Modelos lineales

Regresión simple y múltiple

Universitat d'Alacant
Universidad de Alicante

Dept. of Marine Science and Applied Biology
Jose Jacobo Zubcoff

Modelos de Regresión Simple

- Que tipo de relación existe entre 2 variables
- Predicción de valores a partir de una de ellas

- Variable Explicativa, Predictor o Independiente
- Variable Dependiente

Estudio conjunto de dos variables

- Datos de dos variables de una muestra.
  - En cada **fila** tenemos los datos de **un individuo**
  - Cada **columna** representa los valores que toma una variable sobre los mismos.
  - Las individuos no se muestran en **ningún orden** particular.

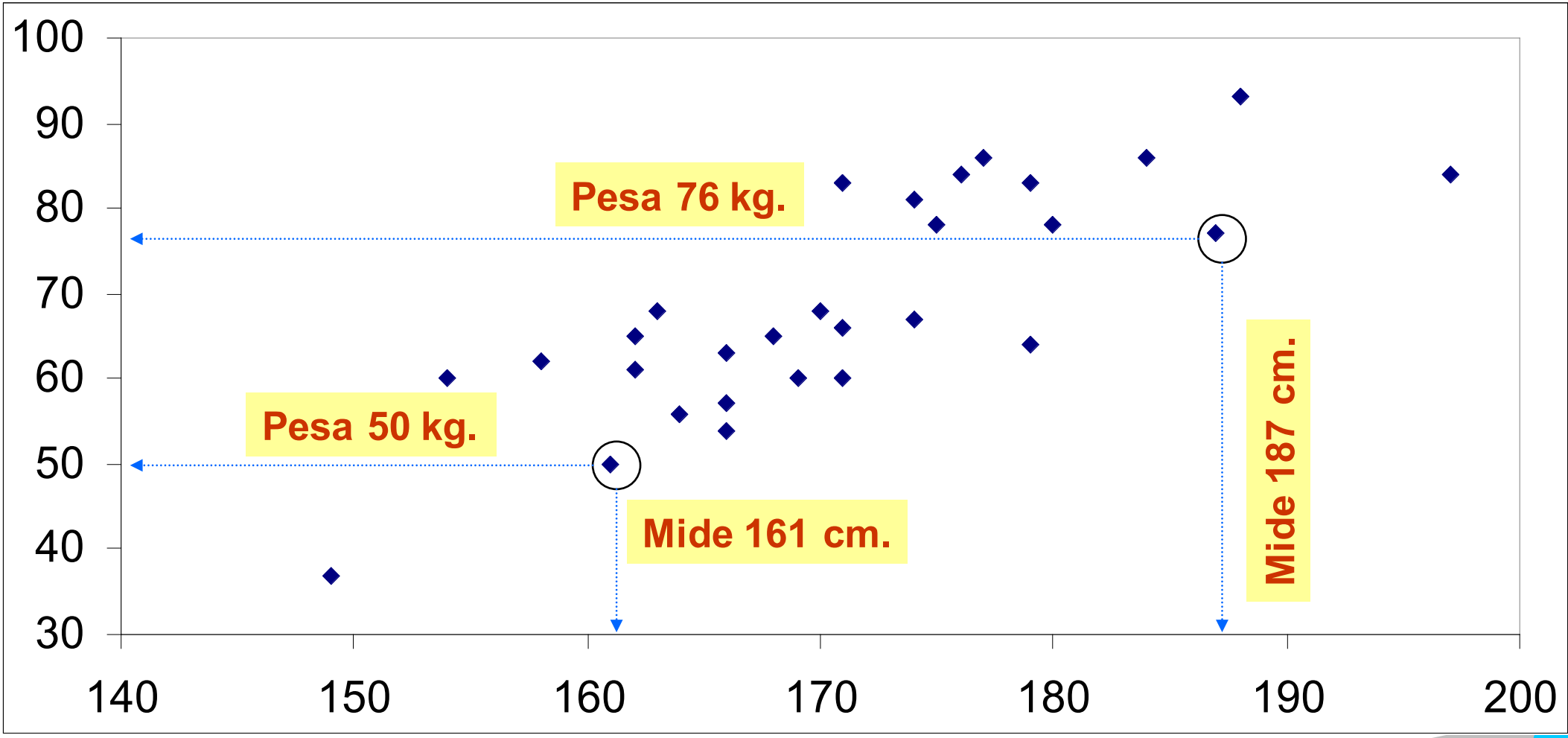
- Las observaciones pueden ser representadas en un **diagrama de dispersión**

- Nuestro objetivo será intentar **reconocer** a partir del mismo si hay **relación** entre las variables, de qué **tipo**, y si es posible **predecir** el valor de una de ellas en función de la otra.

Altura en cm.	Peso en Kg.
162	61
154	60
180	78
158	62
171	66
169	60
166	54
176	84
163	68
...	...

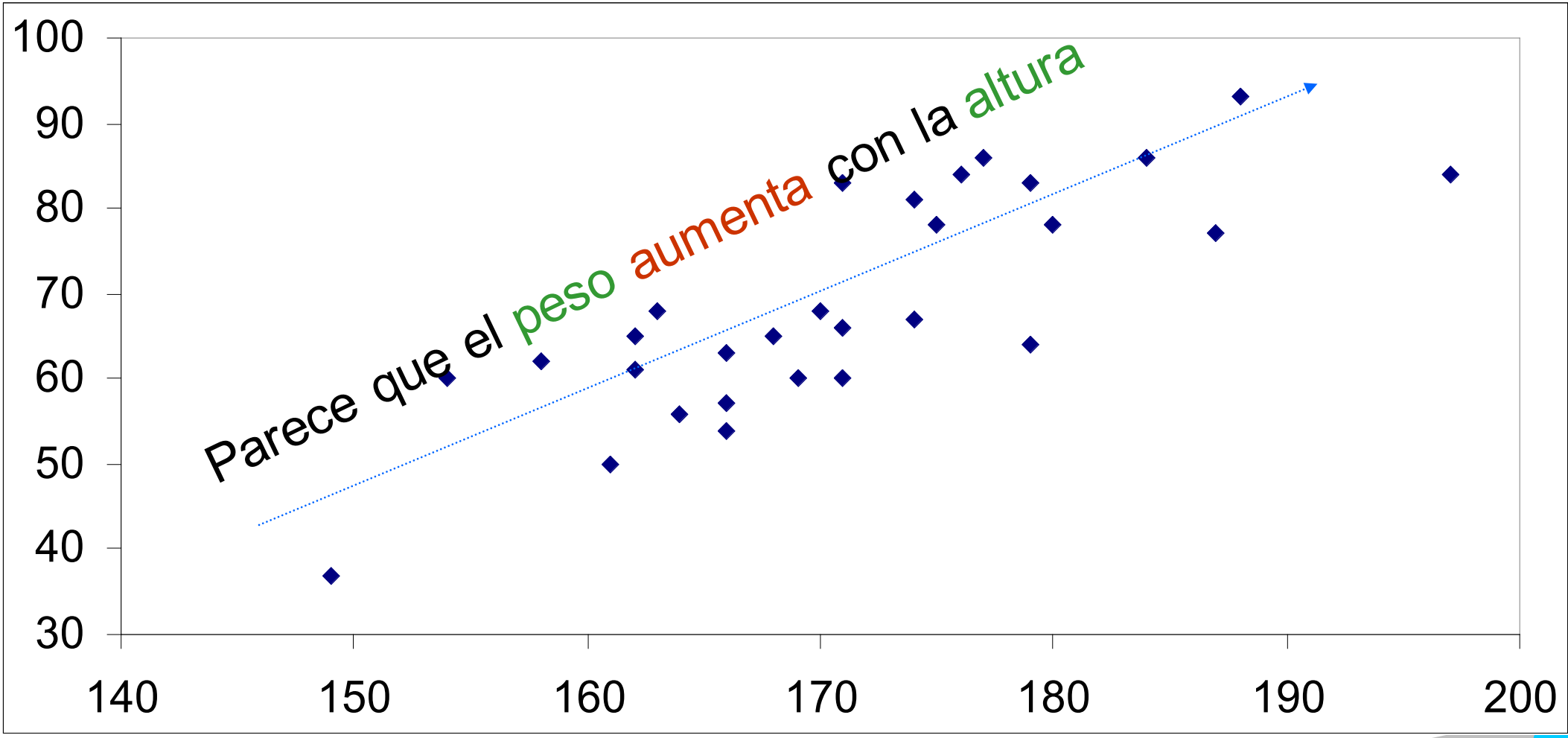
Diagramas de dispersión o nube de puntos

Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



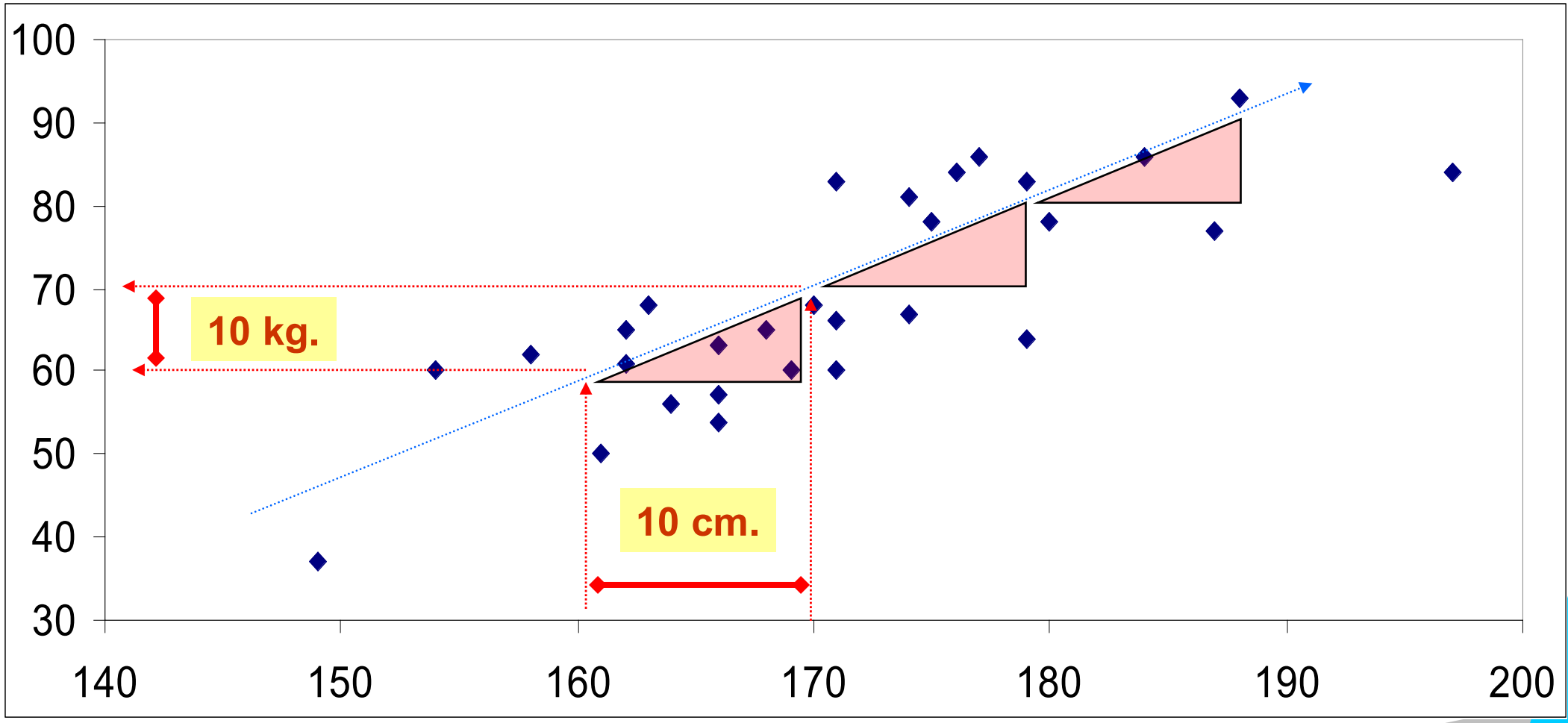
Relación entre variables

Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.

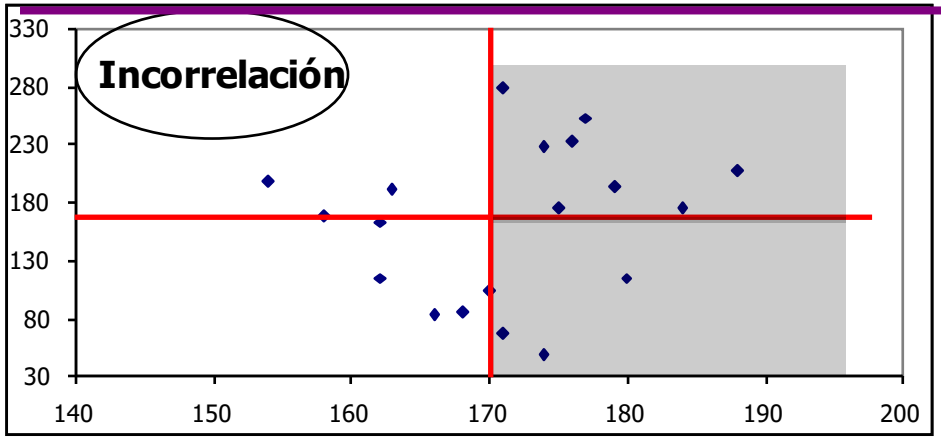


Predicción de una variable en función de la otra

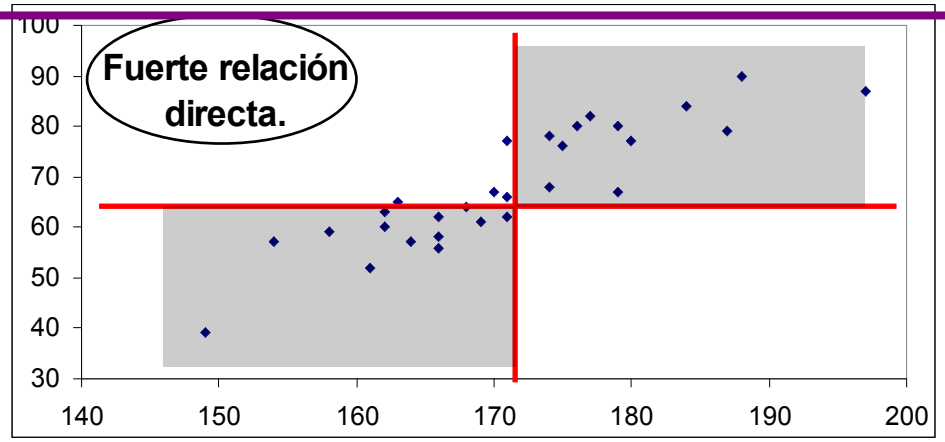
Aparentemente el peso aumenta 10Kg por cada 10 cm de altura... o sea, el peso aumenta en una unidad por cada unidad de altura.



Relación directa e inversa

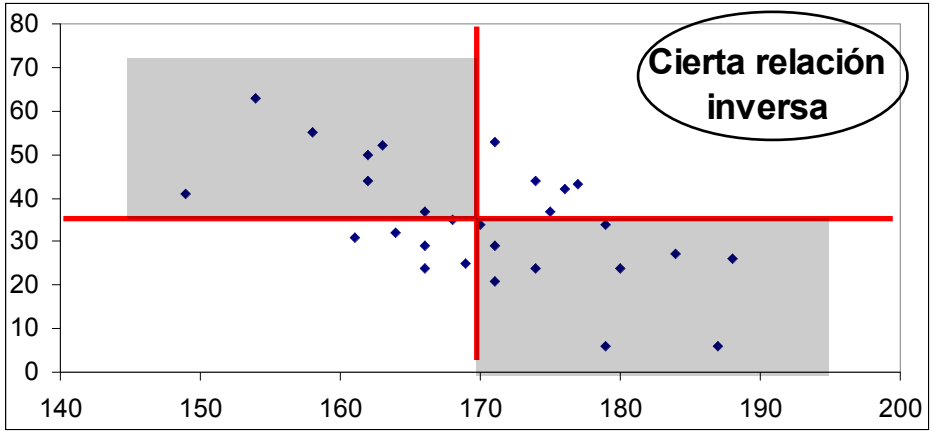


Para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares. Incorrelación.



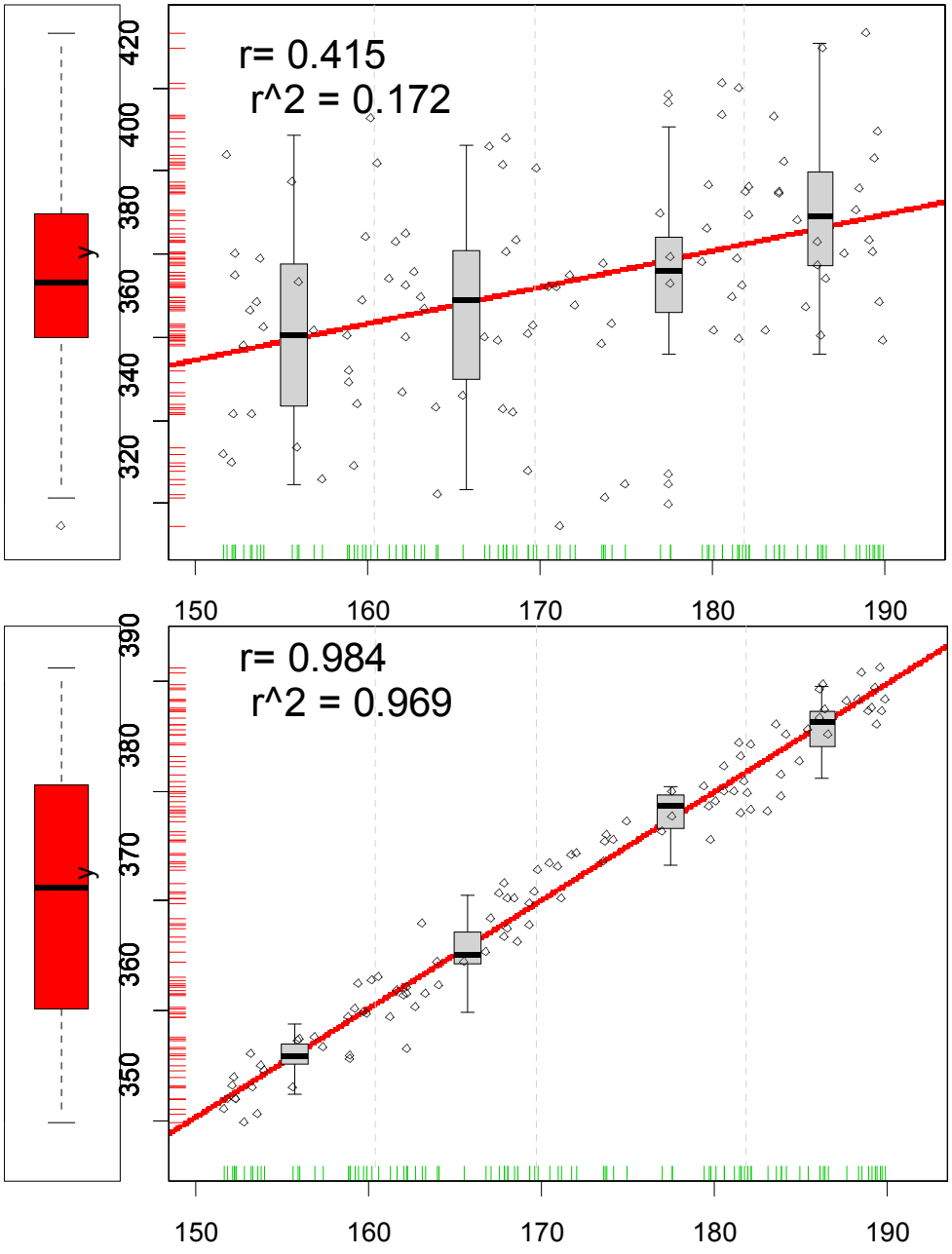
- Para los valores de X mayores que la media le corresponden valores de Y mayores también.
- Para los valores de X menores que la media le corresponden valores de Y menores también.

Esto se llama relación directa.



Para los valores de X mayores que la media le corresponden valores de Y menores. Esto es relación inversa o decreciente.

¿Cuándo es bueno un modelo de regresión?



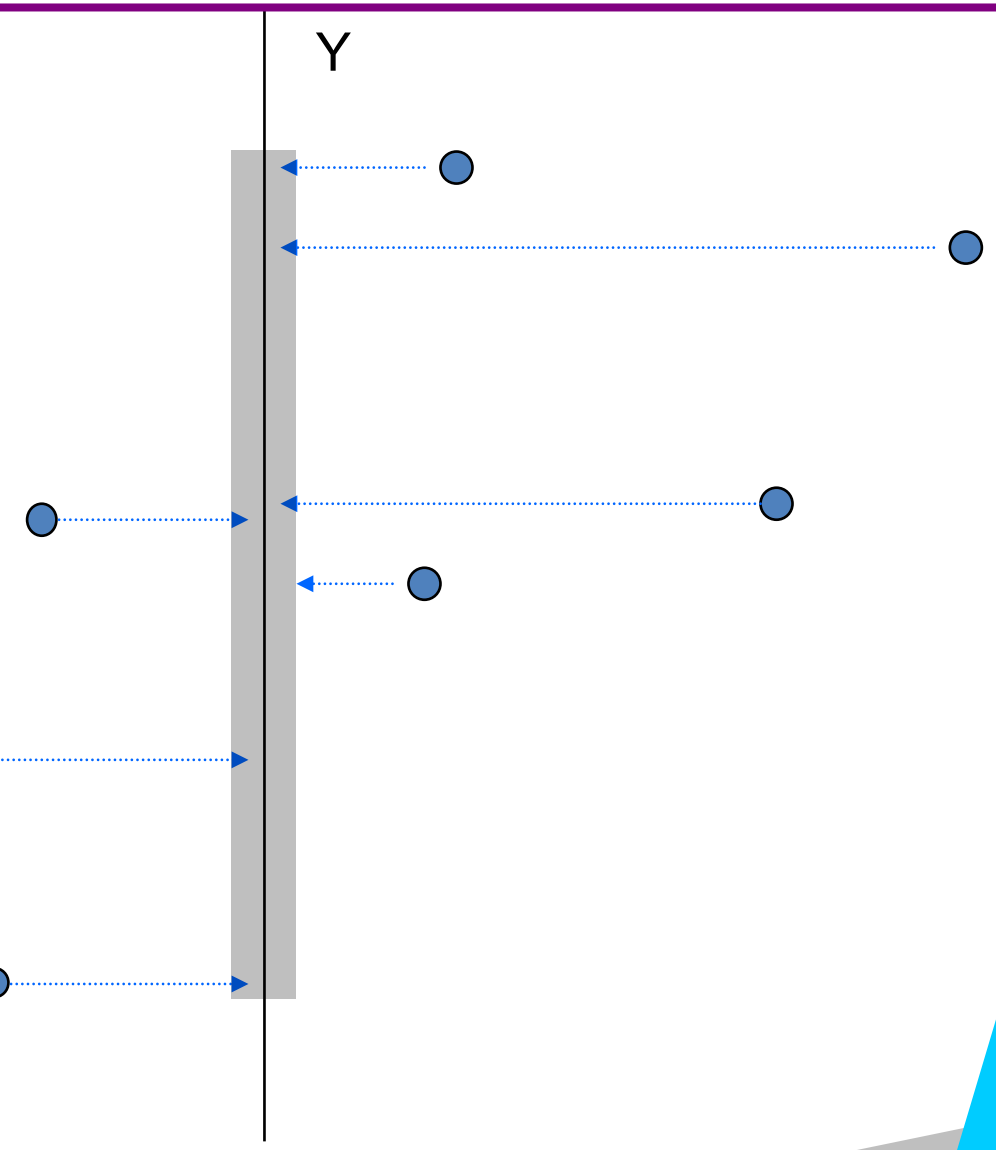
- Lo adecuado del modelo depende de la relación entre:
  - la dispersión marginal de Y
  - La dispersión de Y condicionada a X
- Es decir, fijando valores de X, vemos cómo se distribuye Y
  - La distribución de Y, para valores fijados de X, se denomina distribución condicionada.
  - La distribución de Y, independientemente del valor de X, se denomina distribución marginal.
- Si la dispersión se reduce notablemente, el modelo de regresión será adecuado.

Interpretación de la variabilidad en Y

En primer lugar olvidemos que existe la variable X. Veamos cuál es la variabilidad en el eje Y.

La franja sombreada indica la zona donde varían los valores de Y.

Proyección sobre el eje Y = olvidar X

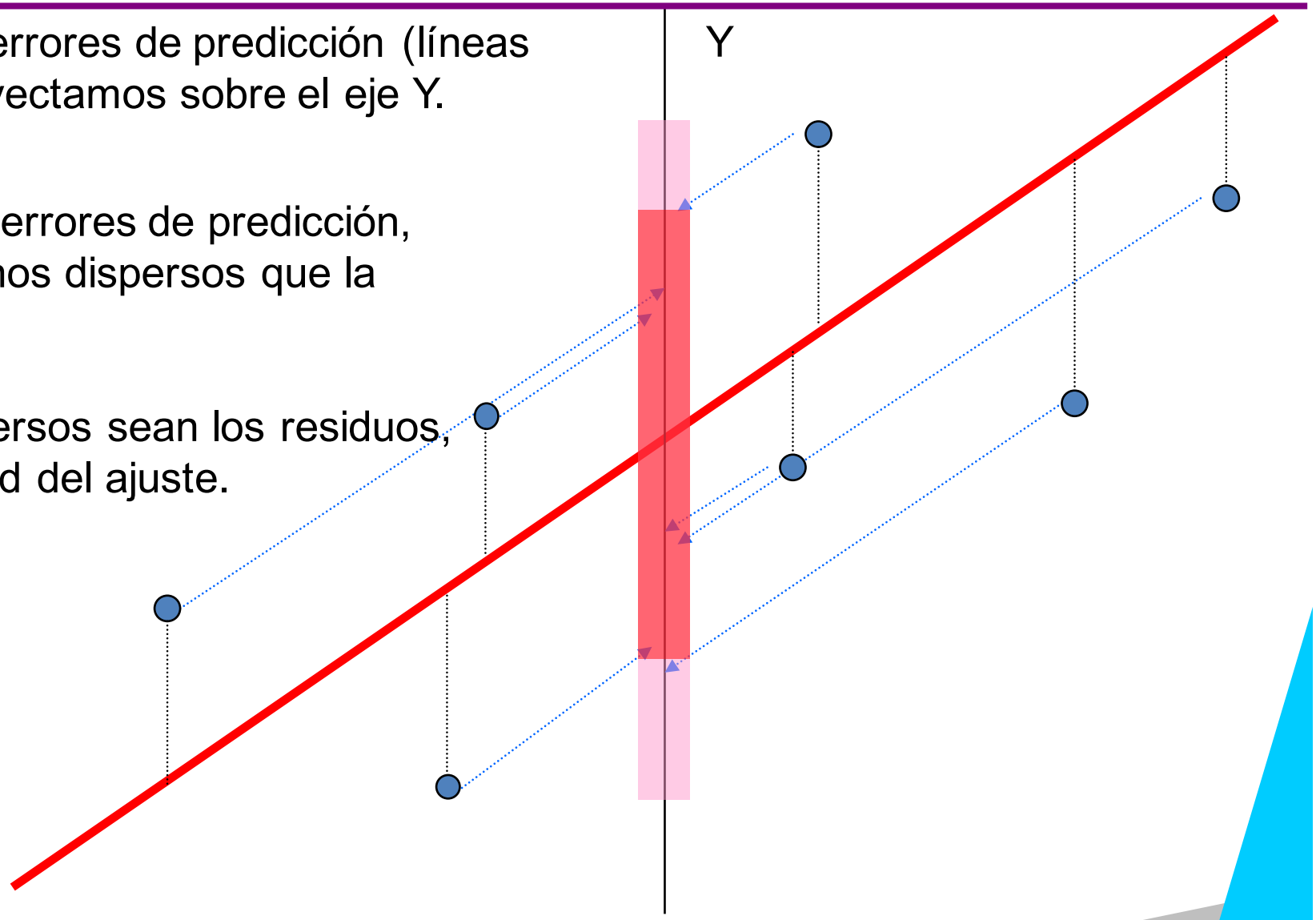


Interpretación del residuo

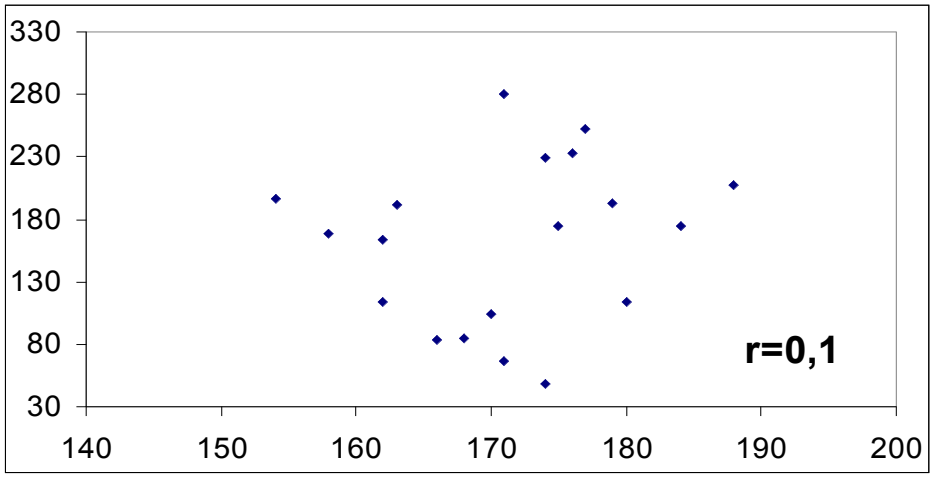
Miremos ahora los errores de predicción (líneas verticales). Los proyectamos sobre el eje Y.

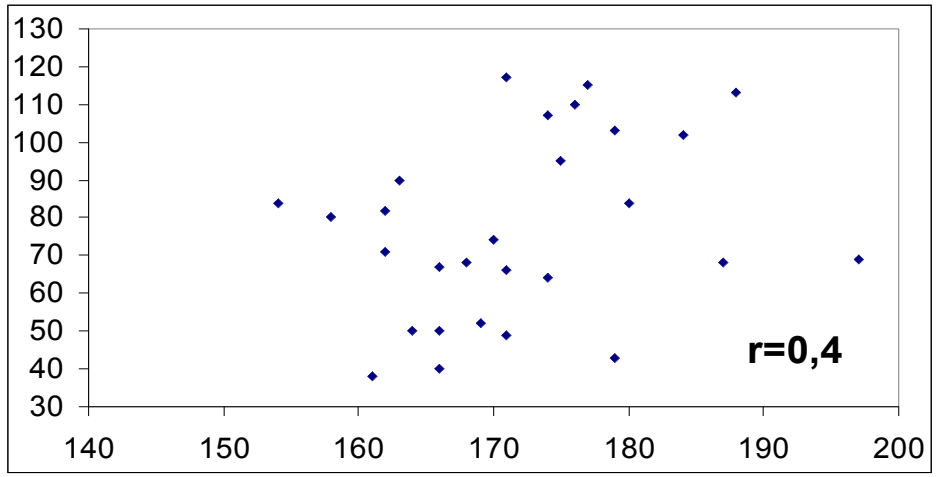
Se observa que los errores de predicción, residuos, están menos dispersos que la variable Y original.

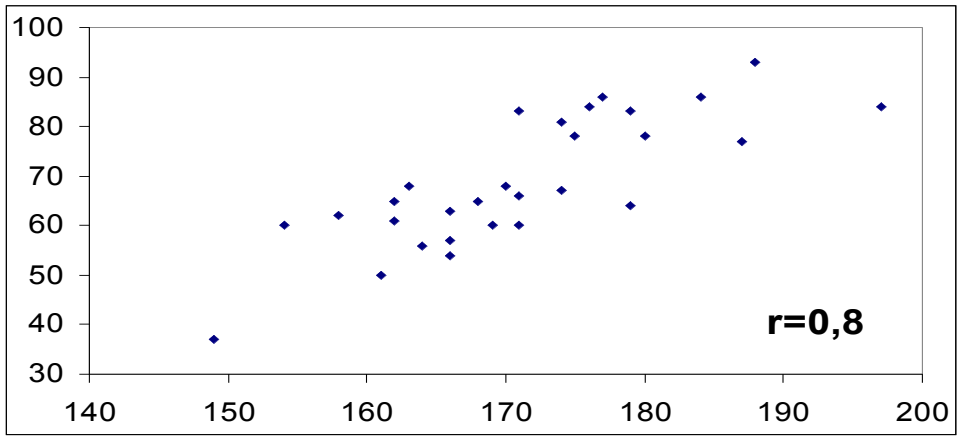
Cuanto menos dispersos sean los residuos, mejor será la bondad del ajuste.

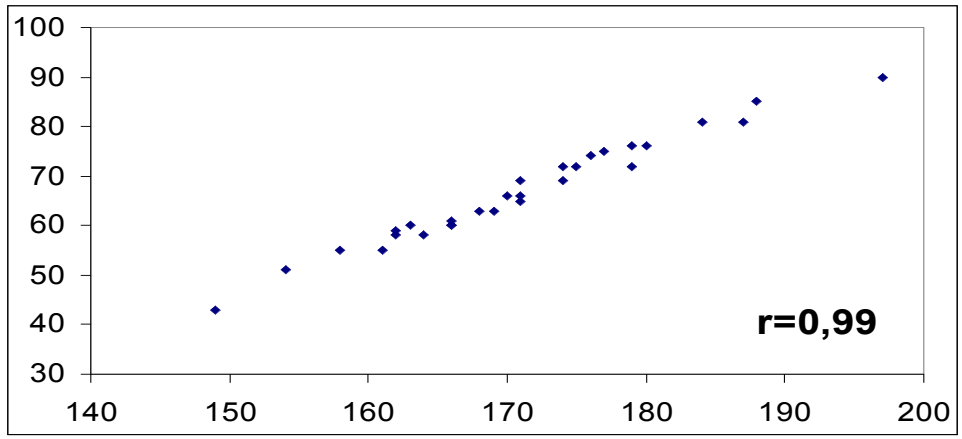


Ejemplos de correlaciones positivas

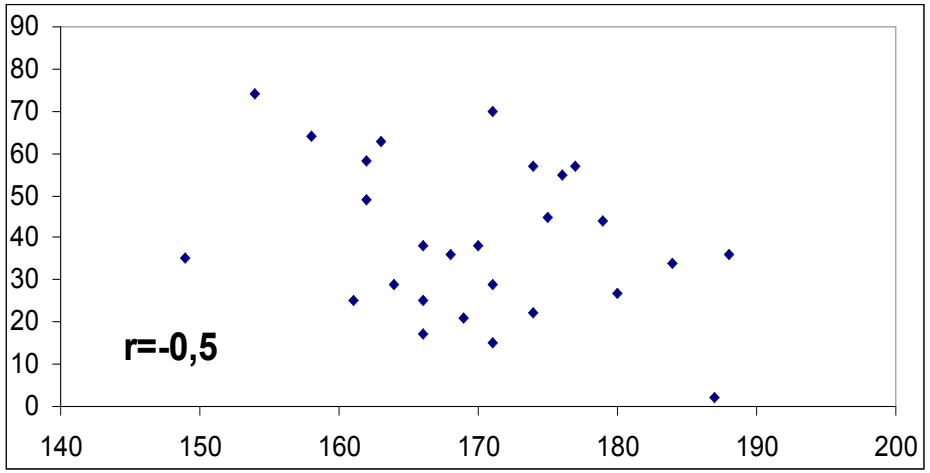


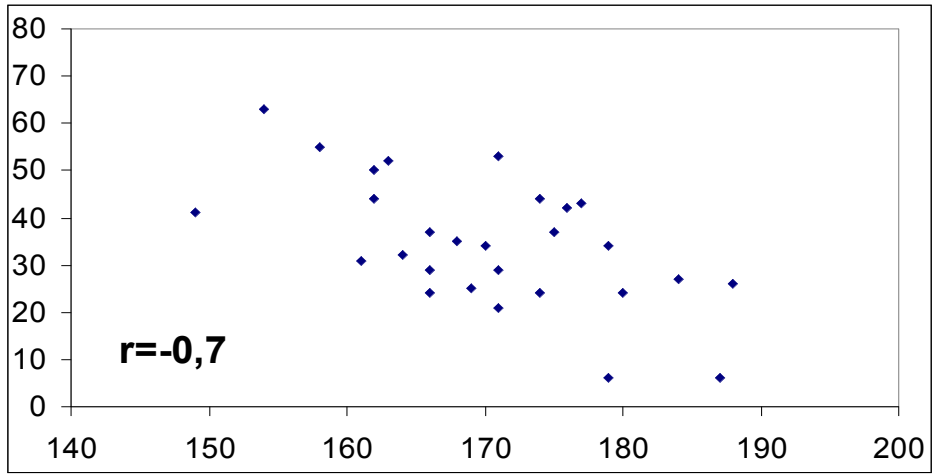


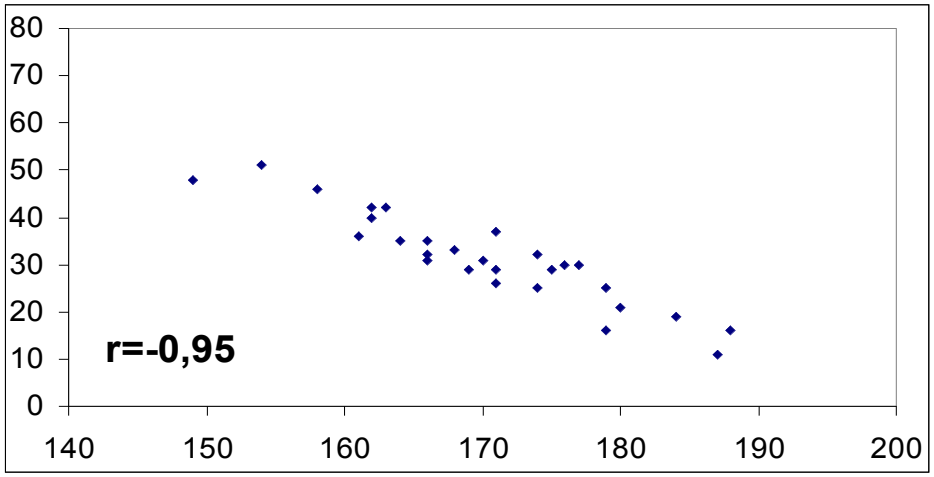


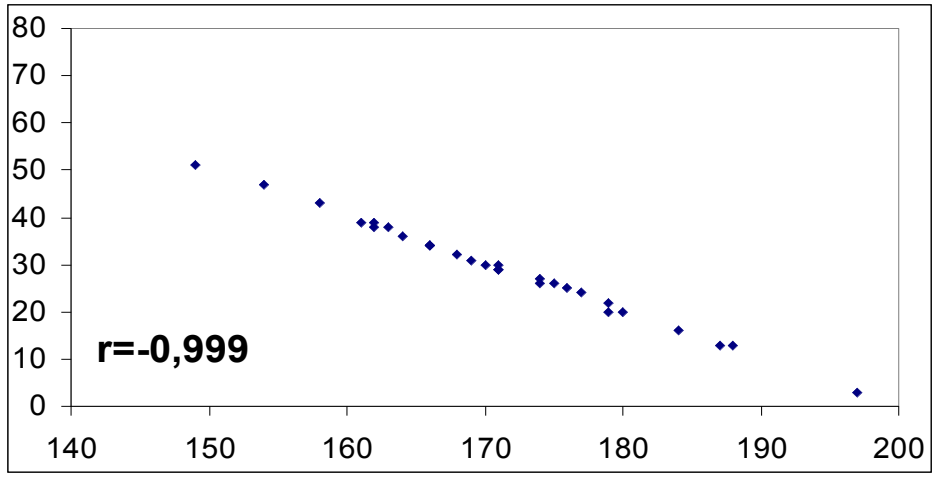


Ejemplos de correlaciones negativas

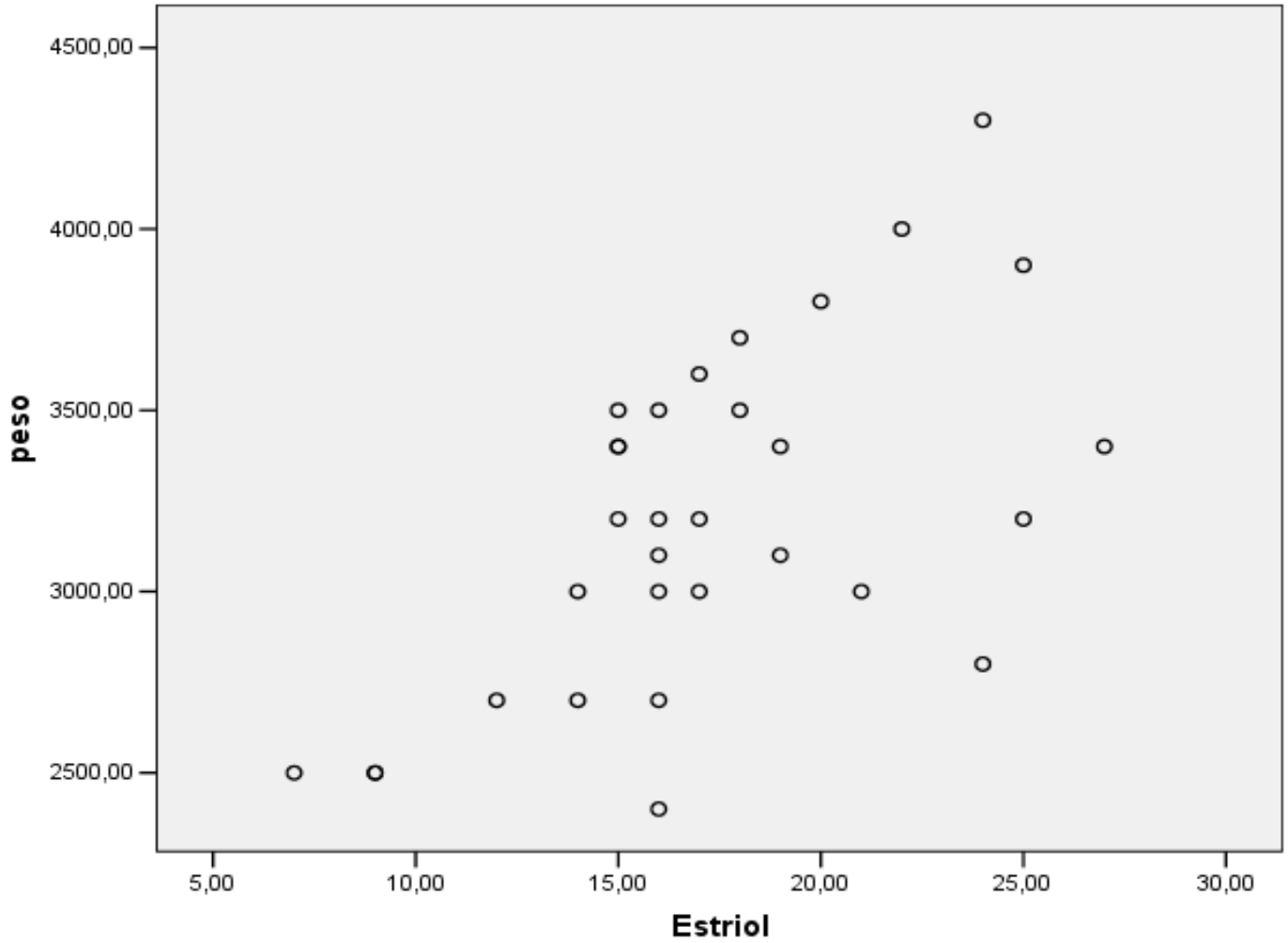








Regresion Lineal



Asociación entre variables continuas

Correlations

Descriptive Statistics

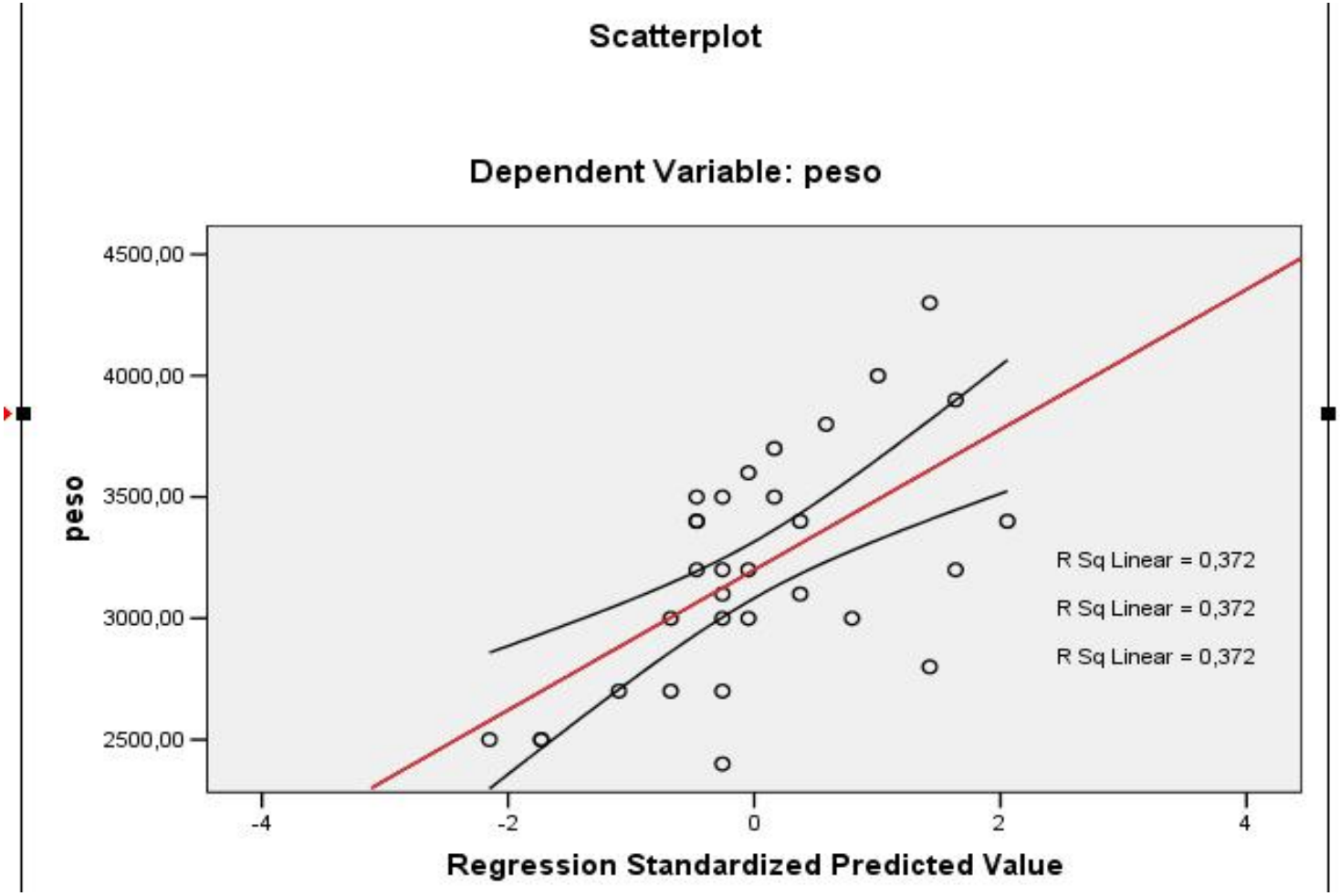
	Mean	Std. Deviation	N
Estriol	17.2258	4.75191	31
peso	3200.0000	473.99015	31

Correlations

		Estriol	peso
Estriol	Pearson Correlation	1	.610**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	677.419	41200.000
	Covariance	22.581	1373.333
	N	31	31
peso	Pearson Correlation	.610**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	41200.000	6740000
	Covariance	1373.333	224666.7
	N	31	31

**. Correlation is significant at the 0.01 level

Regresion Lineal



Regresión lineal

Medidas de Bondad de ajuste

- **Coeficiente de correlación**

El coeficiente de correlación (r) es una medida de la intensidad de la relación entre dos variables

$$-1 <= r <= 1$$

- **Coeficiente de determinación**

Es la proporción de la varianza de la variable dependiente que está explicada por una variable independiente en un modelo estadístico.

$$0 <= r^2 <= 1$$

- **Pendiente y constante**

Modelo de regresión lineal simple

- En el modelo de regresión lineal simple, dado dos variables
  - Y (dependiente)
  - X (independiente, explicativa, predictora)

- buscamos encontrar una función de X ***muy simple (lineal)*** que nos permita aproximar Y mediante
  - $\hat{Y} = b_0 + b_1 X$
    - b_0 (ordenada en el origen, constante)
    - b_1 (pendiente de la recta)

- Y e $\hat{Y}$ rara vez coincidirán por muy bueno que sea el modelo de regresión. A la cantidad
  - $e = (Y-\hat{Y})$ se le denomina residuo o error residual.

Regresión lineal

- El modelo lineal de regresión se construye utilizando la técnica de estimación mínimo cuadrática:
  - Buscar b_0, b_1 de tal manera que se minimice la cantidad
    - $\Sigma_i e_i^2$
- Se comprueba que para lograr dicho resultado basta con elegir:

$$b_1 = r\frac{S_Y}{S_X} \qquad b_0 = \bar{y} - b_1\bar{x}$$

- Se obtiene además otras ventajas
  - El **error residual medio es nulo**
  - La **varianza del error residual es mínima** para dicha estimación.
  - Traducido: En término medio no nos equivocamos. Cualquier otra estimación que no cometa error en término medio, si es de tipo lineal, será peor por presentar mayor variabilidad con respecto al error medio (que es cero).

Modelos de Regresión Simple

- Modelo Lineal o *Recta de Regresión*

$$Y = \gamma + \beta X$$

$$E(Y|X) = \gamma + \beta X$$

Método de Mínimos Cuadrados

$$\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (Y_i - (\gamma + \beta X))^2$$

Modelos de Regresión Simple

Fórmula para la estimación por Mínimos Cuadrados

$$b = \left[\frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2} \right] = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$a = \overline{Y} - b\overline{X}$$

Varianza Residual de Y para cada valor de X

$$S_{Y..X}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 = \frac{n-1}{n-2} \left(S_Y^2 - b^2 S_X^2 \right)$$

Regresion Lineal Múltiple

Supuestos de la Regresión Lineal

•No debe existir **ninguna relación exacta entre** cualesquiera de las **variables independientes**.
•Cuando el modelo tiene dos variables explicativas se habla de colinealidad, si hay más de dos variables explicativas, entonces, hablamos de multicolinealidad.
•La multicolinealidad es una **cuestión de grado** no de existencia.
•Se supone que es un **problema de la muestra** y no de la población.

Si en un modelo de RLM alguna variable independiente es combinación lineal de otras, el modelo es irresoluble, debido a que, en ese caso, la matriz X'X es singular, es decir, su determinante es cero y no se puede invertir.

Regresion Lineal

Supuestos de la Regresión Lineal

- **Multicolinealidad**: ¿cómo detectarla?

FORMAS DE DETECTAR LA MULTICOLINEALIDAD: Reglas Prácticas
Coeficiente de correlación simple.
R^2 elevado y pocas *t* significativas.
Correlaciones parciales.
Regresiones auxiliares: la multicolinealidad puede ser un problema complicado si el R^2 de una regresión auxiliar es mayor que el R^2 global.

Regresion Lineal

Supuestos de la Regresión Lineal

- **Multicolinealidad**: Soluciones

CAMBIAR DE MODELO	AMPLIAR LA INFORMACIÓN	OTROS MÉTODOS DE ESTIMACIÓN	ACEPTAR LA MULTICOLINEALIDAD Y VIGILAR SUS CONSECUENCIAS
Eliminar la variable menos significativa. Eliminación de tendencias	Ampliar la muestra Utilización de datos mixtos Información inicial de algún parámetro	Interés relativo. Componentes principales Regresión Cresta	SOLUCIÓN HABITUAL

Si el modelo se utiliza de cara a la predicción no será un problema importante si asumimos que la correlación entre las variables se mantiene en el futuro.

Regresion Lineal

Supuestos de la Regresión Lineal

Variabilidad de los valores de *X*

No todos los valores de *X* en una muestra dada deben ser iguales.

La varianza de (X) debe ser un número positivo finito. Gujarati (1997)

La variación en *Y* al igual que en *X* es esencial para utilizar el análisis de regresión como herramienta de investigación.

las variables deben variar!!!

Regresion Lineal

Supuestos de la Regresión Lineal

- **El valor medio de la perturbación u_i es igual a cero. Técnicamente, el valor de la media condicional de u_i es cero:**

$$E(u_i / X_i) = 0$$

Gujarati (1997)

Comentario:

Esto significa que los factores que no están incluidos en el modelo y que, por consiguiente, están incorporados en u_i no afectan sistemáticamente el valor de la media de Y.

Regresion Lineal

Supuestos de la Regresión Lineal

El número total de observaciones debe ser superior al número de parámetros estimados por el modelo. (evitar la micronumerosidad)

La NBR 14.653-2 establece lo siguiente:

- **Nivel de Fundamentación III: 6(K+1)**
- **Nivel de Fundamentación II: 4(K+1)**
- **Nivel de Fundamentación I: 3(K+1)**

K: número de variables independientes

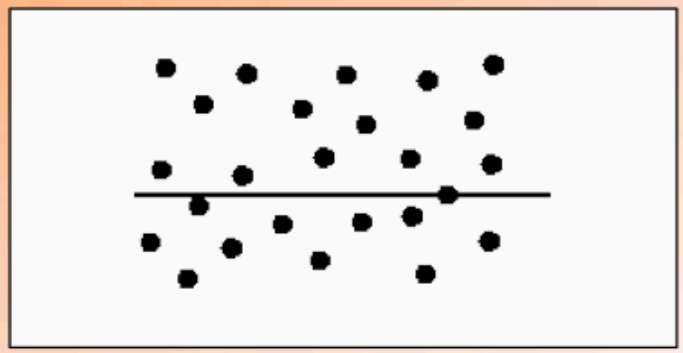
Pelli (2005)

Regresion Lineal

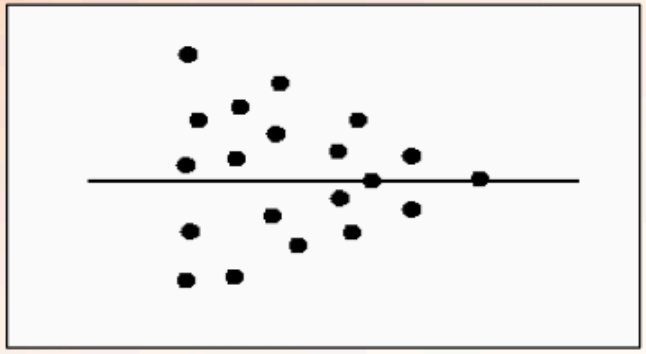
Supuestos de la Regresión Lineal

Los errores son variables aleatorias con valor esperado cero y varianza constante. Principio de la Homocedasticidad [o *igual* (homo) *dispersión* (cedasticidad)].

En un gráfico de e_i versus los valores ajustados por el modelo de regresión ($\hat{Y}_i$):



Homocedástico



Heterocedástico

Regresion Lineal

Heterocedasticidad

Causas	Consecuencias
✓Sesgo de especificación. ✓Existencia de variables endógenas retardadas. ✓Manipulación de los datos. ✓Suele ser un problema cuando se utilizan datos transversales.	✓Los estimadores son insesgados. ✓Los estimadores son ineficientes. ✓Es más fácil aceptar la hipótesis de que algún parámetro sea igual a 0. ✓Las pruebas de hipótesis (*t*,*F*) pueden ofrecer resultados incorrectos.

Regresion Lineal

Supuestos de la Regresión Lineal

Heterocedasticidad

Métodos para detectarla	Solución
✓Método gráfico. ✓Contraste de Goldfeld - Quandt. ✓Pruebas de Glejser.	✓Transformar el modelo hasta que cumpla las hipótesis. ✓Transformación logarítmica.

Regresion Lineal

Supuestos de la Regresión Lineal

Los errores son variables aleatorias con distribución normal.

68% de los residuos deben estar entre -1 e +1

90% de los residuos deben estar entre -1,64 a +1,64

95% de los residuos deben estar entre -1,96 a +1,96

Comentario:

La suposición de normalidad de los residuos simplifica la teoría de análisis de regresión y es necesaria para garantizar la validez de las pruebas de hipótesis y la estimación de intervalos de confianza.

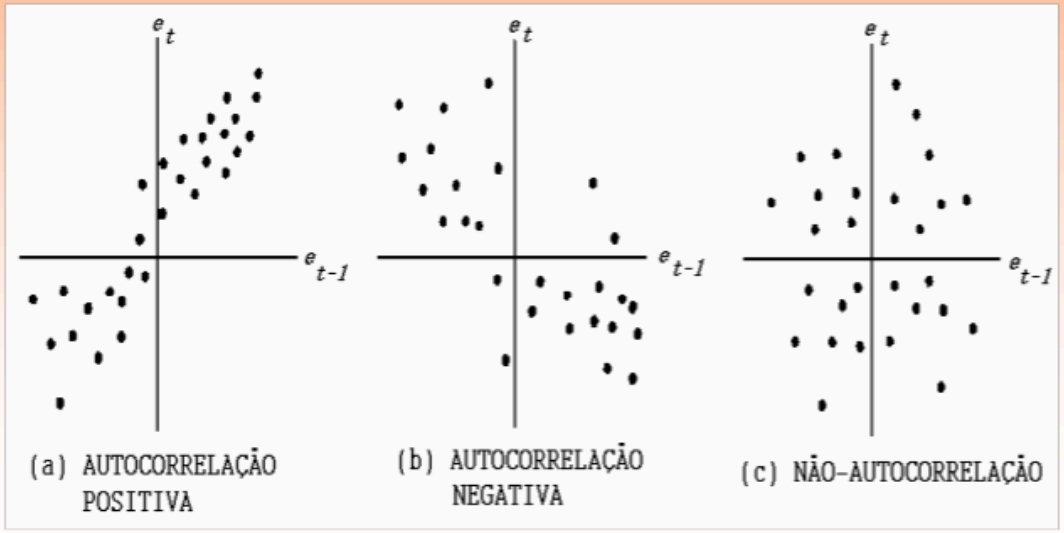
Si se constata la no - normalidad, se puede intentar con la transformación de los datos (Ln o raíz cuadrada).

Pelli (2005)

Regresion Lineal

Supuestos de la Regresión Lineal

Los errores no están correlacionados, es decir, que son independientes bajo la condición de normalidad. No auto-regresión.



González (2006)

Regresion Lineal

Supuestos de la Regresión Lineal: AUTOCORRELACIÓN

Causas	Consecuencias
✓Naturaleza de los datos. ✓Sesgo de especificación. ✓Existencia de variables endógenas retardadas. ✓Manipulación de los datos. ✓Suele ser un problema cuando se utilizan datos temporales.	✓Los estimadores son insesgados. ✓Los estimadores son ineficientes. ✓Es más fácil aceptar la hipótesis de que algún parámetro sea igual a 0. ✓Probable sobrestimación de R^2. ✓En presencia de auto-regresión, las pruebas *t* y *F* señalan conclusiones incorrectas.

Regresion Lineal

Supuestos de la Regresión Lineal

AUTOCORRELACIÓN

Métodos para detectarla	Solución
✓Método gráfico. ✓Prueba de "las Rachas" ✓Contraste de Durbin-Watson.	✓Transformar el modelo hasta que cumpla las hipótesis. ✓Reordenar los datos (caso de datos de índices).

PRUEBA DE CONTRASTE
DURBIN - WATSON

✓**Obtener los residuos del modelo.**

✓**Calcular el estadístico d:**

$$d = \frac{\sum_{t=2}^{t=n} (e_t - e_{t-1})^2}{\sum_{t=2}^{t=n} e_t^2} = 2(1-\hat{\rho})$$

$$\hat{\rho} = \frac{\sum e_i . e_{i-1.}}{\sum e_i^2}$$

H_0: no existe autocorrelación
H_1: existe correlación

✓**Buscar en la tabla los valores d_L y d_U entrando con *n (número de datos)* y con *k' (variables explicativas).***

✓**Regla de Decisión - Rechazar H_0 si:**

d< dL ó d> 4-du

Rodríguez (2005)

REPRESENTACIÓN GRÁFICA DE LA PRUEBA - DURBIN WATSON

-Si $d_u < d < 4 - d_u$, rechaza H_0, o sea, rechaza la hipótesis de que los residuos están correlacionados a favor de la hipótesis de no autocorrelación al nivel de significancia establecido;

-Si $d < d_L$ se acepta la hipótesis de auto-regresión positiva;

-Si $d > 4 - d_L$ se acepta la hipótesis de auto-regresión negativa;

-En los demás casos la prueba no es concluyente.

La representación gráfica de la prueba puede ser visualizada en la figura 5.30.

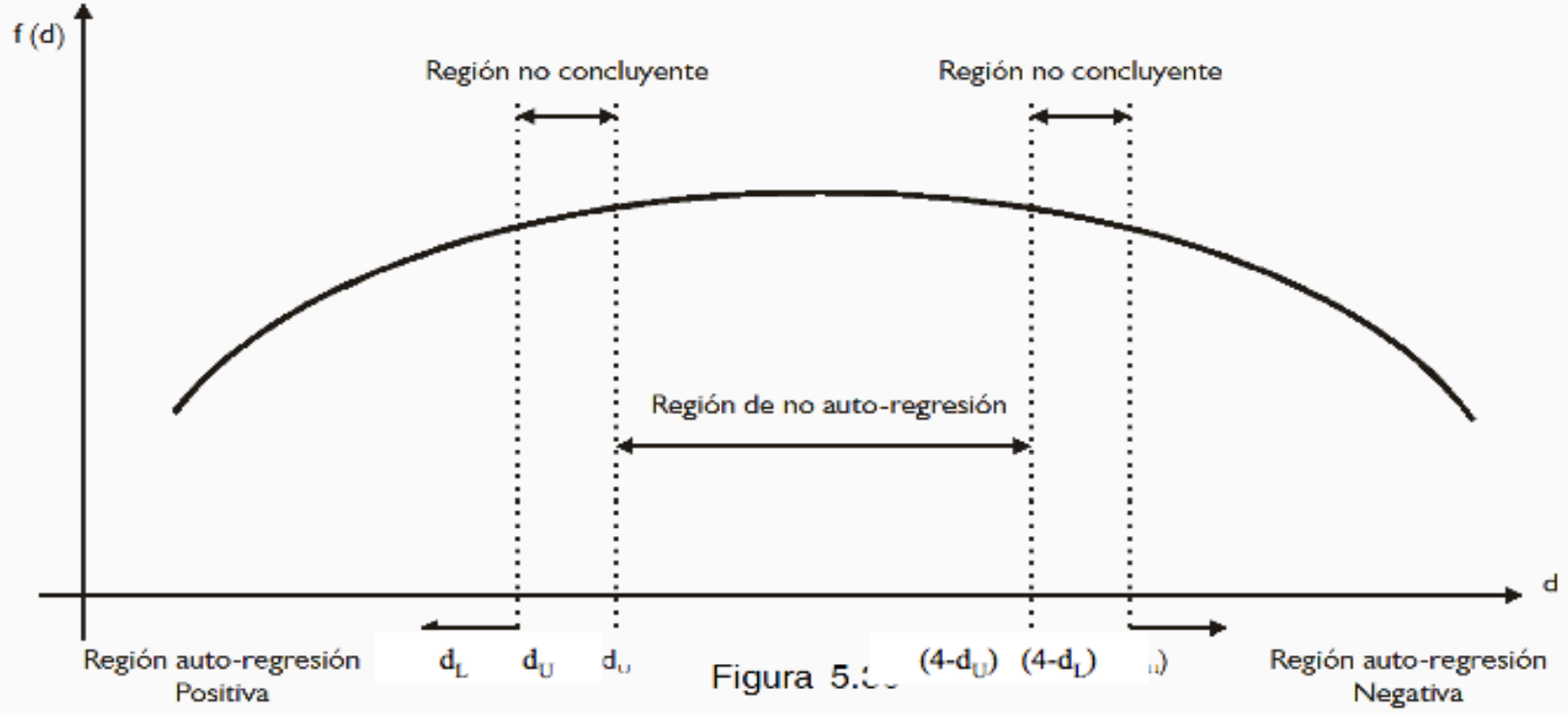


Figura 5.3

Regresion Lineal

OTROS Supuestos de la Regresión Lineal

- **Las variables importantes fueron incluidas en el modelo.**
- **No existen observaciones atípicas: *outliers* o *puntos influyentes*.**
- **Outliers:**

Están representados por un dato que tiene un residuo grande en relación a los demás que conforman la muestra.

Se detectan rápidamente a través de un gráfico de los residuos estandarizados (e^*_i) versus los valores estimados ($\hat{Y}_i$).

Regresion Lineal

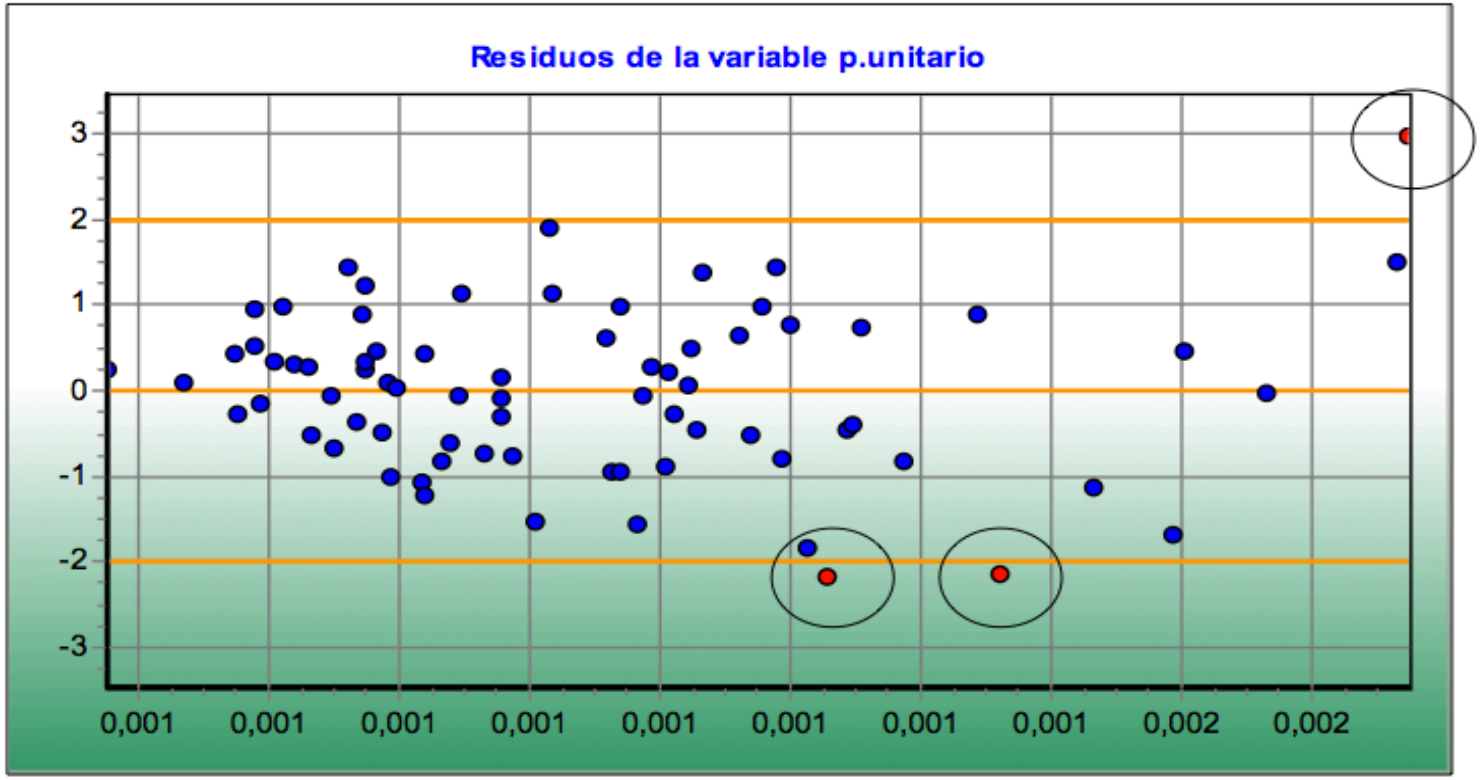
OTROS Supuestos de la Regresión Lineal

- **Si el modelo está correcto, se puede emplear los residuos estandarizados para verificar la existencia de *outliers*, como el 95% de la distribución Normal está contenida en el intervalo (-1,96,1,96), se puede esperar que aproximadamente el 95% de las cantidades $e_i/s(e)$ estén dentro de estos límites (generalmente se aproxima a ±2 desviaciones estándar)**
- **No se deben remover inmediatamente, pero se debe buscar explicaciones para esos valores, manteniendolos si no hay indicios de errores incorregibles. (5%)**

Regresion Lineal

OTROS Supuestos de la Regresión Lineal: Valores atípicos

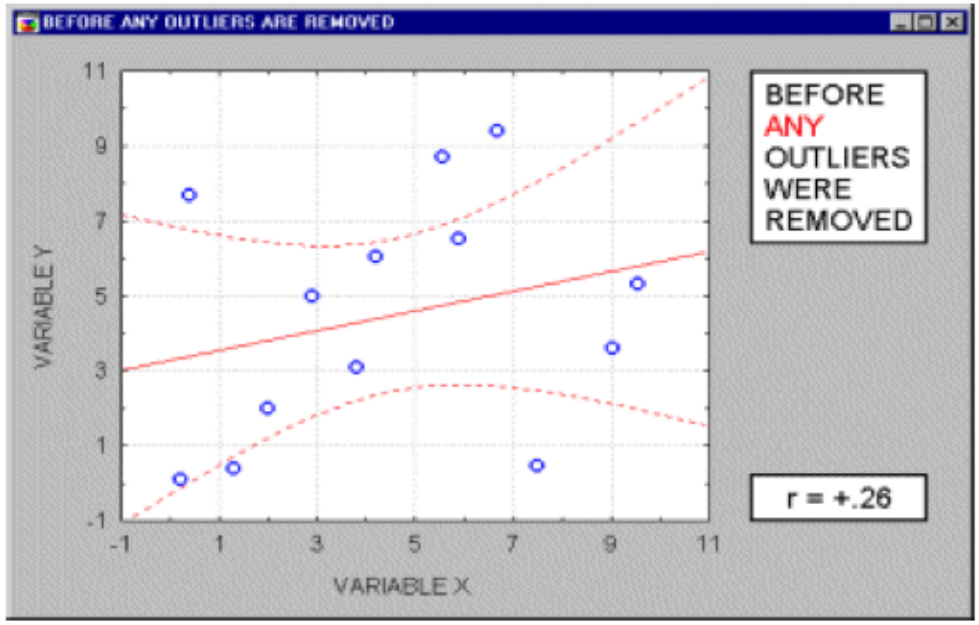
Los outliers pueden ser interpretados como indicios de fallas de recolección, análisis o explicación de los datos y no de los propios datos.



Regresion Lineal

OTROS Supuestos de la Regresión Lineal: Valores atípicos

Crean dificultades en el análisis, principalmente en el uso del Método de las Mínimos Cuadrados, que busca el ajuste de todos los puntos observados. Un punto extremo afecta los coeficientes de regresión, disfrazando los resultados. González (2006)



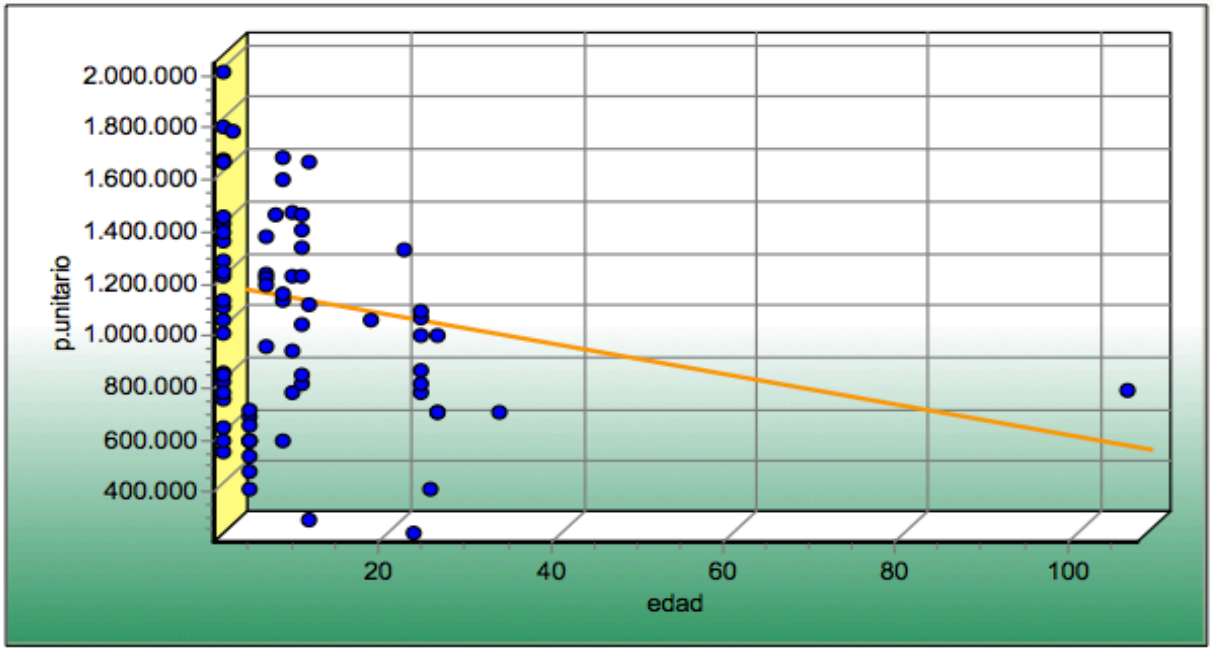
www.statsoft.com

Regresion Lineal

Puntos Influyentes

Son aquellos que tienen pequeños residuos, algunas veces nulos, que se distancian de la masa de datos, pudiendo alterar completamente las tendencias naturales indicadas por el mercado.

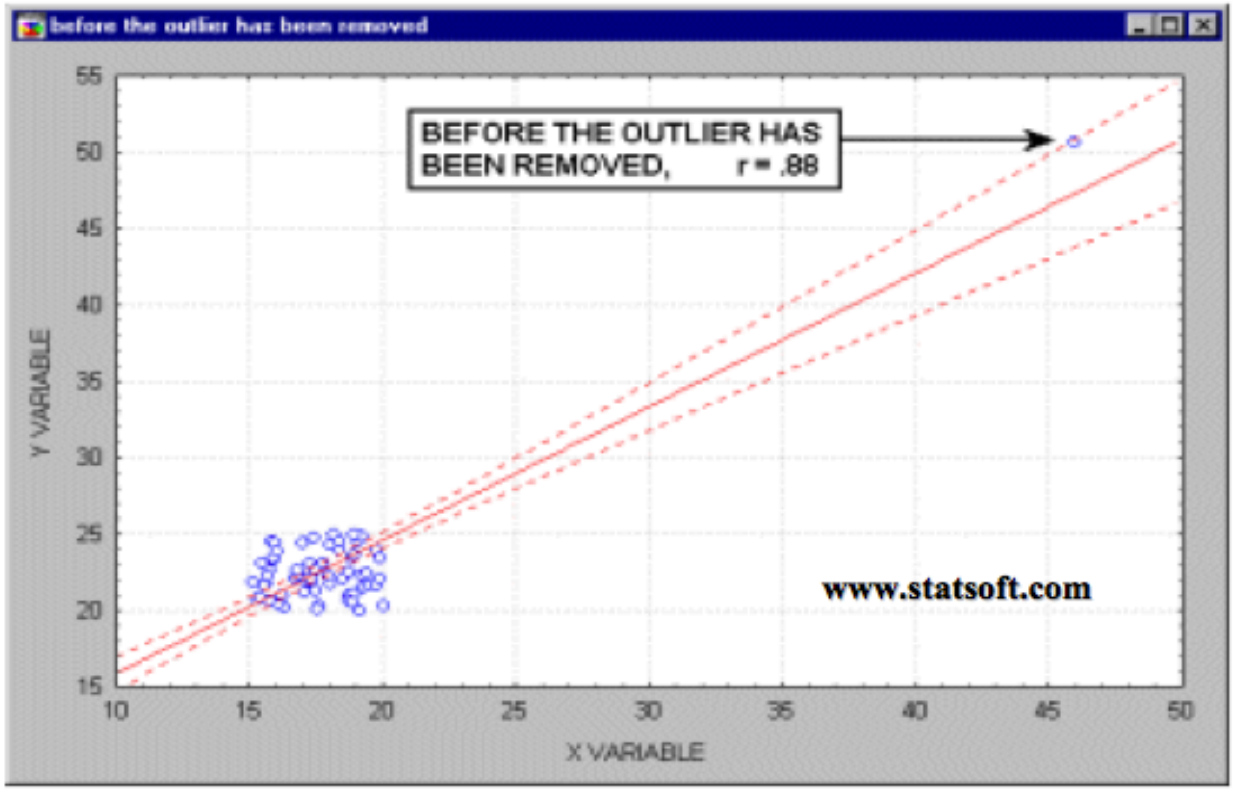
González (2006)



Regresion Lineal

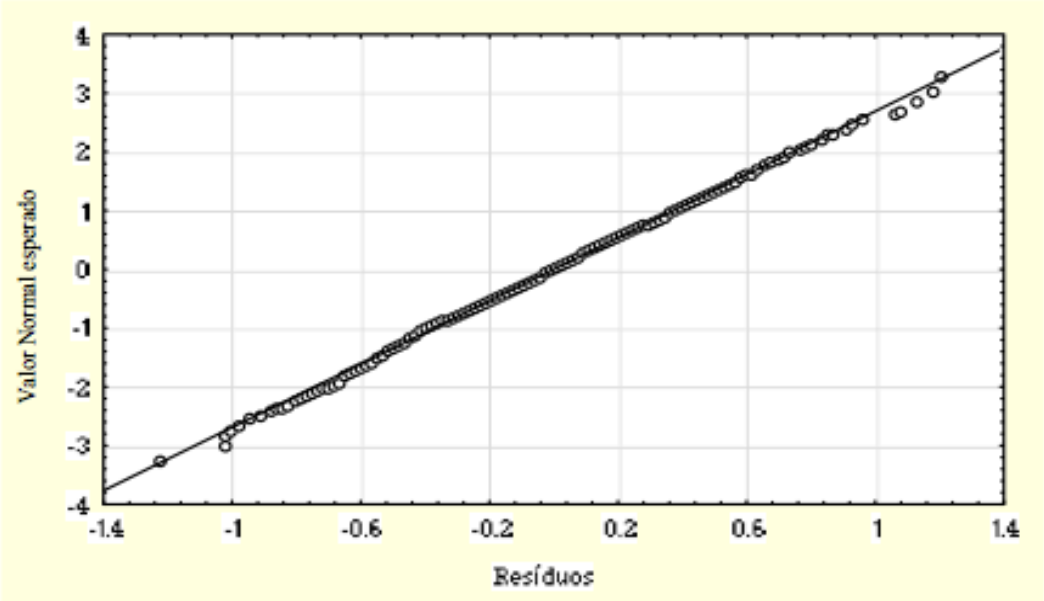
Puntos Influyentes

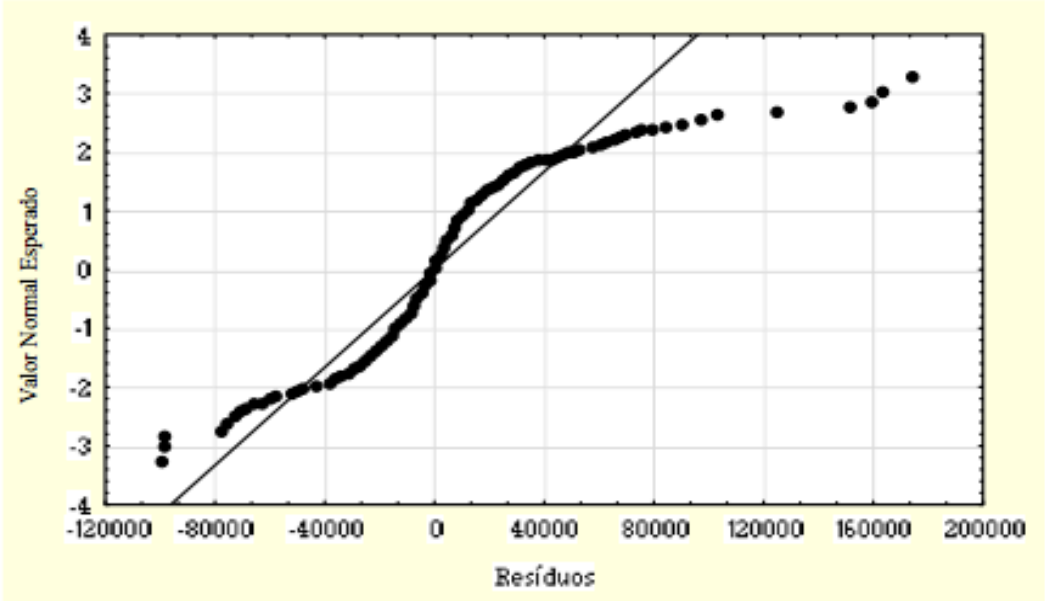
Existen varios criterios estadísticos para detectar estos puntos, como lo es la Distancia de Cook (ver SisReN)



Regresion Lineal

Supuestos de la Regresión Lineal: NORMALIDAD





La forma mas simple de observar la normalidad de los residuos es a través de gráficos. Generalmente los softwares estadísticos presentan gráficos de residuos (X) contra valores esperados (Y). Hay una recta de referencia, representando la curva Normal. En cuanto mas los residuos se aproximen a una recta, coincidente con la de referencia, mayor es la confianza de que se puede tener Normalidad.

Regresion Lineal Múltiple

RESUMEN

Supuestos de la Regresión Lineal

- No debe existir **ninguna relación exacta entre** cualesquiera de las **variables independientes**. Multicolinealidad
- Independencia de los Residuos (ausencia de autocorrelación)
- Normalidad de los Residuos
- Homocedasticidad de los Residuos

Otros requisitos

- Las variables deben variar!!
- Número suficiente de muestras
- Variables NO incluidas en el modelo NO afectan sistemáticamente
- Linealidad (en caso de NO linealidad es un error de especificación)
- Ausencia de outliers y puntos influyentes