaciones más importantes consiste en poder realizar estimaciones y predicciones acerca del valor que tomaría la variable dependiente en el futuro o para una unidad extramuestral.
- Se pueden construir intervalos de confianza sobre dichos valores
- Los pronósticos son válidos en el rango estudiado

- Se puede utilizar el modelo de regresión obtenido para:
  - Estimar el valor promedio de y para un dado X
  - Predecir el valor de Y para un dado X
  - Estimar la pendiente

Estimación y predicción



Estimación y predicción

- El mejor estimador de $E(y)$ y de y es

$$\hat{y} = b_0 + b_1 x_0$$

- En la predicción de un valor particular de y hay más incertidumbre que en la estimación de un promedio, lo que se refleja en una mayor amplitud del intervalo

- ¿Cuál será el contenido medio de proteína a los 7 días de desarrollo embrionario?

$$\hat{y}_i = 0.71 + 0.27 x_i = 0.71 + 0.27 * 7 = 2.6g$$

$$2.6 \pm EM$$

Intervalos de confianza

- **Para estimar la media de y:**

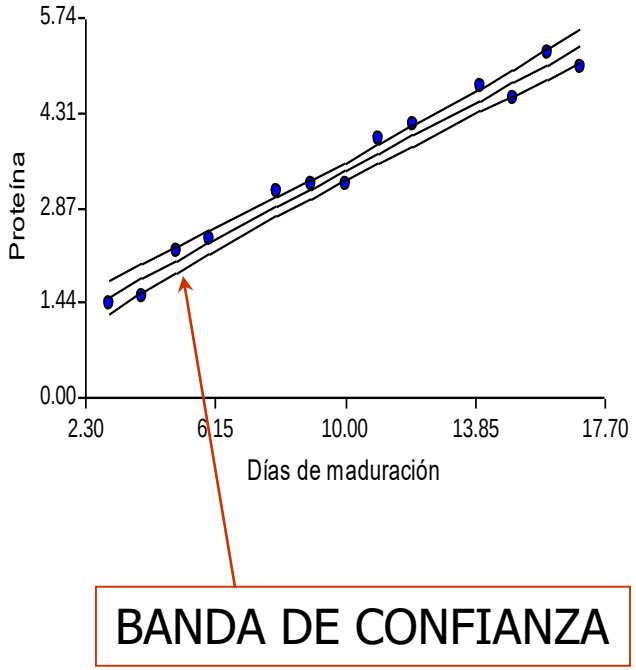
$$\hat{y}_0 \pm t_{n-2;1-\alpha/2}\sqrt{CM\,res\left[\frac{1}{n}+\frac{(x_O-\bar{x})^2}{SC_{xx}}\right]}$$

- **Para predecir un valor de y:**

$$\hat{y}_O \pm t_{n-2;1-\alpha/2}\sqrt{CM_{res}\left[1+\frac{1}{n}+\frac{(x_O-\bar{x})^2}{SC_{xx}}\right]}$$

- **Para la pendiente**

$$b_1 \pm t_{n-2;1-\alpha/2}\sqrt{\frac{CM_{res}}{SC_{xx}}}$$



Ejemplo: Crecimiento en niños

Se estudió el crecimiento de niños y se determinó que sigue una relación lineal en función de la edad.
La recta de regresión hallada para niños de 4 a 9 años fue:

$$\hat{y} = 80 + 6x$$

donde Y es la altura en cm y X es la edad en años

a) Interprete los coeficientes de la recta
b) ¿Cuál es la altura estimada de un niño de 8 años?
c) ¿Cuál sería la altura de una persona de 25 años?

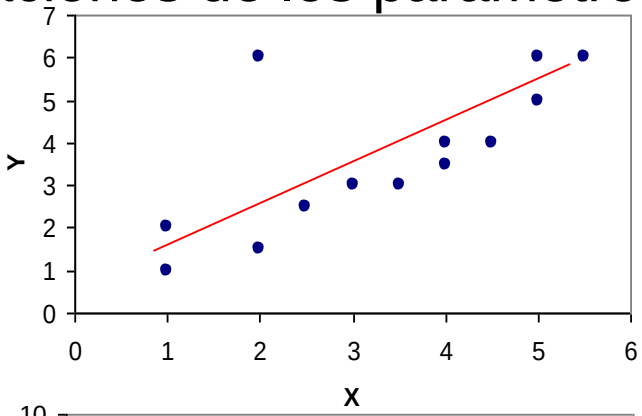
Cuando la recta de regresión se usa para predecir valores de Y para valores de X fuera del rango de los valores observados de X, se dice que se ha hecho una **extrapolación**

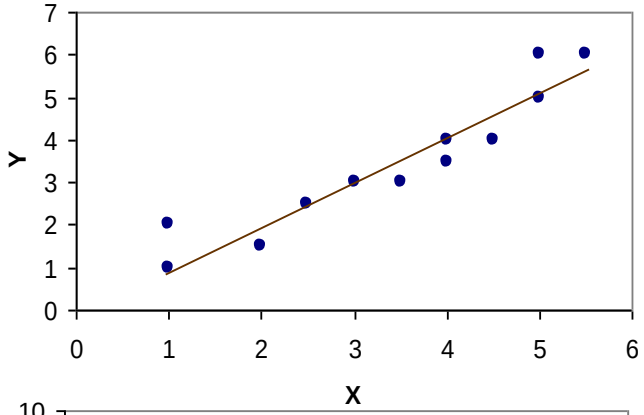
Observaciones atípicas e influyentes

Atípicas: Son observaciones alejadas de la recta, que poseen un residuo significativo

Influyentes: Son observaciones que tienen mucho peso en las estimaciones de los parámetros

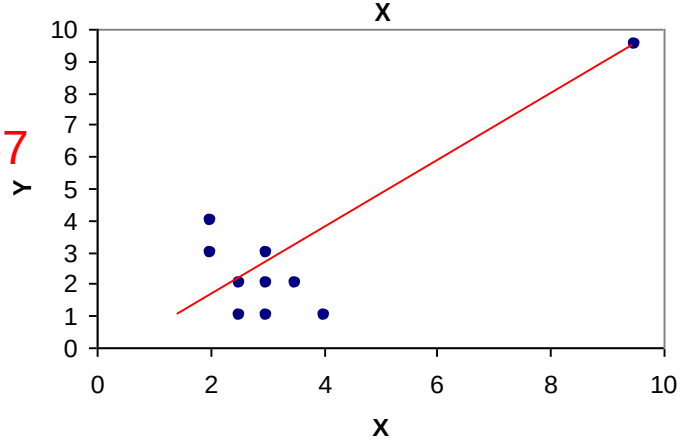
R^2=0.52

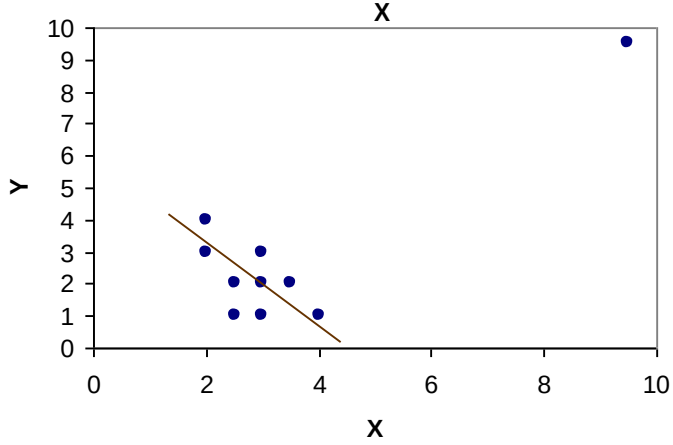




R^2=0.89

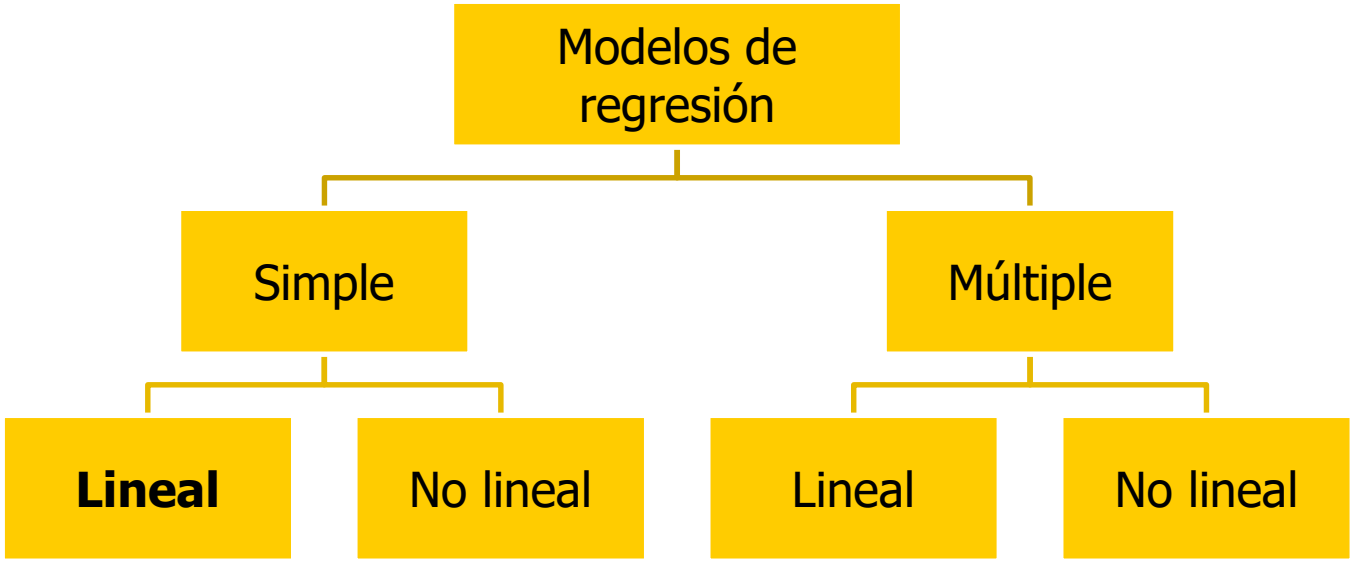
R^2=0.67





R^2=0.36

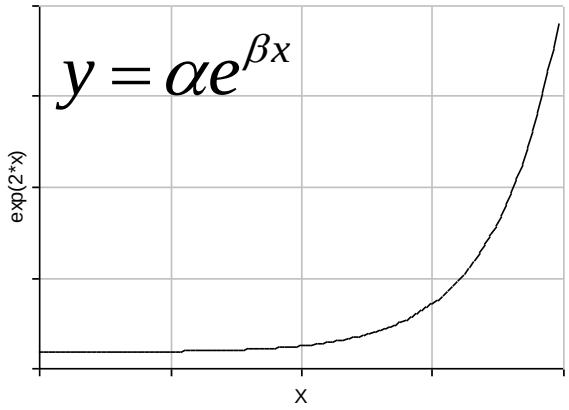
Modelos de análisis de regresión



- Relaciones no lineales pueden convertirse en lineales aplicando transformaciones a las variables:
  - Transformando a *X*
  - Transformando a *Y*
  - Transformando ambas

Funciones linealizables

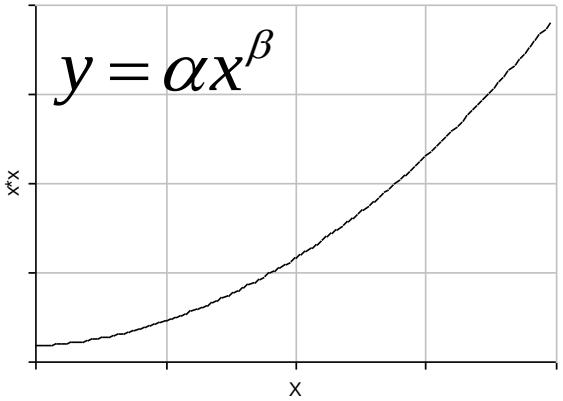
Exponencial



$$y = \alpha e^{\beta x}$$

$$y^* = \ln y$$

Potencia

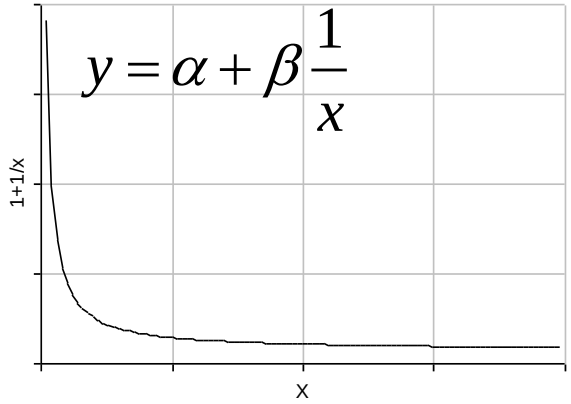


$$y = \alpha x^{\beta}$$

$$y^* = \ln y$$

$$x^* = \log x$$

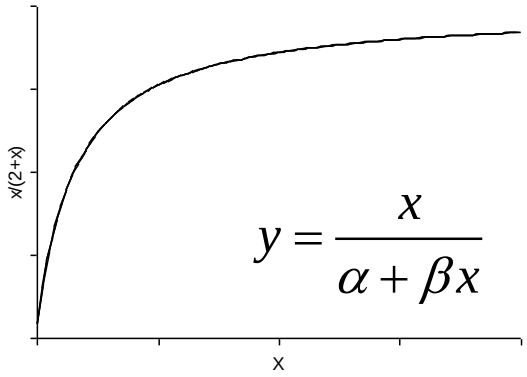
Recíproca



$$y = \alpha + \beta \frac{1}{x}$$

$$x^* = \frac{1}{x}$$

Hiperbólica



$$y = \frac{x}{\alpha + \beta x}$$

$$y^* = \frac{1}{y}$$

$$x^* = \frac{1}{x}$$

Modelo de correlación

El **coeficiente de correlación lineal de Pearson** indica el grado de asociación lineal entre dos variables aleatorias

- Coeficiente de correlación poblacional ρ **(parámetro)**

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(xy)}{\sigma_x \sigma_y}$$

- Coeficiente de correlación muestral **r (estimador)**

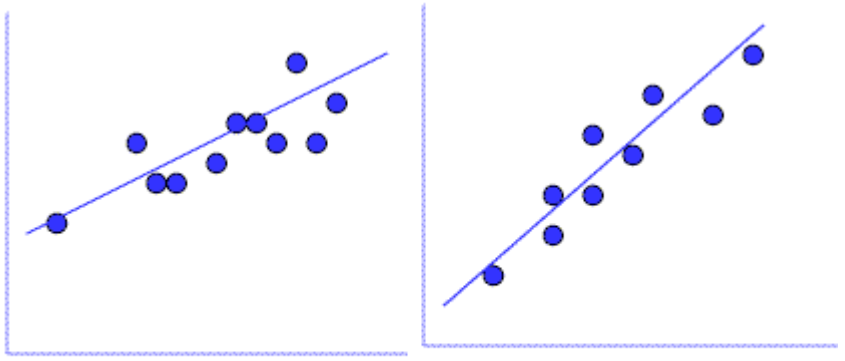
$$r = \frac{S_{xy}}{S_x S_y}$$

Coeficiente de correlación lineal

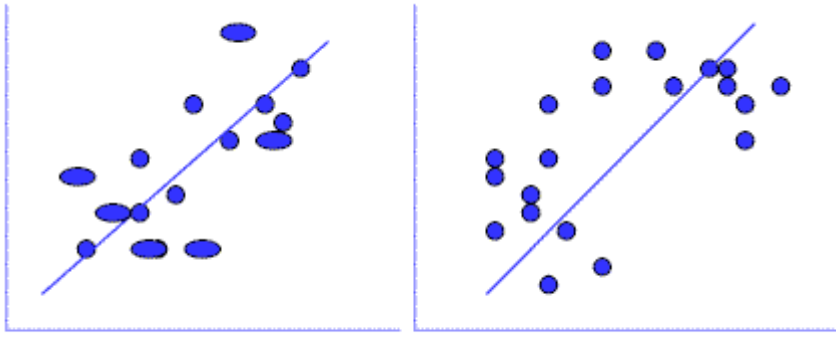
- Solo toma valores entre [-1,1]
- No tiene unidades, ya que surge de una estandarización de la covarianza
- Su **signo** nos indica si la posible relación es directa o inversa:
  - Directa: r >0
  - Inversa: r <0
  - Variables independientes: r =0
- Cuanto más cerca esté de +1 o -1 mejor será el grado de relación lineal (siempre que no existan datos anómalos)
- Es útil para determinar si hay relación **lineal** entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)

¿Qué no mide el coeficiente de correlación?

- no mide la magnitud de la pendiente

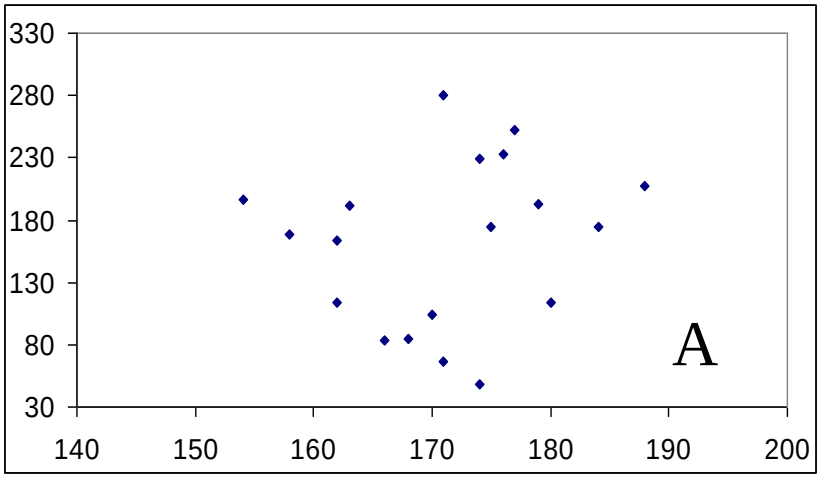


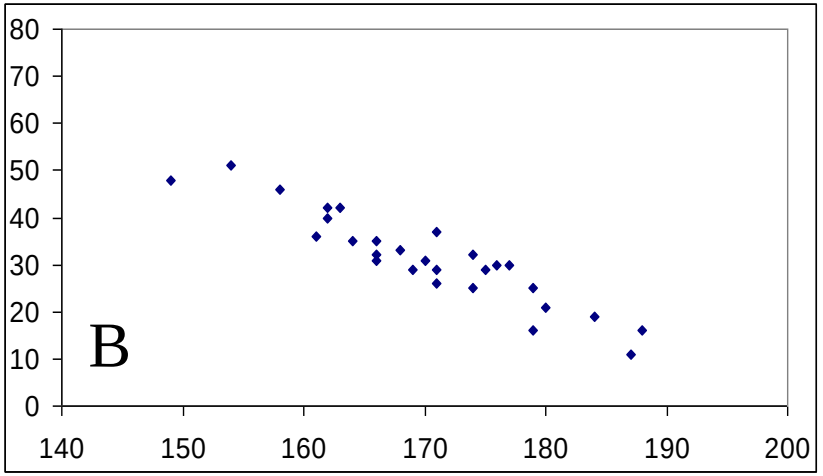
- tampoco mide lo apropiado del modelo lineal

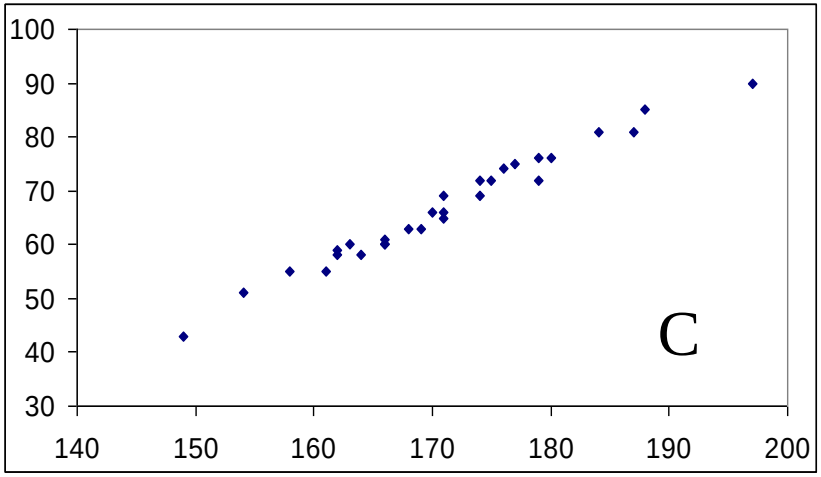


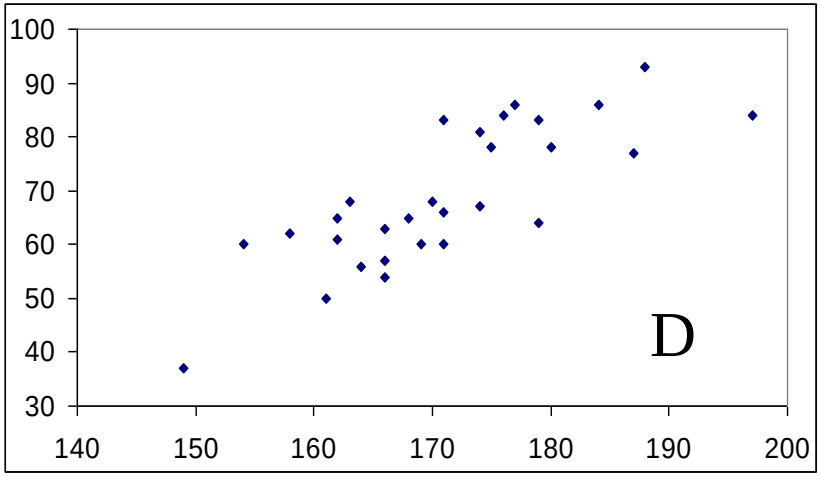
Algunos ejemplos

0.1 0.99 0.8 -0.95









Prueba de hipótesis para ρ

- Ho: ρ = 0 (no existe asociación lineal entre las variables aleatorias X e Y)
- H1: ρ ≠ 0 (sí existe asociación lineal entre las variables aleatorias X e Y)

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Un ejemplo

A possible role of social activity to explain differences in publication output among ecologists

Tomáš Grim

Oikos 2008

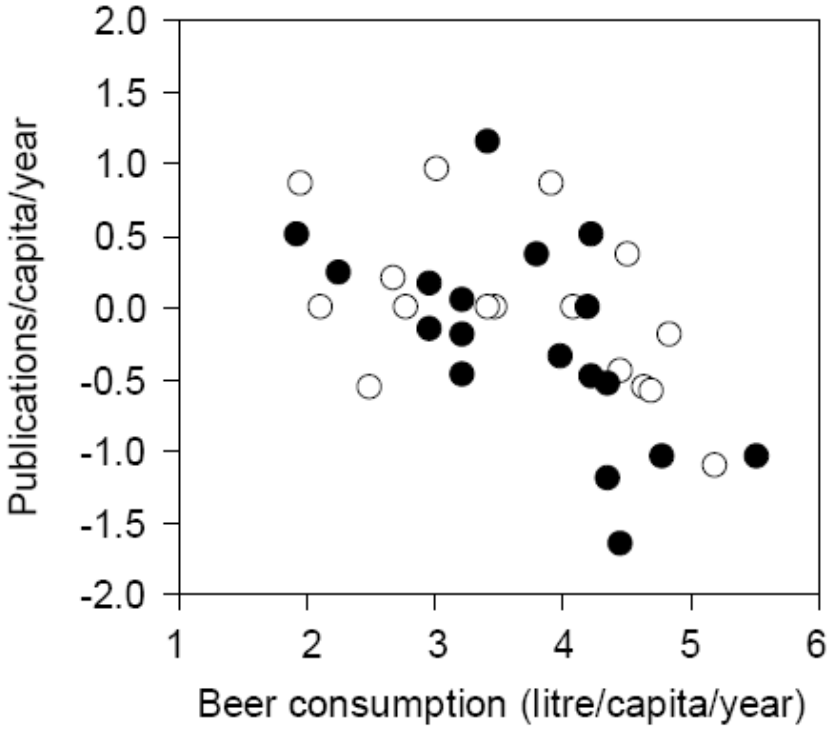


Fig. 1. Number of publications per capita per year published by Czech avian ecologists up to 2006 plotted against their beer consumption per capita per year in litres. Both data sets shown are Box-Cox transformed (thus neither the output score nor the consumption score values enable the identification of particular persons included in this research). The negative relationship between beer consumption and publication success is significant not only for the whole data set ($r_s = -0.55$, $n = 34$, $p = 0.0008$) but also for "past" (included in the first survey in 2002; ●) and "present" researchers (included in 2006; ○) analyzed separately ("past": $r_s = -0.68$, $n = 18$, $p = 0.002$; "present": $r_s = -0.52$, $n = 16$, $p = 0.04$).

Atención:

- Si se observa una relación entre dos variables, eso no implica necesariamente una relación causa-efecto (sobre todo en estudios observacionales)
- La relación entre dos variables puede estar influenciada por una tercer variable no estudiada (variable subyacente o de confusión)
- La relación puede estar influenciada por outliers o por puntos influyentes
- No es correcto extrapolar