CORRELACIÓN Y REGRESIÓN LINEAL

Autor: Clara Laguna

4.1 INTRODUCCIÓN

Después de estudiar cómo hay que organizar, representar gráficamente y analizar un conjunto de datos a partir de algunos parámetros, nos proponemos estudiar las relaciones entre variables. Por ejemplo, podemos determinar si existe alguna relación entre la variables peso y altura de un conjunto de personas.

Pretendemos estudiar una situación muy usual y por tanto de gran interés en la práctica: Si Y es una variable definida sobre la misma población que X, ¿será posible determinar si existe alguna relación entre las modalidades de X y de Y?

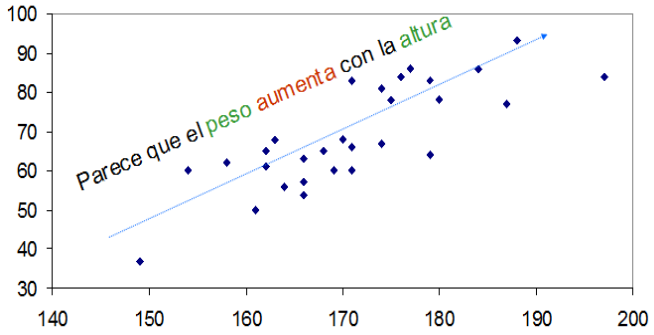


Figura 4.1. Diagrama de dispersión o nube de puntos

En este tema se presentan el **coeficiente de correlación** y la **regresión lineal simple** como las dos técnicas estadísticas más utilizadas para investigar la **relación entre dos variables continuas X e Y**.

Gráficamente el **diagrama de dispersión** o nube de puntos permite obtener información sobre el tipo de relación existente entre X e Y, además de ayudarnos a detectar posibles valores atípicos o extremos.
En el diagrama de dispersión de la figura 4.1 tenemos representadas las alturas y los pesos de 30 individuos. Vemos como a medida que aumenta la variable X=”altura” va aumentando la variable Y=”peso”.

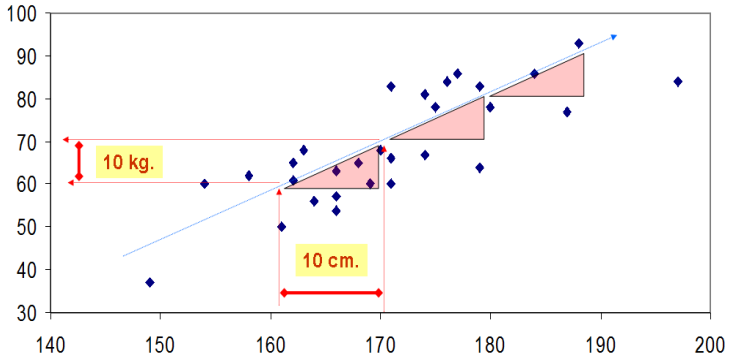


Figura 4.2.

Si nos fijamos en la figura 4.2 aparentemente el peso aumenta 10Kg por cada 10 cm de altura... es decir, el peso aumenta en una unidad por cada unidad de altura.

El diagrama de dispersión se obtiene representando cada observación (x_i, y_i) como un punto en el plano cartesiano XY.

Las técnicas de correlación y las de regresión están estrechamente relacionadas, aunque obedecen a estrategias de análisis un tanto diferentes.

Por un lado, el *coeficiente de correlación* determina el grado de asociación lineal entre X e Y, sin establecer a priori ninguna direccionalidad en la relación entre ambas variables. Por el contrario, la *regresión lineal simple* permite cuantificar el cambio en el nivel medio de la variable Y conforme cambia la variable X, asumiendo implícitamente que **X es la variable explicativa o independiente** e **Y es la variable respuesta o dependiente**.

4.2 CORRELACIÓN

La finalidad de la correlación es examinar la dirección y la fuerza de la asociación entre dos variables cuantitativas. Así conoceremos la intensidad de la relación entre ellas y *si, al aumentar el valor de una variable, aumenta o disminuye el valor de la otra variable*.

Para valorar la asociación entre dos variables, la primera aproximación suele hacerse mediante un diagrama de dispersión.

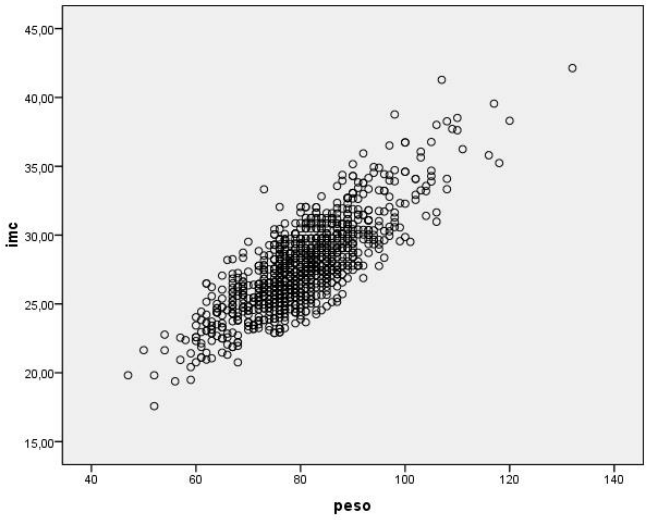


Figura 4.3.

En el diagrama de dispersión de la figura 4.3 parece existir una relación lineal entre el peso y el índice de masa corporal de los pacientes. Además, si nos fijamos parece que existe un dato atípico que se aleja de la nube de puntos.

Con la nube de puntos podemos apreciar si existe o no una tendencia entre las dos variables, pero si queremos cuantificar esta asociación debemos calcular un coeficiente de correlación.

Hay dos **coeficientes de correlación** que se usan frecuentemente: el de **Pearson** (paramétrico) y el de **Spearman** (no paramétrico, se utiliza en aquellos casos donde las variables examinadas no cumplen criterios de normalidad o cuando las variables son ordinales).

El coeficiente de correlación de Pearson evalúa específicamente la adecuación a la recta lineal que defina la relación entre dos variables cuantitativas. El coeficiente no paramétrico de Spearman mide cualquier tipo de asociación, no necesariamente lineal.

*Si se desea **medir o cuantificar el grado de asociación** entre dos variables cuantitativas se debe calcular un **coeficiente de correlación**.*

4.2.1 Coeficiente de Correlación lineal de Pearson

El estimador muestral más utilizado para evaluar la asociación lineal entre dos variables X e Y es el **coeficiente de correlación de Pearson (r).** Se trata de un índice que mide si los puntos tienen tendencia a disponerse en una línea recta. Puede tomar **valores entre -1 y +1**.
Es un método estadístico paramétrico, ya que utiliza la media, la varianza,…y por tanto, requiere criterios de normalidad para las variables analizadas.

Se define como la covarianza muestral entre X e Y dividida por el producto de las desviaciones típicas de cada variable:

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

La expresión matemática para el coeficiente de correlación de Pearson parece compleja, pero esconde un planteamiento que en el fondo, es sencillo: "r" estará próximo a 1 (en valor absoluto) cuando las dos variables X e Y estén intensamente relacionadas, es decir, al aumentar una aumenta otra y viceversa. A este concepto de variación al unísono se le llama *covarianza*.

Covarianza

El numerador del coeficiente de correlación es la **covarianza muestral S_{XY}** entre X e Y, que nos indica ***si la posible relación entre dos variables es directa o inversa***. Es una medida que nos habla de la variabilidad conjunta de dos variables cuantitativas.

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

Así, si valores altos (o bajos) de X tienden a asociarse con valores altos (o bajos) de Y, el producto de las desviaciones tenderá a ser positivo y la covarianza será positiva. Por el contrario, si valores altos de una variable se relacionan con valores bajos de la otra variable, el producto de las desviaciones tenderá a ser negativo y la covarianza será negativa.

De tal modo que:

- Si S_{XY} >0 las dos variables crecen o decrecen a la vez (nube de puntos creciente).
- Si S_{XY} <0 cuando una variable crece, la otra tiene tendencia a decrecer (nube de puntos decreciente).
- Si los puntos se reparten con igual densidad alrededor del centro de gravedad $(\bar{x}, \bar{y})$, S_{XY} =0 (no hay relación lineal).

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.

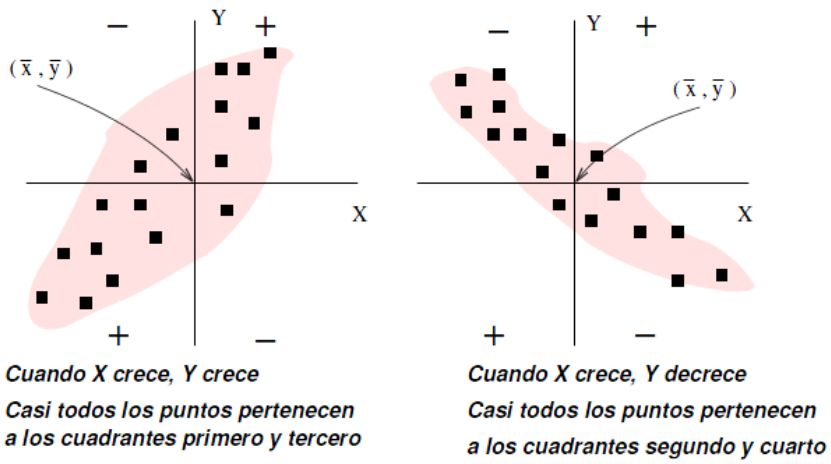


Figura 4.4. Interpretación geométrica de S_{XY}

Resulta complicado determinar el grado de asociación lineal entre dos variables a partir de la magnitud de la covarianza, ya que ésta depende de las unidades de medida de las variables.

Volviendo al **coeficiente** de **correlación lineal r**, veamos qué **propiedades** tiene:

- Carece de unidades de medida (adimensional).
- Sólo toma valores comprendidos entre [-1,1].
- Cuando |r| esté **próximo a uno**, r= +1 (recta lineal creciente de izquierda a derecha) o r= -1 (recta lineal decreciente), se tiene que existe una **relación lineal muy fuerte** entre las variables.
- Cuando **r≈0**, puede afirmarse que **no existe relación lineal** entre ambas variables. Se dice en este caso que las variables son **incorreladas**.

Para entenderlo mejor, veamos los siguientes diagramas de dispersión:

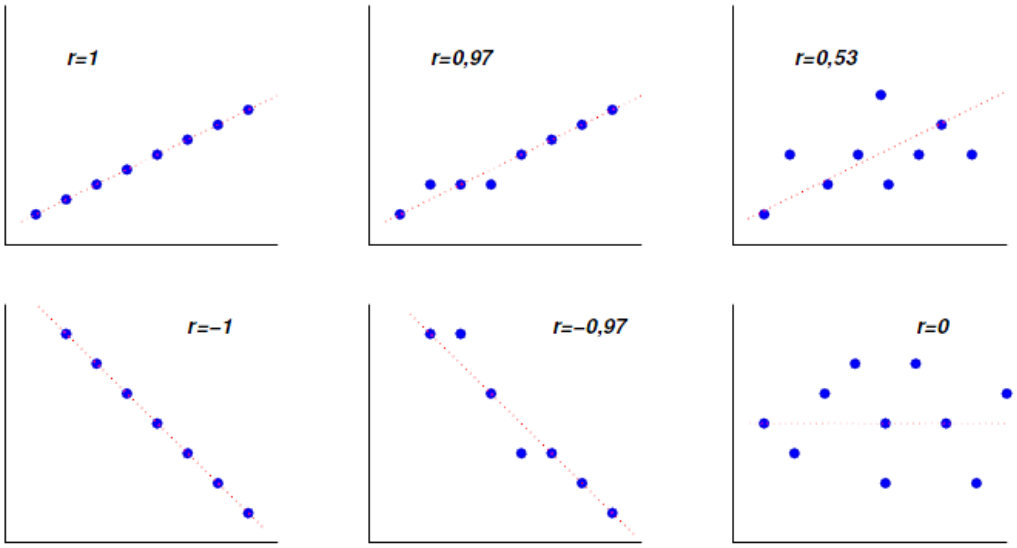


Figura 4.5.

En la figura 4.5 vemos que r = ±1 es lo mismo que decir que las observaciones de ambas variables están perfectamente alineadas. **El signo de r, es el mismo que el de S_{XY}, por tanto nos indica el crecimiento o decrecimiento de la recta**. La relación lineal es tanto más perfecta cuanto r está cercano a ±1.

En la correlación ***no se distingue la variable dependiente de la independiente***, la correlación de X con respecto a Y es la misma que la correlación de Y con respecto a X.

Aunque la interpretación de la magnitud del coeficiente de correlación depende del contexto particular de aplicación, en términos generales se considera que una correlación es baja por debajo de 0,30 en valor absoluto, que existe una asociación moderada entre 0,30 y 0,70, y alta por encima de 0,70.

Condiciones de aplicación de la correlación:

- *Variables cuantitativas*: Ambas variables examinadas han de ser cuantitativas. Para variables ordinales se puede usar el coeficiente de Spearman.
- *Normalidad*: La normalidad de ambas variables es un requisito en el caso del coeficiente de correlación de Pearson, pero no en el de Spearman.
- *Independencia*: Las observaciones han de ser independientes, es decir, sólo hay una observación de cada variable para cada individuo. No tendría sentido, aplicar la correlación en un estudio que relacionase la ingesta diaria de sal y la tensión intraocular si se tomaran mediciones en ambos ojos de cada individuo. En este caso hay dos observaciones por paciente que están

autocorrelacionadas, no son independientes; habría que considerar *N* como el número de pacientes y no el de ojos.

Ejemplo 4.1

En la Figura 4.6 se presenta el diagrama de dispersión entre el índice de masa corporal, medida de obesidad que se obtiene de dividir el peso en kilogramos por la altura en metros al cuadrado, y el colesterol HDL en un estudio realizado a 533 individuos.

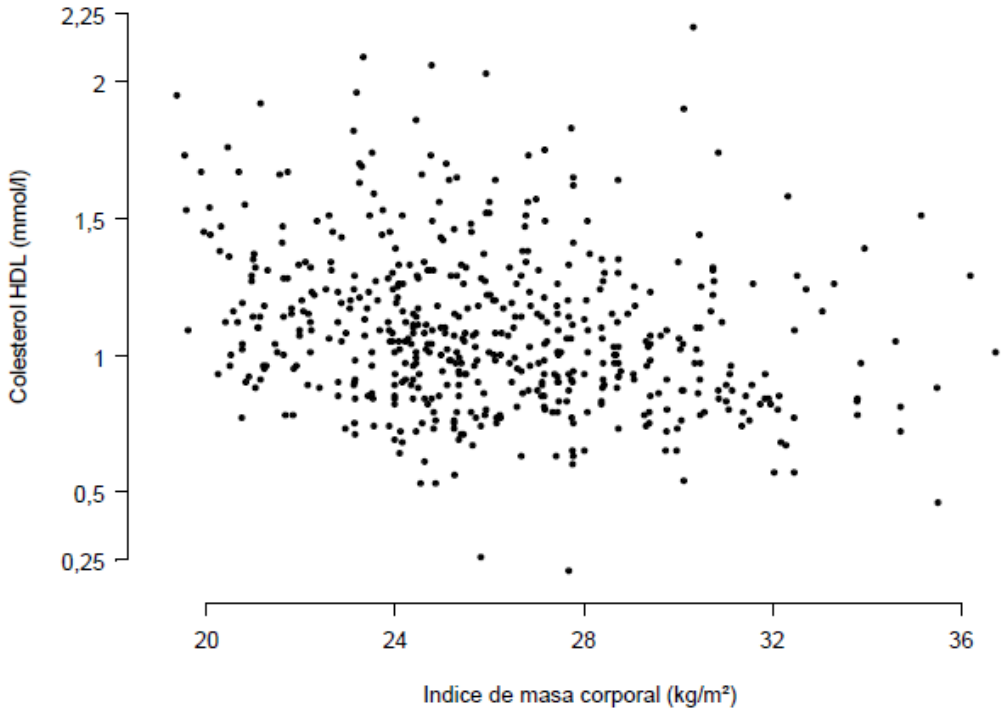


Figura 4.6.

A simple vista, se aprecia un cierto grado de dependencia lineal negativa entre ambas variables; esto es, el colesterol HDL tiende a decrecer conforme aumenta el índice de masa corporal. Esta apreciación visual se confirma mediante el cálculo del coeficiente de correlación muestral de Pearson que indica una asociación lineal negativa moderada entre el índice de masa corporal y el colesterol HDL.

$$r = \frac{\frac{1}{532}\sum_{i=1}^{533}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{-0{,}285}{3{,}50 \cdot 0{,}295} = -0{,}276$$

4.3 REGRESIÓN LINEAL SIMPLE

La **regresión** está dirigida a describir como es la ***relación entre dos variables X e Y***, de tal manera que incluso se pueden ***hacer predicciones*** sobre los valores de la variable Y, a partir de los de X. Cuando la asociación entre ambas variables es fuerte,

la regresión nos ofrece un modelo estadístico que puede alcanzar finalidades predictivas.

La *regresión* supone que hay una *variable fija*, controlada por el investigador (es la variable independiente o predictora), y otra que *no está controlada* (variable respuesta o dependiente). La *correlación* supone que ninguna es fija: las dos variables están fuera del control de investigador.

La regresión es su forma más sencilla se llama **regresión lineal simple**. Se trata de una técnica estadística que analiza la relación entre **dos variables cuantitativas**, tratando de verificar si dicha ***relación es lineal***.
Si tenemos dos variables hablamos de *regresión simple*, si hay más de dos variables *regresión múltiple*.

Su objetivo es **explicar el comportamiento de una variable Y**, que denominaremos variable explicada (o **dependiente** o endógena), **a partir de otra variable X**, que llamaremos variable explicativa (o **independiente** o exógena).

Una vez que hemos hecho el diagrama de dispersión y después de observar una posible relación lineal entre las dos variables, nos proponemos encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos. Esta recta se denomina **recta de regresión**.

Si sobre un grupo de personas observamos los valores que toman las variables X = altura medida en centímetros, Y=altura medida en metros, sabemos que la relación que hay entre ambas es: Y = X/100.

Obtener esta relación es menos evidente cuando lo que medimos sobre el mismo grupo de personas es X = altura medida en centímetros e Y= peso en kilogramos. La razón es que no es cierto que conocida la altura x_i de un individuo, podamos determinar de modo exacto su peso y_i (dos personas que miden 1, 70m pueden tener pesos de 60 y 65 kilos). Sin embargo, alguna relación entre ellas debe existir, ya que parece más probable pensar que un individuo de 2m pese más que otro que mida 1,20m.

A la deducción, a partir de una serie de datos, de este tipo de relaciones entre variables, es lo que denominamos regresión.

Mediante las técnicas de regresión inventamos una variable $\hat{Y}$ como función de otra variable X (o viceversa).
El criterio para construir esta función es que la **diferencia entre Y e $\hat{Y}$**, denominada **error o residuo**, sea **pequeña**.

$$\hat{Y} = f(X), \qquad Y - \hat{Y} = \text{error},$$

Los residuos o errores $\mathbf{e_i}$ son la **diferencia entre los valores observados** (verdadero valor de Y) **y los valores pronosticados** por el modelo: $e_i = Y - \hat{Y}$. Recogen la *parte de la variable Y que no es explicada por el modelo de regresión.*

A partir de la definición de residuo, podemos escribir **Y = f(X) + error**.

El término que hemos denominado *error* debe ser tan pequeño como sea posible. El objetivo será **buscar la función** (modelo de regresión) **$\hat{Y}$= f(X)** que lo **minimice**.

4.3.1 Ajuste de una recta por mínimos cuadrados

La **regresión lineal** consiste en encontrar (aproximar) los valores de una variable a partir de los de otra, usando una relación funcional de tipo lineal, es decir, buscamos cantidades **a** (ordenada en el origen) y **b** (pendiente de la recta lineal) tales que se pueda escribir $\hat{Y} = a + bX$, con el menor error posible entre Ŷ e Y.

Para cada valor observado de la variable independiente x_i podemos considerar dos valores de la variable dependiente, el observado y_i y el estimado a partir de la ecuación de la recta, $\hat{y}_i = a + bx_i$

Para cada observación podemos definir el **error** o residuo como la **distancia vertical** entre el punto (x_i, y_i) y la recta, es decir: $y_i - (a + bx_i)$

Por cada recta que consideremos, tendremos una colección diferente de residuos.

Se trata de **buscar la recta que dé lugar a los residuos más pequeños,** es decir la recta que hace mínima la suma de cuadrados de las distancias verticales entre cada punto y la recta, de tal manera que se **minimice la suma de los errores al cuadrado.**

$$SC\,\mathrm{Re}\,s = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Para determinar la recta de regresión, utilizaremos el **método de los mínimos cuadrados**[1].

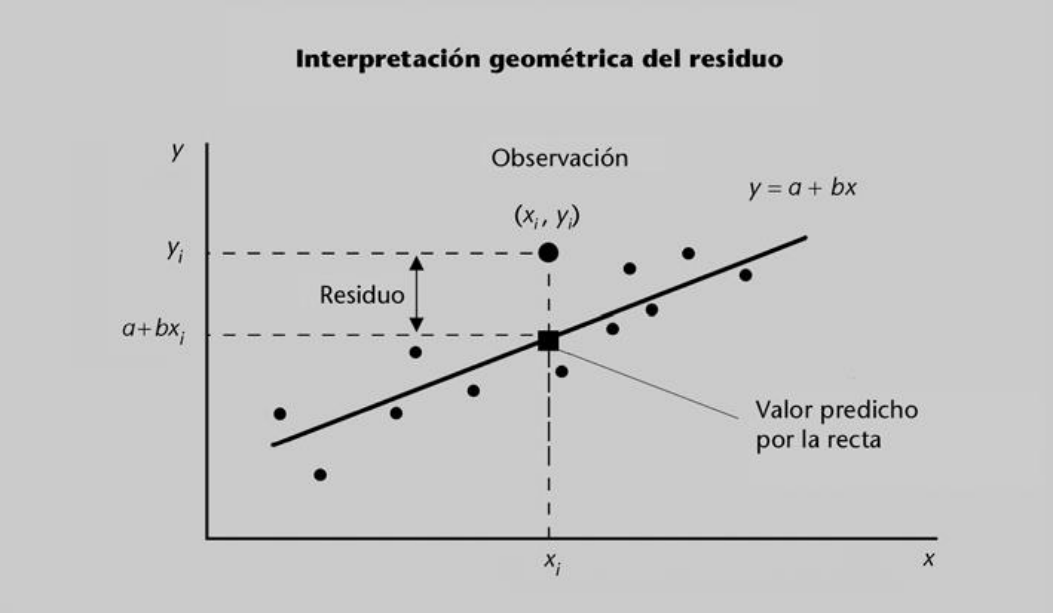


Figura 4.7.

[1] No entramos en el desarrollo matemático del método

Las cantidades a y b que minimizan dicho error son los llamados **coeficientes de regresión**:

$$b = \frac{S_{XY}}{S_X^2} \qquad a = \bar{y} - b\bar{x}$$

La cantidad b se denomina *"coeficiente de regresión de Y sobre X"*.

Interpretación de la ordenada en el origen a:
Este parámetro representa la estimación del valor de Y cuando X es igual a cero.

Interpretación de la pendiente de la recta b:

El coeficiente de regresión es muy importante, porque **mide el cambio de la variable Y por cada unidad de cambio de X**. Este parámetro nos informa de cómo están relacionadas las dos variables en el sentido de que nos indica en qué cantidad (y si es positiva o negativa) varían los valores de *Y* cuando varían los valores de la *X* en una unidad. De hecho el coeficiente de regresión *b* y el coeficiente de correlación *r* siempre tendrán el mismo signo.

- Si b > 0, cada aumento de X se corresponde con un aumento de Y;
- Si b < 0, Y decrece a medida que aumenta X.

*El **método de los mínimos cuadrados** consiste en buscar los valores de los parámetros a y b de manera que la suma de los cuadrados de los residuos sea mínima. Esta recta es la **recta de regresión** por mínimos cuadrados.*

Ejemplo 4.2

En el estudio de la relación entre el índice de masa corporal y el colesterol HDL, resulta natural considerar el índice de masa corporal como variable independiente X y el colesterol HDL como variable dependiente Y. El objetivo es, estimar los cambios en el nivel medio del colesterol HDL conforme aumenta el índice de masa corporal utilizando un modelo de regresión lineal simple.

Las estimaciones de la pendiente y la constante de la recta de regresión por el método de mínimos cuadrados son:

$$b = \frac{S_{XY}}{S_X^2} = \text{-}0{,}023 \qquad a = \bar{y} - b\bar{x} = 1{,}69$$

La constante $a = 1{,}69$ *mmol/l es una estimación del valor esperado del colesterol HDL para un sujeto con un imc igual a 0 kg/m^2, extrapolación que carece de sentido biológico.*

*La **pendiente b= -0,023** estima que, por cada **incremento** de 1kg/m^2 en el índice de masa corporal, el nivel medio de colesterol HDL disminuye en 0,023 mmol/l.*

*La **recta de regresión** (figura 4.8) estimada del colesterol HDL sobre el índice de masa corporal es:*

$$\hat{y} = 1{,}69 - 0{,}023x$$

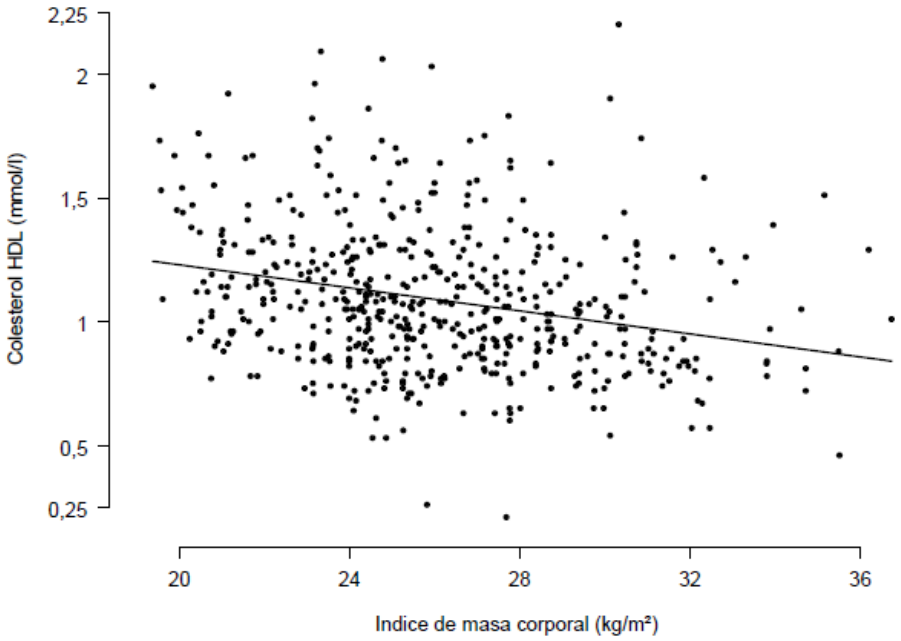


Figura 4.8.

*Esta recta de regresión puede utilizarse para **estimar o predecir** el valor esperado del colesterol HDL en función del índice de masa corporal.*
Por ejemplo, para un índice de masa corporal de 25 kg/m^2, el modelo estima un nivel medio de colesterol HDL de

$$\hat{y}(25) = 1{,}69 - 0{,}023 \cdot 25 = 1{,}11 mmol/l$$

Interpolación y extrapolación:

Como acabamos de ver, uno de los objetivos más importantes de la regresión es la aplicación del modelo para el **pronóstico del valor de la variable dependiente (Y)** para un valor de la variable independiente (X) no observado en la muestra.

Ejemplo 4.3

A partir de la recta de regresión que relaciona los pesos y las alturas de una muestra de 10 personas, podemos estar interesados en conocer el peso de una persona de altura de 1,60 m

$$\hat{y} = -96{,}11 + 0{,}979x = -96{,}11 + 0{,}979 \cdot 160 = 60{,}53$$

para un valor de X de 160 cm, tenemos un valor estimado para la Y de 60,53 kg.

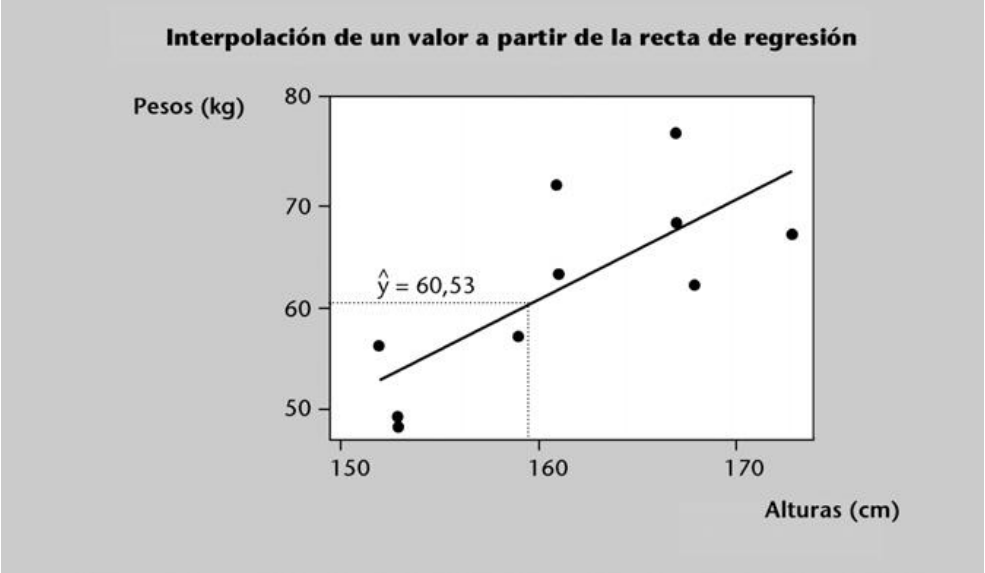


Figura 4.9.

Un aspecto importante a la hora de aplicar el modelo de regresión obtenido es el riesgo de la extrapolación. Es decir, cuando queremos conocer el valor que presentará la variable *Y* para un determinado valor de *X* que se encuentre fuera del intervalo de valores que toma la muestra. Entonces tenemos que ir con mucho cuidado:

a. Hemos determinado el modelo con la información contenida en la muestra, de manera que no hemos tenido ninguna información del comportamiento de la variable *Y* para valores de *X* de fuera del rango de la muestra.

b. Es posible que no tenga sentido la extrapolación que queremos hacer. Antes de utilizar el modelo de regresión, debemos preguntarnos por lo que estamos haciendo. Por ejemplo, no tendría ningún sentido utilizar el modelo de regresión para calcular el peso de personas de diez centímetros o tres metros de altura. El modelo nos dará un resultado numérico que, en todo caso, hay que interpretar.

Supuestos del modelo de regresión:

Linealidad: El valor esperado de la variable dependiente Y es una función lineal de la variable explicativa X, de tal forma que cambios de magnitud constante a distintos niveles de X se asocian con un mismo cambio en el valor medio de Y.

Homogeneidad de la varianza: La varianza de la variable dependiente Y es la misma para cualquier valor de la variable explicativa X; es decir, a diferencia de la media, la varianza de Y no está relacionada con X.

Normalidad: Para un valor fijo de la variable explicativa X, la variable dependiente Y sigue una distribución normal.

Independencia: Cada observación de la variable Y debe ser independiente de las demás.

4.3.2 Bondad de un ajuste

La recta de regresión por mínimos cuadrados minimiza la suma de los cuadrados de los residuos. Ahora **nos preguntamos si este ajuste es lo bastante bueno**.

Mirando si en el diagrama de dispersión los puntos experimentales quedan muy cerca de la recta de regresión obtenida, podemos tener una idea de si la recta se ajusta o no a los datos, pero nos hace falta un valor numérico que nos ayude a precisarlo.

El coeficiente de determinación, R^2

Queremos evaluar en qué grado el modelo de regresión lineal que hemos encontrado a partir de un conjunto de observaciones explica las variaciones que se producen en la variable dependiente de éstas.

La medida más importante de la bondad del ajuste es el **coeficiente de determinación R^2**.
Este coeficiente nos indica el grado de ajuste de la recta de regresión a los valores de la muestra, y se define como el **porcentaje de la variabilidad total de la variable dependiente Y que es explicada por la recta de regresión**.

Cuanto *menos dispersos sean los residuos* (recordad que lo residuos o errores son la diferencia entre los valores observados y los valores estimados por la recta de regresión), *mejor será la bondad del ajuste*[2].

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

Las características de este coeficiente son:

- R^2 es una cantidad adimensional que sólo puede tomar valores en [0, 1]
- Cuando un ajuste es bueno, R^2 será cercano a uno (mayor será la fuerza de asociación entre ambas variables)
- Cuando un ajuste es malo, R^2 será cercano a cero (la recta no explica nada, no existe asociación entre X e Y)

[2] Para entender mejor cómo se mide la bondad de un ajuste de un modelo de regresión, os aconsejo que veáis con detenimiento la presentación disponible en material de apoyo

Puesto que R^2 nos explica la proporción de variabilidad de los datos que queda explicada por el modelo de regresión, cuanto más cercano a la unidad esté, mejor es el ajuste.

Volviendo al ejemplo 4.3 de las alturas y los pesos, hemos obtenido un coeficiente de determinación R^2 = 0,5617 que nos informa de que la altura sólo nos explica el 56,17% de la variabilidad del peso.

Relación entre R^2 y r

Es muy importante tener clara la diferencia entre el coeficiente de correlación y el coeficiente de determinación:

- R^2: mide la proporción de variación de la variable dependiente explicada por la variable independiente.
- r: mide el grado de asociación entre las dos variables.

No obstante, en la **regresión lineal simple** tenemos que $\mathbf{R^2 = r^2}$.

Esta relación nos ayuda a comprender por qué antes considerábamos que un valor de r = 0,5 era débil. Este valor representará un R^2 = 0,25, es decir, el modelo de regresión sólo nos explica un 25% de la variabilidad total de las observaciones.

A diferencia de R^2 que siempre es positivo, *r* puede ser positivo o negativo (tendrá el mismo signo que la pendiente de la recta que hemos llamado *b*).
Por tanto, es importante tener presente que *r* nos da más información que R^2. El signo de *r* nos informa de si la relación es positiva o negativa. Así pues, con el valor de *r* siempre podremos calcular el valor de R^2, pero al revés siempre nos quedará indeterminado el valor del signo a menos que conozcamos la pendiente de la recta.

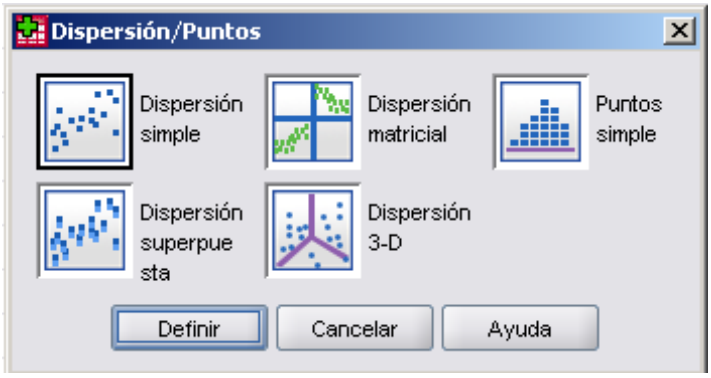
Por ejemplo, dado un R^2 = 0,81, si sabemos que la pendiente de la recta de regresión es negativa, entonces podremos afirmar que el coeficiente de correlación será *r* = -0,9.

Una correlación puede parecer impresionante, por ejemplo r = 0,7, y sin embargo el modelo lineal explicaría menos del 50% de lo observado (R^2=0,49).

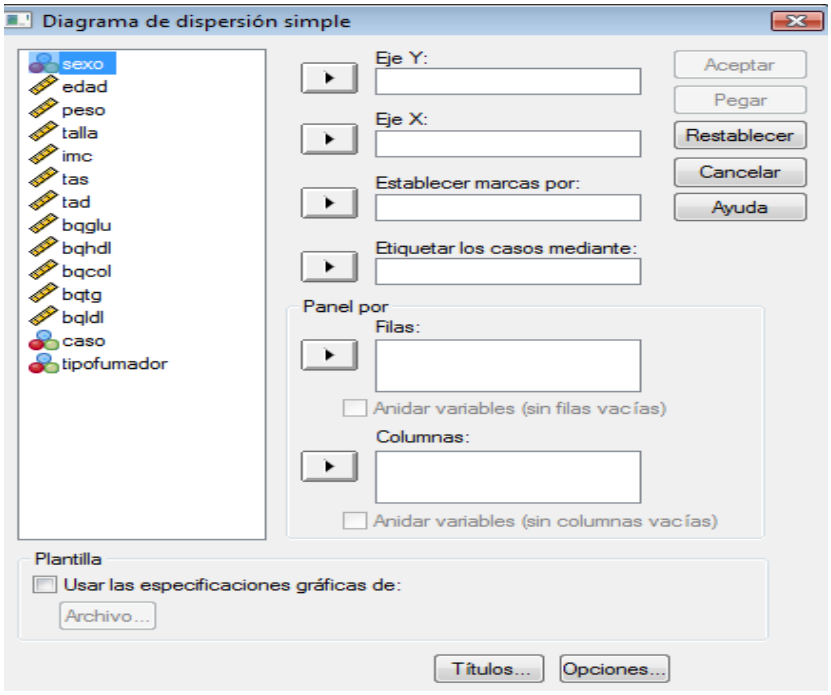
REGRESIÓN LINEAL SIMPLE EN SPSS

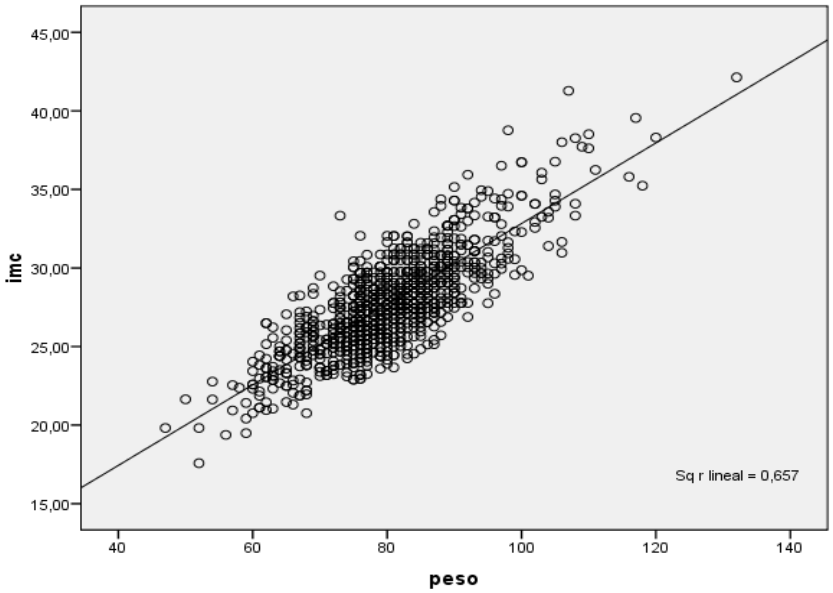
El primer paso debe ser siempre pedir a SPSS un gráfico de dispersión para apreciar visualmente si se puede asumir un modelo lineal entre ambas variables.
Como hemos visto el diagrama de dispersión o nube de puntos permite obtener información sobre el tipo de relación existente entre dos variables y sirve para detectar posibles datos atípicos o valores extremos.

Para representar nubes de puntos, se selecciona en la barra del menú principal **GRÁFICOS>DISPERSIÓN**.



OPCIÓN DISPERSIÓN SIMPLE: Seleccionando esta opción podremos representar la nube de puntos para **un par de variables**, distinguiendo (si queremos) los puntos según los valores de una tercera variable.

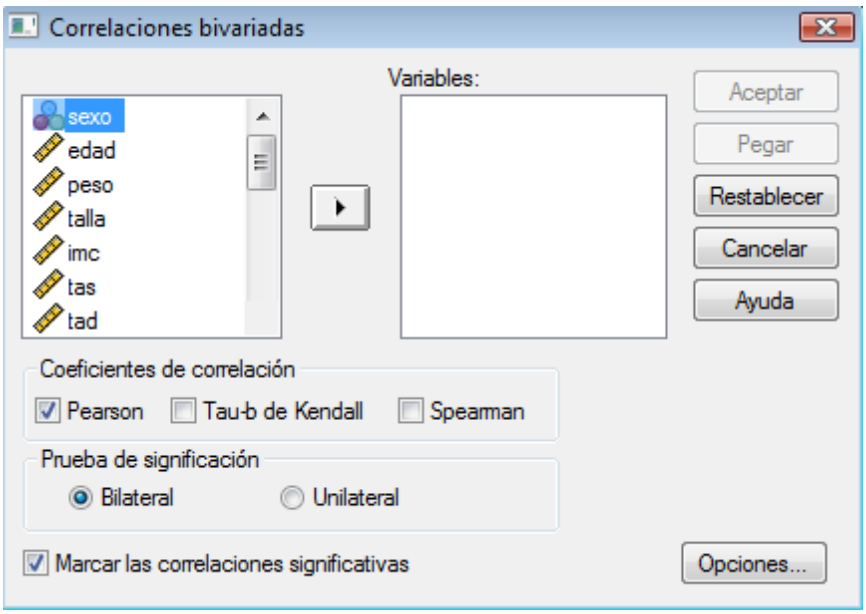




*Interpretación: Aparentemente parece existir una **relación lineal entre el peso y el índice de masa corporal de los pacientes**. Si nos fijamos parece que existe un dato atípico.*

Una vez dibujada la nube de puntos es posible **representar la recta de regresión**, la parábola o la función cúbica que mejor se ajusta y **obtener el valor del coeficiente de determinación que mide la bondad del ajuste**.

Para cuantificar el grado de relación LINEAL entre dos variables con mayor precisión de la que nos permite el diagrama de dispersión utilizamos los **Coeficientes de Correlación**.
Seleccionando en el menú **ANALIZAR>CORRELACIONES>BIVARIADAS** se obtiene:



Utilizaremos el coeficiente de correlación lineal de Pearson (r) entre dos variables cuantitativas X e Y cuando ambas sean Normales.

Y los coeficientes de correlación No Paramétricos: Rho de Spearman y Tau-b de Kendall. También se pueden utilizar con variables ordinales.

Correlaciones

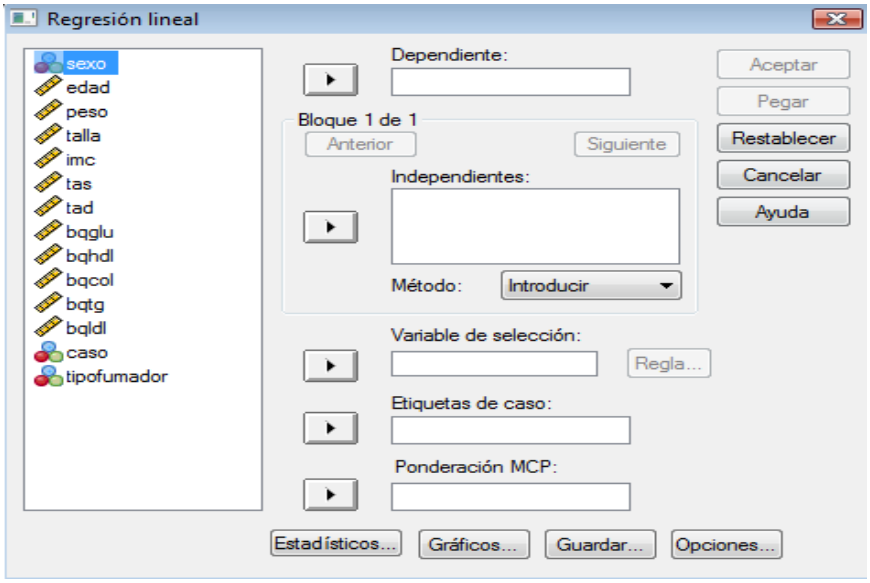
			peso	imc	tas	tad	bqcol	bqldl
Rho de Spearman	peso	Coeficiente de correlación	1,000	,759**	,163**	,230**	,044	,022
		Sig. (bilateral)	.	,000	,000	,000	,155	,481
		N	1024	1024	1024	1024	1023	1007
	imc	Coeficiente de correlación	,759**	1,000	,270**	,313**	,060	,015
		Sig. (bilateral)	,000	.	,000	,000	,055	,633
		N	1024	1024	1024	1024	1023	1007
	tas	Coeficiente de correlación	,163**	,270**	1,000	,744**	,089**	,065*
		Sig. (bilateral)	,000	,000	.	,000	,004	,039
		N	1024	1024	1024	1024	1023	1007
	tad	Coeficiente de correlación	,230**	,313**	,744**	1,000	,099**	,081*
		Sig. (bilateral)	,000	,000	,000	.	,002	,010
		N	1024	1024	1024	1024	1023	1007
	bqcol	Coeficiente de correlación	,044	,060	,089**	,099**	1,000	,930**
		Sig. (bilateral)	,155	,055	,004	,002	.	,000
		N	1023	1023	1023	1023	1023	1007
	bqldl	Coeficiente de correlación	,022	,015	,065*	,081*	,930**	1,000
		Sig. (bilateral)	,481	,633	,039	,010	,000	.
		N	1007	1007	1007	1007	1007	1007

**. La correlación es significativa al nivel 0,01 (bilateral).

*. La correlación es significativa al nivel 0,05 (bilateral).

<u>Interpretación</u>: Observa que existe una fuerte correlación positiva entre el PESO y el IMC, entre las variables TAS y TAD y entre el valor total del colesterol BQCOL y BQLDL. Entre el resto de las variables la correlación es débil.

Una vez elegida la función a ajustar, se estiman los valores de los parámetros, se calcula la bondad del ajuste y se analizan los residuos con la opción **ANALIZAR> REGRESIÓN>LINEAL**.



Interpretación:

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	peso[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: imc

La tabla **Resumen del Modelo**, muestra el valor del **coeficiente de determinación general** que sirve para medir la bondad del ajuste: **R^2=0,657** indica que el 65,7% de la variabilidad del imc está explicada por el peso.

El error típico de la estimación (S_e) es la desviación típica de los residuos. A mayor R^2 menor S_e

Uno de los **supuestos** básicos del modelo de regresión lineal es la **independencia entre los residuos**. El estadístico de **Durbin-Watson** (DW) oscila entre 0 y 4, toma el valor 2 cuando los residuos son independientes. Suele aceptarse que los residuos son independientes cuando el DW toma valores comprendidos entre 1,5 y 2,5.
En nuestro ejemplo, **podemos aceptar que los residuos son independientes** (DW=1,608).

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,810[a]	,657	,657	1,82745	1,608

a. Variables predictoras: (Constante), peso

b. Variable dependiente: imc

La tabla ANOVA de la Regresión informa **si existe o no relación significativa entre X e Y**. F contrasta la H_0 de que el valor poblacional de R es cero (pendiente de la recta de regresión es cero).[3]
En este caso **ambas variables están linealmente relacionadas**.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	6533,998	1	6533,998	1956,539	,000[a]
	Residual	3413,041	1022	3,340		
	Total	9947,039	1023			

a. Variables predictoras: (Constante), peso

b. Variable dependiente: imc

La última tabla muestra los estimadores mínimo-cuadráticos de los **coeficientes de la recta de regresión**. El modelo obtenido tiene de ecuación: **IMC=7,175+ 0,257·PESO**

[3] Este contraste lo entenderéis mejor una vez que estudiemos los temas correspondientes a inferencia estadística

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	7,175	,469		15,301	,000
	peso	,257	,006	,810	44,233	,000

a. Variable dependiente: imc

El *coeficiente no tipificado correspondiente a PESO es 0,257*. Es el **cambio medio que aumenta el IMC por cada unidad de cambio de PESO**.

$$IMC=7,175+ 0,257 \cdot PESO$$

En el *gráfico de dispersión entre ZPRED y ZRESID* están representados en el eje horizontal los **valores pronosticados** y en el eje vertical los **residuos**, ambos tipificados. Si la nube de puntos no muestra ningún patrón y los valores de los residuos se encuentran mayoritariamente entre -2 y 2, se concluye que **el modelo recoge toda la información necesaria para predecir el valor de la variable dependiente**.