Curso Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

CEU
Universidad
San Pablo

Licencia creative commons

Curso básico de estadística

Alfredo Sánchez Alberca (asalber@gmail.com).

Con esta licencia eres libre de:

- Copiar, distribuir y mostrar este trabajo.
- Realizar modificaciones de este trabajo.

Bajo las siguientes condiciones:

Reconocimiento. Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).

No comercial. No puede utilizar esta obra para fines comerciales.

Compartir bajo la misma licencia. Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- Estas condiciones pueden no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Relaciones entre variables

Hasta ahora se ha visto como describir el comportamiento de una variable, pero en los fenómenos naturales normalmente aparecen más de una variable que suelen estar relacionadas. Por ejemplo, en un estudio sobre el peso de las personas, deberíamos incluir todas las variables con las que podría tener relación: altura, edad, sexo, dieta, tabaco, ejercicio físico, etc.

Para comprender el fenómeno no basta con estudiar cada variable por separado y es preciso un estudio conjunto de todas las variables para ver cómo interactúan y qué relaciones se dan entre ellas. El objetivo de la estadística en este caso es dar medidas del grado y del tipo de relación entre dichas variables.

Generalmente, se considera una *variable dependiente* Y que se supone relacionada con otras variables $X_1, \ldots, X_n$ llamadas *variables independientes*.

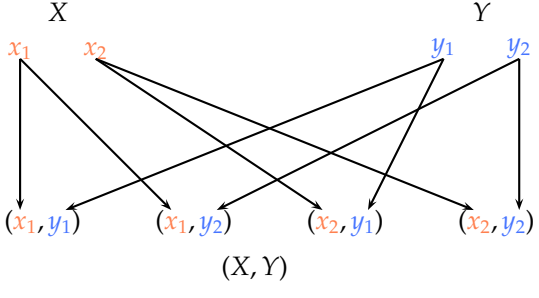
El caso más simple es el de una sola variable independiente, y en tal caso se habla de *estudio de dependencia simple*. Para más de una variable independiente se habla de *estudio de dependencia múltiple*.

En este tema se verán los estudios de dependencia simple que son más sencillos.

Variables bidimensionales

Al estudiar la dependencia simple entre dos variables X e Y, no se pueden estudiar sus distribuciones por separado, sino que hay que estudiarlas en conjunto.

Para ello, conviene definir una **variable estadística bidimensional** (X, Y), cuyos valores serán todos los pares formados por los valores de las variables X e Y.



Frecuencias de una variable bidimensional

Definición (Frecuencias muestrales de una variable bidimensional)

Dada una muestra de tamaño n de una variable bidimensional (X, Y), para cada valor de la variable (x_i, y_j) observado en la muestra se define:

- Frecuencia absoluta n_{ij}: Es el número de individuos de la muestra que presentan simultáneamente el valor x_i de la variable X y el valor y_j de la variable Y.
- Frecuencia relativa f_{ij}: Es la proporción de individuos de la muestra que presentan simultáneamente el valor x_i de la variable X y el valor y_j de la variable Y.

$$f_{ij} = \frac{n_{ij}}{n}$$

¡Ojo! Para las variables bidimensionales no tienen sentido las frecuencias acumuladas.

Distribución de frecuencias bidimensional

Al conjunto de valores de la variable bidimensional y sus respectivas frecuencias muestrales se le denomina **distribución conjunta**.

La distribución conjunta de una variable bidimensional se suele representar mediante una **tabla de frecuencias bidimensional**.

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_q
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}

Distribución de frecuencias bidimensional

Ejemplo con estaturas y pesos

Se ha medido la estatura (en cm) y el peso (en Kg) de 30 universitarios obteniendo:

(179,85), (173,65), (181,71), (170,65), (158,51), (174,66), (172,62), (166,60), (194,90), (185,75),(162,55), (187,78), (198,109), (177,61), (178,70), (165,58), (154,50), (183,93),(166,51), (171,65), (175,70), (182,60), (167,59), (169,62), (172,70), (186,71), (172,54), (176,68), (168,67), (187,80).

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$
$(150,160]$	2	0	0	0	0	0
$(160,170]$	4	4	0	0	0	0
$(170,180]$	1	6	3	1	0	0
$(180,190]$	0	1	4	1	1	0
$(190,200]$	0	0	0	0	1	1

Diagrama de dispersión

A menudo, la información de la tabla de frecuencias bidimensional se representa también gráficamente.

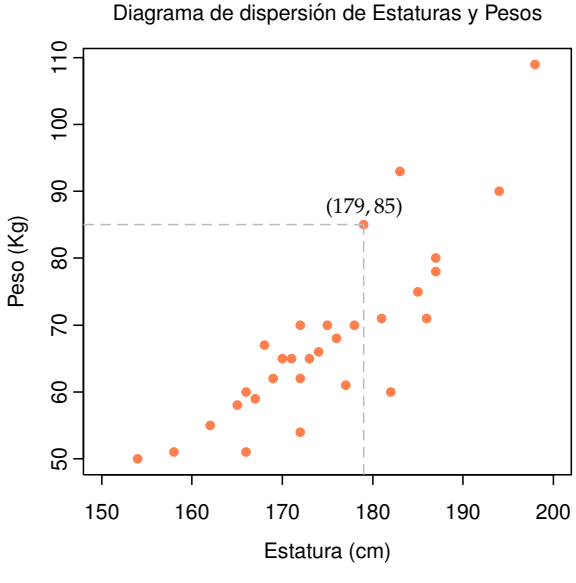
La representación gráfica que más se utiliza en el estudio de la dependencia de dos variables es el **diagrama de dispersión**, que consiste en representar sobre un plano cartesiano los puntos que se corresponden con los valores (x_i, y_j) de la variable bidimensional.

El conjunto de todos estos puntos recibe el nombre de *nube de puntos*.

En un diagrama de dispersión sólo se recogen los valores observados en la muestra, no las frecuencias de los mismos. Para reflejar las frecuencias tendríamos que recurrir a otro tipo de representación como un *diagrama de burbujas* o *histograma tridimensional*.

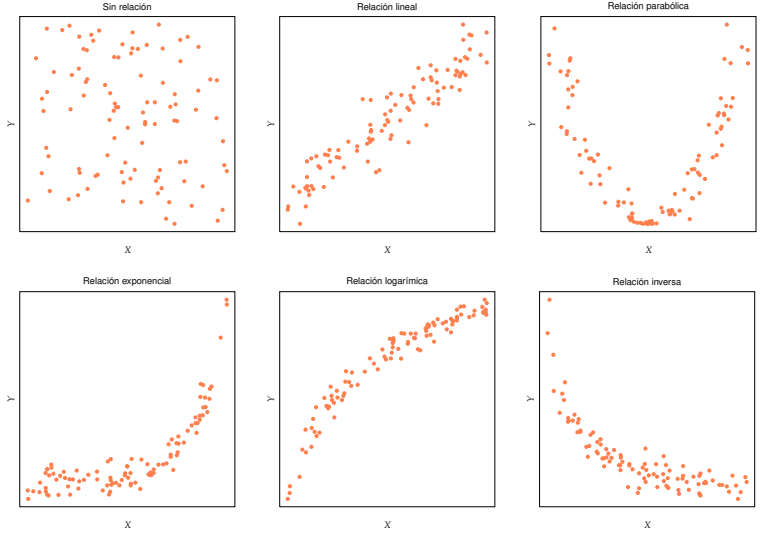
¡Ojo! No tiene sentido cuando alguna de las variables es un atributo.

Diagrama de dispersión



Interpretación del diagrama de dispersión

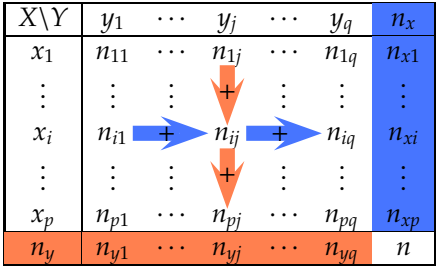
El diagrama de dispersión da información visual sobre el tipo de relación entre las variables.



Distribuciones marginales

A cada una de las distribuciones de las variables que conforman la variable bidimensional se les llama **distribuciones marginales**.

Las distribuciones marginales se pueden obtener a partir de la tabla de frecuencias bidimensional, sumando las frecuencias por filas y columnas.



Distribuciones marginales

Ejemplo con estaturas y pesos

En el ejemplo anterior de las estaturas y los pesos, las distribuciones marginales son

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$	n_x
$(150,160]$	2	0	0	0	0	0	2
$(160,170]$	4	4	0	0	0	0	8
$(170,180]$	1	6	3	1	0	0	11
$(180,190]$	0	1	4	1	1	0	7
$(190,200]$	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

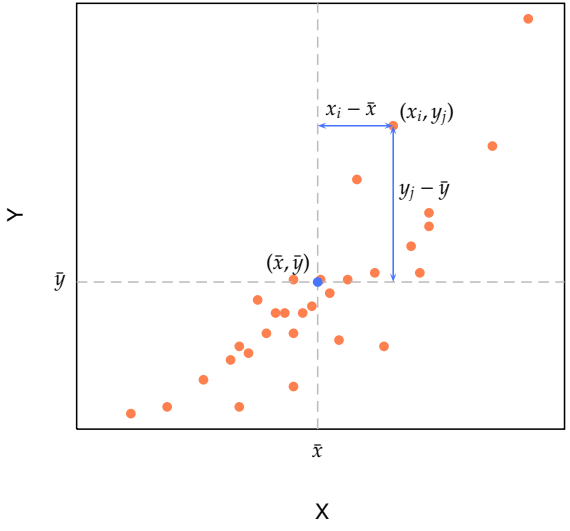
y los estadísticos asociados:

$\bar{x} = 174{,}67$ cm $\quad s_x^2 = 102{,}06$ cm^2 $\quad s_x = 10{,}1$ cm

$\bar{y} = 69{,}67$ Kg $\quad s_y^2 = 164{,}42$ Kg^2 $\quad s_y = 12{,}82$ Kg

Desviaciones respecto de las medias

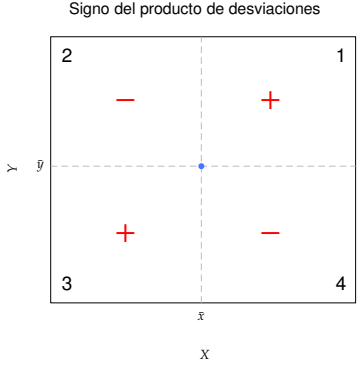
Para analizar la relación entre dos variables cuantitativas es importante hacer un estudio conjunto de las desviaciones respecto de la media de cada variable.



Estudio de las desviaciones respecto de las medias

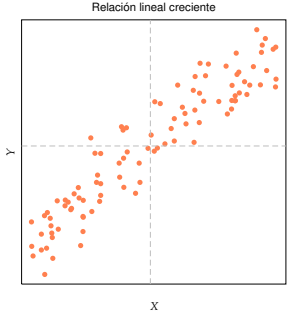
Si dividimos la nube de puntos del diagrama de dispersión en 4 cuadrantes centrados en el punto de medias $(\bar{x}, \bar{y})$, el signo de las desviaciones será:

Cuadrante	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	−	+	−
3	−	−	+
4	+	−	−



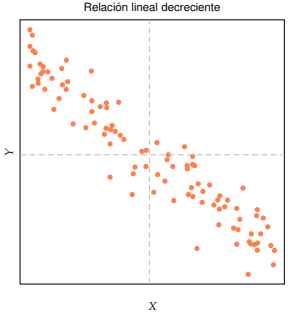
Estudio de las desviaciones respecto de las medias

Si la relación entre las variables es *lineal y creciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 1 y 3 y la suma de los productos de desviaciones será positiva.



$$\sum (x_i - \bar{x})(y_j - \bar{y}) = +$$

Si la relación entre las variables es *lineal y decreciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 2 y 4 y la suma de los productos de desviaciones será negativa.



$$\sum (x_i - \bar{x})(y_j - \bar{y}) = -$$

Covarianza

Del estudio conjunto de las desviaciones respecto de la media surge el siguiente estadístico de relación lineal:

Definición (Covarianza muestral)

La *covarianza muestral* de una variable aleatoria bidimensional (X, Y) se define como el promedio de los productos de las respectivas desviaciones respecto de las medias de X e Y.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_j - \bar{y}) n_{ij}}{n}$$

También puede calcularse de manera más sencilla mediante la fórmula

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y}.$$

La covarianza sirve para estudiar la relación lineal entre dos variables:

- Si $s_{xy} > 0$ existe una relación lineal creciente entre las variables.
- Si $s_{xy} < 0$ existe una relación lineal decreciente entre las variables.
- Si $s_{xy} = 0$ no existe relación lineal entre las variables.

Cálculo de la covarianza

Ejemplo con estaturas y pesos

En el ejemplo de las estaturas y pesos, teniendo en cuenta que

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$	n_x
$(150,160]$	2	0	0	0	0	0	2
$(160,170]$	4	4	0	0	0	0	8
$(170,180]$	1	6	3	1	0	0	11
$(180,190]$	0	1	4	1	1	0	7
$(190,200]$	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

$$\bar{x} = 174{,}67 \text{ cm} \qquad \bar{y} = 69{,}67 \text{ Kg}$$

la covarianza vale

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \cdots + 195 \cdot 105 \cdot 1}{30} - 174{,}67 \cdot 69{,}67 =$$

$$= \frac{368200}{30} - 12169{,}26 = 104{,}07 \text{ cm} \cdot \text{ Kg},$$

lo que indica que existe una relación lineal creciente entre la estatura y el peso.

Regresión

En muchos casos el objetivo de un estudio no es solo detectar una relación entre variables, sino explicarla mediante alguna función matemática.

La **regresión** es la parte de la estadística que trata de determinar la posible relación entre una variable numérica dependiente Y, y otro conjunto de variables numéricas independientes, $X_1, X_2, \ldots, X_n$, de una misma población. Dicha relación se refleja mediante un modelo funcional

$$y = f(x_1, \ldots, x_n).$$

El objetivo es determinar una ecuación mediante la que pueda estimarse el valor de la variable dependiente en función de los valores de las independientes.

El caso más sencillo se da cuando sólo hay una variable independiente X, entonces se habla de *regresión simple*. En este caso el modelo que explica la relación de Y como función de X es una función de una variable $y = f(x)$ que se conoce como **función de regresión**.

Modelos de regresión simple

Dependiendo de la forma de función de regresión, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Familia de curvas	Ecuación genérica
Lineal	$y = a + bx$
Parabólica	$y = a + bx + cx^2$
Polinómica de grado n	$y = a_0 + a_1 x + \cdots + a_n x^n$
Potencial	$y = a \cdot x^b$
Exponencial	$y = a \cdot e^{bx}$
Logarítmica	$y = a + b \log x$
Inverso	$y = a + \frac{b}{x}$
Curva S	$y = e^{a + \frac{b}{x}}$

La elección de un tipo u otro depende de la forma que tenga la nube de puntos del diagrama de dispersión.

Residuos o errores predictivos

Una vez elegida la familia de curvas que mejor se adapta a la nube de puntos, se determina, dentro de dicha familia, la curva que mejor se ajusta a la distribución.

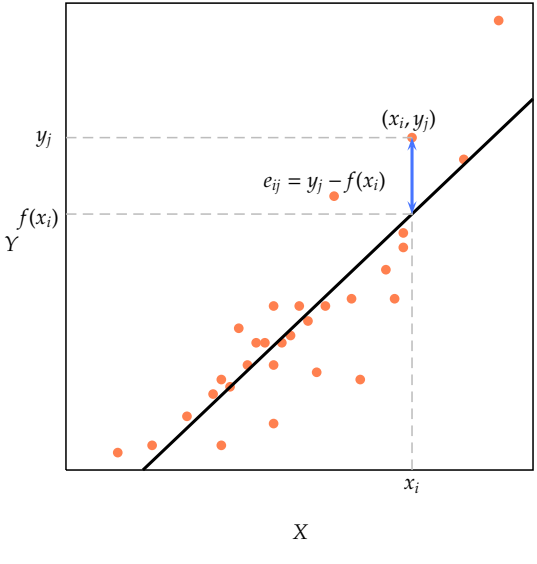
El objetivo es encontrar la función de regresión que haga mínimas las distancias entre los valores de la variable dependiente observados en la muestra, y los predichos por la función de regresión. Estas distancias se conocen como *residuos* o *errores predictivos*.

Definición (Residuos o Errores predictivos)

Dado el modelo de regresión $y = f(x)$ para una variable bidimensional (X, Y), el *residuo* o *error predictivo* de un valor (x_i, y_j) observado en la muestra, es la diferencia entre el valor observado de la variable dependiente y_j y el predicho por la función de regresión para x_i:

$$e_{ij} = y_j - f(x_i).$$

Residuos o errores predictivos en Y



Método de mínimos cuadrados

Una forma posible de obtener la función de regresión es mediante el método de *mínimos cuadrados* que consiste en calcular la función que haga mínima la suma de los cuadrados de los residuos

$$\sum e_{ij}^2.$$

En el caso de un modelo de regresión lineal $f(x) = a + bx$, como la recta depende de dos parámetros (el término independiente a y la pendiente b), la suma también dependerá de estos parámetros

$$\theta(a, b) = \sum e_{ij}^2 = \sum (y_j - f(x_i))^2 = \sum (y_j - a - bx_i)^2.$$

Así pues, todo se reduce a buscar los valores a y b que hacen mínima esta suma.

Cálculo de la recta de regresión

Método de mínimos cuadrados

Considerando la suma de los cuadrados de los residuos como una función de dos variables $\theta(a, b)$, se pueden calcular los valores de los parámetros del modelo que hacen mínima esta suma derivando e igualando a 0 las derivadas:

$$\frac{\partial \theta(a,b)}{\partial a} = \frac{\partial \sum (y_j - a - bx_i)^2}{\partial a} = 0$$
$$\frac{\partial \theta(a,b)}{\partial b} = \frac{\partial \sum (y_j - a - bx_i)^2}{\partial b} = 0$$

Tras resolver el sistema se obtienen los valores

$$a = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} \qquad b = \frac{s_{xy}}{s_x^2}$$

Estos valores hacen mínimos los residuos en Y y por tanto dan la recta de regresión.

Recta de regresión

Definición (Recta de regresión)

Dada una variable bidimensional (X, Y), la *recta de regresión* de Y sobre X es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

La recta de regresión de Y sobre X es la recta que hace mínimos los errores predictivos en Y, y por tanto es la recta que hará mejores predicciones de Y para cualquier valor de X.

Cálculo de la recta de regresión

Ejemplo con estaturas y pesos

Siguiendo con el ejemplo de las estaturas (X) y los pesos (Y) con los siguientes estadísticos:

$$\bar{x} = 174{,}67 \text{ cm} \qquad s_x^2 = 102{,}06 \text{ cm}^2 \qquad s_x = 10{,}1 \text{ cm}$$
$$\bar{y} = 69{,}67 \text{ Kg} \qquad s_y^2 = 164{,}42 \text{ Kg}^2 \qquad s_y = 12{,}82 \text{ Kg}$$
$$s_{xy} = 104{,}07 \text{ cm} \cdot \text{ Kg}$$

Entonces, la recta de regresión del peso sobre la estatura es:

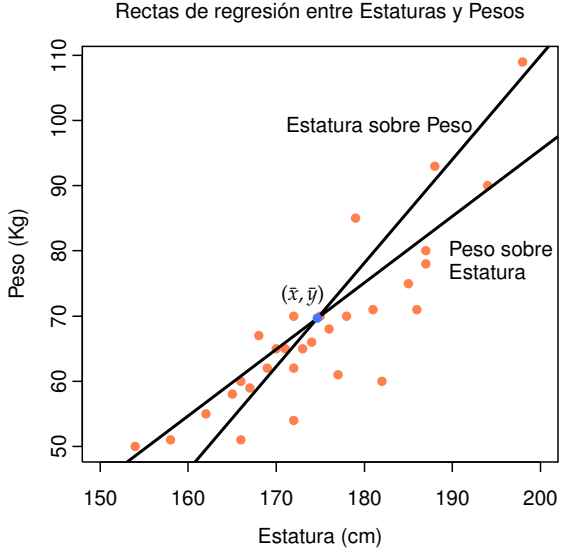
$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 69{,}67 + \frac{104{,}07}{102{,}06}(x - 174{,}67) = 1{,}02x - 108{,}49.$$

De igual modo, si en lugar de considerar el peso como variable dependiente, tomamos la estatura, entonces la recta de regresión de la estatura sobre el peso es:

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 174{,}67 + \frac{104{,}07}{164{,}42}(y - 69{,}67) = 0{,}63y + 130{,}78.$$

Rectas de regresión

Ejemplo de estaturas y pesos



Posición relativa de las rectas de regresión

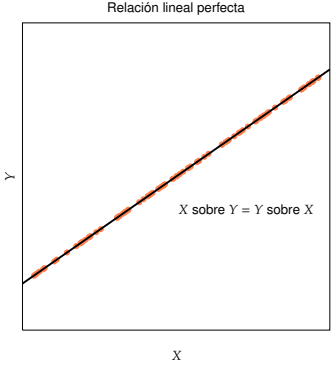
Las rectas de regresión siempre se cortan en el punto de medias $(\bar{x}, \bar{y})$.

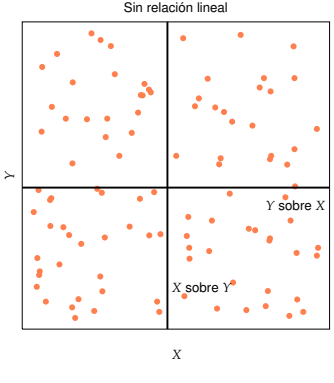
Si entre las variables la relación lineal es perfecta, entonces ambas rectas coinciden ya que sus residuos son nulos.

Si no hay relación lineal, entonces las ecuaciones de las rectas son

$$y = \bar{y}, \qquad x = \bar{x},$$

y se cortan perpendicularmente





Coeficiente de regresión

Definición (Coeficiente de regresión b_{yx})

Dada una variable bidimensional (X, Y), el *coeficiente de regresión* de la recta de regresión de Y sobre X es su pendiente,

$$b_{yx} = \frac{s_{xy}}{s_x^2}$$

El coeficiente de regresión siempre tiene el mismo signo que la covarianza y refleja el crecimiento de la recta de regresión, ya que da el número de unidades que aumenta o disminuye la variable dependiente por cada unidad que aumenta la variable independiente, según la recta de regresión.

En el ejemplo de las estaturas y los pesos, el coeficiente de regresión del peso sobre la estatura es $b_{yx} = 1{,}02$ Kg/cm, lo que indica que, según la recta de regresión del peso sobre la estatura, por cada cm más de estatura, la persona pesará $1{,}02$ Kg más.

Predicciones con las rectas de regresión

Ejemplo con estaturas y pesos

Las rectas de regresión, y en general cualquier modelo de regresión, suele utilizarse con fines predictivos.

¡Ojo! Para predecir una variable, esta siempre debe considerarse como dependiente en el modelo de regresión que se utilice.

Así, en el ejemplo de las estaturas y los pesos, si se quiere predecir el peso de una persona que mide 180 cm, se debe utilizar la recta de regresión del peso sobre la estatura:

$$y = 1{,}02 \cdot 180 - 108{,}49 = 75{,}11 \text{ Kg.}$$

Y si se quiere predecir la estatura de una persona que pesa 79 Kg, se debe utilizar la recta de regresión de la estatura sobre el peso:

$$x = 0{,}63 \cdot 79 + 130{,}78 = 180{,}55 \text{ cm.}$$

Ahora bien, ¿qué fiabilidad tienen estas predicciones?

Correlación

Una vez construido un modelo de regresión, para saber si se trata de un buen modelo predictivo, se tiene que analizar el grado de dependencia entre las variables según el tipo de dependencia planteada en el modelo. De ello se encarga la parte de la estadística conocida como **correlación**.

Para cada tipo de modelo existe el correspondiente tipo de correlación.

La correlación se basa en el estudio de los residuos. Cuanto menores sean éstos, más se ajustará la curva de regresión a los puntos, y más intensa será la correlación.

Varianza residual muestral

Una medida de la bondad del ajuste del modelo de regresión es la *varianza residual*.

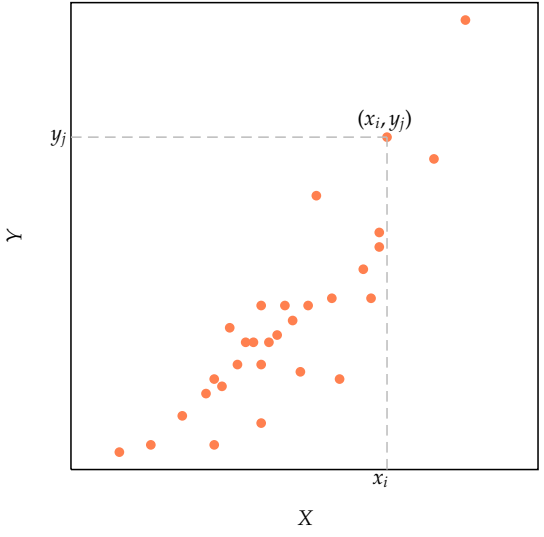
Definición (Varianza residual s^2_{ry})

Dado un modelo de regresión simple $y = f(x)$ de una variable bidimensional (X, Y), su *varianza residual muestral* es el promedio de los cuadrados de los residuos para los valores de la muestra,

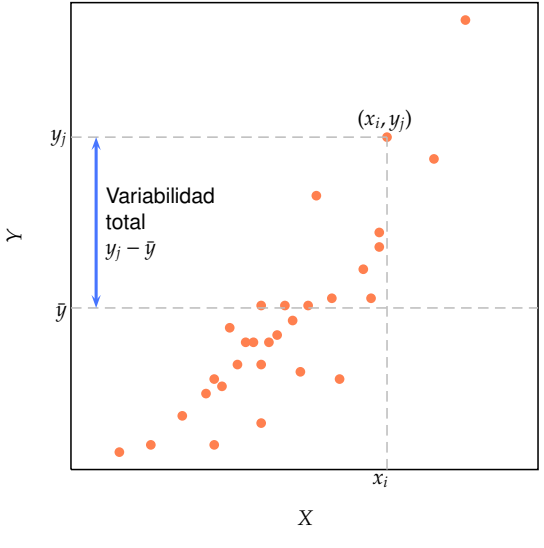
$$s^2_{ry} = \frac{\sum e^2_{ij} n_{ij}}{n} = \frac{\sum (y_j - f(x_i))^2 n_{ij}}{n}.$$

Cuanto más alejados estén los puntos de la curva de regresión, mayor será la varianza residual y menor la dependencia.

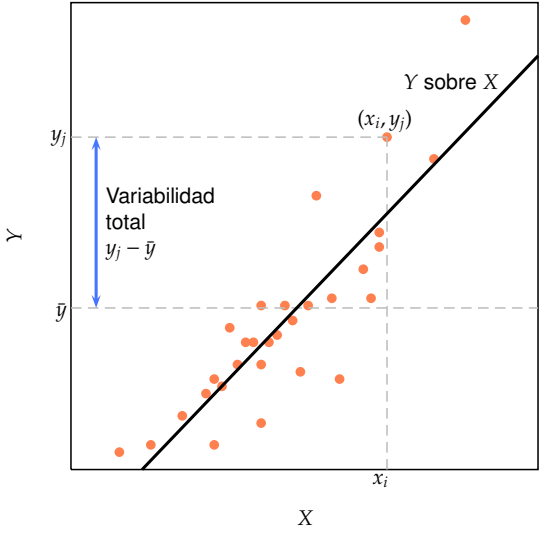
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



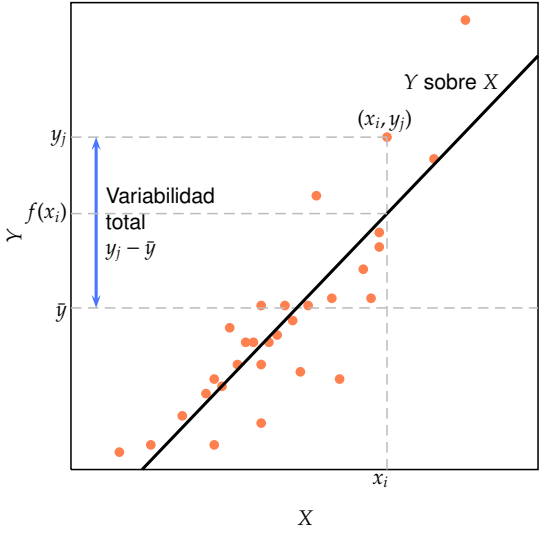
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



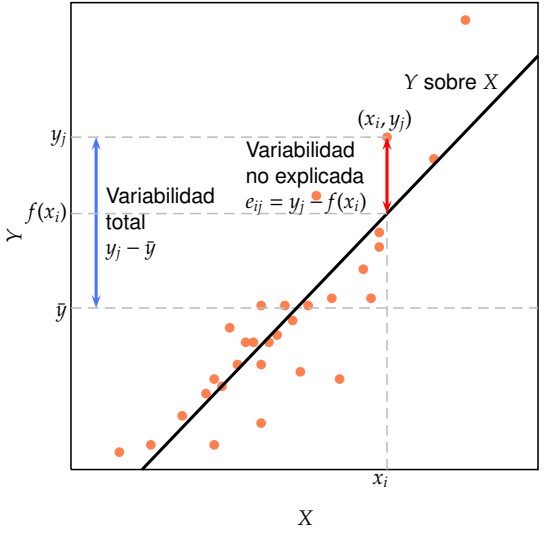
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



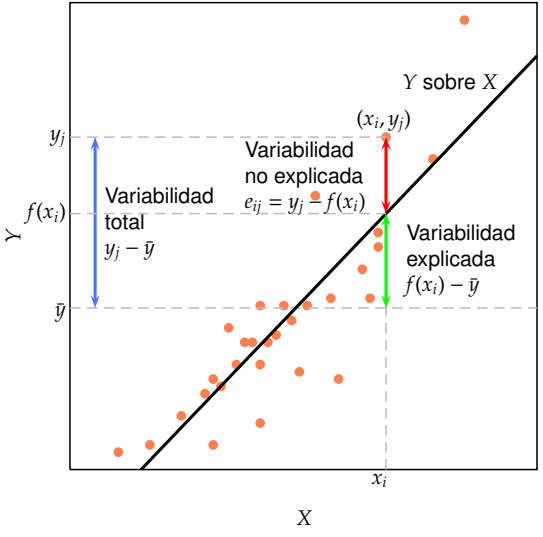
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



Descomposición de la variabilidad total: Variabilidad explicada y no explicada



Descomposición de la variabilidad total: Variabilidad explicada y no explicada



Coeficiente de determinación

A partir de la varianza residual se puede definir otro estadístico más sencillo de interpretar.

Definición (Coeficiente de determinación muestral)

Dado un modelo de regresión simple $y = f(x)$ de una variable bidimensional (X, Y), su *coeficiente de determinación muestral* es

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2}$$

El coeficiente de determinación mide la proporción de variabilidad de la variable dependiente explicada por el modelo de regresión, y por tanto,

$$0 \leq r^2 \leq 1$$

Cuanto mayor sea r^2, mejor explicará el modelo de regresión la relación entre las variables, en particular:

- Si $r^2 = 0$ entonces no existe relación del tipo planteado por el modelo.
- Si $r^2 = 1$ entonces la relación que plantea el modelo es perfecta.

Coeficiente de determinación lineal

En el caso de las rectas de regresión, la varianza residual vale

$$s_{ry}^2 = \sum e_{ij}^2 f_{ij} = \sum (y_j - f(x_i))^2 f_{ij} = \sum \left(y_j - \bar{y} - \frac{s_{xy}}{s_x^2}(x_i - \bar{x}) \right)^2 f_{ij} =$$

$$= \sum \left((y_j - \bar{y})^2 + \frac{s_{xy}^2}{s_x^4}(x_i - \bar{x})^2 - 2\frac{s_{xy}}{s_x^2}(x_i - \bar{x})(y_j - \bar{y}) \right) f_{ij} =$$

$$= \sum (y_j - \bar{y})^2 f_{ij} + \frac{s_{xy}^2}{s_x^4} \sum (x_i - \bar{x})^2 f_{ij} - 2\frac{s_{xy}}{s_x^2} \sum (x_i - \bar{x})(y_j - \bar{y}) f_{ij} =$$

$$= s_y^2 + \frac{s_{xy}^2}{s_x^4} s_x^2 - 2\frac{s_{xy}}{s_x^2} s_{xy} = s_y^2 - \frac{s_{xy}^2}{s_x^2}.$$

y, por tanto, el coeficiente de determinación lineal vale

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2} = 1 - \frac{s_y^2 - \frac{s_{xy}^2}{s_x^2}}{s_y^2} = 1 - 1 + \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Cálculo del coeficiente de determinación lineal

Ejemplo de estaturas y pesos

En el ejemplo de las estaturas y pesos se tenía

$$\bar{x} = 174{,}67 \text{ cm} \qquad s_x^2 = 102{,}06 \text{ cm}^2$$
$$\bar{y} = 69{,}67 \text{ Kg} \qquad s_y^2 = 164{,}42 \text{ Kg}^2$$
$$s_{xy} = 104{,}07 \text{ cm} \cdot \text{ Kg}$$

De modo que el coeficiente de determinación lineal vale

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(104{,}07 \text{ cm} \cdot Kg)^2}{102{,}06 \text{ cm}^2 \cdot 164{,}42 \text{ Kg}^2} = 0{,}65.$$

Esto indica que la recta de regresión del peso sobre la estatura explica el 65 % de la variabilidad del peso, y de igual modo, la recta de regresión de la estatura sobre el peso explica el 65 % de la variabilidad de la estatura.

Coeficiente de correlación lineal

Definición (Coeficiente de correlación lineal)

Dada una variable bidimensional (X, Y), el *coeficiente de correlación lineal muestral* es la raíz cuadrada de su coeficiente de determinación lineal, con signo el de la covarianza

$$r = \sqrt{r^2} = \frac{s_{xy}}{s_x s_y}$$

.

Como r^2 toma valores entre 0 y 1, el coeficiente de correlación lineal tomará valores entre -1 y 1:

$$-1 \leq r \leq 1$$

El coeficiente de correlación lineal también mide el grado de dependencia lineal:

- Si $r = 0$ entonces no existe relación lineal.
- Si $r = 1$ entonces existe una relación lineal creciente perfecta.
- Si $r = -1$ entonces existe una relación lineal decrececiente perfecta.

Coeficiente de correlación lineal

Ejemplo

En el ejemplo de las estaturas y los pesos, el coeficiente de correlación lineal vale

$$r = \frac{s_{xy}}{s_x s_y} = \frac{104{,}07 \text{ cm}\cdot Kg}{10{,}1 \text{ cm} \cdot 12{,}82 \text{ Kg}} = +0{,}8.$$

lo que indica que la relación lineal entre el peso y la estatura es fuerte, y además creciente.

Fiabilidad de las predicciones de un modelo de regresión

Aunque el coeficiente de determinación o el de correlación hablan de la bondad de un modelo de regresión, no es lo único que hay que tener en cuenta a la hora de hacer predicciones.

La fiabilidad de las predicciones que hagamos con un modelo de regresión depende de varias cosas:

- El coeficiente de determinación: Cuanto mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.
- La variabilidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.

Además, hay que tener en cuenta que un modelo de regresión es válido únicamente para el rango de valores observados en la muestra. Fuera de ese rango no hay información del tipo de relación entre las variables, por lo que no deben hacerse predicciones para valores lejos de los observados en la muestra.