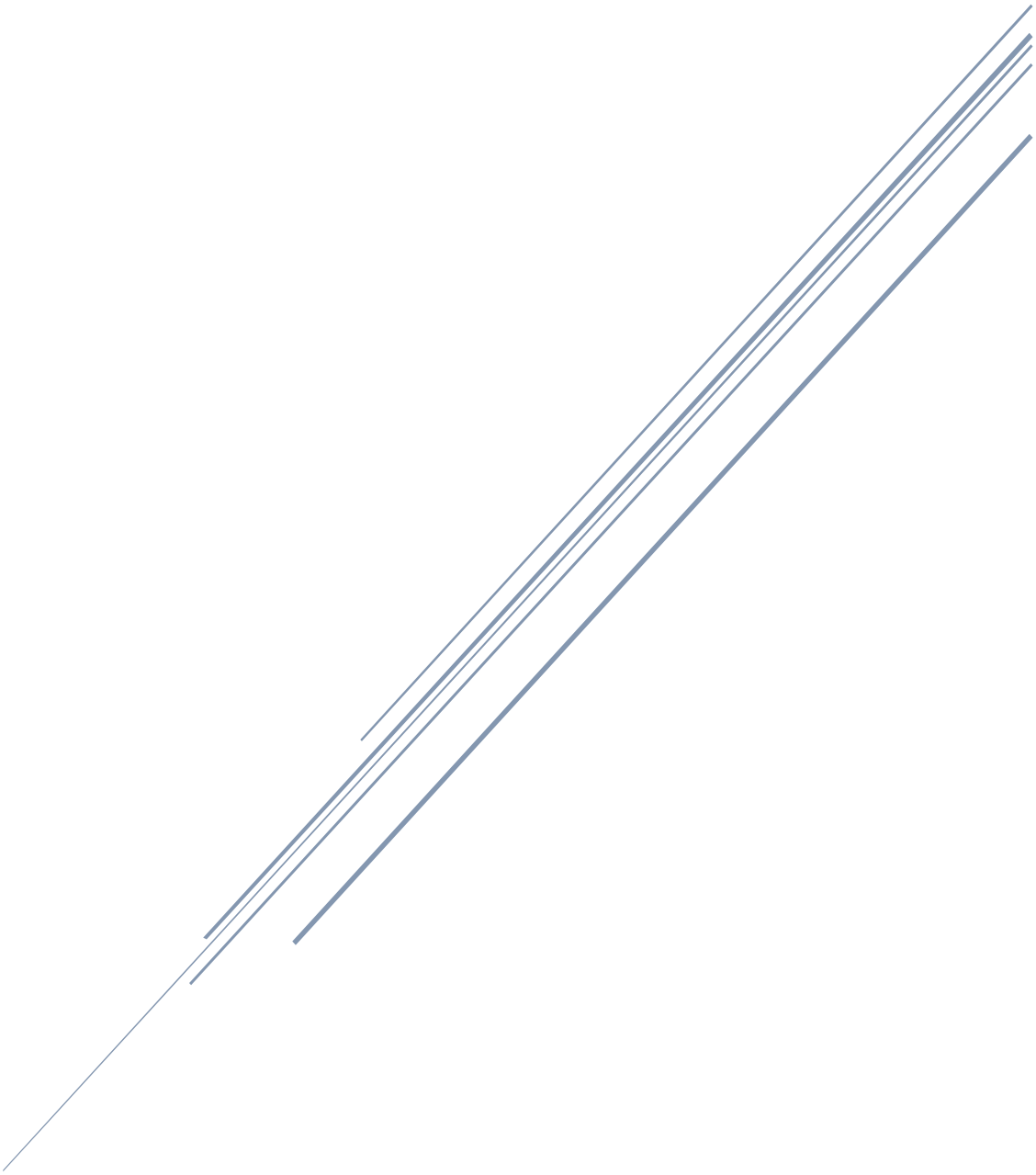
# TEMA 7. REGRESIÓN Y CORRELACIÓN LINEAL SIMPLE

## Curso OCW de “Estadística Descriptiva con Excel para Grados de Ciencias Sociales”



Grupo Decisión Multicriterio Zaragoza (GDMZ)

## 1. INTRODUCCIÓN

En el capítulo anterior se han presentado situaciones reales que estudian fenómenos en los que intervienen dos variables conjuntamente, buscando analizar la relación entre ambas. Así, por ejemplo, se puede estudiar la relación existente entre la experiencia profesional de los trabajadores y sus respectivos sueldos o entre la producción agraria y la cantidad de fertilizantes utilizados, etc.

Sabemos también que siempre que sea posible predecir con exactitud los valores de una variable a partir de los de la otra, se dice que ambas variables están en relación funcional o que cuando las dos variables no tienen ninguna relación se dice que son independientes y podemos estudiarlas por separado.

No obstante, en la mayoría de los problemas económico-empresariales, entre dos variables no se puede establecer una relación funcional ni tampoco afirmar que exista interrelación. Se dice que existe relación o dependencia estadística entre las dos variables y su análisis se puede abordar desde dos enfoques distintos y complementarios:

a) La determinación de una función matemática que mejor explique las variaciones de la variable dependiente (**endógena**), en función de las fluctuaciones que experimente la variable independiente (**exógena**).

b) El estudio del grado de dependencia existente entre las variables estudiadas.

De la determinación de la función matemática que explica las fluctuaciones de una variable en función de la otra se encarga la denominada **Teoría de la Regresión**. Mientras que el estudio del grado de dependencia que pueda existir entre las variables es propio de la denominada **Teoría de la Correlación**.

## 2. DIAGRAMAS DE DISPERSIÓN

Con el fin de analizar visualmente el tipo de relación existente entre las dos variables consideradas X e Y, una de las representaciones gráficas más usuales de una distribución de frecuencias conjunta es la nube de puntos o **diagrama de dispersión**, que consiste en trazar un plano de coordenadas sobre cuyo eje de abscisas representamos los valores de la variable X, reservando el eje de ordenadas para los valores de Y. Es decir, se construye representando cada elemento observado por un punto en el plano de manera que las coordenadas sobre los dos ejes cartesianos sean los valores que toman las dos variables en ese elemento.

**Ejemplo:**

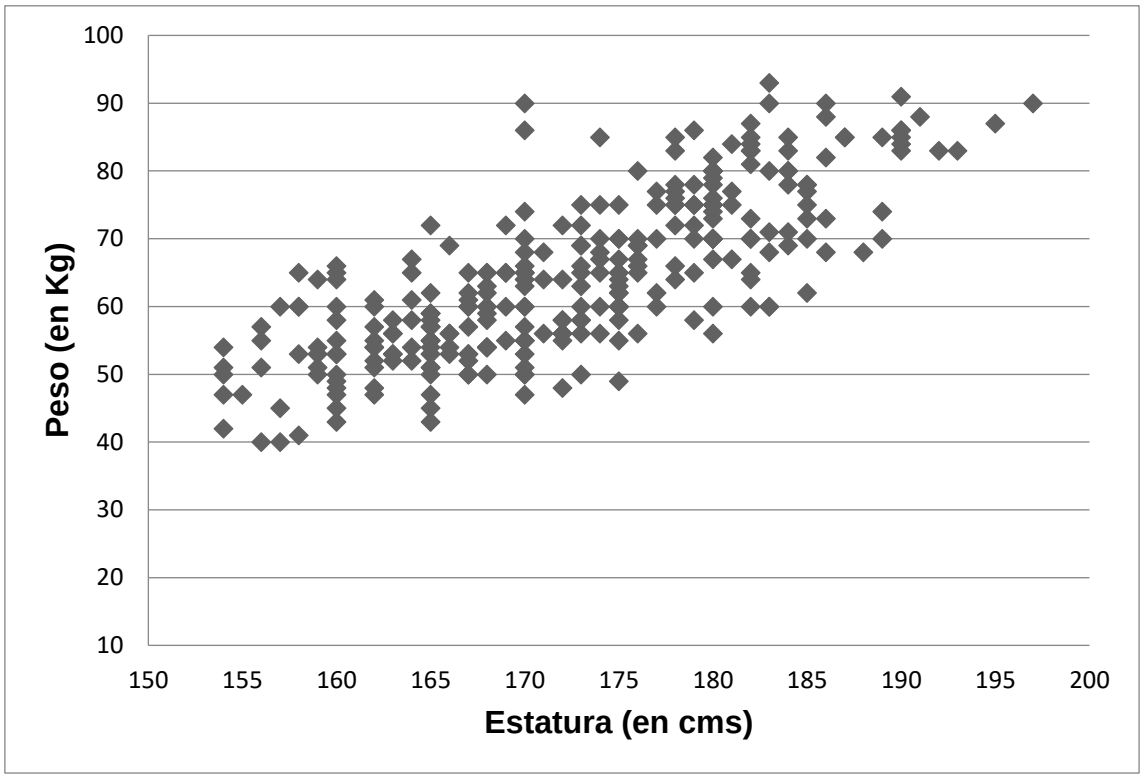



**Figura 1.** Diagrama de dispersión de la estatura frente al peso de un grupo de estudiantes

Estos diagramas nos permiten apreciar si existe relación entre las variables, si esta relación es lineal o no lineal, si es directa o inversa y la intensidad de la misma.

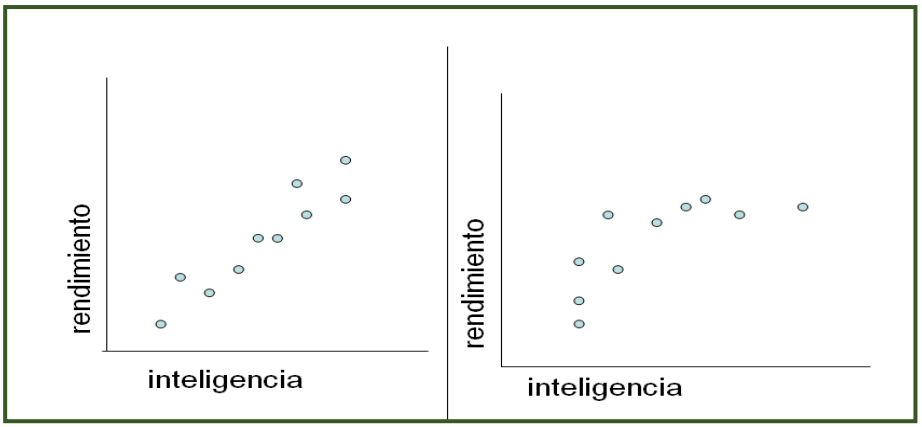



**Figura 2.** Relación lineal vs. Relación no lineal (ambas son directas)

## 3. COVARIANZA Y CORRELACIÓN

La relación entre dos variables también puede expresarse de forma numérica. La **covarianza** es una medida de la asociación lineal entre dos variables que resume la información existente en un gráfico de dispersión. Así, la covarianza entre X e Y se denotará por $S_{XY}$ y viene dada por la expresión:

$$S_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

### 3.1. Interpretación:

La covarianza es una medida del grado de variación conjunta de dos variables en el sentido de que su signo informa acerca de la existencia o no de la tendencia de la nube de puntos a situarse preferentemente en los cuadrantes primero y tercero (si $S_{XY} > 0$), segundo y cuarto (si $S_{XY} < 0$) o en ninguno de ellos (si $S_{XY} = 0$) en el diagrama que toma como origen el centro de gravedad de la distribución de frecuencias, $(\bar{x}, \bar{y})$. Así, los puntos situados en el primer y tercer cuadrantes verifican que $(x_i - \bar{x}) \times (y_i - \bar{y}) > 0$ mientras que los situados en el segundo y el cuarto cuadrantes $(x_i - \bar{x}) \times (y_i - \bar{y}) < 0$.

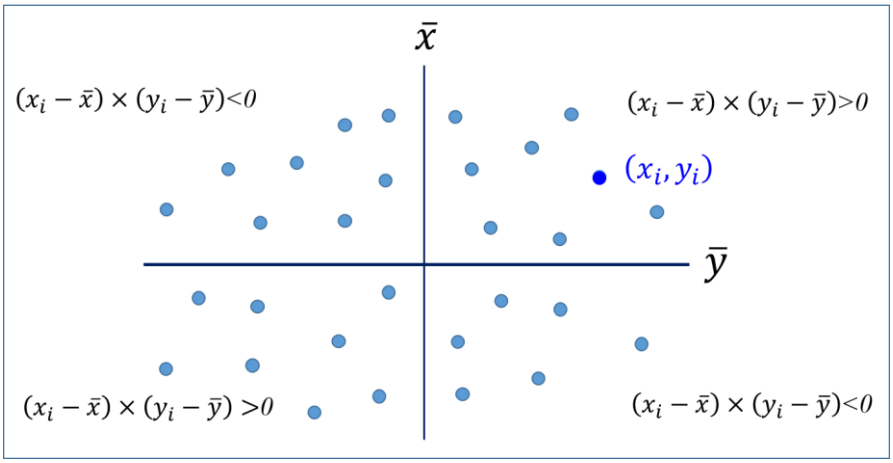



**Figura 3. Signo de los términos de la covarianza**

La covarianza calcula la media de los productos anteriores de forma que si $S_{XY} > 0$ es porque el número de pares $(x_i, y_i)$ que verifican que $(x_i - \bar{x}) \times (y_i - \bar{y}) > 0$ es mayor que el que los verifican que $(x_i - \bar{x}) \times (y_i - \bar{y}) < 0$ (ver Figura 4a) ocurriendo lo contrario si $S_{XY} < 0$ (ver Figura 4b) y estando equilibrados ambos conjuntos si $S_{XY} \sim 0$ (ver Figura 4c) y d). Por tanto, si $S_{XY} > 0$ ello indica que la relación existente entre *X* e *Y* es directa, es decir, que cuanto mayor es el valor de *X*, mayor tiende a ser el valor de *Y* (ver Figura 4a), mientras que si $S_{XY} < 0$ la relación es inversa, es decir, que cuanto mayor es el valor de *X*, menor tiende a ser el valor de *Y* (ver Figura 4b). Si $S_{XY} \sim 0$ no se puede

decir nada acerca del tipo de relación: puede no existir (ver Figura 4c) o no ser lineal (ver Figura 4d).

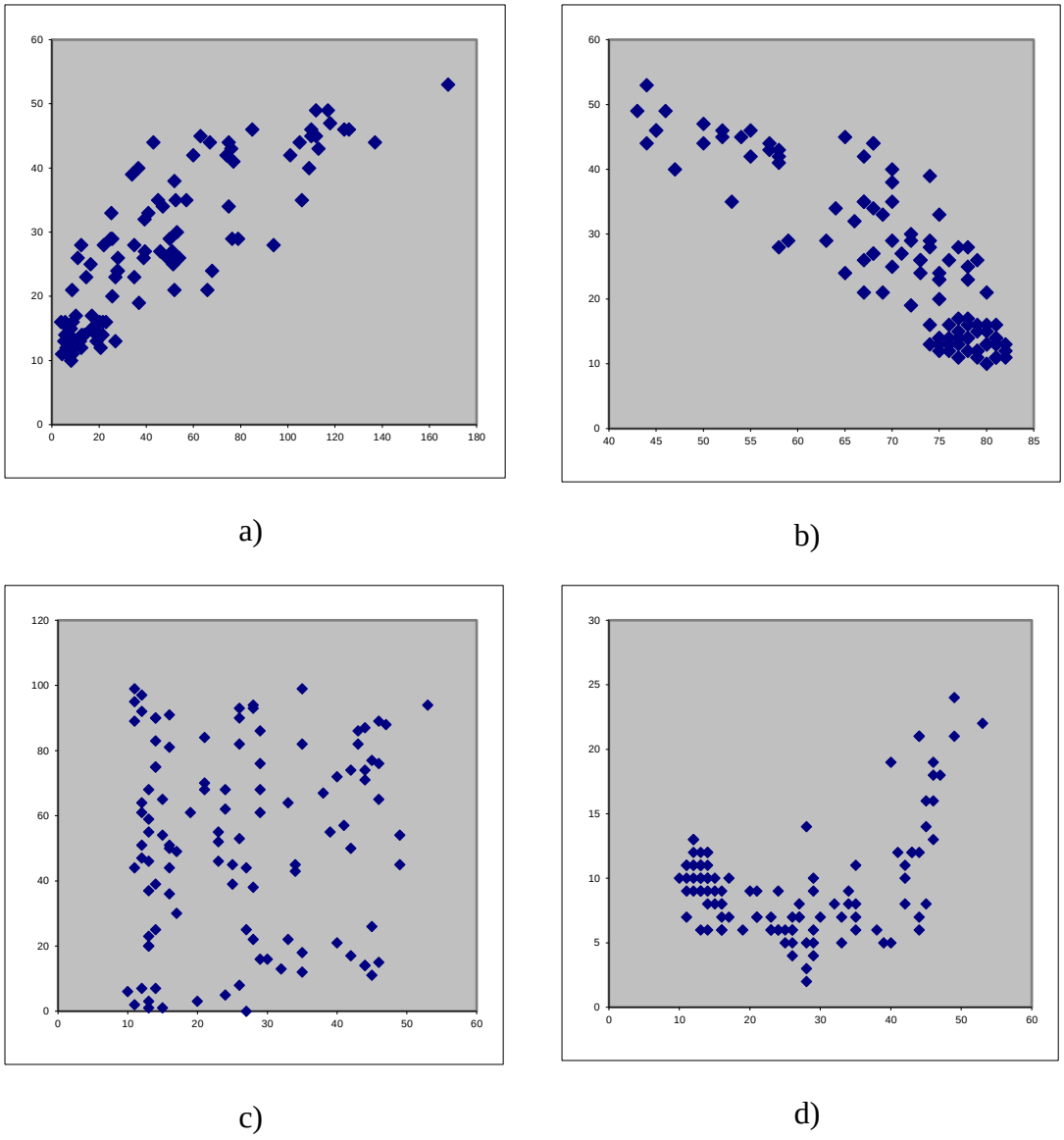



**Figura 4.** Diagramas de dispersión correspondientes a variables (X,Y) con $S_{XY} > 0$ (a), $S_{XY} < 0$ (b) o $S_{XY} \sim 0$ (c), (d)

### 3.2. Propiedades de la Covarianza

- Para su cálculo se puede emplear la fórmula abreviada:

$$S_{XY} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

- Si las variables son estadísticamente independientes entonces son incorreladas ($S_{XY} = 0$). El recíproco, sin embargo, no es cierto, es decir,

puede ocurrir que $S_{XY} = 00$ y, sin embargo, puede existir una relación de dependencia entre *X* e *Y*.

- La covarianza es invariante ante cambios de origen, pero no ante cambios de escala:

$$\left.\begin{matrix} U = a + b \times X \\ V = c + d \times Y \end{matrix}\right\} \Rightarrow S_{UV} = bdS_{XY}$$

- Dado que la covarianza no es invariante por cambios de origen y escala, es poco recomendable utilizarlo para medir el grado de asociación existente entre *X* e *Y* dado que dicha asociación debe ser, en particular, independiente de las unidades de medida. Para ello se utiliza el **coeficiente de correlación** r que viene dado por:

$$r = \frac{S_{XY}}{S_X S_Y}$$

que es adimensional y que está relacionado con la medición de la bondad de ajuste de la recta de regresión tal y como se verá más adelante.

## 4. REGRESIÓN LINEAL SIMPLE: CRITERIO DE LOS MÍNIMOS CUADRADOS

El problema de la **Regresión** consiste en la obtención de la ecuación de una curva que pase "cerca" de los puntos representados en el diagrama de dispersión, y que se adapte lo mejor posible al conjunto de los mismos, cumpliendo determinadas condiciones. Por lo tanto, cuando se pretende hacer un ajuste nos encontramos con dos problemas:

(a) Elegir el *tipo de curva* que mejor se adapte a los datos disponibles, es decir, que mejor represente la relación entre las variables endógena y exógena. En esta fase suele ser de gran utilidad la representación gráfica como orientación para la elección.

(b) Fijado el tipo de curva a través de su ecuación en forma explícita con un cierto número de parámetros, determinar éstos mediante las condiciones que se impongan según el procedimiento de ajuste empleado.

En este tema, vamos a reducir nuestro problema inicial a buscar la relación lineal que mejor explique una variable a partir de la otra. El hecho de ceñirnos a rectas no es una limitación sustancial ya que son muchas las variables que originalmente, o tras sencillas transformaciones, se ajustan en cierto grado a una recta, o lo que es lo mismo, su nube de puntos tiene una forma más o menos lineal.

La ecuación de la recta Y = a + bX depende de dos coeficientes a y b que deben calcularse a partir de los datos observados. El parámetro b es la pendiente de la recta, se denomina coeficiente de regresión y nos dice cuanto aumenta la variable dependiente cuando la independiente aumenta una unidad. El parámetro a es la ordenada en el origen y representa el valor de la variable dependiente cuando la independiente toma el valor cero. En muchos problemas no tiene sentido considerar un valor cero para X. En este caso debe tomarse como un valor de referencia necesario para describir la recta sin atribuirle una interpretación lógica en el problema. De las infinitas rectas del plano, tendremos que buscar aquella que mejor se ajusta a la nube de puntos.

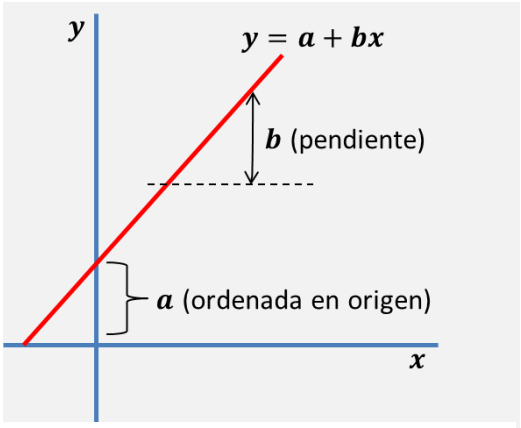



**Figura 5.** Recta de regresión

### 4.1. Obtención de las rectas de regresión

Con el fin de simplificar la notación supondremos de ahora en adelante, que los datos observados son de la forma $\{(x_i, y_i), i = 1 \ldots N$. 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, es decir, $\hat{y}_i = a + bx_i$. La diferencia entre el valor observado y el valor teórico recibe el nombre de error o residuo: $e_i = y_i - \hat{y}_i$

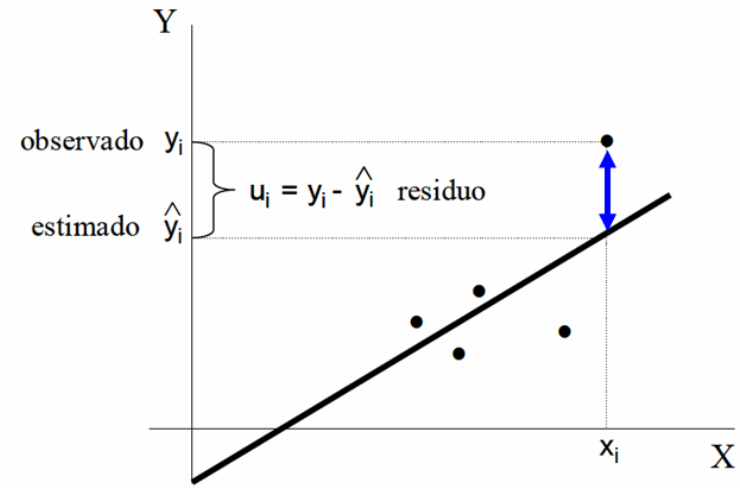



**Figura 6.** Residuos de la regresión lineal de Y sobre X

Parece razonable considerar como recta de regresión estimada aquella que proporciona los errores más pequeños. Una primera posibilidad podría ser minimizar la suma de los residuos:

$$SRes = \sum_{i=1}^{N} e_i = \sum_{i=1}^{N} (y_i - \hat{y}_i)$$

pero no es un criterio muy adecuado ya que los errores pueden ser positivos o negativos y al sumarlos pueden cancelarse unos con otros proporcionando una idea errónea.

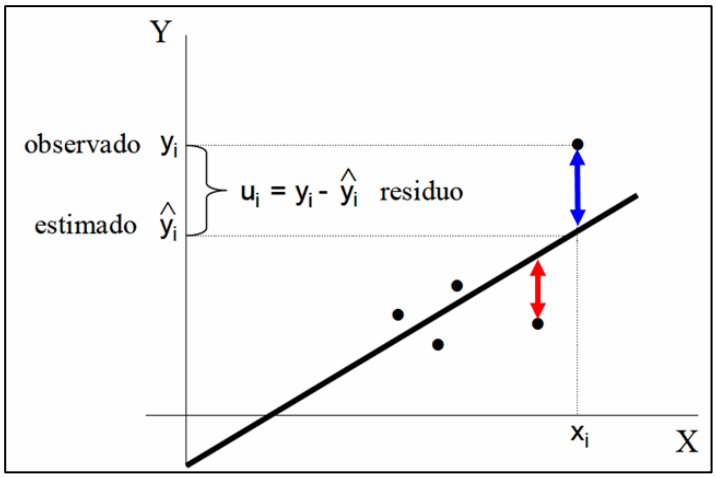



**Figura 7.** Residuos de la regresión lineal

Para eliminar este problema se puede utilizar el criterio de minimizar la suma de los valores absolutos de los residuos:

$$SARes = \sum_{i=1}^{N} |e_i| = \sum_{i=1}^{N} |(y_i - \hat{y}_i)|$$

Con este criterio no se presenta el problema de que los errores positivos se compensen con los negativos. Sin embargo, tiene el inconveniente de que no se presta a manipulaciones algebraicas, por ser una suma de valores absolutos.

Por último, un tercer criterio consiste en minimizar la suma de los errores al cuadrado. Este método recibe el nombre de *Método de los Mínimos Cuadrados*, consistente en calcular los coeficientes de la función a estimar, con la condición de que la suma de los cuadrados de las desviaciones de cada ordenada observada frente a la ordenada estimada sea mínima.

$$ECM = \frac{1}{N}\sum_{i=1}^{N} {e_i}^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

En este criterio como los errores están elevados al cuadrado, no se compensan los residuos positivos con los negativos y, además, la expresión es susceptible de manipulaciones algebraicas

Los valores de *a* y *b* que minimizan dicha función vienen dados por las siguientes expresiones:

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

de forma que la ecuación de la recta de regresión puede escribirse

$$Y - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x})$$

De la expresión anterior, fácilmente se comprueba que la ecuación se satisface para el punto $(\bar{x}, \bar{y})$, es decir, la recta de regresión pasa siempre por el punto de las medias de la distribución bidimensional, llamado centro de gravedad.

Intercambiando los papeles de las variables y realizando el mismo procedimiento, obtenemos la ecuación de la recta de regresión de *X* sobre *Y*, *X=a'+b'Y:*

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

### 4.2 Signo de la dependencia lineal

Dada la recta de regresión $Y = a + b\,X$, si el coeficiente de regresión *b* es positivo, la nube de puntos tiene una disposición tal que aumentan los valores de *Y* al aumentar los de *X*, y en este caso se dice que estamos ante una dependencia lineal entre las variables *X* e *Y* de tipo directo (ver Figura 8). Si, por el contrario, el coeficiente de regresión, *b*, es negativo, la nube de puntos está configurada de modo que disminuyen los valores de *Y* al aumentar los de *X* (ver Figura 8), y diremos que se trata de una dependencia lineal inversa. Finalmente, si $b = 0$ no existe dependencia lineal entre *X* e *Y* y las variables se dice que son **incorreladas**. En este caso $S_{XY} = 0$ pero, como se ha comentado anteriormente, este hecho no implica la inexistencia de otro tipo de dependencia (ver Figura 4d).

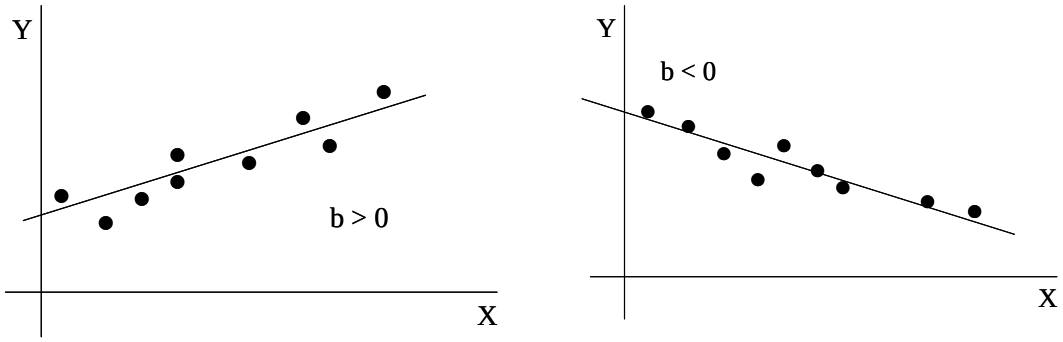



**Figura 8.** Signo de la dependencia lineal de X e Y

Notar que, como $b = \frac{S_{XY}}{S_X^2}$, el signo de *b*, coincide con el signo de la covarianza, $S_{XY}$.

## 5. BONDAD DE AJUSTE Y CORRELACIÓN

El objetivo fundamental de la Teoría de la Regresión es poder expresar una variable en función de otra y hacer uso de dicha función para realizar predicciones, de manera que, si se quiere predecir un valor de Y, y el ajuste realizado ha sido lineal, se deberá considerar la recta de regresión de Y/X, mientras que si la predicción que se quiere efectuar es de X se hará uso de la recta de regresión de X/Y. Ahora bien, cualquier predicción debería ir acompañada de un estudio de la fiabilidad que ésta merece. En este sentido, resulta obvio que cuanto mejor sea el ajuste, mayor será la fiabilidad de la predicción efectuada a partir de él. Del estudio de la fiabilidad y del grado de dependencia que pueda existir entre las dos variables se encarga la denominada Teoría de la Correlación a la que vamos a dedicar este apartado definiendo distintos indicadores que serán de utilidad para analizar la bondad del ajuste efectuado.

El grado de bondad del ajuste se deducirá a partir de los residuos o errores. Si cada residuo es nulo, la línea pasa por todos los puntos de la nube, y, en este caso, se dice que entre las variables existe una dependencia funcional y el ajuste es perfecto. Cuando estos residuos son pequeños, el ajuste es bueno y la recta tiene una gran representatividad; en cambio, si son grandes, el ajuste no es fino y la línea ajustada explica el fenómeno observado con dificultad.

A partir de estos residuos, el problema que se plantea ahora es la medición de la intensidad con que dos variables pueden estar relacionadas mediante la construcción de coeficientes numéricos. A continuación, vamos a presentar algunas de las medidas que existen para cuantificar esta intensidad.

## 5.1 Varianza residual y varianza explicada de Y sobre X

Una vez obtenida la ecuación de la recta de regresión de Y sobre X, un indicador que puede ser utilizado para evaluar la bondad del ajuste es la varianza residual $S_{r_Y}^2$ o varianza de los residuos,

$$S_{r_Y}^2 = \frac{1}{N}\sum_{i=1}^{N} {e_i}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Cuanto menor sea la varianza residual mejor será el ajuste, pues más próximos estarán los valores teóricos a los observados. El ajuste será óptimo en el caso de que la varianza residual valga cero, pues en ese caso los valores observados y predichos coincidirán. Sin embargo, cuanto mayor sea la varianza residual peor será el ajuste, puesto que los valores observados y predichos serán más dispares.

Otro coeficiente interesante, es la varianza explicada o varianza de los valores predichos, es decir:

$$S_{\exp Y}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{\hat{y}_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Y se puede probar que:

$$S_Y^2 = S_{r_Y}^2 + S_{exp_Y}^2$$

es decir, la varianza total, $S_Y^2$ , se descompone como suma de la varianza explicada por la regresión y la varianza residual. De modo que si $S_Y^2 = S_{exp_Y}^2$ o equivalentemente $S_{r_Y}^2 = 0$, implica que toda la varianza de *Y* queda explicada por la regresión, es decir, el ajuste es perfecto. El caso opuesto es cuando $S_Y^2 = S_{r_Y}^2$ o $S_{exp_Y}^2 = 0$, la varianza de *Y* no puede explicarse por la regresión.

Un inconveniente de ambas varianzas es su dependencia de las unidades de medida, por lo que resulta aconsejable obtener otro indicador que solvente esta desventaja.

## 5.2. Coeficiente de determinación y coeficiente de correlación lineal de Pearson

Dado que las varianzas anteriores vienen afectadas por la unidad de medida, un indicador objetivo de la bondad explicativa de la recta de regresión es el *coeficiente de determinación*, cociente entre la varianza explicada y la varianza total y, por tanto, independiente de cambios de origen y escala. Este coeficiente mide la proporción de varianza que se explica con la regresión.

$$r^2 = \frac{S^2_{exp_Y}}{S^2_Y} = \frac{S^2_Y - S^2_{r_Y}}{S^2_Y} = 1 - \frac{S^2_{r_Y}}{S^2_Y}$$

De su expresión se deduce que toma valores entre 0 y 1. Vale 1 en el caso extremo de que todos los errores o residuos sean nulos, existiendo entonces dependencia lineal perfecta. Vale 0 cuando $S^2_Y = S^2_{r_Y}$ y, en este caso, las variaciones de *Y* no son debidas a la variable *X*, sino que son debidas a los errores. Para otros valores intermedios del coeficiente de determinación, podemos concluir que valores próximos a 0,9 son indicativos de ajustes muy aceptables, mientras que valores del mismo inferiores a 0,6 tienen escasa fiabilidad y sugieren la búsqueda de otra línea de ajuste más adecuada.

Se puede demostrar además que:

$$r^2 = \frac{S^2_{XY}}{S^2_X S^2_Y} = \left(\frac{S_{XY}}{S_X S_Y}\right)^2$$

Es decir, el coeficiente de determinación es también el cuadrado del *coeficiente de correlación lineal de Pearson*. Se comprueba que al estar el coeficiente de determinación comprendido entre 0 y 1, este coeficiente está acotado entre –1 y 1. Recordemos que el signo de *r* dependía del signo de la covarianza, puesto que las desviaciones típicas son siempre positivas. Así, si $r > 0$, la relación lineal entre las variables es directa, y tanto más fuerte cuanto mayor sea el coeficiente de correlación, siendo perfecta cuando $r = 1$. De forma análoga, si $r < 0$, la relación entre las variables es inversa, siendo ésta más intensa a medida que *r* se aproxima a –1. Finalmente, un *r* igual a cero significa que no existe relación lineal entre las variables o, equivalentemente, que las variables están incorreladas, lo cual no quiere decir que no exista algún otro tipo de dependencia no lineal.

## 6. PREDICCIÓN

La *predicción* constituye la aplicación más interesante de la técnica de la regresión y consiste en determinar a partir del modelo estimado el valor que toma la variable endógena para un valor dado de la variable exógena.

Cuando se quieren predecir valores de la variable *Y* para un valor concreto de la variable independiente *X*, se considera la recta de regresión de *Y* sobre *X*. Es decir, esta *predicción* de *Y* para un valor concreto $x_h$ de la variable *X* se obtendría sin más que sustituir el valor $x_h$ en la ecuación de la recta:

$$\hat{y}_h = a_{y|x} + b_{y|x} x_h$$

Análogamente, cuando se quieren predecir valores de la variable *X*, se considera la recta de regresión de *X* sobre *Y*. Esta predicción de *X* para un valor concreto $y_h$ de la variable *Y* se obtendría sin más que sustituir el valor $y_h$ en la ecuación de la recta. Es decir:

$$\hat{x}_h = a_{x|y} + b_{x|y} y_h$$

Cuando la predicción se hace para valores de la variable exógena situados dentro del intervalo de variación de los datos con los que se ha calculado la regresión, recibe el nombre de *interpolación.* Si la predicción se efectúa para un valor de la variable exógena situado fuera de ese intervalo se denomina *extrapolación* y su fiabilidad disminuirá. En general, las predicciones para valores muy alejados del centro de gravedad de la distribución pueden no ser muy fiables puesto que se corre el peligro de que no sea válido el modelo utilizado. Además, la bondad de los valores pronosticados para la variable endógena será tanto mayor cuanto mejor sea el ajuste, es decir cuanto mayor sea el valor del coeficiente de determinación. Resumiendo:

- Predicciones fiables $\Leftrightarrow$ $r^2$ alto e interpolación
- Predicciones poco fiables $\Leftrightarrow$ $r^2$ bajo o extrapolación