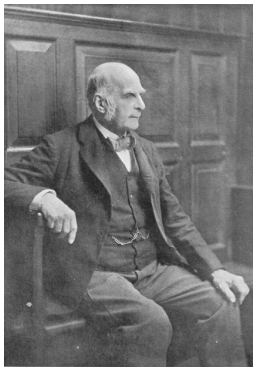
Bioestadística

Tema 3: Estadística descriptiva bivariante y regresión lineal.

Relaciones entre variables y regresión

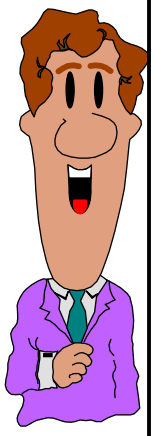
- El término regresión fue introducido por Galton en su libro "*Natural inheritance*" (1889) refiriéndose a la "ley de la regresión universal":
  - "Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor."
    - Regresión a la media
  - Su trabajo se centraba en la descripción de los rasgos físicos de los descendientes (una variable) a partir de los de sus padres (otra variable).
  - Pearson (un amigo suyo) realizó un estudio con más de 1000 registros de grupos familiares observando una relación del tipo:
    - Altura del hijo = 85cm + **0,5** altura del padre (aprox.)
    - Conclusión: los padres muy altos tienen tendencia a tener hijos que heredan parte de esta altura, aunque tienen tendencia a acercarse (*regresar*) a la media. Lo mismo puede decirse de los padres muy bajos.
- Hoy en día el sentido de regresión es el de predicción de una medida basándonos en el conocimiento de otra.



Francis Galton
•Primo de Darwin
•Estadístico y aventurero
•Fundador (con otros) de la estadística moderna para explicar las teorías de Darwin.

Qué vamos a estudiar

- En este capítulo vamos a tratar diferentes formas de describir la relación entre dos variables cuando estas son numéricas.
  - Estudiar si hay relación entre la altura y el peso.
- Haremos mención de pasada a otros casos:
  - Alguna de las variables es ordinal.
    - Estudiar la relación entre el sobrepeso y el dolor de espalda (ordinal)
  - Hay más de dos variables relacionadas.
    - ¿Conocer el peso de una persona conociendo su altura y contorno de cintura?
- El estudio conjunto de dos variables cualitativas lo aplazamos hasta que veamos contrastes de hipótesis (X^2).
  - ¿Hay relación entre fumar y padecer enfermedad de pulmón?



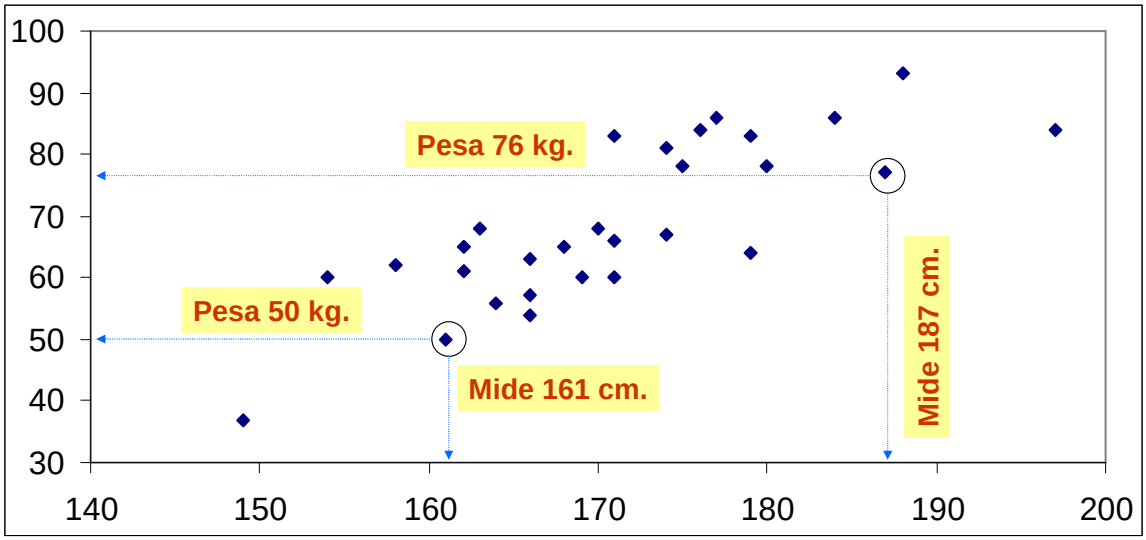
Estudio conjunto de dos variables

- A la derecha tenemos una posible manera de recoger los datos obtenido observando dos variables en varios individuos 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.
- Dichas observaciones pueden ser representadas en un **diagrama de dispersión** ('scatterplot'). En ellos, cada individuos es un punto cuyas coordenadas son los valores de las variables.
- 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
...	...

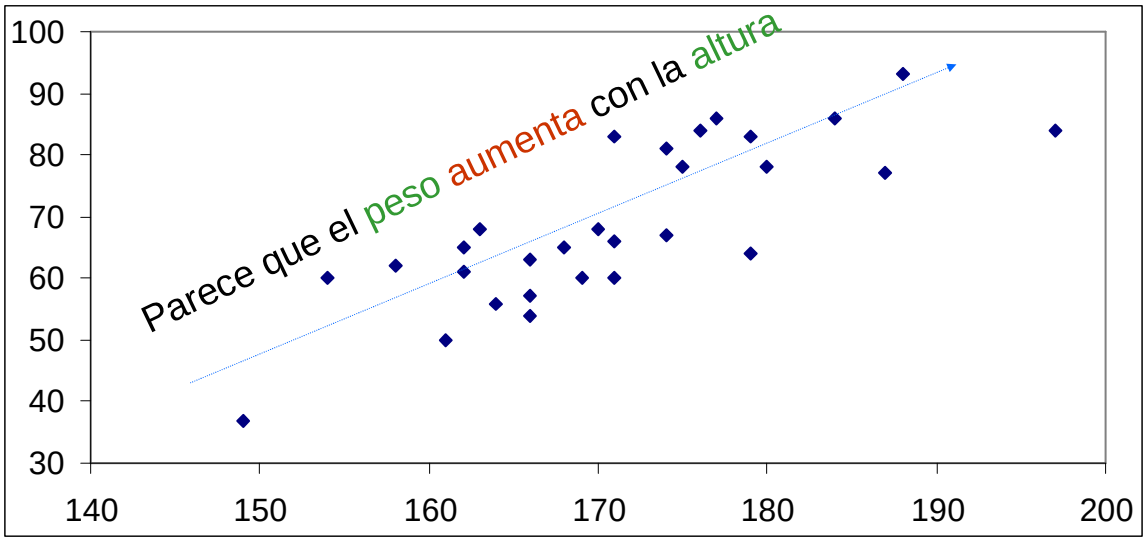
Diagramas de dispersión o nube de puntos

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



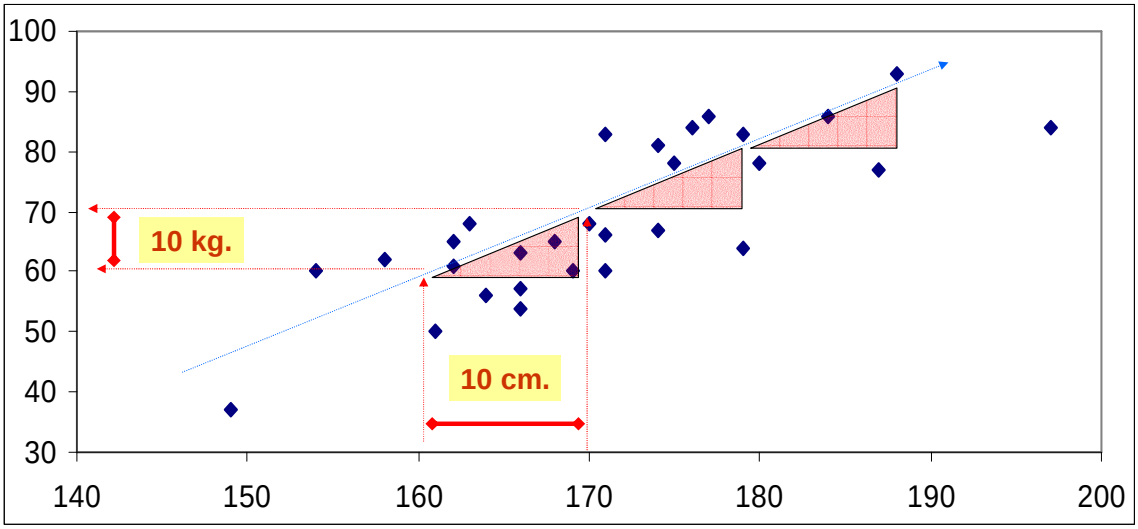
Relación entre variables.

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

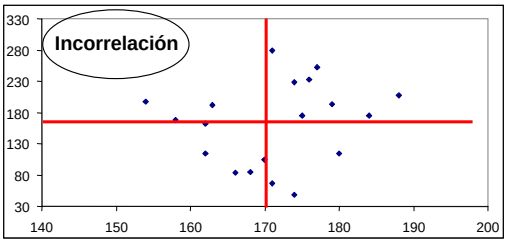


Predicción de una variable en función de la otra

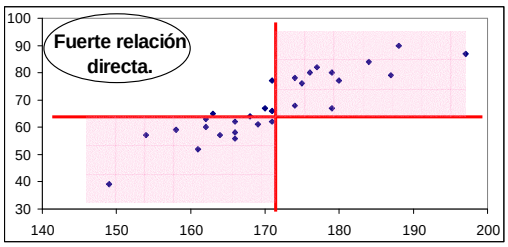
Aparentemente el peso aumenta 10Kg por cada 10 cm de altura... o sea, el peso aumenta en una unidad por cada unidad de altura.



Relación directa e inversa



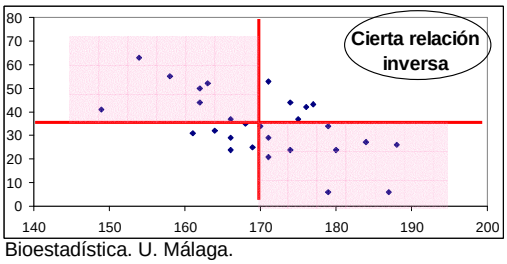
Para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares. Incorrelación.



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

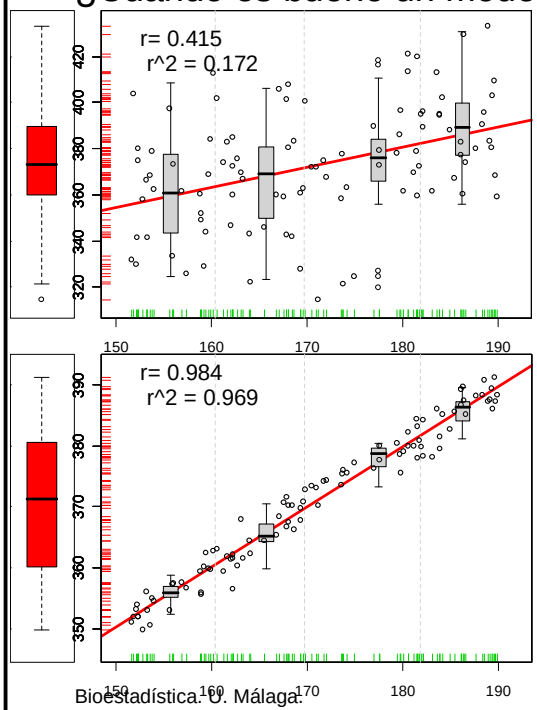
•Para los valores de X menores que la media le corresponden valores de Y menores también.

•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.

Covarianza de dos variables X e Y

- La covarianza entre dos variables, $\mathbf{S_{xy}}$, nos indica si la posible relación entre dos variables es directa o inversa.
  - Directa: $S_{xy} > 0$
  - Inversa: $S_{xy} < 0$
  - Incorreladas: $S_{xy} = 0$

$$S_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- El signo de la covarianza nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el grado de relación entre las variables.

Coef. de correlación lineal de Pearson

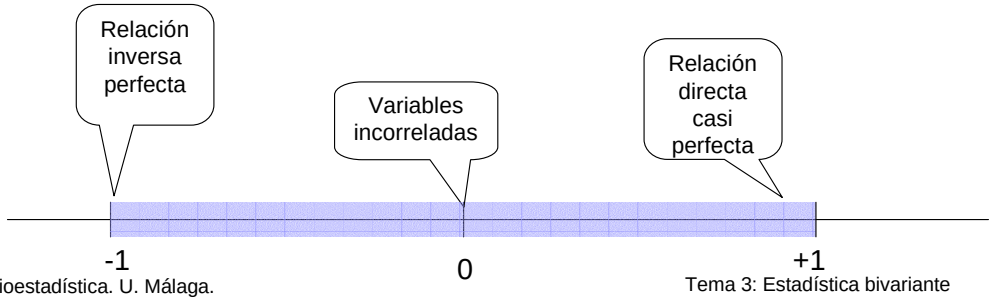
- La coeficiente de correlación lineal de Pearson de dos variables, **r**, nos indica si los puntos tienen una tendencia a disponerse alineadamente (excluyendo rectas horizontales y verticales).

- tiene el mismo signo que S_{xy} por tanto de su signo obtenemos el que la posible relación sea directa o inversa.
- r es útil para determinar si hay relación lineal entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)

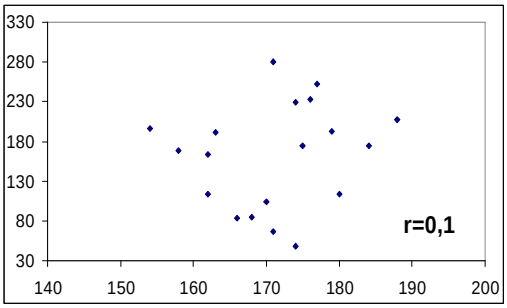
$$r = \frac{S_{xy}}{S_x S_y}$$

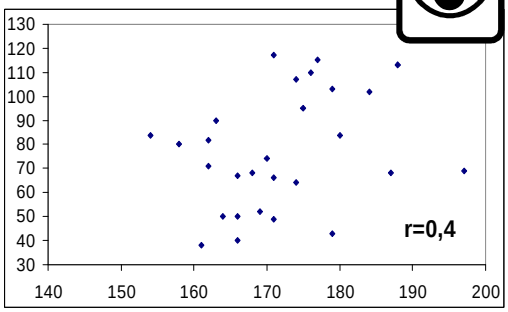
Propiedades de r

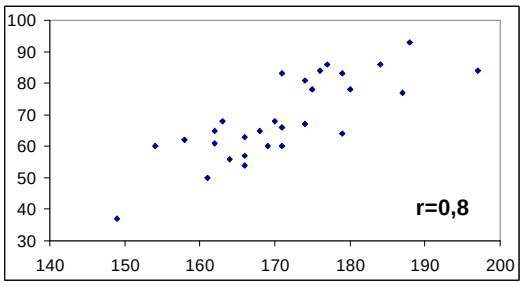
- Es adimensional
- Sólo toma valores en [-1,1]
- Las variables son incorreladas ⇔ r=0
- Relación lineal perfecta entre dos variables ⇔ r=+1 o r=-1
  - Excluimos los casos de puntos alineados horiz. o verticalmente.
- Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.
  - Siempre que no existan observaciones anómalas.

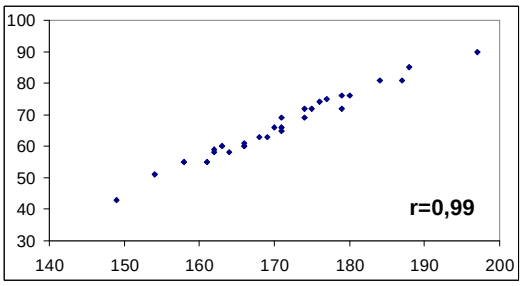


Entrenando el ojo: correlaciones positivas

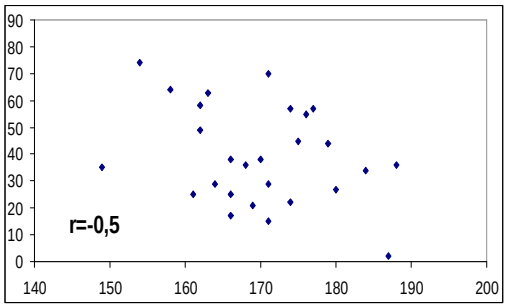


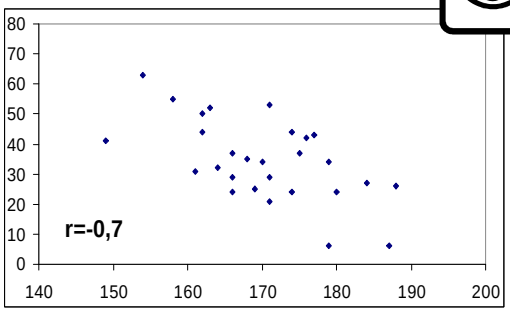


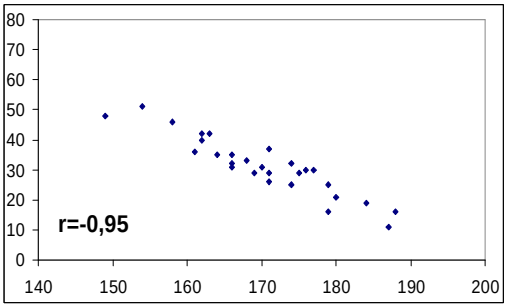


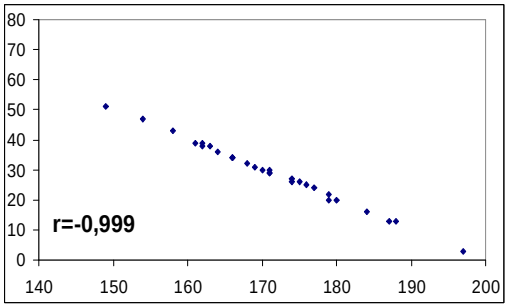


Entrenando el ojo: correlaciones negativas

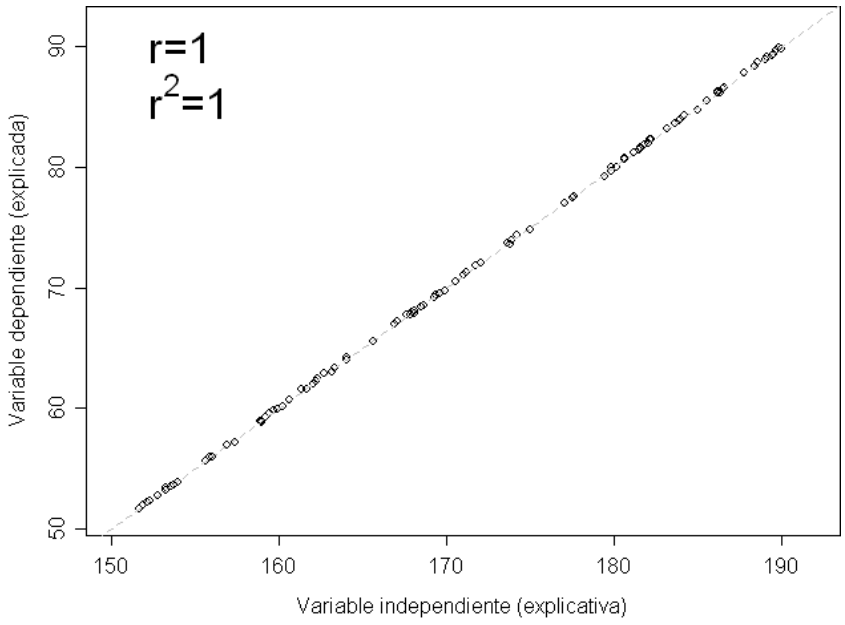




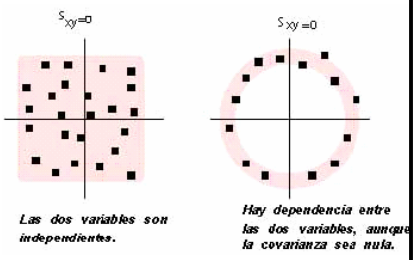




Animación: Evolución de r y diagrama de dispersión



Preguntas frecuentes

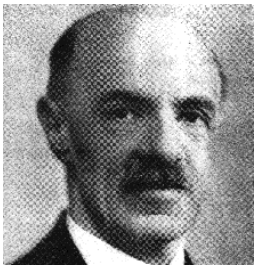


- **¿Si r=0 eso quiere decir que no las variables son independientes?**
  - En la práctica, casi siempre sí, pero no tiene por qué ser cierto en todos los casos.
  - Lo contrario si es cierto: Independencia implica incorrelación.
- **Me ha salido r=1'2 ¿la relación es "superlineal"**[sic]**?**
  - ¿Superqué? Eso es un error de cálculo. Siempre debe tomar un valor entre -1 y +1.
- **¿A partir de qué valores se considera que hay "buena relación lineal"?**
  - Imposible dar un valor concreto (mirad los gráficos anteriores). Para este curso digamos que si |r|>0,7 hay buena relación lineal y que si |r|>0,4 hay cierta relación (por decir algo... la cosa es un poco más complicada… observaciones atípicas, homogeneidad de varianzas...)

Otros coeficientes de correlación

- Cuando las variables en vez de ser numéricas son ordinales, es posible preguntarse sobre si hay algún tipo de correlación entre ellas.
- Disponemos para estos casos de dos estadísticos, aunque no los usaremos en clase:
  - ρ ('ro') de Spearman
  - τ ('tau') de Kendall
- No tenéis que estudiar nada sobre ellos en este curso. Recordad sólo que son estadísticos análogos a r y que los encontrareis en publicaciones donde las variables no puedan considerarse numéricas.

Maurice George Kendall



Charles Edward Spearman

Regresión

- El análisis de regresión sirve para predecir una medida en función de otra medida (o varias).
  - Y = Variable dependiente
    - predicha
    - explicada
  - X = Variable independiente
    - predictora
    - explicativa
  - ¿Es posible descubrir una relación?
    - Y = f(X) + error
      - f es una función de un tipo determinado
      - el error es aleatorio, pequeño, y no depende de X

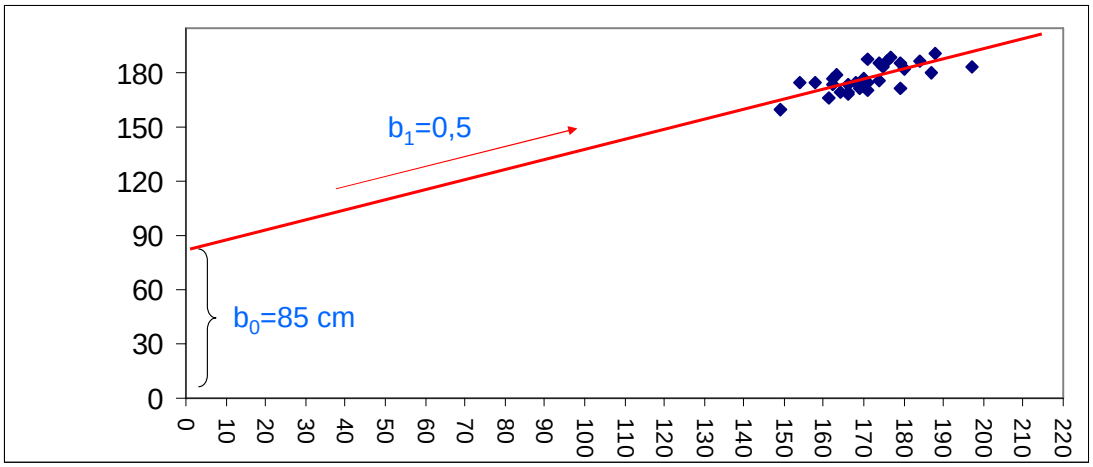
Regresión

- El ejemplo del estudio de la altura en grupos familiares de Pearson es del tipo que desarrollaremos en el resto del tema.
  - Altura del hijo = 85cm + **0,5** altura del padre (Y = 85 + 0,5 X)
    - Si el padre mide 200cm ¿cuánto mide el hijo?
      - Se espera (predice) 85 + 0,5x200=185 cm.
        - Alto, pero no tanto como el padre. Regresa a la media.
    - Si el padre mide 120cm ¿cuánto mide el hijo?
      - Se espera (predice) 85 + 0,5x120=145 cm.
        - Bajo, pero no tanto como el padre. Regresa a la media.
- Es decir, nos interesaremos por modelos de regresión lineal simple.

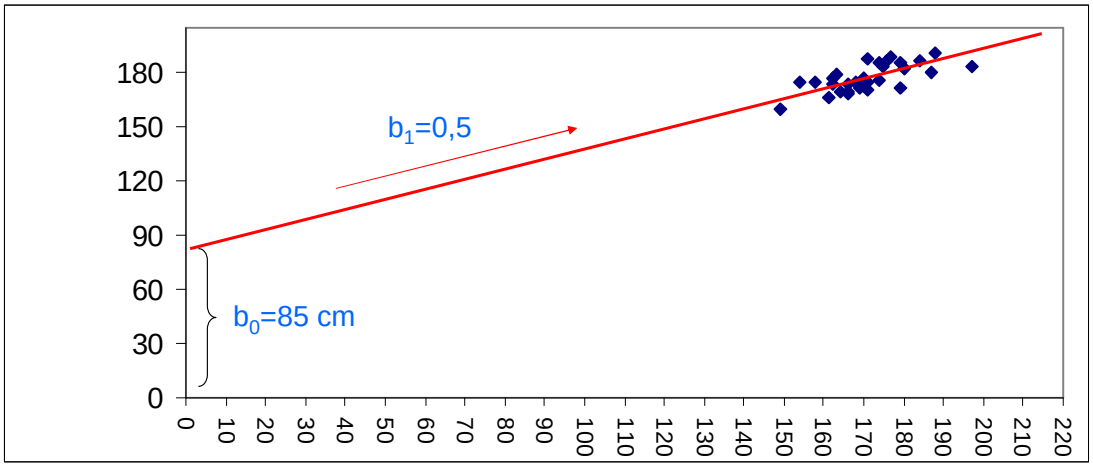
Modelo de regresión lineal simple

- 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.

- En el ejemplo de Pearson y las alturas, él encontró:
  - $\hat{Y} = b_0 + b_1 X$
    - b_0=85 cm (No interpretar como altura de un hijo cuyo padre mide 0 cm ¡Extrapolación salvaje!
    - b_1=0,5 (En media el hijo gana 0,5 cm por cada cm del padre.)



- La relación entre las variables no es exacta. Es natural preguntarse entonces:
  - Cuál es la mejor recta que sirve para predecir los valores de Y en función de los de X
  - Qué error cometemos con dicha aproximación (residual).

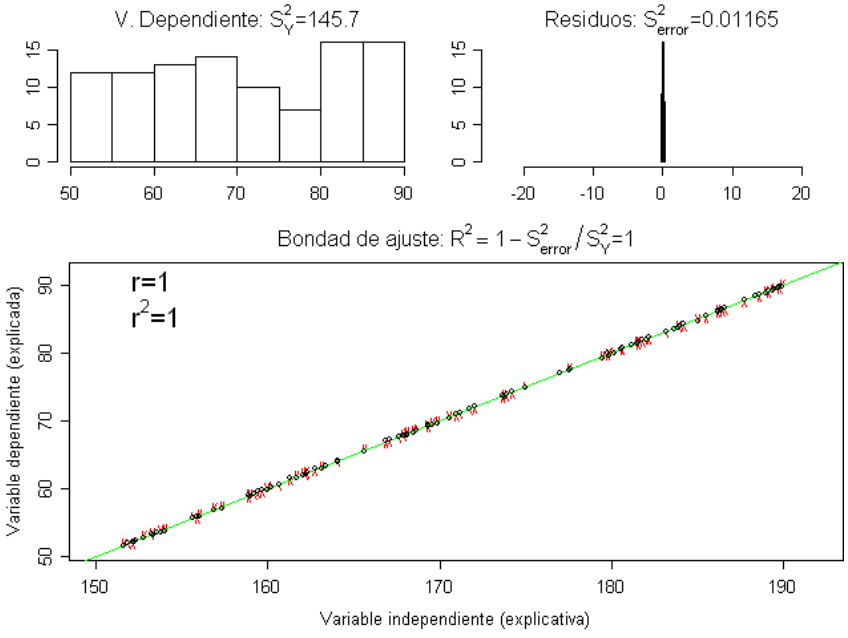


- El modelo lineal de regresión se construye utilizando la técnica de estimación mínimo cuadrática:
  - Buscar b_0, b_1 de tal manera que se minimice la cantidad
    - $\Sigma_i e_i^2$
- Se comprueba que para lograr dicho resultado basta con elegir:

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

- Se obtiene además unas ventajas "de regalo"
  - 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).

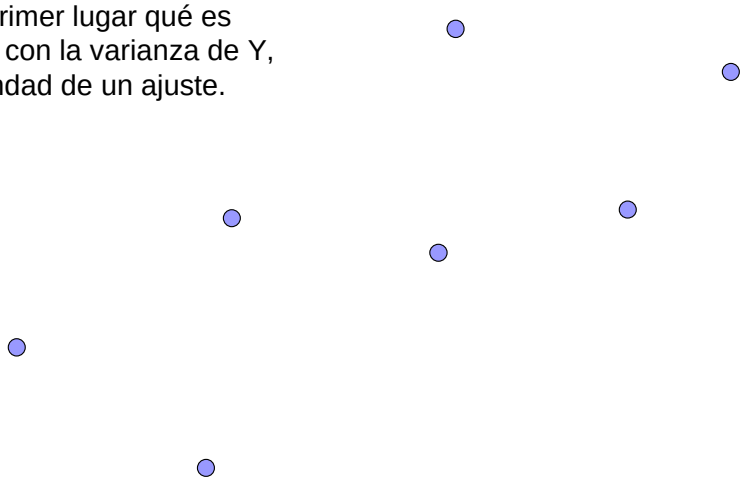
Animación: Residuos del modelo de regresión



- Que el error medio de las predicciones sea nulo **no** quiere decir que las predicciones sean buenas.
- Hay que encontrar un medio de expresar la bondad del ajuste (bondad de la predicción)

¿Cómo medir la bondad de una regresión?

Imaginemos un diagrama de dispersión, y vamos a tratar de comprender en primer lugar qué es el error residual, su relación con la varianza de Y, y de ahí, cómo medir la bondad de un ajuste.

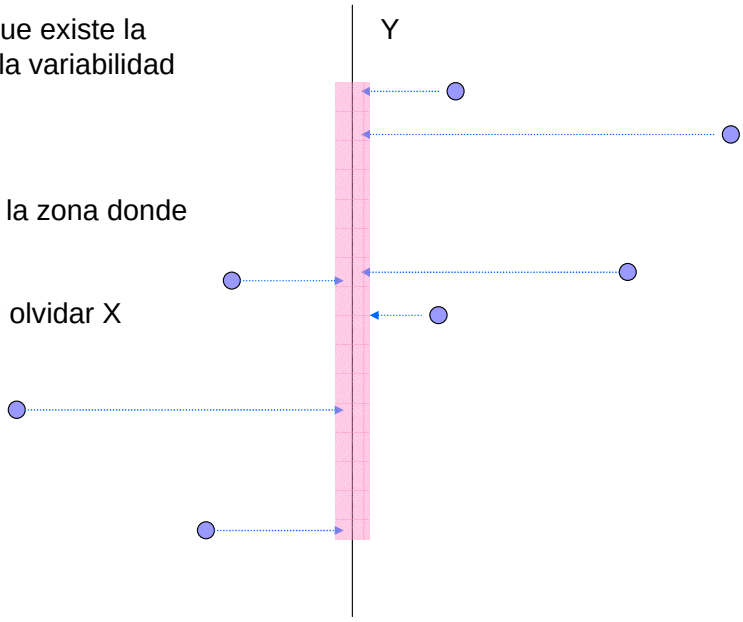


Interpretación de la variabilidad en Y

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

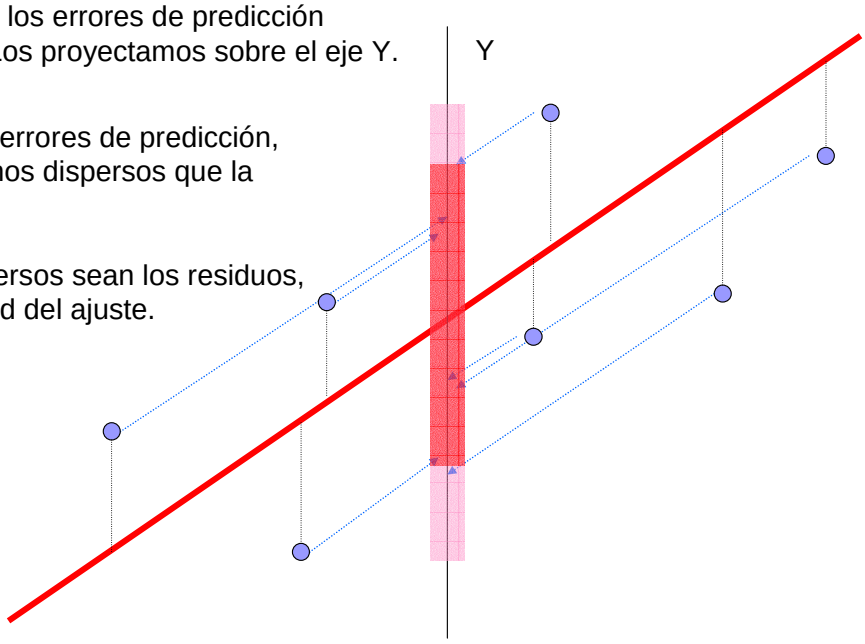


Interpretación del residuo

Fijémonos ahora en 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.



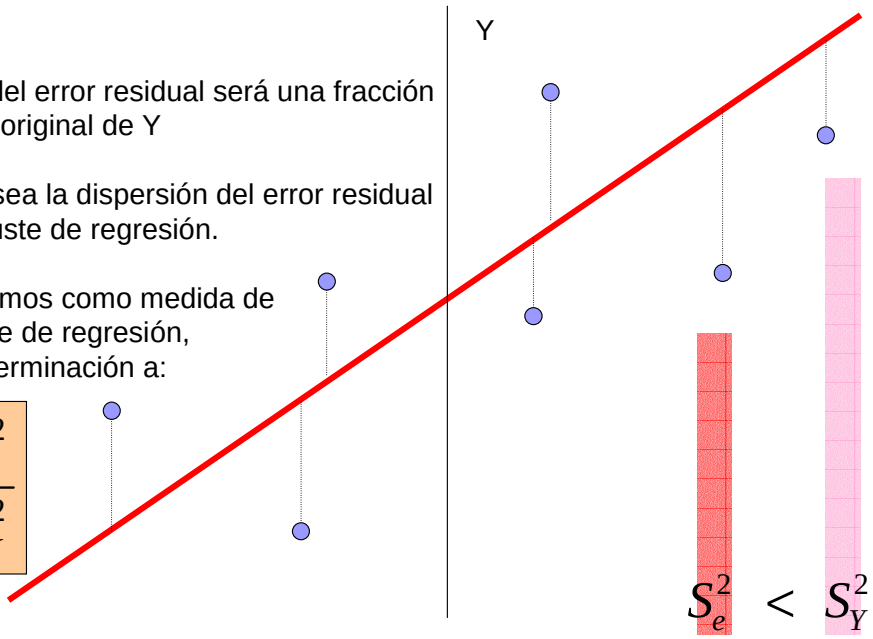
Bondad de un ajuste

Resumiendo:

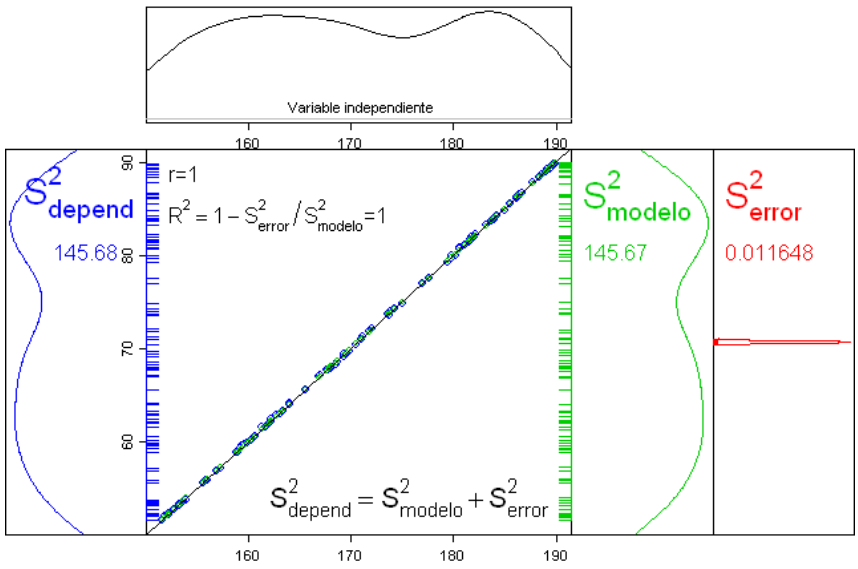
- La dispersión del error residual será una fracción de la dispersión original de Y
- Cuanto menor sea la dispersión del error residual mejor será el ajuste de regresión.

Eso hace que definamos como medida de bondad de un ajuste de regresión, o coeficiente de determinación a:

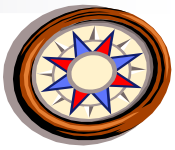
$$R^2 = 1 - \frac{S_e^2}{S_Y^2}$$



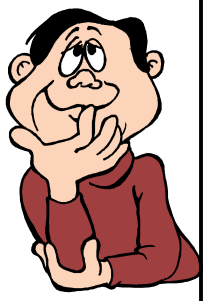
Animación: Descomposición de la varianza



Resumen sobre bondad de un ajuste

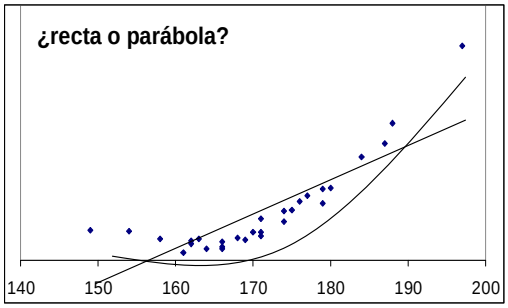


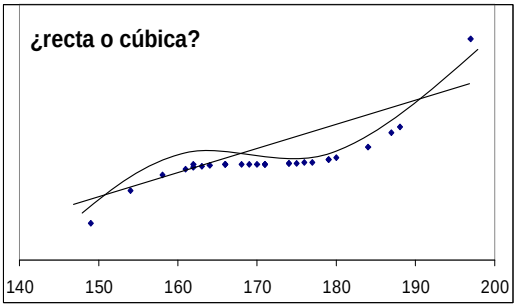
- La bondad de un ajuste de un modelo de regresión se mide usando el coeficiente de determinación R^2
- R^2 es una cantidad adimensional que sólo puede tomar valores en [0, 1]
  - Para el alumno astuto: ¿por qué?
- Cuando un ajuste es bueno, R^2 será cercano a uno.
  - ¿por qué?
- Cuando un ajuste es malo R^2 será cercano a cero.
  - ¿por qué?
- A R^2 también se le denomina porcentaje de variabilidad explicado por el modelo de regresión.
  - ¿por qué? Difícil.
- R^2 puede ser pesado de calcular en modelos de regresión general, pero en el modelo lineal simple, la expresión es de lo más sencilla: $\mathbf{R^2=r^2}$
  - ¿Es coherente lo dicho entonces sobre los valores de R^2?

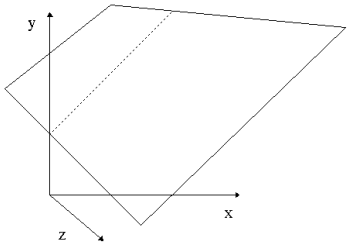


Otros modelos de regresión

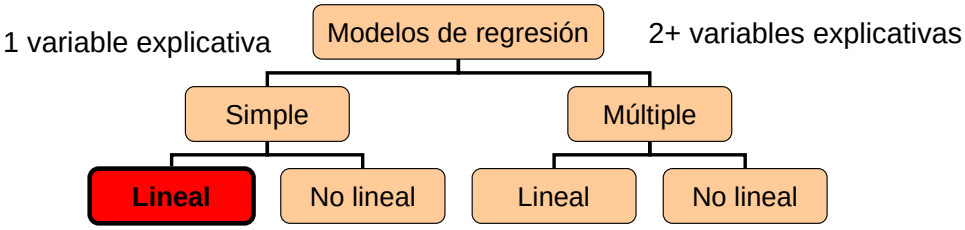
- Se pueden considerar otros tipos de modelos, en función del aspecto que presente el diagrama de dispersión (regresión no lineal)
- Incluso se puede considerar el que una variable dependa de varias (regresión múltiple).







Modelos de análisis de regresión

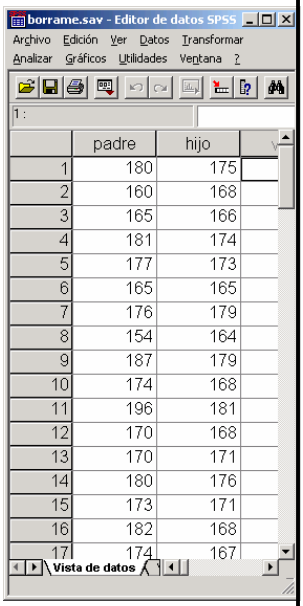


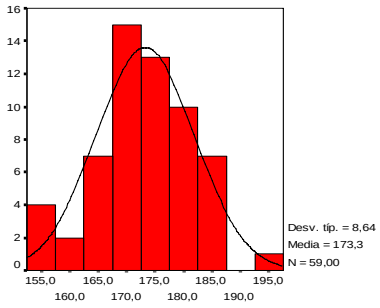
En clase sólo tratamos el modelo de regresión lineal simple.
En todos los demás la bondad del ajuste se mide usando R^2

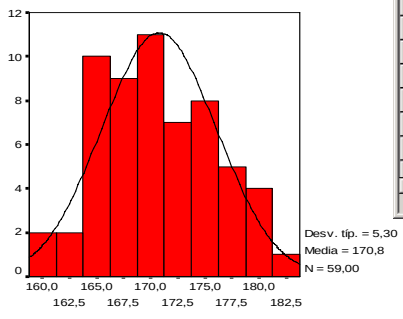
No ajustaremos modelos a mano. Usaremos para ello SPSS.

Ejemplo con SPSS

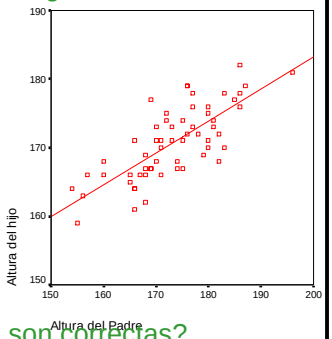
- A continuación vamos a analizar un ejemplo realizado con datos simulados, de lo que podría parecer el estudio sobre alturas de hijos y padres, realizado con SPSS.
- Suponemos que hemos recogido la altura de 60 varones, junto a las de su padre.
- El estudio descriptivo univariante de ambas variables por separado no revela nada sobre una posible relación.







- En el diagrama de dispersión se aprecie una clara relación lineal directa.
  - ¿Aprecias regresión a la media en el sentido de Galton en la gráfica?
- La tabla de correlaciones nos muestra que r=0,759
  - ¿Por qué se ven algunos r=1?
- El modelo de regresión lineal simple es
  - Altura hijo = b_0 + b_1 Altura del padre
    - b_0=89,985
    - b_1=0,466
    - ¿Aprecias regresión a la media?
- La bondad del ajuste es de R^2=0,577= 57,7%
  - ¿Eso significa que el 57% de las predicciones del modelo son correctas?
  - ¿Cómo lo interpretas?



Correlaciones

		Altura del hijo	Altura del Padre
Correlación de Pearson	Altura del hijo	1,000	,759
	Altura del Padre	,759	1,000

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,759[a]	,577	,569	3,480

a. Variables predictoras: (Constante), Altura del Padre

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Coeficientes no estandarizados Error típ.
1	(Constante)	89,985	9,180
	Altura del Padre	,466	,053

a. Variable dependiente: Altura del hijo

¿Qué hemos visto?

- Relación entre variables
- Diagrama de dispersión
- Covarianza
  - Relación directa, inversa e incorrelación
- Correlación lineal
  - Relación directa, inversa e incorrelación
  - grado de relación lineal entre variables
- Regresión, predicción
  - Variable dependiente
  - Variable(s) independientes
  - Modelo lineal de regresión
    - Ordenada en el origen
    - Pendiente
  - Residuo, error
  - Bondad del ajuste, coef. determinación
    - En el modelo lineal simple: r^2