Correlación

Correlación y Regresión

- En estadística es común analizar la relación entre variables y determinar si tal relación es real o se debe al azar.
- El caso más simple es la relación entre 2 variables.
- Se necesita especificar que tipo de análisis se quiere realizar:
  - ¿Es el propósito explorar el tipo de relación entre ambas variables?
  - ¿Es el propósito usar una variable para explicar la variación en otra variable?
  - Existe una diferencia en el examinar la relación entre peso y altura, y usar altura para predecir el peso.
- Es necesario distinguir entre:
  - Un análisis de correlación, donde se mide la intensidad de la relación entre 2 variables
  - Una variable es utilizada para predecir los valores de una segunda variable.

Correlación

- El término correlación generalmente se confunde como cualquier tipo de correlación entre variables.
- Sin embargo, la definición técnica de correlación es:
  - La intensidad de una aparente relación lineal entre 2 variables continuas.
- La correlación no distingue entre variable respuesta (dependiente) y variable explicatoria (independiente). La correlación de X e Y es igual a la correlación de Y y X.

Gráfico de Dispersión (Scatterplot)

- El gráfico de dispersión es la primera herramienta para explorar la relación entre 2 variables continuas.
- Generalmente se gráfica la variable respuesta (Y) en el eje vertical y la variable exploratoria en el eje horizontal (X). Si no hay una clara distinción entre ambas variables entonces no importa que variable va en cada eje.

Gráfico de Dispersión (Scatterplot)

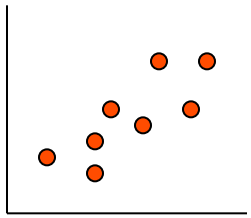
Que se debe buscar en un gráfico de dispersión

- Patrón general
- Forma de la relación
- Intensidad de la relación
- Observaciones atípicas (outliers)

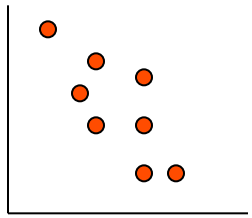
Gráfico de Dispersión (Scatterplot)

Que se debe buscar en un gráfico de dispersión

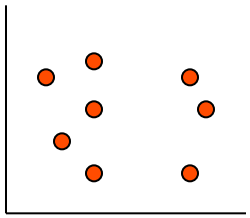
Patrón general



Asociación positiva



Asociación negativa

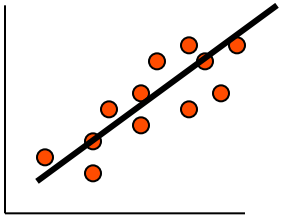


No asociación

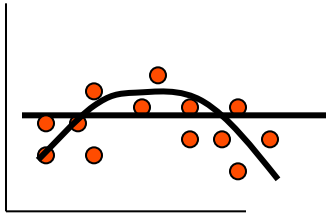
Gráfico de Dispersión (Scatterplot)

Que se debe buscar en un gráfico de dispersión

Forma de la relación



Relación lineal

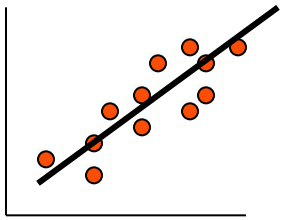


Relación curvilineal

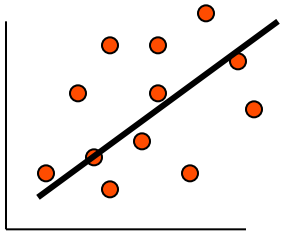
Gráfico de Dispersión (Scatterplot)

Que se debe buscar en un gráfico de dispersión

Intensidad de la relación



Asociación intensa

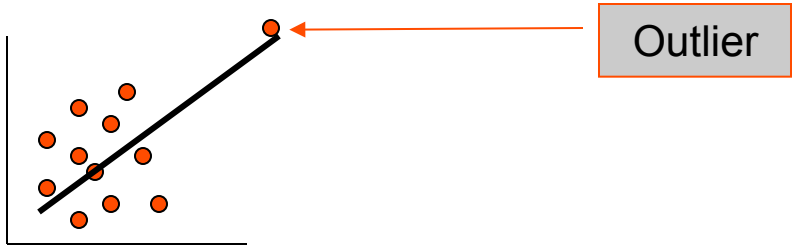


Asociación débil

Gráfico de Dispersión (Scatterplot)

Que se debe buscar en un gráfico de dispersión

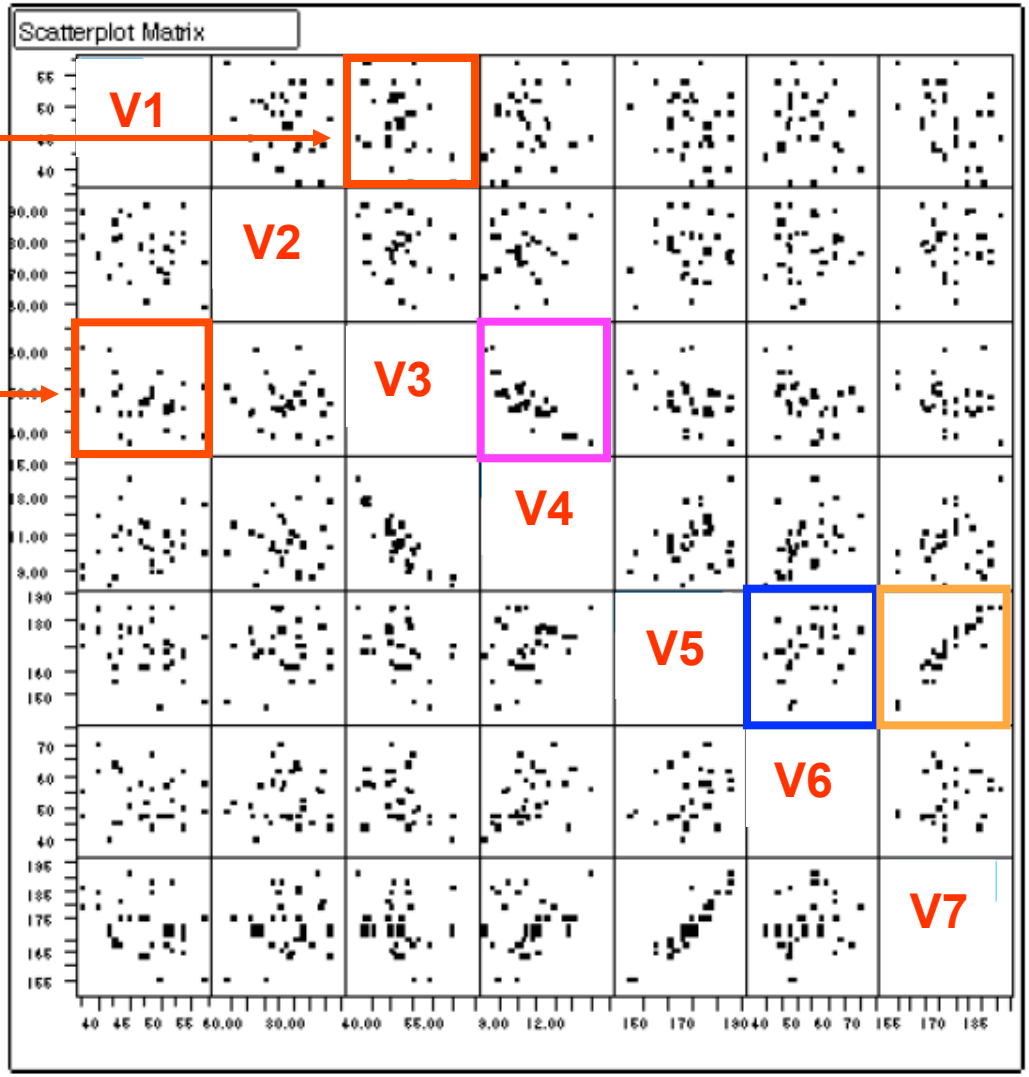
Observaciones atípicas (outliers)



Interpretando una matriz de dispersión

Por ejemplo, fila 1 columna 3 es la variable V1 en eje Y contra V3 en eje X.

En cambio, fila 3 columna 1 es la variable V1 en el eje X contra variable V3 en el eje Y.



Intensidad de asociación:

V5 contra V7 fuerte
V5 contra V6 débil

Dirección de la asociación

V5 contra V7 positiva
V3 contra V4 negativa

Covarianza y Correlación

Una medida de la intensidad de la relación lineal entre 2 variables continuas es determinar cuanto covarían las variables.

Covarianza (σ_{Y1Y2}) :

$$s_{Y_1Y_2} = \frac{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{n-1}$$

Si una variable no cambia mientras la otra aumenta o disminuye, entonces las variables no covarían.

También es posible cuantificar la intensidad de la correlación entre 2 variables mediante el coeficiente de correlación.

Coeficiente de correlación muestral (r):

$$r_{y_1y_2} = \frac{\sum_{i=1}^{n}[(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n}(y_{i2} - \bar{y}_2)^2}}$$

El coeficiente de correlación de la población es ρ (letra griega Rho)

Covarianza y Correlación

- La covarianza va de -∞ a ∞.
- El coeficiente de correlación producto-momento de Pearson r estandariza la covarianza al dividirla por el desvío estándar de las dos variables, así la relación lineal entre 2 variables va de -1 a 1.
- r mide la intensidad de la relación lineal (línea recta) entre 2 variables.
- La relación puede ser positiva o negativa, y si no hay correlación, entonces r = 0
- Si el muestreo no es aleatorio, r no debería ser utilizado.
- r es una medida sin unidad.

Covarianza y Correlación

Prueba de hipótesis para ρ = 0

Se obtiene un valor de t que se compara con el valor crítico de t para grados de libertad = n-2

$$t = \frac{r}{s_r}$$

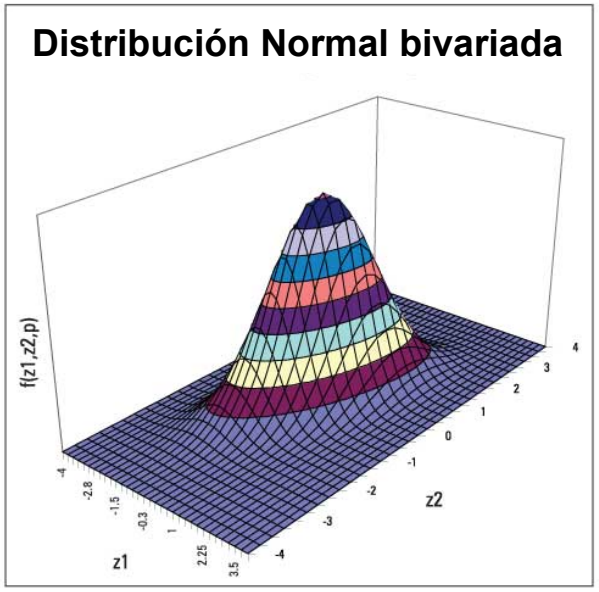
Error estándar

$$s_r = \sqrt{\frac{\left(1 - r^2\right)}{\left(n - 2\right)}}$$

También es posible obtener intervalos de confianza para r.

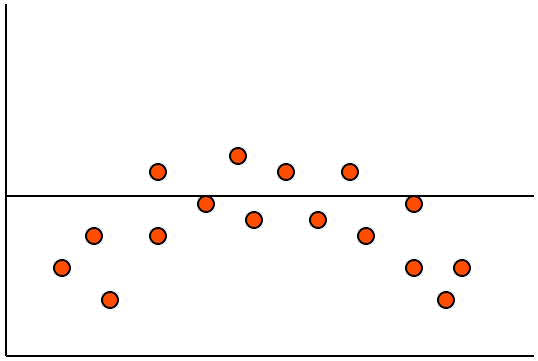
Supuestos del coeficiente de Correlación parámetrico de Pearson

- Para que r no sea severamente sesgado se tiene que cumplir con los siguientes supuestos:
- Muestras obtenidas al azar
- Independencia de las observaciones
- La distribución conjunta de ambas variables es normal bivariada.



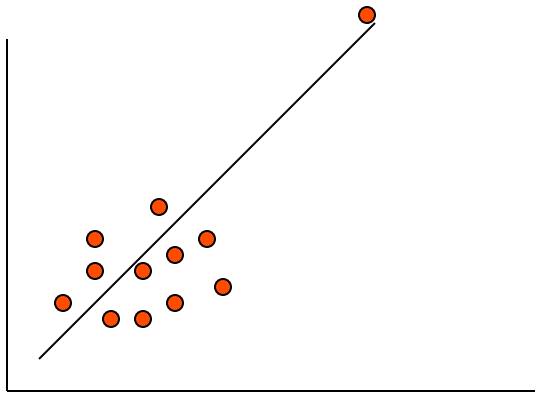
Precauciones en la interpretación de la correlación

La correlación mide la intensidad de relaciones lineales (línea recta), una relación curvilínea puede tener una correlación 0, sin embargo, la correlación entre ambas variables puede ser intensa.



Precauciones en la interpretación de la correlación

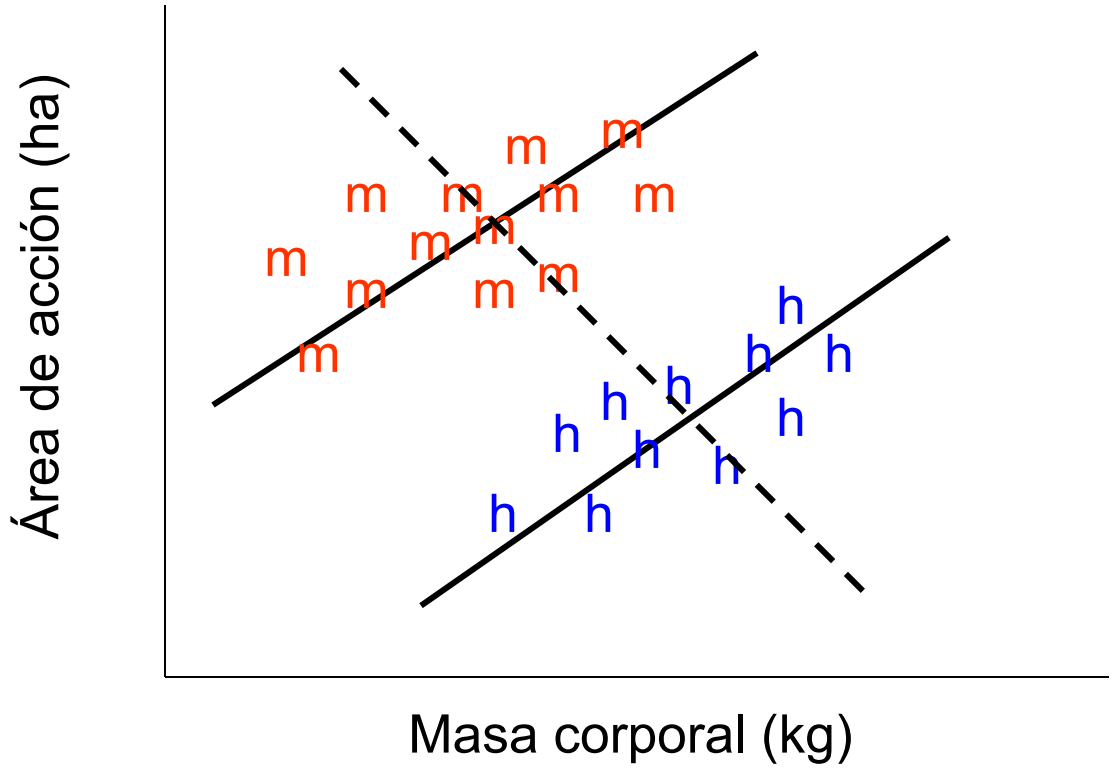
Efecto de outliers



Precauciones en la interpretación de la correlación

Efecto “escondido” de variables (lurking variables)

Por ejemplo, el área de acción puede tener una asociación positiva con el peso corporal de los machos, lo que también se da en la hembras, sin embargo, el área de acción de los machos es mayor que el área de acción de las hembras. Hay una relación positiva dentro de cada sexo, pero en general hay una relación negativa cuando los datos son agrupados.

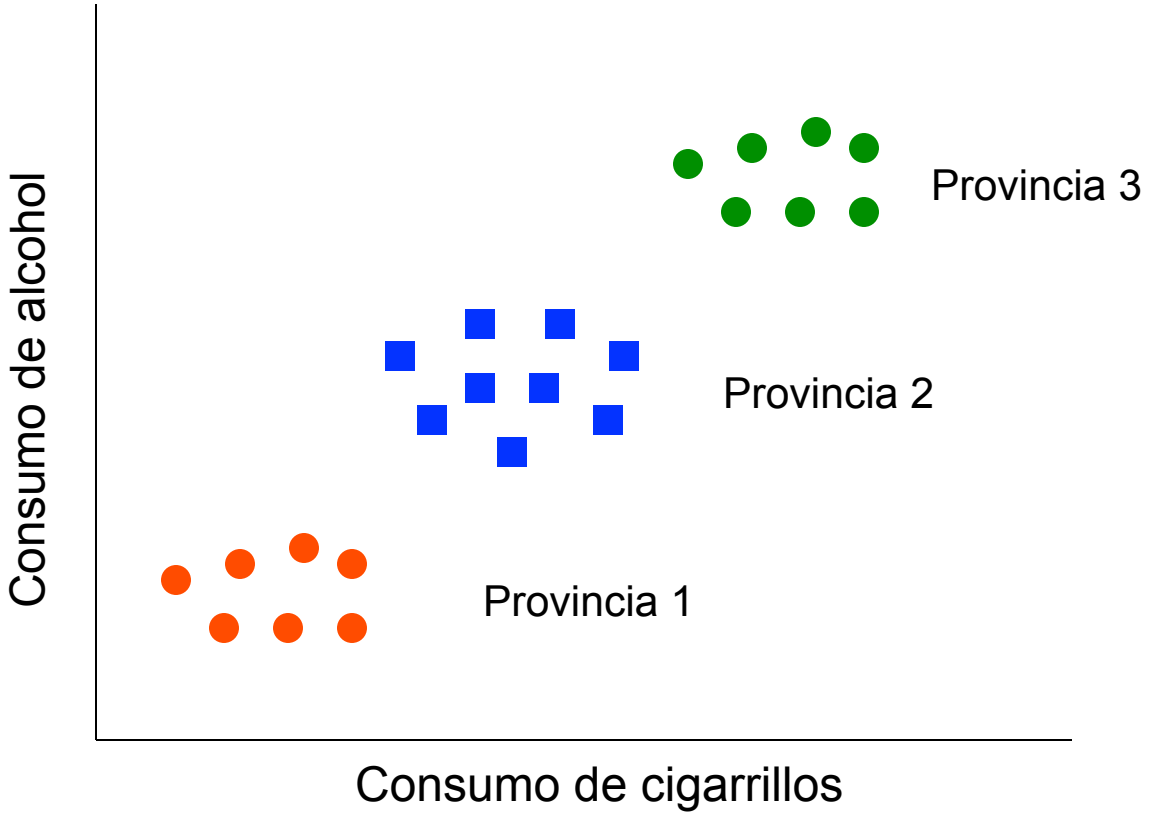


Precauciones en la interpretación de la correlación

La falacia ecológica

Cuando la inferencia sobre individuos se hace sobre estadísticos que resultan del agrupamiento de las observaciones.

Aún cuando hay alta correlación entre los promedios de las variables, no significa que haya correlación entre los valores individuales de cada variable



Los estereotipos son una forma de Falacia Ecológica

Precauciones en la interpretación de la correlación

- Correlación no implica causalidad.
- Se deben satisfacer una serie de principios de inferencia causal para establecer causa y efecto

Métodos de correlación no paramétricos

Cuando no se cumple el supuesto de normalidad:

1. Coeficiente de correlación de Spearman
2. Kendall tau

Precauciones en la interpretación de la correlación

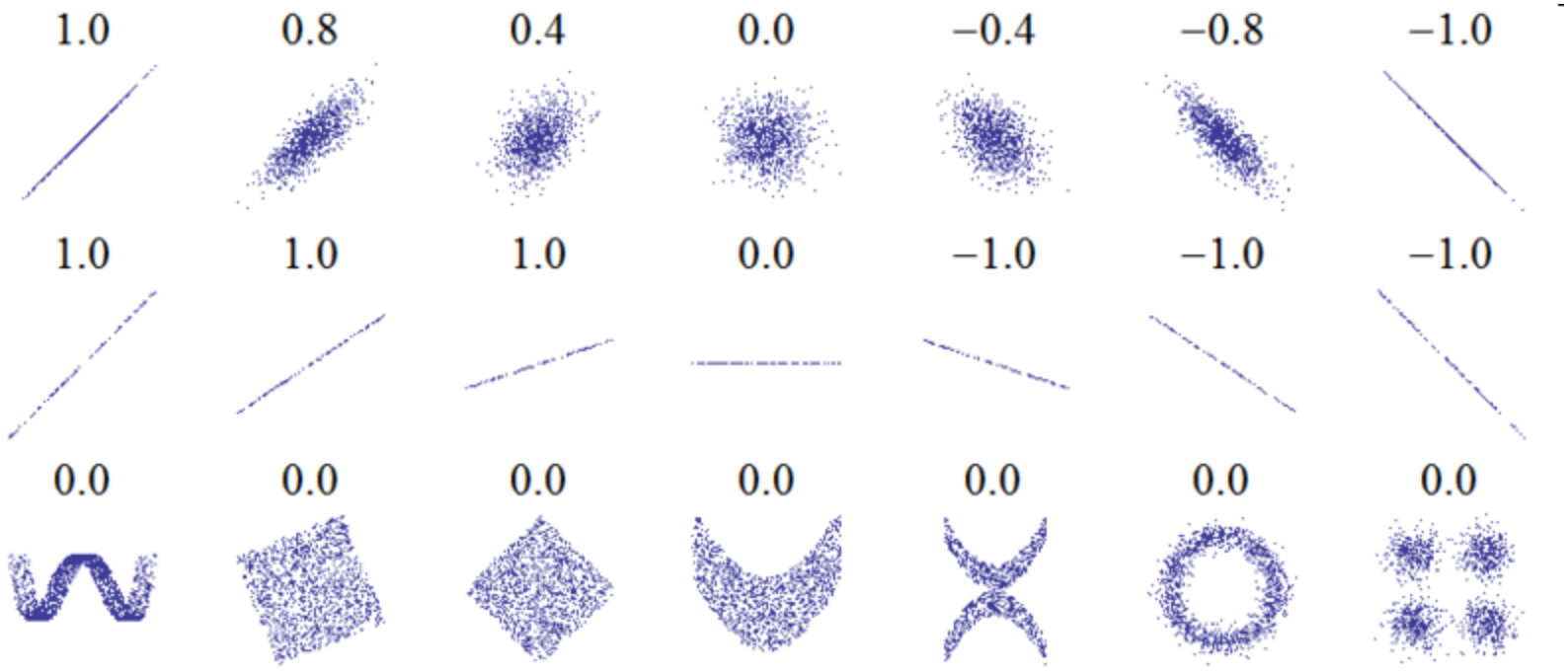
Tipos de asociaciones

Se puede encontrar una asociación entre 2 variables por muchas razones.

- Puede haber causalidad directa
  - Ejemplo: fumar causa cáncer de pulmón.
- Puede haber una causa común
  - Ejemplo: venta de helados y número de ahogados incrementa con la temperatura.
- Puede haber un “confounding factor”
  - Ejemplo: accidentes en la carretera disminuye cuando el límite de velocidad es reducido a 60 km/h, al mismo tiempo que el aumento de la gasolina hace que la gente viaje menos.
- Puede haber una coincidencia
  - Ejemplo: la población de Costa Rica se ha incrementado, al mismo tiempo que la luna se ha acercado unos pocos kilómetros a la tierra.

Precauciones en la interpretación de la correlación

Valores de correlación diferentes



La correlación refleja el “ruido” y la dirección de la relación lineal (fila superior) pero no la pendiente de la relación (fila del medio), ni muchos aspectos de relaciones no lineales.

Modelos Lineales

Regresión Lineal Simple

Modelos Estadísticos

¿Qué es un modelo?

- Un modelo es una descripción generalizada para analizar o explicar algo.
- Comparamos (confrontamos) modelos con datos.
- Evaluamos el ajuste de modelos a los datos observados.

Modelos Estadísticos

- Variable respuesta (dependiente): Y. Se pretende explicar la variación en Y.

  Y puede ser cuantitativa (continua o discreta) o categórica.

- Variables explicativa (independiente): variables medidas (X1, X2, etc.) que pueden explicar Y.

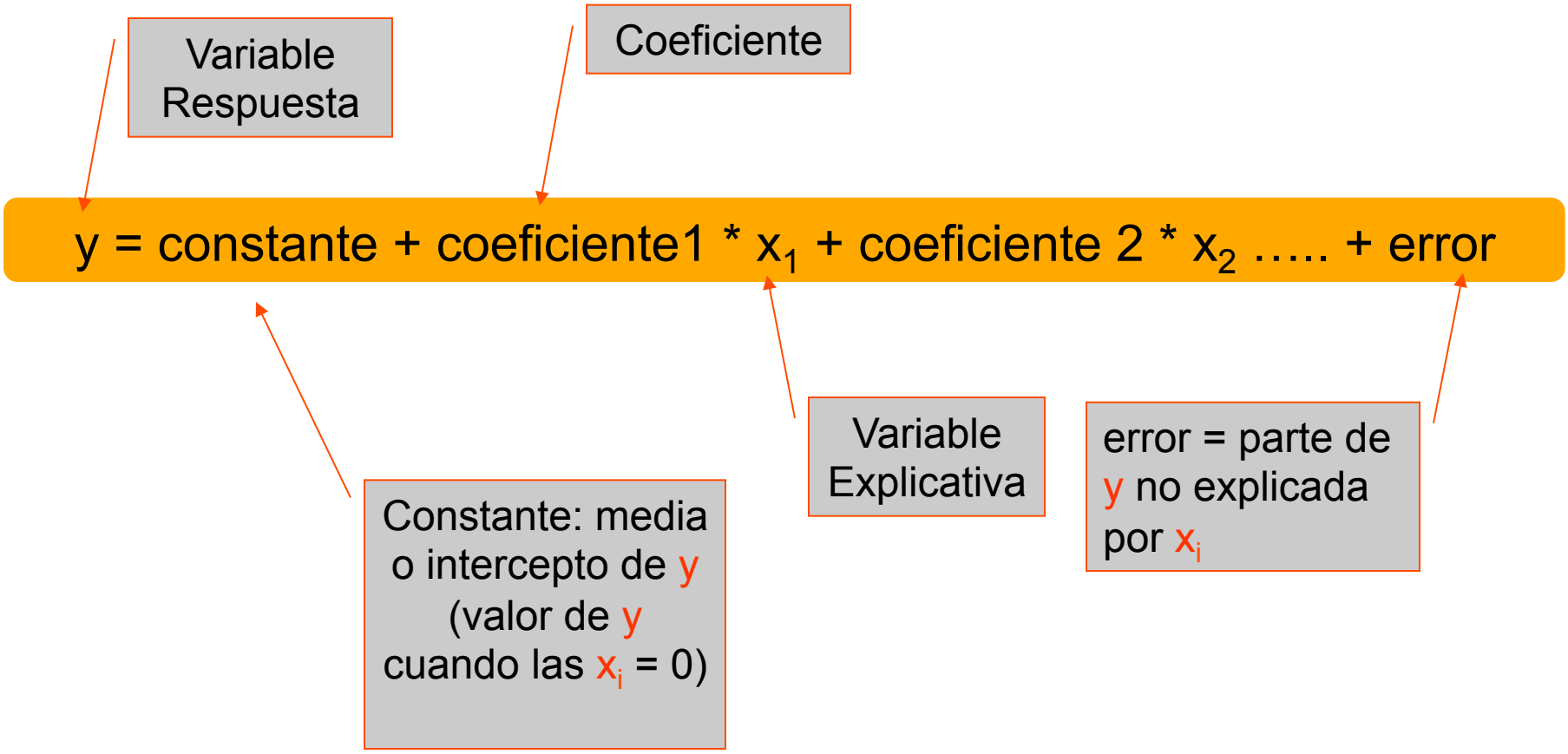
  Xs pueden ser cuantitativas o categóricas (o ambas).

Variable respuesta = modelo + error

Valor observado = componente sistemático + componente aleatorio

Valor observado = señal + ruido

Modelos Lineales



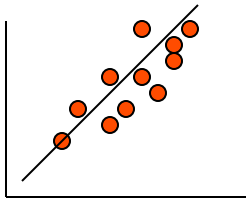
El principal objetivo es ajustar modelos a los datos observados

Modelos Lineales

¿Que significa lineal?

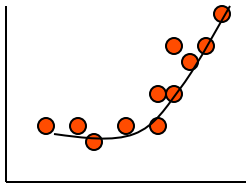
El término lineal se utiliza de 2 maneras diferentes:

1. Modelo de una relación de línea recta entre 2 variables (mayoría de los biólogos lo interpretan de esta manera).
2. Modelo en el cual cualquier valor de la variable dependiente o respuesta (yi) es descrita por una combinación lineal de una serie de parámetros (coeficientes de la pendiente, intercepto) y ningún parámetro aparece como un exponente o es multiplicado o dividido por otro parámetro.
   - Definición más correcta.
   - Se refiere a la combinación de parámetros no a la forma de la relación.



Modelos lineales

$y = a + bx$



$y = a + bx^2$

Modelo no lineal

$y = x1^{\alpha}x2^{\beta} + \varepsilon$

Modelos Lineales

Componente sistemático

- Combinación de variables explicativas
  - 1 o más variables explicativas
  - cuantitativas (continuas y discretas) o categóricas
  - Puede incluir términos de interacción (x1 + x2 + x1*x2) y polinomial ($x1 + x1^2$)
  - Parámetros (coeficientes) son fijos, pero los valores poblacionales son desconocidos y se tratan de estimar a través de los datos.

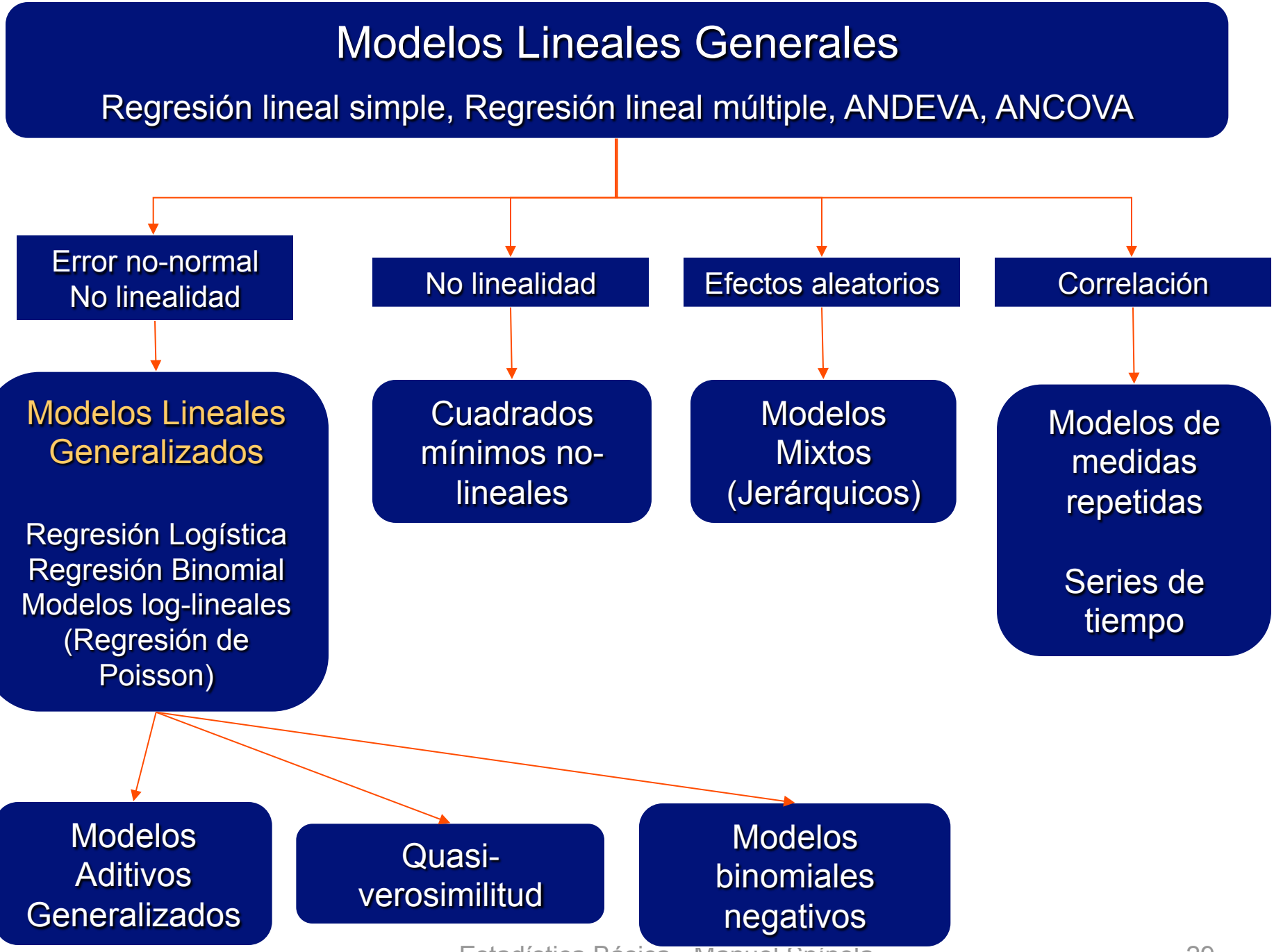
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Parámetro (coeficiente)

Modelos Lineales

Componente aleatorio

- Variable respuesta Y
  - una sola variable cuantitativa o categórica.
- Distribución probabilística de Y
  - Distribución probabilística del término de error del modelo.
- Familia de distribuciones
  - Normal, binomial, Poisson, etc.
  - Definida por media y varianza.


Modelos Lineales Generales
Regresión lineal simple, Regresión lineal múltiple, ANDEVA, ANCOVA
Error no-normal
No linealidad
No linealidad
Efectos aleatorios
Correlación
Modelos Lineales Generalizados
Regresión Logística
Regresión Binomial
Modelos log-lineales
(Regresión de Poisson)
Cuadrados mínimos no-lineales
Modelos Mixtos (Jerárquicos)
Modelos de medidas repetidas
Series de tiempo
Modelos Aditivos Generalizados
Quasi-verosimilitud
Modelos binomiales negativos

Modelos Lineales Generales

- Regresión Lineal Simple
- Regresión Lineal Múltiple
- ANOVA
- ANCOVA

Regresión Lineal

La regresión lineal asume una relación lineal entre la variable respuesta y las variables explicatorias.

Propósitos de la regresión lineal

Descripción

Describir la relación entre la variable respuesta (Y) y la variable explicatoria (X).

Explicación

Cuantificar la variación de Y explicada por la relación lineal con X.

Predicción

Predecir nuevos valores de Y basados en nuevos valores de X.

La mayor parte del tiempo los ecólogos usan la regresión para descripción y explicación.

Regresión Lineal

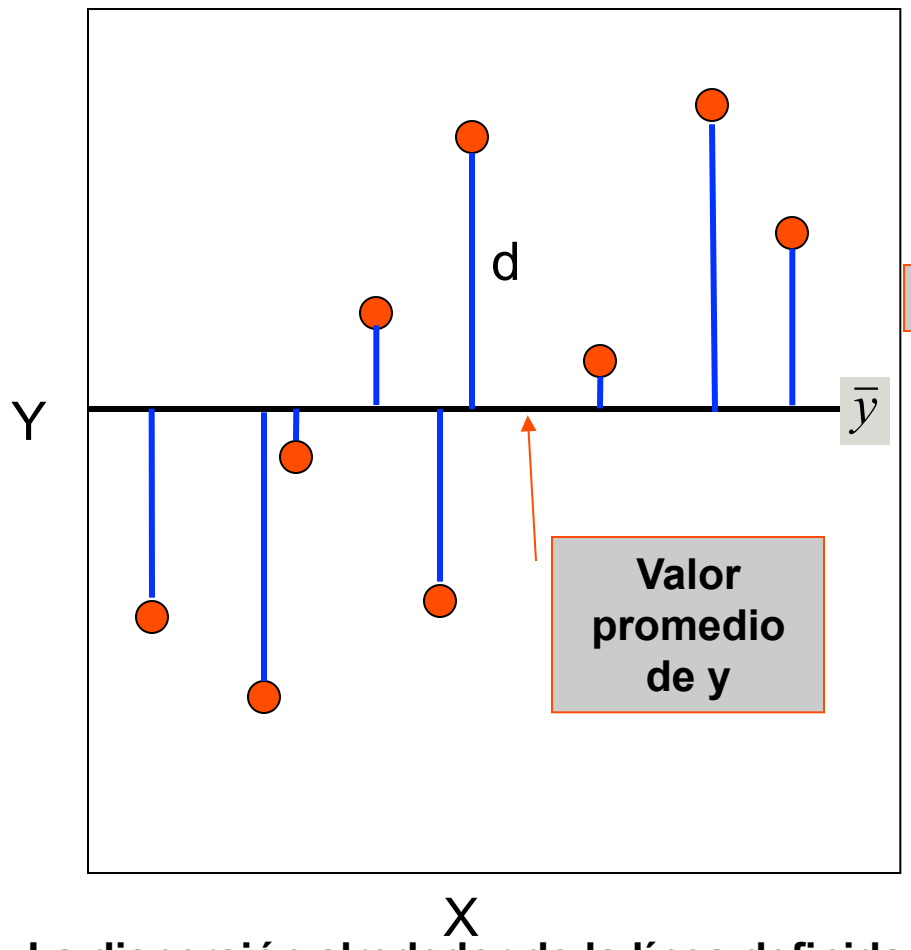
Regresión lineal simple

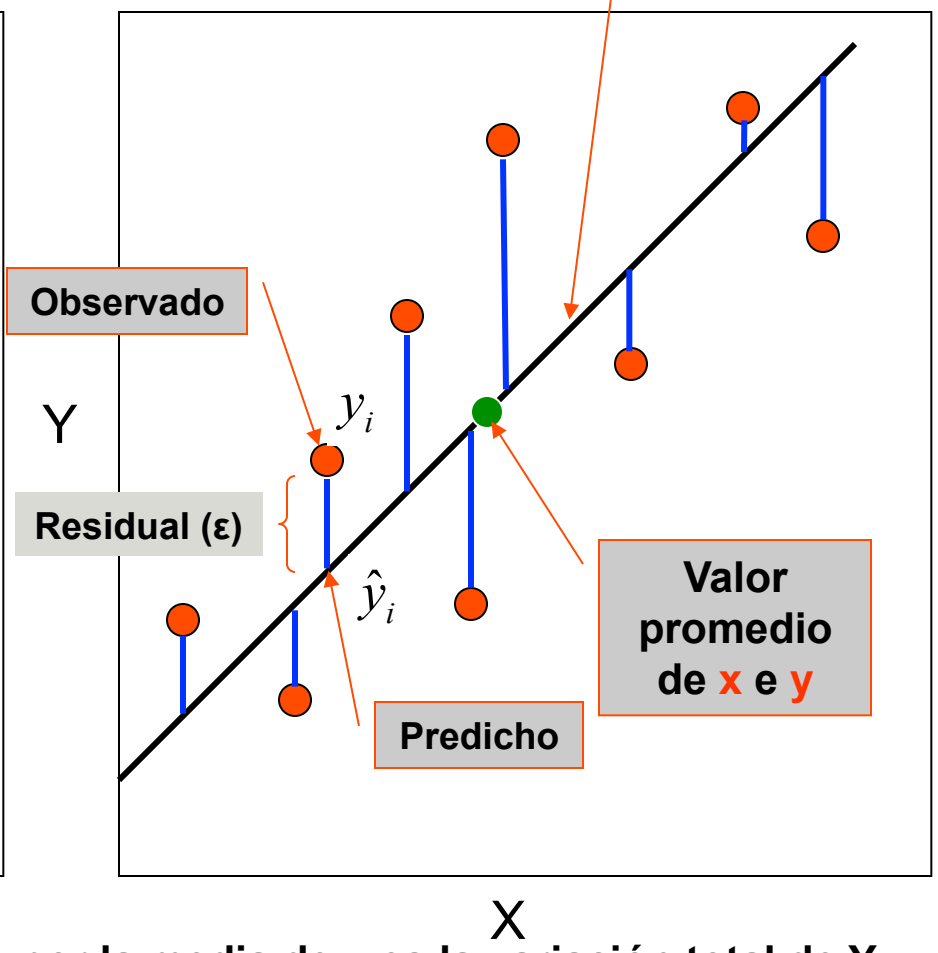
Cuando se tiene una variable respuesta (Y) y una variable explicativa (X).

Formato de datos para una regresión

Observación	Y	X
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
4	y_4	x_4
i	y_i	x_i
Media muestral	$\bar{y}$	$\bar{x}$
Media poblacional	μ_Y	μ_X

Regresión Lineal





La dispersión alrededor de la línea definida por la media de y es la variación total de Y.

$$d = (y - \bar{y})^2$$

$$SCT = \sum (y - \bar{y})^2$$

Suma de cuadrados total

$$\hat{y} = a + bx$$

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

$$SCE = \sum (y - \hat{y})^2$$

Suma de cuadrados de error

Regresión Lineal

Modelo de regresión simple

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

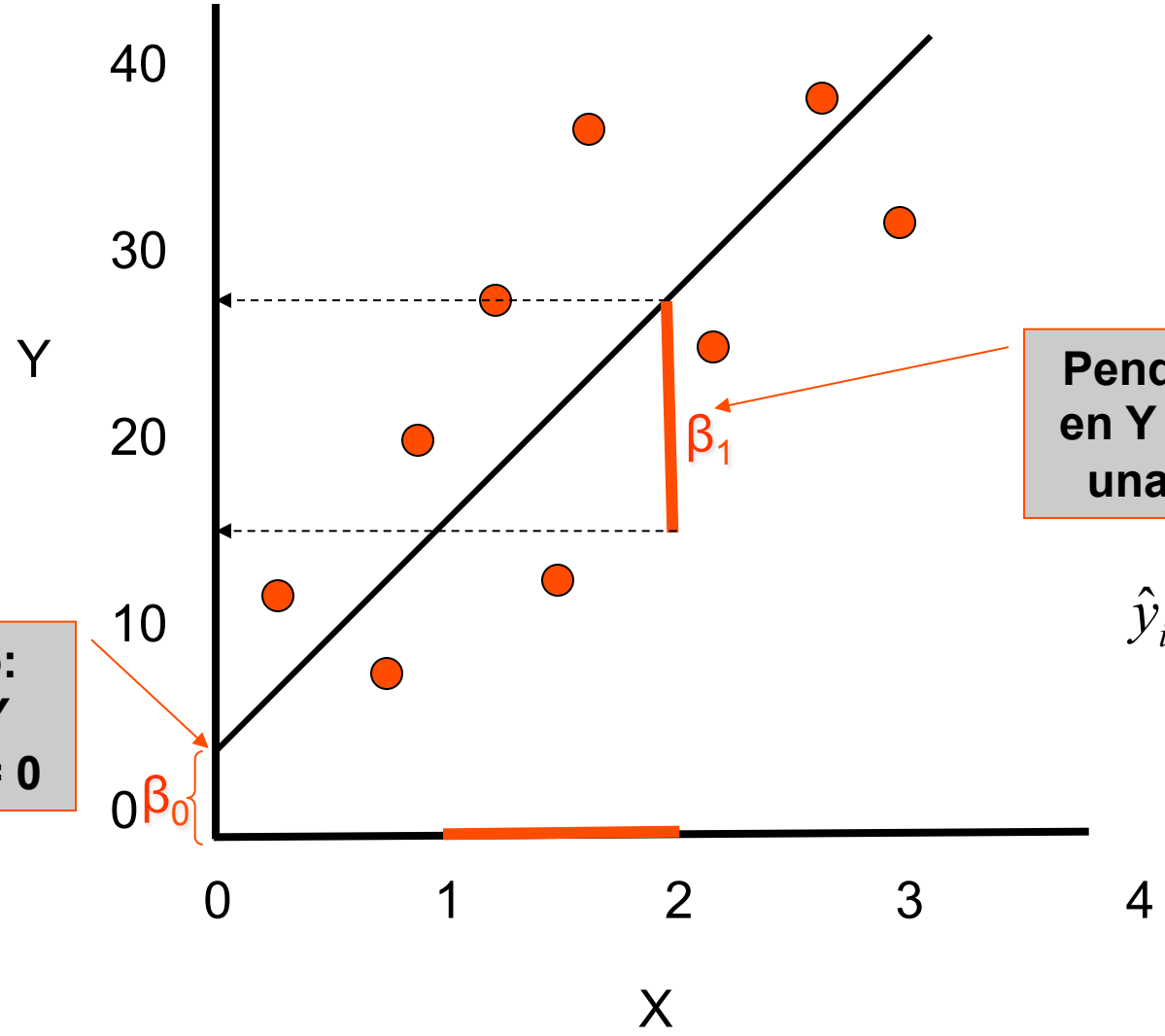
- y_i: Valor de Y para la enésima observación
- β_0: Intercepto de la población
- β_1: Pendiente de la población
- X_i: Valor de X para la enésima observación
- ε_i: Término de error para la enésima observación

Ecuación de regresión de la muestra

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

- $\hat{y}_i$: Valor predicho de Y para x_i
- b_0: Estimado del intercepto (β_0)
- b_1: Estimado de la pendiente de la regresión (β_1)
- x_i: Valor de la variable X

Regresión Lineal



Pendiente: Cambio en Y por cambio de una unidad en X.

Intercepto: Valor de Y cuando X = 0

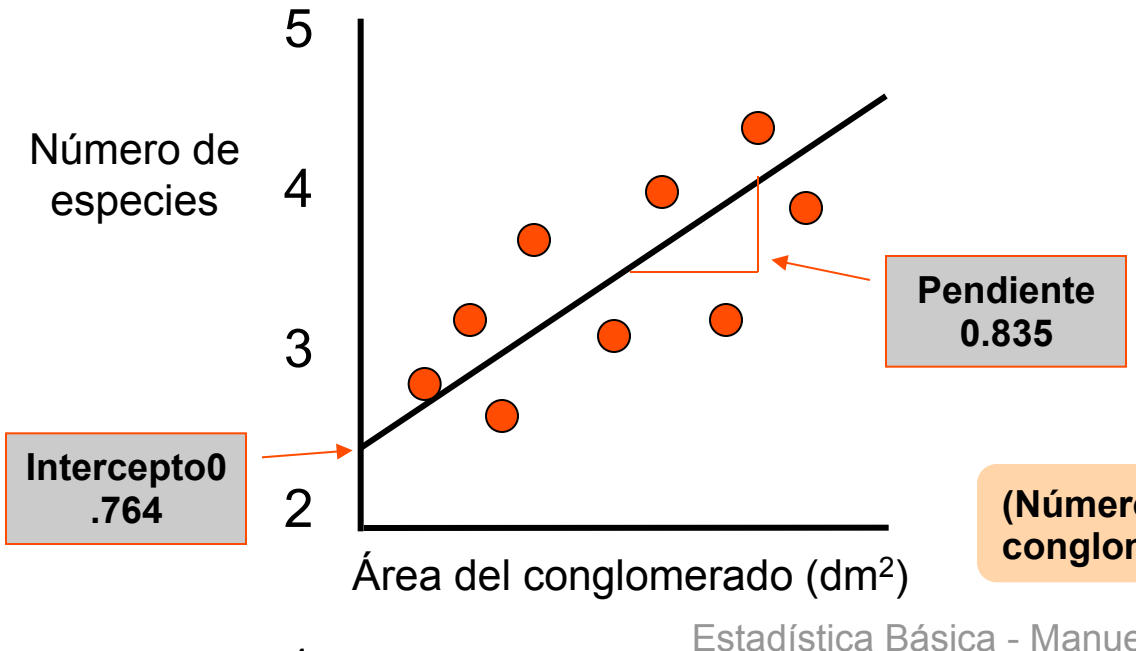
$$\hat{y}_i = b_0 + b_1 x_i$$

Estimación de parámetros en Regresión Lineal

Ejemplo

Número de especies de macro-invertebrados por área de conglomerados de mejillones.

Coeficiente	Estimado	Error Estándar
Intercepto (b_0)	0.764	0.025
Pendiente (b_1)	0.835	0.071



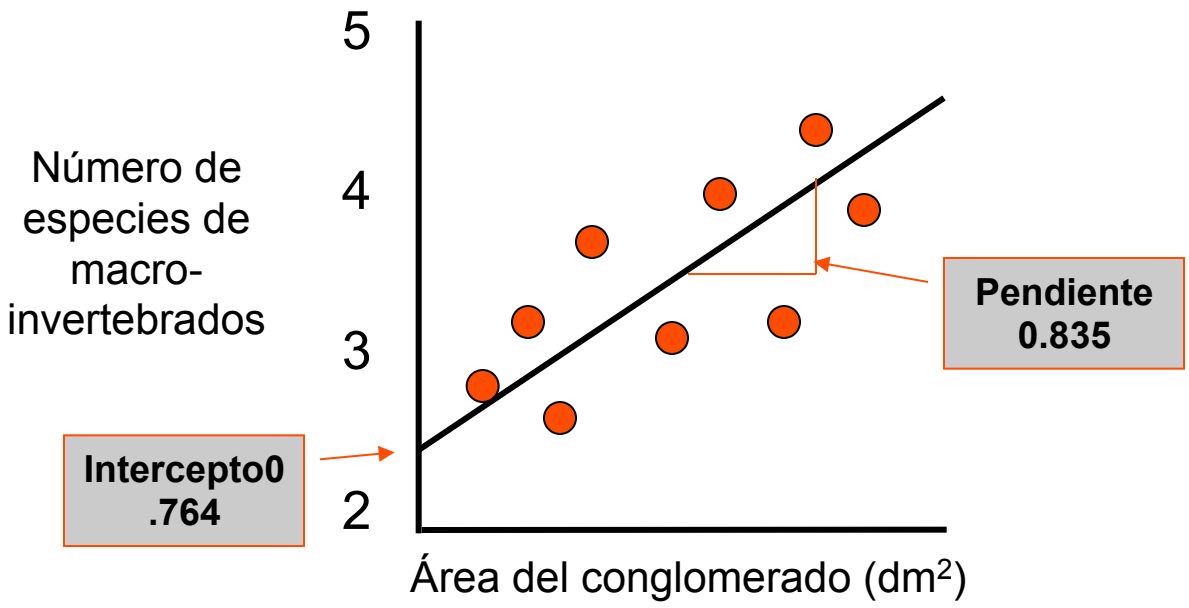
$$\hat{y}_i = b_0 + b_1 x_i$$

(Número de especies)i = β_0 + β_1(área de conglomerado de mejillones)i + ε_i

Estimación de parámetros en Regresión Lineal

Ejemplo

Número de especies de macro-invertebrados en área de conglomerados de mejillones.



(Número de especies)i = β_0 + β_1(área de conglomerado de mejillones)i + ε_i

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

Número predicho de especies para cada valor de área del conglomerado

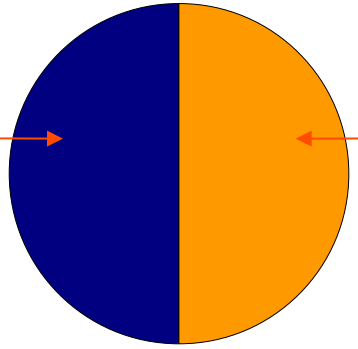
Número predicho de especies cuando área es 0

El cambio estimado en el número de especies cuando aumenta el área del conglomerado en 1 dm²

Partición de la variación en Regresión Lineal

La variación en la variable respuesta (Y) se divide en 2 componentes:

Variación explicada por la variable explicatoria (Modelo de Regresión)



Variación no explicada por la variable explicatoria (Variación residual o error)

$$\sum (y_i - \bar{y})^2$$

Variación Total (Suma de Cuadrados Total) en Y

$$\sum (\hat{y}_i - \bar{y})^2$$

Variación en Y explicada por la regresión ($SC_{Regresión}$)

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

Variación en Y no explicada por la regresión ($SC_{Residuales}$)

Análisis de Varianza en Regresión Lineal

Cuadrados medios

Promedios de las sumas de los cuadrados de las desviaciones.

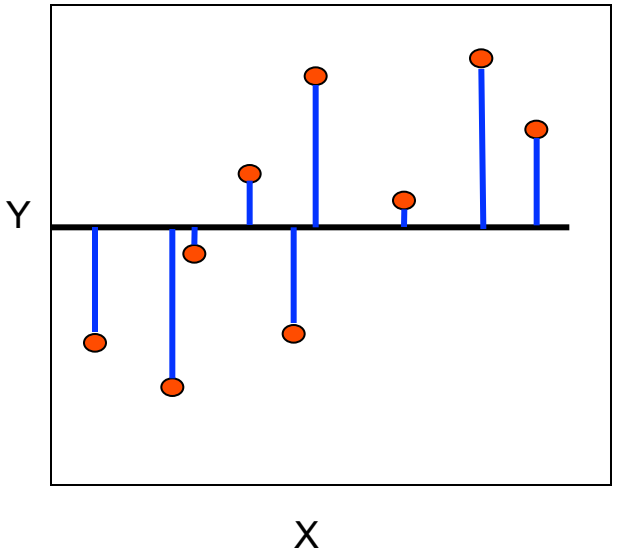
Grados de libertad: g.l. total (n-1) = g.l. regresión (1) + g.l. residual (n-2).

El cuadrado medio es una varianza: Suma de los cuadrados de las desviaciones dividido por grados de libertad.

Tabla de ANOVA de la Regresión

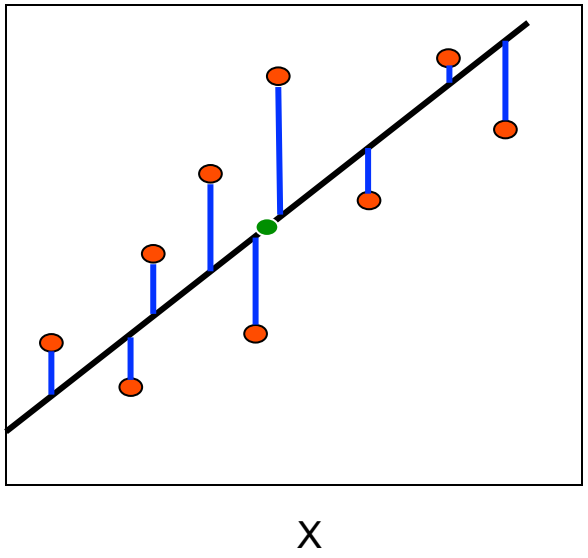
Fuente de Variación	Suma de Cuadrados (de las desviaciones)	gl	Varianza (Cuadrados medios)
Regresión	$\sum(\hat{y}_i - \bar{y})^2$	1 (# de par. -1)	$SC_{Regresión}$ / 1
Residual	$\sum(y_i - \hat{y}_i)^2$	n-2	$SC_{Residual}$ / n-2
Total	$\sum(y_i - \bar{y})^2$	n-1	

Regresión Lineal



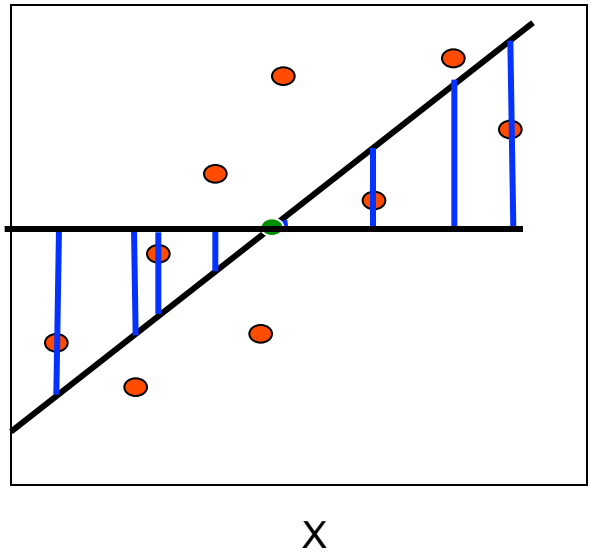
$$SC_{Total} = \sum \left(y - \bar{y}\right)^2$$

Suma de cuadrados total



$$SC_{\text{Re}sidual} = \sum \left(y - \hat{y}\right)^2$$

Suma de cuadrados de residuales o error

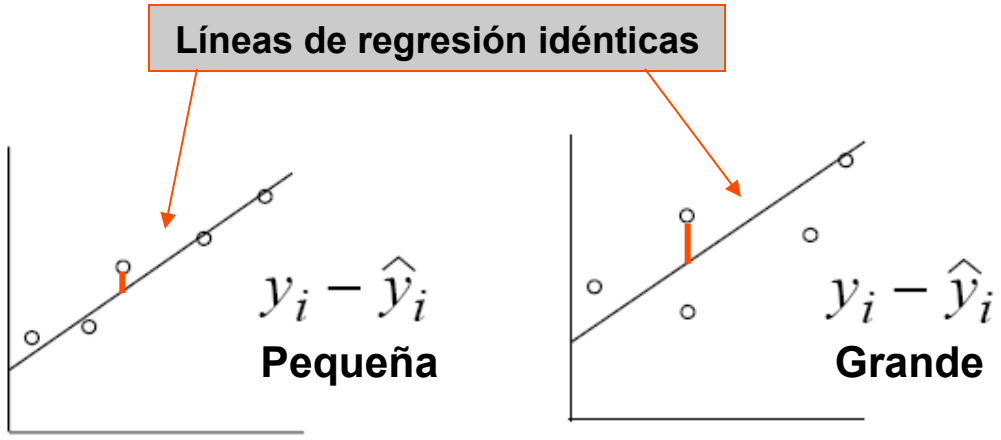


$$SC_{\text{Re}gresión} = \sum \left(\hat{y}_i - \bar{y}\right)^2$$

Suma de cuadrados de la regresión

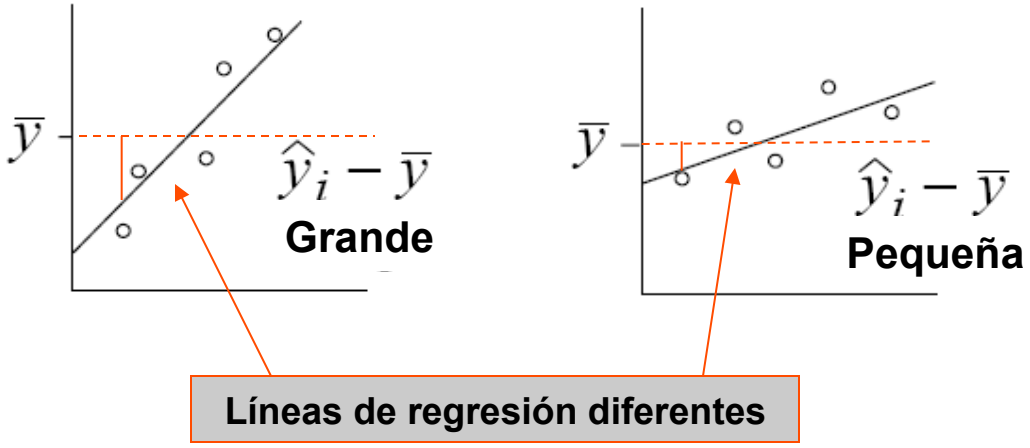
Variación explicada y no explicada en Regresión Lineal

Residual
(No explicada)



Diferencia entre observaciones observadas y predichas por el modelo de regresión

Regresión
(Explicada)



Variación explicada y no explicada en Regresión Lineal

$SC_{Regresión} + SC_{Residual} = SC_{Total}$

$gl_{Regresión} + gl_{Residual} = gl_{Total}$

Las sumas de cuadrados y grados de libertad son aditivos

$CM_{Regresión} + CM_{Residual} \neq CM_{Total}$

Los cuadrados medios (Varianza) no son aditivos).

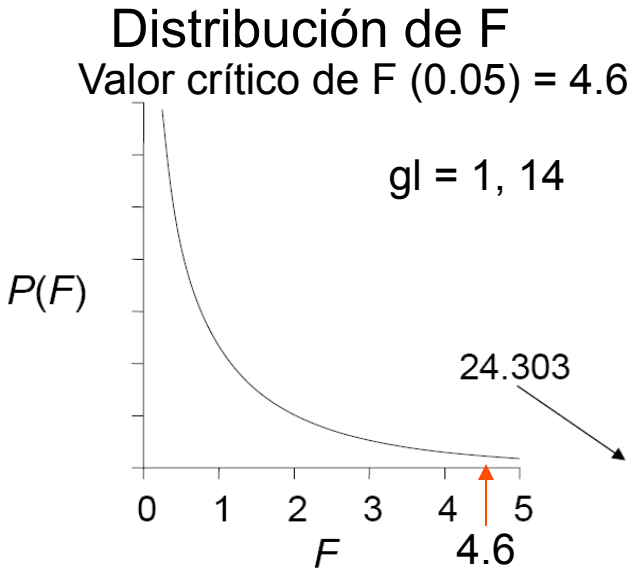
Pruebas de Hipótesis Nulas en Regresión Lineal

- La hipótesis nula mas comúnmente examinada en regresión es: $\beta_1 = 0$ y por lo tanto, no hay una relación lineal entre Y y X.
- β_1 es la pendiente del modelo de regresión.
- Hay 2 formas para efectuar la prueba de hipótesis en regresión:
  - A través del estadístico F de la tabla de ANOVA de la Regresión
    - $F = CM_{Regresión} / CM_{Residual}$
  - Con una prueba de t
    - $t = b_1 / EE(b_1)$
  - $F = t^2$ Por ejemplo, si F = 24.295, t = 4.929, y t2 = 24.295

Pruebas de Hipótesis Nulas en Regresión Lineal

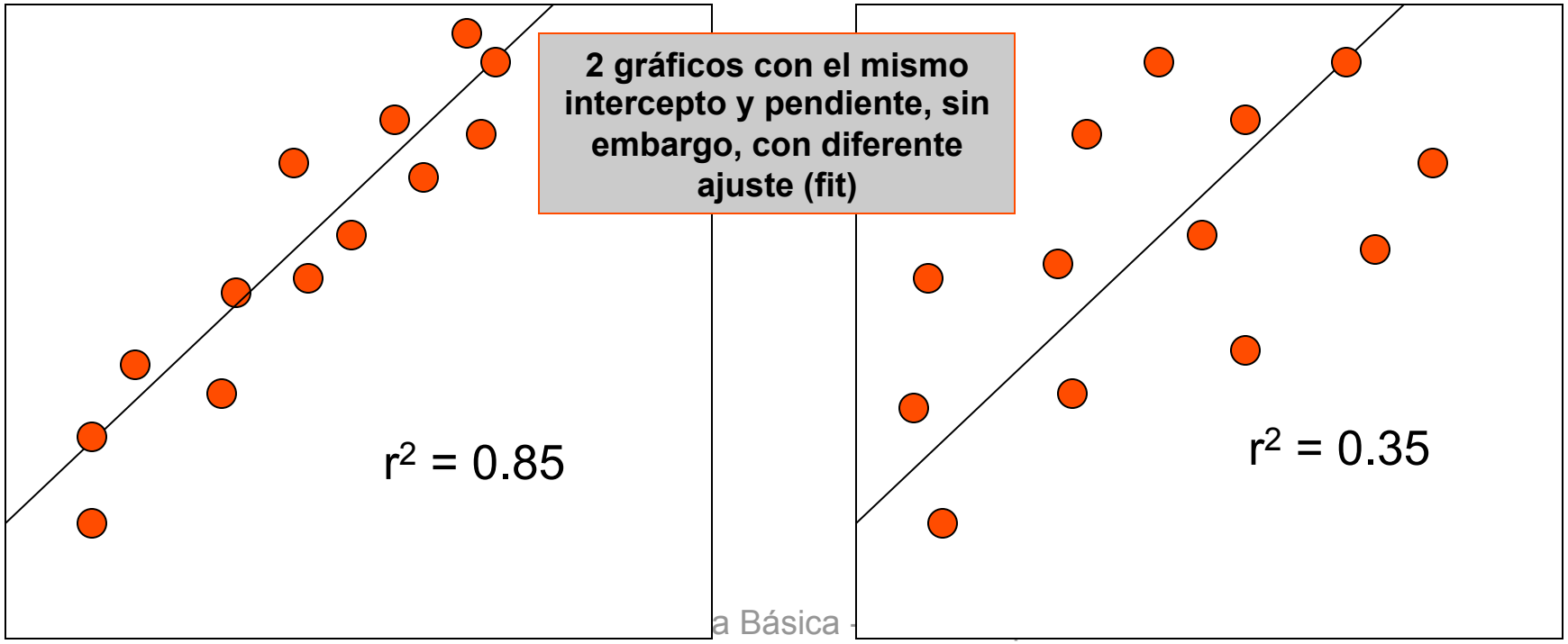
- El estadístico F es la razón entre 2 varianzas (esto es, 2 cuadrados medios)
- La distribución probabilística de F es conocida (como también lo vimos para la distribución de t).
  - La distribución es diferente para cada combinación de los grados de libertad de las 2 varianzas.

Fuente de Variación	gl	Varianza (Cuadrados medios)	F	P
Regresión	1	3.205×10^4	24.303	<0.001
Residual	14	1318.969		
Total	15			



Variación Explicada en Regresión Lineal

- Proporción de la variación en Y explicada por una relación lineal con X.
- Se le denomina coeficiente de determinación $r^2 = SC_{Regresión} / SC_{Total}$
- r^2 es el coeficiente de correlación (r) al cuadrado.
- El análisis del número de especies en conglomerados de mejillones arrojó un $r^2 = 0.634$. Lo que significa que aproximadamente el 63% de la variación en el número de especies en conglomerados de mejillones está explicada por el área de los mismos.



Ajuste de un modelo de regresión simple

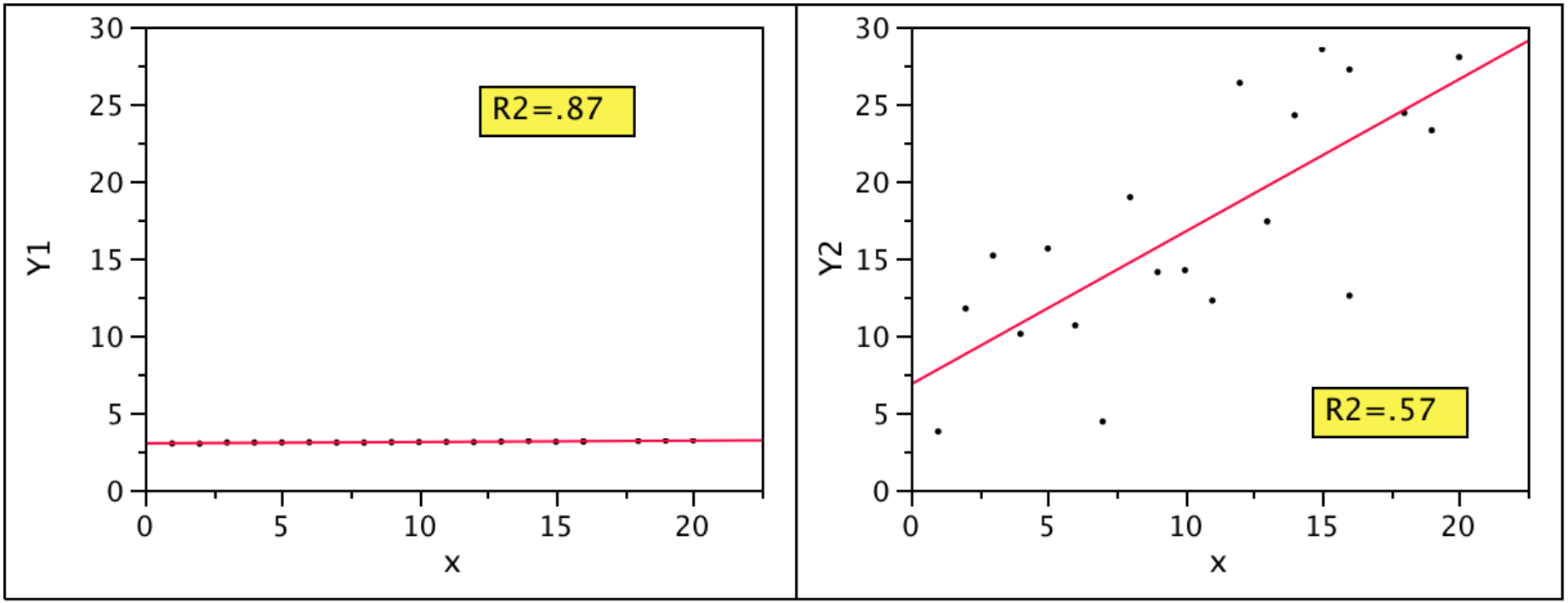
1) $$r^2 = \frac{SC_{\text{Regresión}}}{SC_{Total}} = 1 - \frac{SC_{\text{Residual}}}{SC_{Total}}$$

2) $$r^2 = 1 - \frac{SC_{\text{Residual(Completo)}}}{SC_{\text{Residual(Reducido)}}}$$

Ecuación 1) y 2) son iguales para modelos con un intercepto

r^2 no es una medida absoluta de cuán bien un modelo ajusta a los datos, es una medida de cuánto mejor ajusta un modelo con β1 (β0 y β1) relativo a un modelo sin β1 (sólo con β0).

R^2



Supuestos de una Regresión Lineal

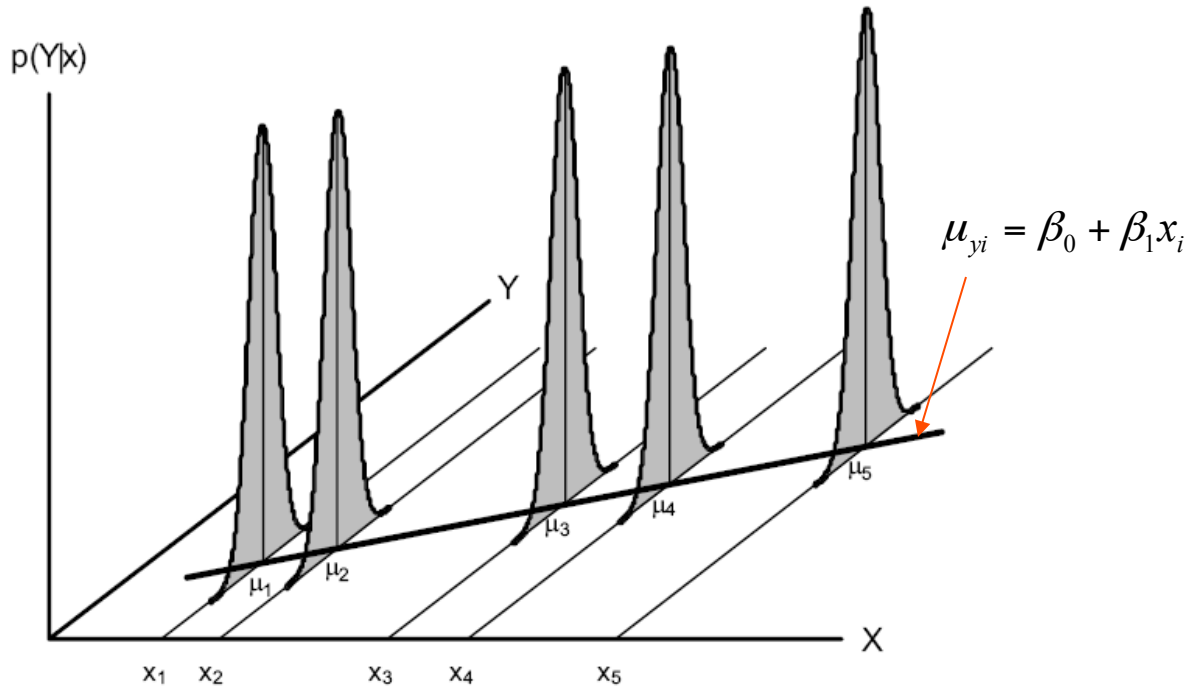
- Linealidad
- Normalidad
- Homogeneidad de varianza
- Independencia
- X es fija

Aplican a la variable respuesta

Aplican al término de error del modelo

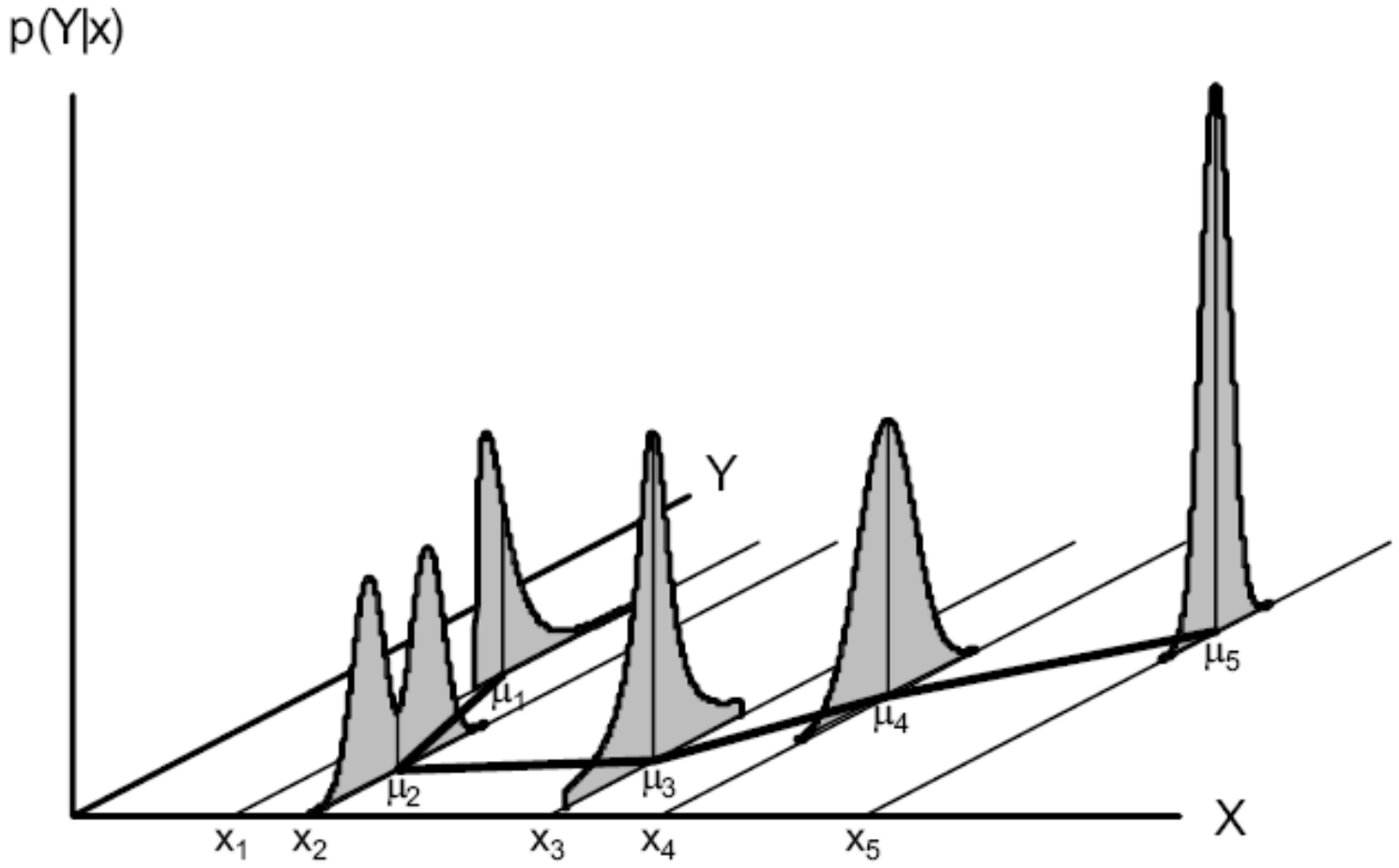
Supuestos de una Regresión Lineal

Los supuestos de linealidad, normalidad, homogeneidad de varianza.



Supuestos de una Regresión Lineal

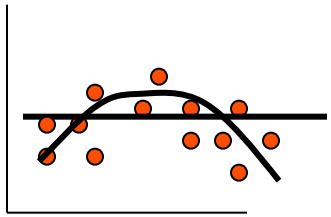
Cuando los supuestos de linealidad, normalidad, homogeneidad de varianza no se cumplen.



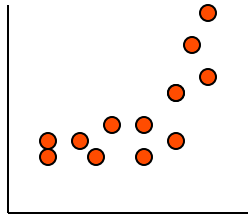
Supuestos de una Regresión Lineal

Linealidad

- Relación entre Y y X es lineal.
- Gráfico de dispersión de Y contra X debería mostrar tal relación.
- Examinar por relaciones asintóticas o exponenciales.
- Transformaciones de Y o Y y X generalmente ayudan a mejorar la relación lineal.



Relación asintótica



Relación exponencial

Supuestos de una Regresión Lineal

Normalidad

- Y y los términos de error están normalmente distribuidos para cada valor de X.
- Gráficos de caja para la variable Y deberían ser simétricos, examinar por outliers y asimetría.
- Gráficos de normalidad (qq plots).
- Pruebas de normalidad.
- Transformar variable ayuda.
- Análisis de regresión generalmente bastante robustos a una moderada falta de normalidad.

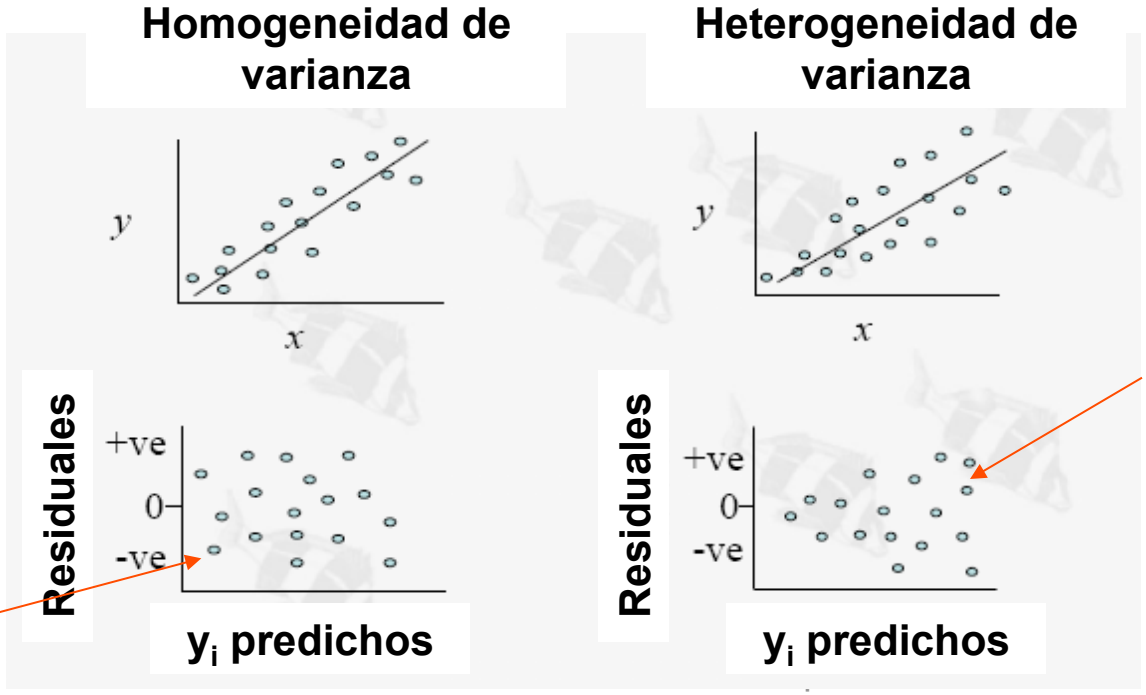
Supuestos de una Regresión Lineal

Homogeneidad de Varianza

- Varianza (dispersión) de Y y términos de error (ε_i) debería ser constante para cada valor de x_i
- Distribuciones asimétricas o outliers producen heterogeneidad de varianza.
- La regresión lineal no es robusta a la violación de este supuesto.
- Transformaciones de Y que mejoran normalidad también hacen que la varianzas sean más homogéneas.

Gráficos de Residuales

Se elaboran con los residuales contra x o valores predichos de y



No hay patrón

Patrón de embudo o cono

Supuestos de una Regresión Lineal

Independencia

- Valores de Y deben ser independientes el uno del otro.
- Ninguna replica es usada más de una vez.
- La falta de independencia puede conducir a la subestimación de los errores estándares.
- Se debe tener cuidado con datos que provienen de series de tiempo de la misma unidad de muestreo o experimental.
- Se debe considerar la independencia en el diseño del estudio.
- Análisis de muestreos repetidos y modelos jerárquicos son apropiados cuando no hay independencia de los datos.

Supuestos de una Regresión Lineal

X es fija o X es medida sin error

- La regresión lineal asume que xi es una constante de valores conocidos (son valores fijos sin varianza asociada).
- Un modelo lineal con variables explicativas fijas es un Modelo I o un modelo de efectos fijos (Modelo II: modelo de efectos aleatorios)
- Este es el caso de experimentos donde los niveles de X son escogidos específicamente (se tienen varios valores de Y para cada valor de xi.
- Este supuesto no se cumple para la mayoría de los análisis de regresión en ecología porque ambas variables son variables aleatorias.
- Sin embargo, la estimación de parámetros y pruebas hipótesis mediante un Modelo I son satisfactorios aún cuando X no es fija.

Diagnósticos de la Regresión

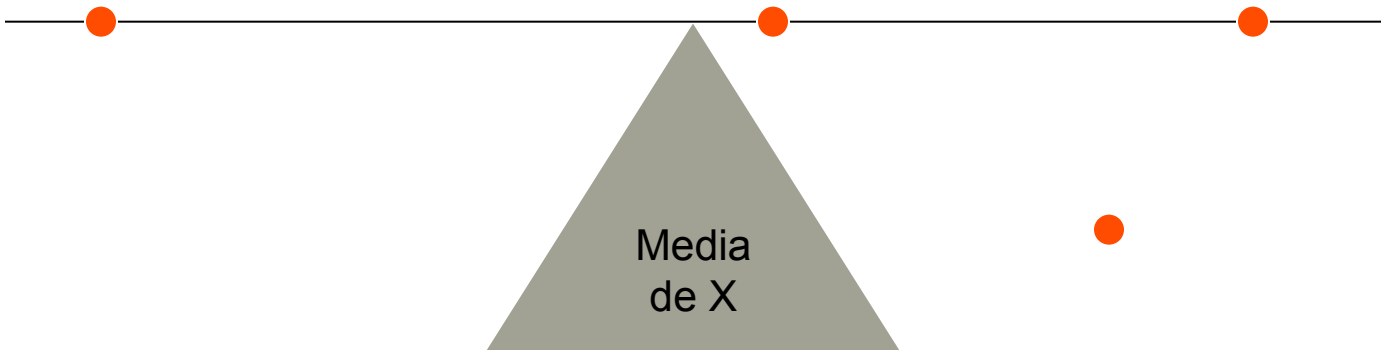
- Una vez que el modelo fue ajustado debemos examinar si el modelo tiene un buen ajuste a los datos.
- Esto también nos ayuda a examinar si los supuestos requeridos se cumplen.

1. ¿Es una línea recta apropiada para definir la relación de Y con X?
2. ¿Existe alguna observación inusual que podría considerarse un outlier y tener una influencia no deseada sobre la estimación de parámetros y el modelo de regresión ajustado?

Diagnósticos de la Regresión

La influencia en una regresión lineal puede tener 2 orígenes:

1. Observación lejos de la media: puede desbalancear más fácilmente una línea de regresión.
2. Observación lejos de la línea ajustada: tiene un residual mayor



- Leverage
- Residuales
- Influencia

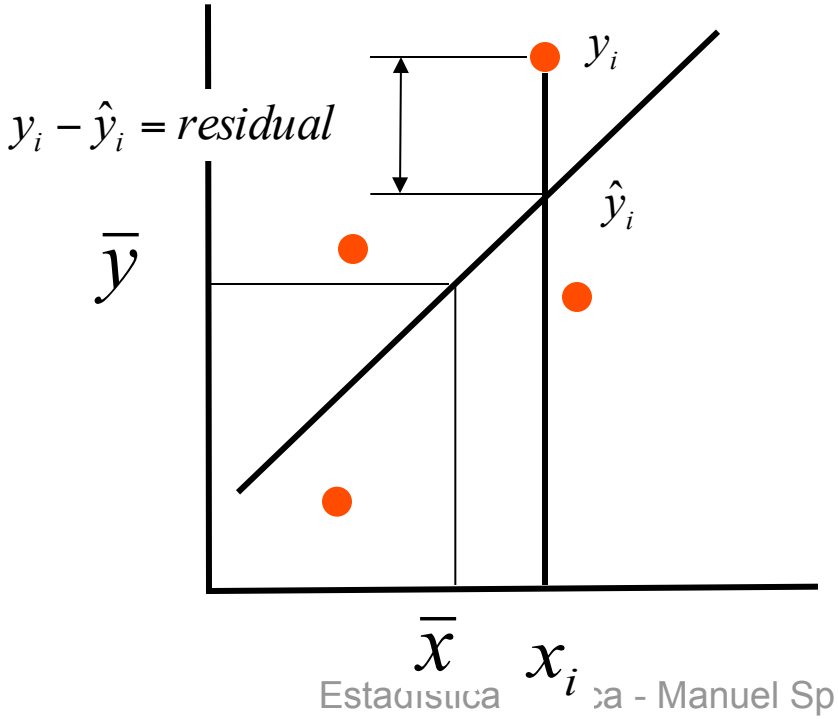
Leverage (hi)

- Leverage es una medida de cuán extremo es una observación para la variable X.
- Es un outlier en el espacio de de X.
- Leverage básicamente mide cuanto influencia cada x_i sobre el valor predicho de Y.
- Generalmente el leverage va de 1/n a 1 y se recomienda que una observación con un valor de leverage mayor a 2(p/n) debería ser examinado (p es el número de parámetros, en regresión lineal simple = 2).

Residuales

- Un residual es un estimado del error (que no puede ser observado).
- El patrón de los residuales es un elemento muy importante como diagnóstico de un modelo de regresión.

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



Influencia

- Distancia de Cook (D_i): mide la influencia de cada observación sobre la línea de regresión ajustada y los estimados de los parámetros de la regresión.
- D_i considera el valor de leverage y el residual para cada observación y mide la influencia de cada observación sobre los estimados de la pendiente de regresión.
- Un D_i grande significa que eliminar dicha observación cambiará los estimados de los parámetros considerablemente.
- 2 formas de usar D_i: 1) examinar los D_i para cada observación y ver cuál de ellos son más grandes que el resto; 2) comparar el D_i con una distribución de $F_{1,n}$. Generalmente una observación con D_i >1 es influyente.
- Una medida de influencia alternativa es $DFITS_i$, que también considera el valor del leverage y el residual.

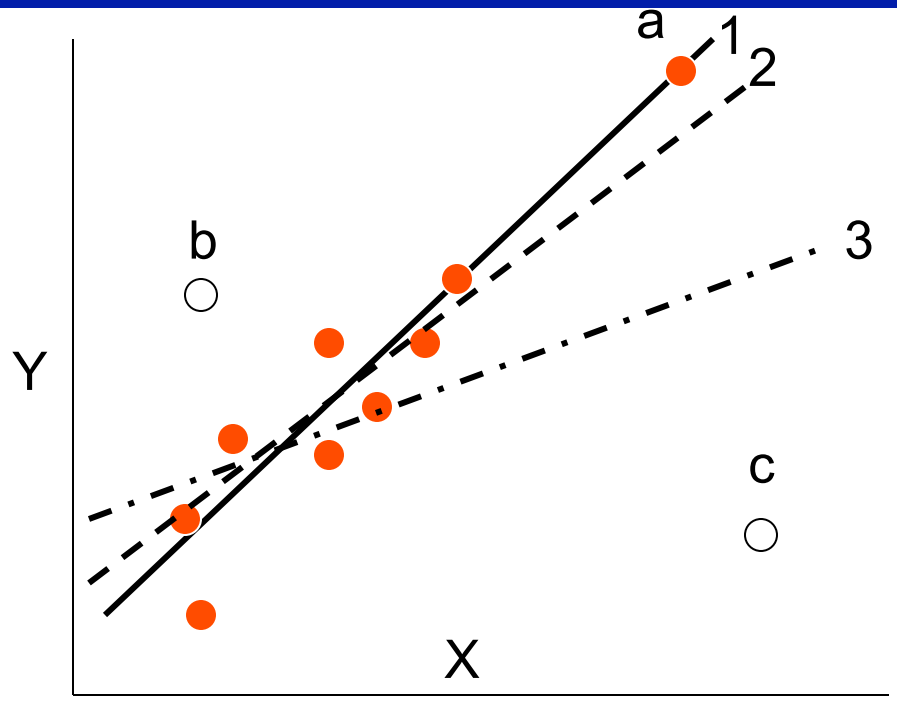
Residuales, Leverage e Influencia

La línea 1 es ajustada a las observaciones (círculos rojos).

La observación a es un outlier para Y y X (leverage alto) pero no para el modelo ajustado, no es influyente.

La observación b no es un outlier para Y y X pero es un outlier para el modelo ajustado (residual alto).

La línea 2 incluye b y cambia apenas la pendiente que tenía la línea 1 por lo que la observación b no es particularmente influyente (Di Cook bajo)).



Observación c no es un outlier para Y pero tiene un leverage alto y es un outlier para el modelo ajustado (residual alto).

La línea 3 incluye la observación c y la pendiente es marcadamente diferente de la línea 1 original (línea 1), así la observación c es muy influyente (Di Cook alto, combinando leverage y residual)

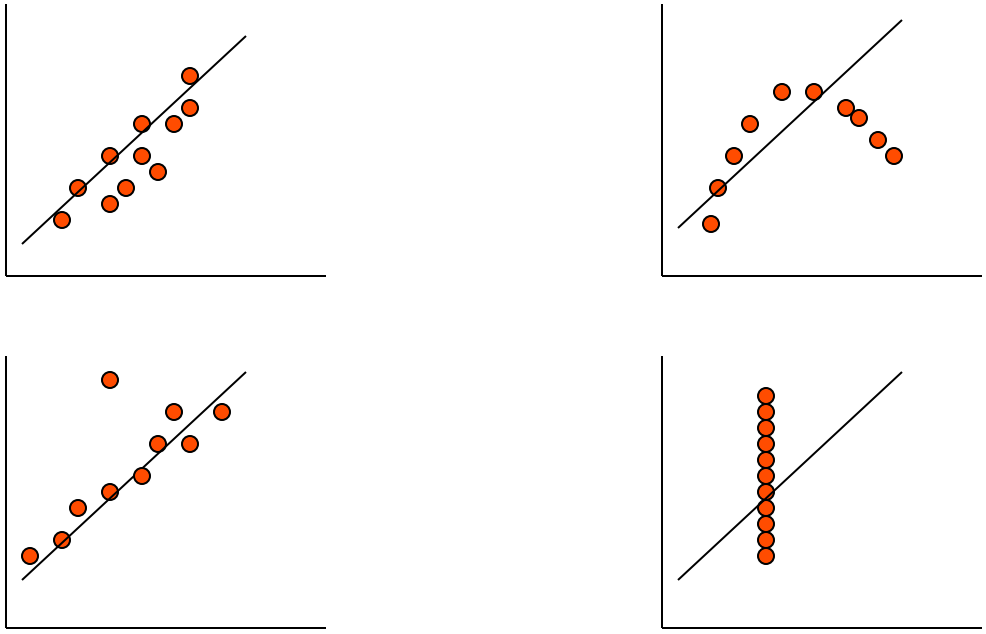
En resumen, observaciones a y c tienen leverage alto, y observaciones b y c tienen residuales altos, pero solo la observación c es muy influyente (leverage y residual alto)

Gráficos Diagnósticos

Los 2 gráficos más útiles para hacer diagnósticos de regresiones lineales son el gráfico de dispersión y el gráfico de residuales.

Gráfico de dispersión

Los gráficos de dispersión pueden indicar heterogeneidad de varianza, no-linealidad y outliers.

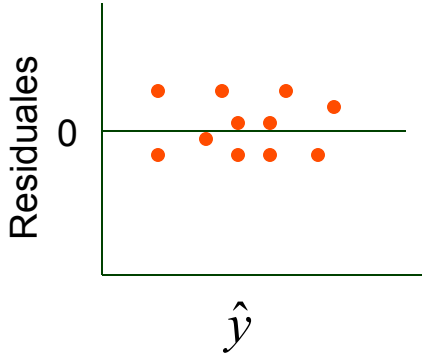


$y_i = 3.0 + 0.5x_i$, $n = 11$, $r^2 = 0.68$, $H_0 : \beta 1 = 0$, $t = 4.24$, $P = 0.002$.

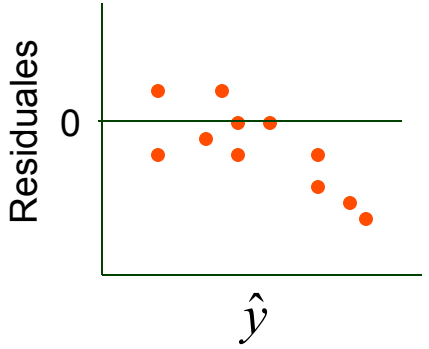
Gráficos Diagnósticos

Gráfico de residuales

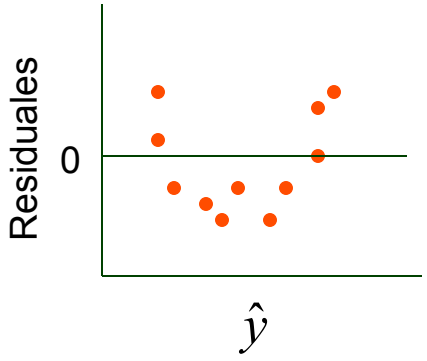
El gráfico de residuales contra los valores ajustados o predichos o x_i.



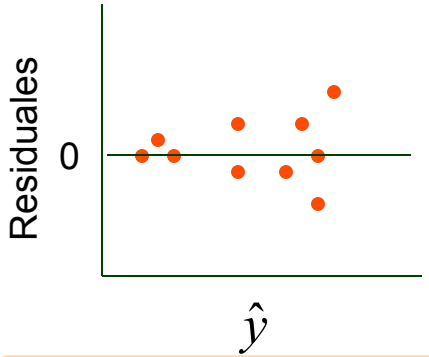
Distribución normal de los errores (no hay patrón obvio)



Residuales para un ajuste no-lineal. Se sobrestima y cuando x aumenta.



Residuales para un ajuste cuadrático o polinomial



Residuales con varianza incrementando cuando aumenta X (heterogeneidad de varianza)

Cuando los supuestos de la regresión lineal no se cumplen

- Transformaciones
- Pruebas no-paramétricas
  - Regresión Robusta
  - Pruebas de aleatorización
- Funciones suavizadoras

Transformaciones

- Cuando hay ceros en distribuciones asimétricas.
  - Log (y + constante) o transformaciones de potencia
- Transformaciones de potencia.
  - Raíz 4ta es útil en datos de abundancia cuando el rango de valores es grande
- Transformación angular (arcoseno) para proporciones.
  - Efecto pequeño excepto para valores que están cerca de 0 o 100.
- Presentación de resultados.
  - Se debería retransformar medias y errores
- Considerar modelos lineales generalizados para errores con distribuciones no-normales.

Modelos Lineales

Funciones de enlace (link function)

- La función de enlace: enlaza el componente aleatorio con el sistemático
  - Enlaza variable respuesta a las variables explicativas.

Enlace de identidad (identity link)

- Modela la media (valor esperado) de Y

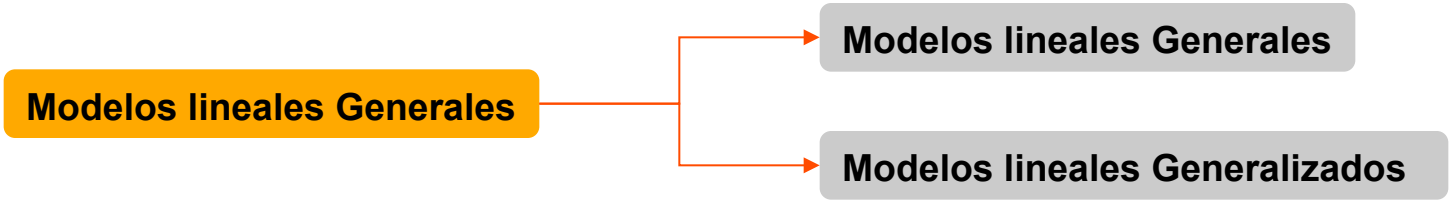
Enlace log (log link)

- Modela la media logarítmica de Y.
- Datos de conteos, no-negativos (Poisson)

Enlace logit (logit link)

- Modela proporciones y datos binarios (muerto/vivo, presencia/ausencia)

Modelos Lineales Generales



Modelos Lineales Generales	Modelos Lineales Generalizados
▪ Y es una variable continua.	▪ Y es cuantitativa (continua o discreta) y categórica.
▪ Distribución normal para Y y el término de error.	▪ Cualquier distribución exponencial para Y y el término de error.
▪ La media de Y se modela con el enlace de identidad.	▪ Utilizan enlace log o logit (u otros).
▪ La estimación de parámetros se hace mediante el método de los cuadrados mínimos (Ordinary Least Squares).	▪ La estimación de parámetros se hace mediante el método de máxima verosimilitud (MLE).
▪ Regresión Lineal y ANOVA	▪ Modelos logístico, logit, Poisson y log-linear (También se les denomina regresión).