Regresión lineal simple

Josep Gibergans Bàguena

P08/75057/02311

Universitat Oberta de Catalunya
www.uoc.edu

Índice

El modelo de regresión simple

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.

Origen de los modelos de regresión

Estos modelos fueron utilizados por Laplace y Gauss en sus trabajos de astronomía y física desarrollados durante el siglo XVIII, pero el nombre de *modelos de regresión* tiene su origen en los trabajos de Galton en biología de finales del siglo XIX. La expresión de Galton:

"*regression towards mediocrity*"

dio nombre a la regresión.

Por ejemplo, podemos estudiar las distribuciones de los pesos y de las alturas de un conjunto de personas por separado. Ahora el objetivo es determinar si existe alguna relación entre estas variables.

Queremos construir modelos que describan la relación entre las variables con el propósito, principalmente, de predecir los valores de una variable a partir de los valores de la otra. Lo haremos con el modelo de regresión lineal simple.

2. Relaciones entre dos variables

El modelo de regresión lineal simple nos permite construir un modelo para explicar la relación entre dos variables.

El 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**).

Ejemplo de relación entre dos variables

Si las dos variables son los ingresos mensuales y los gastos en actividades de ocio, entonces podríamos escoger la segunda como variable explicada Y y la primera como variable explicativa X, ya que, en principio, los gastos en ocio dependerán mucho de los ingresos: cuanto más dinero ganemos, mayor será la parte que gastaremos en ocio.

Es importante observar que también podríamos escoger las variables a la inversa, es decir, los gastos en ocio como variable explicativa X y los ingresos como variable explicada Y. Cuanto más dinero gastemos en ocio, más ingresos tendremos.

No es fácil la decisión de elegir cuál es la variable explicativa y cuál es la variable explicada. Como veremos más adelante, dependerá en gran medida de las características de los datos que tengamos.

Las relaciones entre dos variables pueden ser de dos tipos:

Ejemplo de relación funcional

Podemos conocer el área de un cuadrado a partir de la longitud de su lado.

1) Funcionales (o **deterministas**): cuando hay una fórmula matemática que permite calcular los valores de una de las variables a partir de los valores que toma la otra.

2) Estadísticas (o **estocásticas**): cuando no existe una expresión matemática que las relacione de forma exacta.

Ejemplo de relación estadística

Sabemos que hay una relación entre la altura y el peso de las personas: en general, cuanta más altura, más peso. Pero no existe ninguna fórmula matemática que nos dé una en función de la otra, ya que esto significaría que todas las personas que tienen la misma altura tendrían el mismo peso, y eso sabemos que no es cierto.

En la relación entre el peso y la altura es evidente que existen muchos factores, como pueden ser factores genéticos, la actividad física, la alimentación, etc. que hacen que una persona de una determinada altura tenga un peso u otro. Todos estos factores y otros que no conocemos hacen que la relación entre estas dos variables sea estadística y no funcional.

3. Diagramas de dispersión y curvas de regresión

A partir de un conjunto de observaciones de dos variables X e Y sobre una muestra de individuos, el primer paso en un análisis de regresión es representar estos datos sobre unos ejes coordenados x-y. Esta representación es el llamado *diagrama de dispersión*. Nos puede ayudar mucho en la búsqueda de un modelo que describa la relación entre las dos variables.

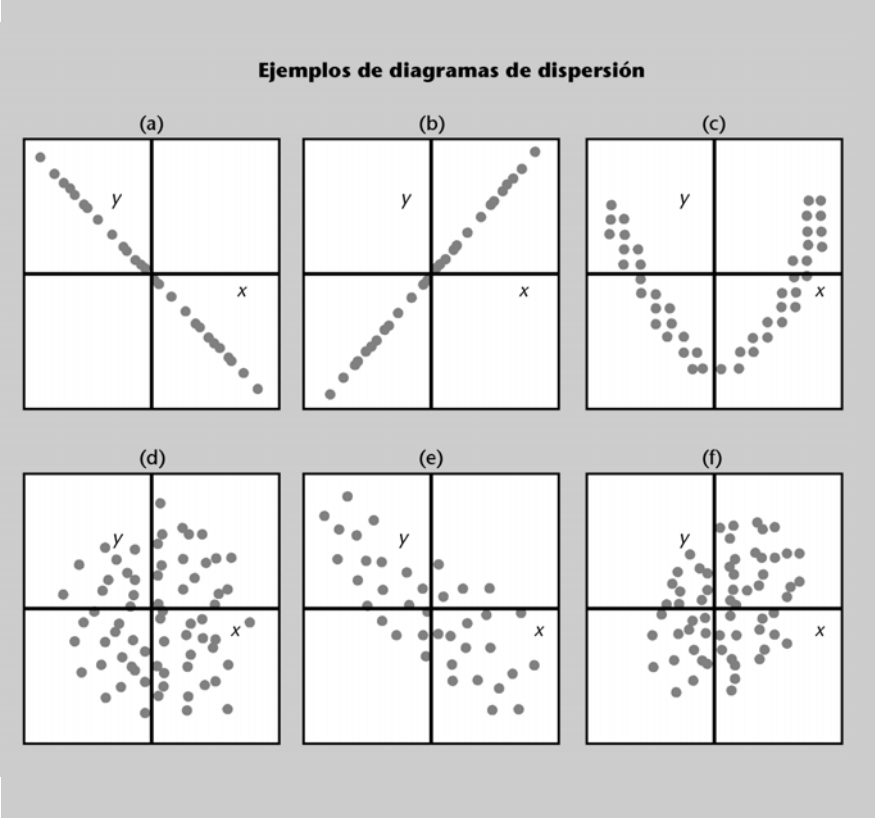
El **diagrama de dispersión** se obtiene representando cada observación (x_i, y_i) como un punto en el plano cartesiano XY.

Terminología

El diagrama de dispersión también se conoce como *nube de puntos*.

Ejemplo de diagramas de dispersión

El diagrama de dispersión puede presentar formas diversas:



En los casos (a) y (b) tenemos que las observaciones se encuentran sobre una recta. En el primer caso, con pendiente negativa, que nos indica que a medida que X aumenta, la Y es cada vez menor y lo contrario en el segundo caso, en el que la pendiente es positiva. En estos dos casos los puntos se ajustan perfectamente sobre la recta, de manera que tenemos una relación funcional entre las dos variables dada por la ecuación de la recta.

En el caso (c) los puntos se encuentran situados en una franja bastante estrecha que tiene una forma bien determinada. No será una relación funcional, ya que los puntos no se sitúan sobre una curva, pero sí que es posible asegurar la existencia de una fuerte relación entre las dos variables. De todos modos, vemos que no se trata de una relación lineal (la nube de puntos tiene forma de parábola).

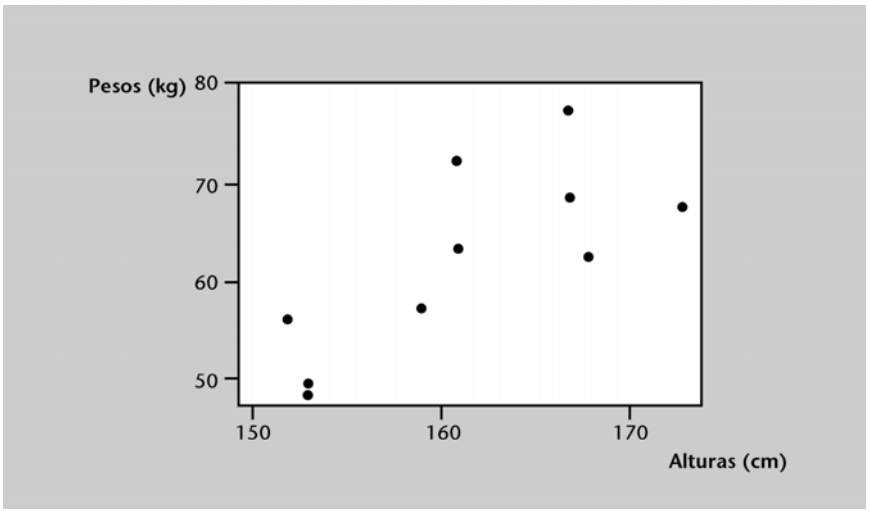
En el caso (d) no tenemos ningún tipo de relación entre las variables. La nube de puntos no presenta una forma "tubular" bien determinada; los puntos se encuentran absolutamente dispersos.

En los casos (e) y (f) podemos observar que sí existe algún tipo de relación entre las dos variables. En el caso (e) podemos ver un tipo de dependencia lineal con pendiente negativa, ya que a medida que el valor de X aumenta, el valor de Y disminuye. Los puntos no están sobre una línea recta, pero se acercan bastante, de manera que podemos pensar en una fuerte relación lineal. En el caso (f) observamos una relación lineal con pendiente positiva, pero no tan fuerte como la anterior.

Ejemplo de las alturas y los pesos

Consideremos las observaciones de los pesos y alturas de un conjunto de 10 personas: el individuo 1 tiene 161 cm de altura y 63 kg de peso, el individuo 2 tiene 152 cm de altura y 56 kg de peso, etc., tal como se ve en la tabla siguiente:

Individuo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
***X* altura (cm)**	161	152	167	153	161	168	167	153	159	173
***Y* peso (kg)**	63	56	77	49	72	62	68	48	57	67



El diagrama de dispersión también nos puede ayudar a encontrar algún valor atípico entre los datos de la muestra que pueda tener su origen en una mala observación o en el hecho de ser una observación correspondiente a un individuo excepcional dentro de la muestra. Cuando tenemos un valor atípico, debemos controlar las influencias que pueda tener en el análisis.

Definición y ejemplo de valor atípico

Por *valor atípico* entendemos un valor muy diferente de los otros y que muy posiblemente es erróneo. Por ejemplo, una persona de 150 cm de altura y 150 kg de peso. En el diagrama de dispersión saldrá como un punto solitario alejado de los otros.

4. Recta de regresión

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**.

4.1. Estimación de los parámetros: método de los mínimos cuadrados

Una recta queda bien determinada si el valor de su **pendiente** (b) y de su **ordenada en el origen** (a) son conocidos. De esta manera la ecuación de la recta viene dada por:

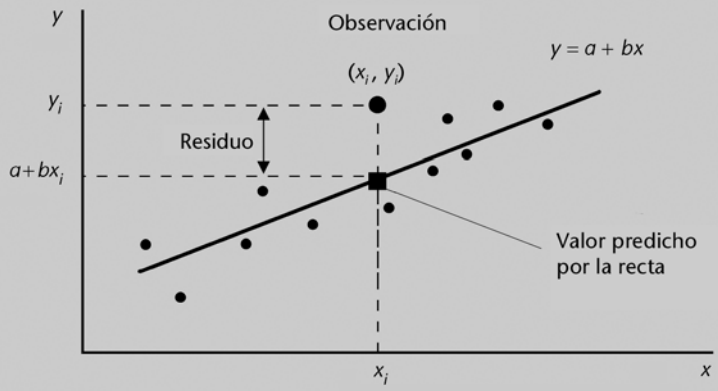
$$y = a + bx$$

A partir de la fórmula anterior definimos para cada observación (x_i, y_i) 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. Buscaremos la recta que dé lugar a los residuos más pequeños en cuanto a la suma de los cuadrados.

Interpretación geométrica del residuo



Para determinar una recta de regresión, utilizaremos el método de los mínimos cuadrados.

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**.

Siendo la suma de los cuadrados la expresión:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Terminología

La suma de los cuadrados de los residuos también se denomina *suma de los errores cuadráticos*.

para encontrar los valores de a y b, sólo hay que determinar las derivadas parciales con respecto a los parámetros a y b:

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n} (y_i - a - bx_i)^2\right) = -2\sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\frac{\partial}{\partial b}\left(\sum_{i=1}^{n} (y_i - a - bx_i)^2\right) = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i$$

La resolución de este sistema de ecuaciones se encuentra en el anexo 1 de esta sesión.

y las igualamos a cero. Así obtenemos el sistema de ecuaciones siguiente, conocido como s*istema de ecuaciones normales*:

$$\begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 \end{cases}$$

Las soluciones de este sistema de ecuaciones son:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{y} \quad a = \bar{y} - b\bar{x}$$

En rigor...

... habría que probar que, efectivamente, estos valores de los parámetros hacen mínima la suma de los cuadrados de los residuos.

en las que:

- $s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ es la **covarianza muestral** de las observaciones (x_i, y_i)
- $s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ es la **varianza muestral** de las observaciones x_i

Notación

Hemos hecho un cambio en la notación para distinguir de manera clara entre una recta cualquiera:

$$y = a + bx$$

y la recta de regresión por mínimos cuadrados:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

obtenida al determinar a y b.

Es muy importante obsevar que, de todas las rectas, la recta de regresión lineal por mínimos cuadrados es aquella que hace mínima la suma de los cuadrados de los residuos.

A partir de ahora, la **recta de regresión** la escribiremos de la manera siguiente:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

donde los **parámetros de la recta** $\hat{\beta}_0$ y $\hat{\beta}_1$ vienen dados por:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{y} \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

De ahora en adelante, a los **residuos calculados** con la recta de regresión los llamaremos e_i, es decir:

$$e_i = y_i - \hat{y}_i$$

donde $\hat{y}_i$ es el **valor estimado** para la recta de regresión.

5. Interpretación de los parámetros de la recta de regresión

Una vez determinada la recta de regresión, es muy importante interpretar los parámetros de la ecuación en el contexto del fenómeno que se estudia.

- **Interpretación de la ordenada en el origen**, $\hat{\beta}_0$:

$\hat{\beta}_0$ en el ejemplo de los pesos y las alturas

En el ejemplo de los pesos y las alturas, el valor de la ordenada en el origen no tendrá sentido, ya que correspondería al peso que tendrían las personas de altura nula.

Este parámetro representa la estimación del valor de Y cuando X es igual a cero:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 0 = \hat{\beta}_0$$

No siempre tiene una interpretación práctica. Para que sea posible, es preciso que:

1. realmente sea posible que X tome el valor $x = 0$
2. se tengan suficientes observaciones cercanas al valor $x = 0$

- **Interpretación de la pendiente de la recta**, $\hat{\beta}_1$

Pendiente en el ejemplo de los pesos y las alturas

En el ejemplo de los pesos y las alturas, en el diagrama de dispersión habíamos observado que, en general, aumenta el peso de las personas a medida que aumenta su altura.

Este parámetro representa la estimación del incremento que experimenta la variable Y cuando X aumenta en una unidad. 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.

6. Construcción de la tabla para determinar los parámetros

Veamos ahora cómo debemos determinar, en la práctica, la recta de regresión. Lo ilustraremos a partir de los datos del ejemplo de los pesos y las alturas.

Ejemplo de las alturas y los pesos

Continuemos con el anterior ejemplo de las alturas y pesos de un grupo de diez personas. Para determinar la recta de regresión, calculamos la covarianza muestral s_{xy}, la varianza muestral s_x^2 y las medias $\bar{x}$ y $\bar{y}$.

Podemos calcular todas estas cantidades a partir de la tabla de cálculos de la recta de regresión.

i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	161	63	–0,4	1,1	0,16	–0,44
2	152	56	–9,4	–5,9	88,36	55,46
3	167	77	5,6	15,1	31,36	84,56
4	153	49	–8,4	–12,9	70,56	108,36
5	161	72	–0,4	10,1	0,16	–4,04
6	168	62	6,6	0,1	43,56	0,66
7	167	68	5,6	6,1	31,36	34,16
8	153	48	–8,4	–13,9	70,56	116,76
9	159	57	–2,4	–4,9	5,76	11,76
10	173	67	11,6	5,1	134,56	59,16
Σ	**1.614**	**619**			**476,40**	**466,40**

Medias muestrales: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 161{,}4$ y $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 61{,}9$

Varianza muestral: $s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{476{,}40}{10-1} = 52{,}933$

Covarianza muestral: $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{466{,}40}{10-1} = 51{,}822$

Los parámetros son:

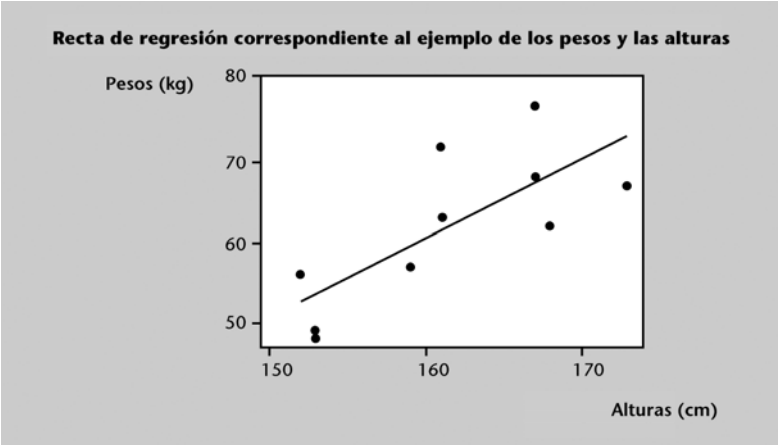
$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{51{,}822}{52{,}933} = 0{,}979009$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 61{,}9 - 0{,}979009 \cdot 161{,}4 = -96{,}1121$$

Tenemos la recta de regresión siguiente:

$$\hat{y} = -96{,}1121 + 0{,}979009x$$

Podemos representar la recta de regresión en el diagrama de dispersión:



Interpretamos los parámetros obtenidos:

- Ordenada en el origen: evidentemente, no tiene sentido pensar que el peso de una persona de altura cero es –96,1121 kg. Ya hemos comentado antes que muchas veces no tiene sentido la interpretación de este parámetro.

- Pendiente: tenemos una pendiente de 0,979009. Un valor positivo que nos informa de que el peso aumenta con la altura a razón de 0,979 kg por cada centímetro.

7. Interpolación y extrapolación

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.

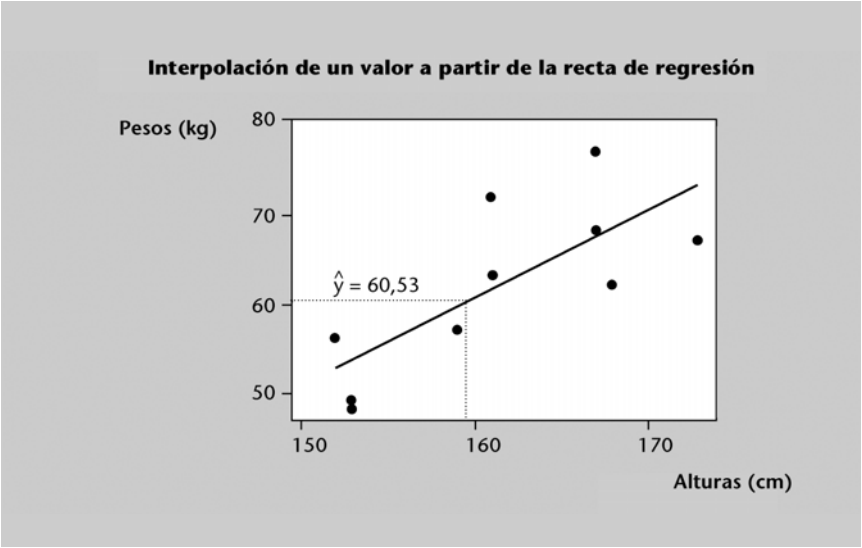
Ejemplo de las alturas y los pesos

En nuestro problema de los pesos y las alturas podríamos estar interesados en conocer el peso de una persona de altura 1,60 m. A partir de nuestra recta de regresión:

$$\hat{y} = -96{,}1121 + 0{,}979009x$$

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

$$\hat{y} = -96{,}1121 + 0{,}979009 \cdot 160 = 60{,}53$$



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:

Extrapolación fuera de rango

Si queremos saber el peso de un bebé que sólo mide cuarenta centímetros, no podremos utilizar la recta de regresión obtenida. Las características biológicas del bebé, muy diferentes de las que presentan las personas adultas, harán que la relación entre el peso y la altura sea diferente. Deberíamos efectuar un análisis de regresión a partir de una muestra de bebés.

1) 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.

Sentido de la extrapolación

No tiene 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.

2) 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.

8. Modelos de regresión no lineales

Aparte de los modelos lineales, se pueden establecer otros, entre los cuales destaca el exponencial.

Curva en un modelo exponencial

En el modelo lineal hemos ajustado la nube de puntos a una recta de ecuación:

$$y = a + bx$$

En el modelo exponencial queremos ajustar a los puntos una curva de ecuación:

$$y = ka^x \text{ con } a > 0 \text{ y } k > 0$$

El **modelo exponencial** es del tipo:

$$y = ka^x \text{ con } a > 0,\ k > 0$$

donde k y a son valores constantes.

Así, puesto que en el caso lineal es muy fácil ver si puede haber una relación lineal entre las variables a partir del diagrama de dispersión, en el caso exponencial es un poco más difícil.

Ejemplos de relaciones exponenciales

Las relaciones entre la variable tiempo (X) y otras variables (Y) como la población, el número de ordenadores infectados por un virus en los primeros días de contaminación, los precios de algunos productos, etc., son exponenciales.

Para tratarlo, linealizamos el problema, es decir, transformamos las variables de manera que el problema se convierta en lineal. Si en la ecuación $y = ka^x$ tomamos logaritmos $\ln y = \ln(ka^x)$, obtenemos, por aplicación de las propiedades de los logaritmos:

$$\ln y = \ln k + x \ln a$$

Propiedades de los logaritmos

$$\ln ab = \ln a + \ln b$$
$$\ln a^x = x \ln a$$

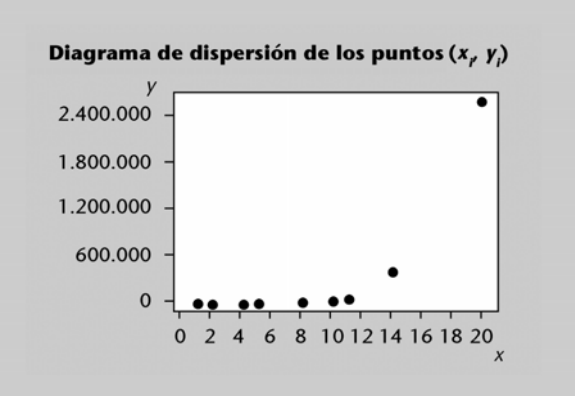
Esta última ecuación nos muestra un modelo lineal entre las variables X y $\ln Y$. Así, si representamos el diagrama de dispersión de los puntos $(x_i, \ln y_i)$ y la nube de puntos presenta una estructura lineal, podemos pensar que entre las variables X e Y hay una relación exponencial.

Ejemplo de la propagación de un virus informático

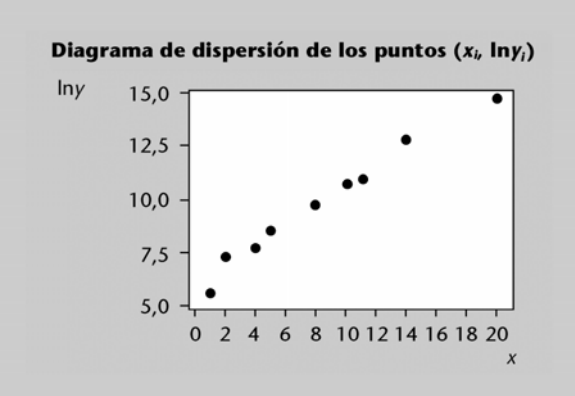
La tabla registra el número de días que han transcurrido desde que se ha detectado un nuevo virus informático y el número de ordenadores infectados en un país.

Número de días x_i	Número de ordenadores infectados y_i	Transformación de Y $\ln y_i$
1	255	5,5413
2	1.500	7,3132
4	2.105	7,6521
5	5.050	8,5271
8	16.300	9,6989
10	45.320	10,7215
11	58.570	10,9780
14	375.800	12,8368
16	1.525.640	14,2379
20	2.577.000	14,7621

El diagrama de dispersión de los puntos siguientes nos hace pensar en la existencia de algún tipo de relación entre las variables que no es lineal. Estudiaremos si se trata de una relación exponencial.



Calculamos el logaritmo de los datos de la variable *Y* y representamos el diagrama de dispersión correspondiente.



Podemos observar que entre las variables *X* y ln *Y* existe una relación lineal; por tanto, entre las variables originales *X* e *Y* habrá una relación exponencial.

Si calculamos la recta de regresión de ln *y* sobre *x*: $\ln\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Obtenemos: $\ln\hat{y} = 5{,}84 + 0{,}482x$, es decir, $\hat{y} = e^{5{,}84 + 0{,}482x}$

De manera que, si queremos estimar el número de ordenadores infectados al cabo de doce días, haremos lo siguiente:

Para $x = 12$: $\ln\hat{y} = 5{,}84 + 0{,}482 \cdot 12 = 11{,}624$

Y tomando exponenciales, podemos aislar $\hat{y}$:

$$\hat{y} = \exp(11{,}624) = 111.747{,}8195$$

Por tanto, al cabo de doce días el número estimado de ordenadores infectados ha sido de 111.748 unidades.

9. Resumen

En esta primera sesión hemos introducido los conceptos de relaciones funcionales y estadísticas, así como el de variables dependientes (o explicadas) y el de variables independientes (o explicativas). A continuación se ha comentado la construcción de un diagrama de dispersión como paso inicial a la hora de bus-

car algún tipo de relación entre dos variables. Si el diagrama nos muestra una estructura lineal, entonces buscamos la línea recta que mejor se ajusta a nuestras observaciones. Lo hacemos mediante el método de los mínimos cuadrados. Hemos puesto de manifesto la importancia de interpretar correctamente los parámetros de la recta. También hemos visto cómo debemos utilizar la recta de regresión para hacer interpolaciones. Finalmente, hemos comentado una relación no lineal tan importante como la relación exponencial y la manera en que podemos transformarla en una lineal.

Ejercicios

1.

El departamento de personal de una empresa informática dedicada a la introducción de datos ha llevado a cabo un programa de formación inicial del personal. La tabla siguiente indica el progreso en pulsaciones por minuto (p.p.m.) obtenido en mecanografía de ocho estudiantes que siguieron el programa y el número de semanas que hace que lo siguen:

Número de semanas	Ganancia de velocidad (p.p.m.)
3	87
5	119
2	47
8	195
6	162
9	234
3	72
4	110

a) Representad el diagrama de dispersión. ¿Creéis que es razonable suponer que existe una relación lineal entre el número de semanas y la ganancia de velocidad?

b) Buscad la recta de regresión. Interpretad los parámetros obtenidos.

c) ¿Qué ganancia de velocidad podemos esperar de una persona que hace siete semanas que va a clase?

2.

Ha salido al mercado un nuevo modelo de grabadora de DVD, un poco más caro que los anteriores, pero con unas prestaciones muy superiores, de manera que la labor de los técnicos de los grandes centros comerciales es muy importante a la hora de presentar este producto al cliente. Con el objetivo de saber si el "número de técnicos comerciales presentes en una tienda" (X) puede tener alguna incidencia en el "número de aparatos vendidos durante una semana" (Y), se observaron quince centros comerciales con los resultados que se muestran a continuación:

$$\sum_{i=1}^{15} x_i = 215\,;\ \sum_{i=1}^{15} x_i^2 = 3.567\,;\ \sum_{i=1}^{15} y_i = 1.700\,;\ \sum_{i=1}^{15} x_i y_i = 28.300$$

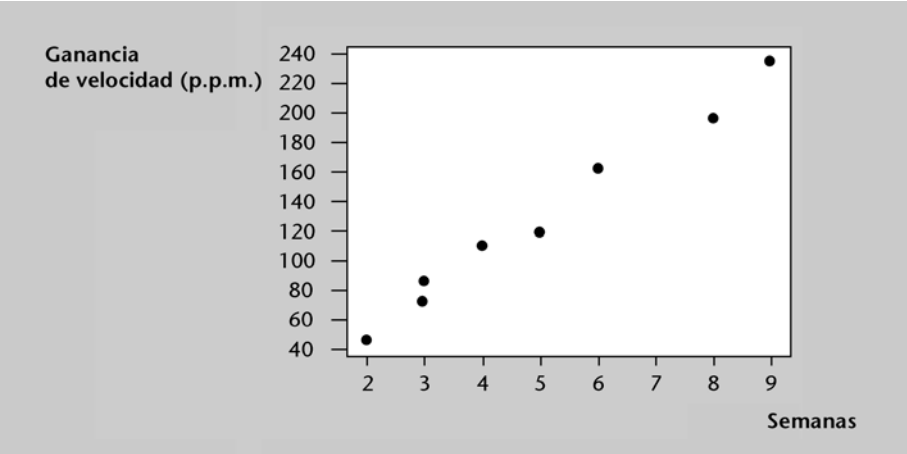
a) Buscad la recta de regresión.

b) ¿Cuál es el número de aparatos que se puede estimar que se venderán en un centro con diecisiete comerciales?

Solucionario

1.

Diagrama de dispersión:



El diagrama de dispersión nos muestra que la relación entre las dos variables es lineal con pendiente positiva, de manera que cuantas más semanas pasan, mayor es la ganancia de velocidad. Por tanto, tiene sentido buscar la recta de regresión. A partir de la tabla de cálculos siguiente:

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})\ (y_i - \bar{y})$
1	3	87	–2	–41,25	4	82,5
2	5	119	0	–9,25	0	0
3	2	47	–3	–81,25	9	243,75
4	8	195	3	66,75	9	200,25
5	6	162	1	33,75	1	33,75
6	9	234	4	105,75	16	423
7	3	72	–2	–56,25	4	112,5
8	4	110	–1	–18,25	1	18,25
Σ	**40**	**1.026**			**44,00**	**1.114,00**

Medias muestrales: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{40}{8} = 5{,}0$ y

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1.206}{8} = 128{,}250$$

Varianza muestral: $s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{44{,}00}{7} = 6{,}286$

Covarianza muestral: $s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1.114{,}00}{7} = 159{,}143$

Ya podemos calcular los coeficientes de la recta de regresión:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{159{,}143}{6{,}286} = 25{,}318 \text{ y}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 128{,}250 - 25{,}318 \cdot 5 = 1{,}659$$

La recta de regresión obtenida es:

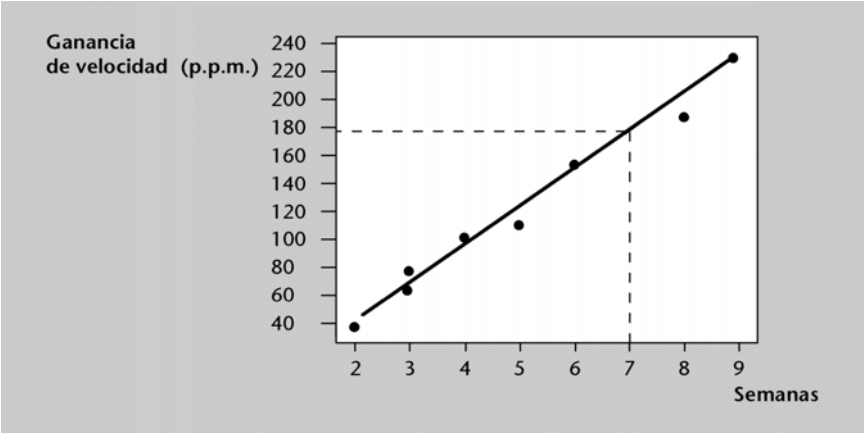
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 1{,}659 + 25{,}318x$$

En este caso la ordenada en el origen no tiene ninguna interpretación con sentido, ya que correspondería a la ganancia de velocidad por cero semanas de clases. Evidentemente, no tiene sentido pensar que sin hacer clases se tiene una ganacia de velocidad de 1,659 p.p.m. La pendiente de la recta sí que nos da una información útil: por cada semana de clase se tiene una ganancia de velocidad de aproximadamente 25 p.p.m.

Para una persona que hace siete semanas que va a clase, podemos calcular la ganancia de velocidad a partir de la recta de regresión, considerando $x = 7$:

$$\hat{y} = 1{,}659 + 25{,}318 \cdot 7 = 178{,}885$$

Es decir, aproximadamente una ganancia de 179 pulsaciones por minuto.



2.

a) Para encontrar la recta de regresión, antes tenemos que encontrar las medias y covarianzas muestrales de las variables X e Y, así como la varianza muestral de X. A partir de los datos que nos da el enunciado:

- Medias muestrales: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{215}{15} = 14{,}333$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1.700}{15} = 113{,}333$$

- Varianza muestral:

Para calcular la varianza muestral a partir de los datos del enunciado, utilizaremos la expresión equivalente:

La deducción de esta fórmula se muestra en el anexo 2 de esta sesión.

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}$$

De manera que:

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1} = \frac{3.567 - 15 \cdot 14{,}333^2}{14} = 34{,}667$$

- Covarianza muestral:

También ahora utilizaremos una nueva expresión para calcular la covarianza muestral:

La deducción de esta fórmula se muestra en el anexo 3 de esta sesión.

$$s_{xy} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1}$$

De manera que:

$$s_{xy} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1} = \frac{28.300 - 15 \cdot 14{,}333 \cdot 113{,}333}{14} = 280{,}952$$

Los parámetros de la recta de regresión son:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{280{,}952}{34{,}667} = 8{,}104$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 113{,}333 - 8{,}104 \cdot 14{,}333 = -2{,}829$$

La recta de regresión obtenida es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2{,}829 + 8{,}104x$$

b) Para un centro con diecisiete comerciales, podemos estimar las ventas de aparatos de DVD mediante la recta de regresión obtenida:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -2{,}829 + 8{,}104 \cdot 17 = 134{,}939$$

Por tanto, en un centro con diecisiete comerciales se habrán vendido aproximadamente unos 135 aparatos.

Anexos

Anexo 1

Resolución del sistema de ecuaciones normales:

$$\begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

A partir de la primera ecuación del sistema:

$$\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = n\bar{y} - n\beta_0 - n\beta_1 \bar{x} = 0$$

Dividiendo por n: $\bar{y} = \beta_0 + \beta_1 \bar{x}$ y aislando la β_0: $\beta_0 = \bar{y} - \beta_1 \bar{x}$

De la segunda ecuación del sistema:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 = \sum_{i=1}^{n} x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i = n\beta_0 \bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2$$, pero tenemos en cuenta que: $\beta_0 = \bar{y} - \beta_1 \bar{x}$

entonces $\sum_{i=1}^{n} x_i y_i = n(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 = n\bar{x}\bar{y} - \beta_1 n\bar{x}^2 + \beta_1 \sum_{i=1}^{n} x_i^2$

Aislando β_1:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

podemos dar una expresión equivalente a partir de la definición de varianza muestral:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s_x^2(n-1) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

y de la definición de covarianza muestral:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy}(n-1) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + n\bar{x}\bar{y} =$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Teniendo en cuenta la varianza y la covarianza, podemos expresar los parámetros de la recta de regresión de la manera siguiente:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \quad \text{i} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Anexo 2

Varianza muestral:

Podemos deducir a partir de la fórmula de su definición:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

una expresión equivalente desarrollando el cuadrado del numerador:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\bar{x}x_i + \sum_{i=1}^{n} (\bar{x})^2 =$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n(\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n(\bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - n(\bar{x})^2$$

De manera que:

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2 \right) - n(\bar{x})^2}{n-1}$$

Anexo 3

Covarianza muestral:

A partir de la definición de la covarianza:

$$s_{xy} = \frac{\sum_{i=1}^{''} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

si desarrollamos el producto del sumatorio del numerador:

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-\bar{x}y_i+\bar{x}\bar{y}) = \sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\bar{y}-\sum_{i=1}^{n}\bar{x}y_i+\sum_{i=1}^{n}\bar{x}\bar{y} =$$

$$= \sum_{i=1}^{n}x_iy_i-\bar{y}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}y_i+\bar{x}\bar{y}n = \left(\sum_{i=1}^{n}x_iy_i\right)-n\bar{x}\bar{y}$$

De manera que:

$$s_{xy} = \frac{\left(\sum_{i=1}^{n}x_iy_i\right)-n\bar{x}\bar{y}}{n-1}$$

La calidad del ajuste

1. Introducción

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.

2. 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 la proporción de varianza explicada por la recta de regresión, es decir:

$$R^2 = \frac{\text{Varianza explicada por la recta de regresi??n}}{\text{Varianza total de los datos}}$$

Notación

La varianza explicada por la recta de regresión es la varianza de los valores estimados $\hat{y}_i$.

La varianza total de los datos es la varianza de los valores observados y_i.

Buscaremos una expresión que nos permita calcular el coeficiente de determinación. Veremos que la varianza de las observaciones se puede descomponer en dos términos: la varianza que queda explicada por el modelo de regresión lineal y una varianza debida a los residuos.

A partir de la definición de residuos (e_i) de la regresión como la diferencia entre los valores observados (y_i) y los valores estimados ($\hat{y}_i$) por la recta de regresión:

Notación

Llamaremos indistintamente *valores estimados* o *valores predichos* ($\hat{y}_i$) a los obtenidos mediante la recta de regresión.

$$e_i = y_i - \hat{y}_i$$

podemos escribir:

$$y_i = \hat{y}_i + e_i$$

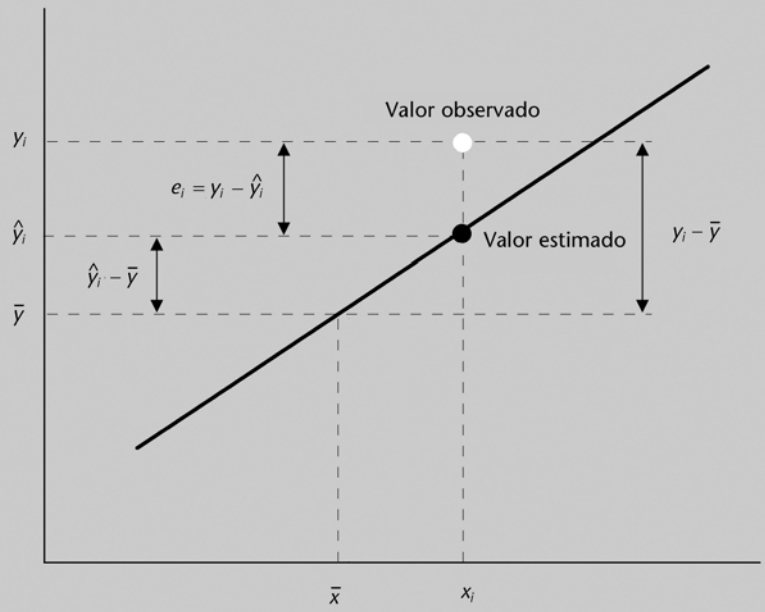
Si ahora restamos a los dos miembros de esta igualdad la media de las observaciones y_i, obtenemos una expresión que nos relaciona las desviaciones con

respecto a la media de las observaciones con las desviaciones con respecto a la media de los valores estimados.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

Representaremos gráficamente las desviaciones con respecto a la media, las observaciones y los valores estimados con la recta de regresión.

Desviaciones con respecto a la media de las observaciones y de los valores estimados con la recta de regresión



Observación

La recta de regresión pasa por $(\bar{x}, \bar{y})$.

Elevando al cuadrado y sumando todos los valores, se puede demostrar que:

Esta deducción matemática se encuentra desarrollada en el anexo 1 de esta sesión.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

Dando nombres a estas cantidades, podemos escribir de una manera más compacta esta expresión:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = SCT$$ Suma de cuadrados totales

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SCR$$ Suma de cuadrados de la regresión

$$\sum_{i=1}^{n} e_i^2 = SCE$$ Suma de cuadrados de los errores

Así, tenemos que:

$$SCT = SCR + SCE$$

Podemos interpretar esta última expresión en el sentido de que la varianza total observada (*SCT*) en la variable *Y* se descompone en dos términos: la varianza explicada por el modelo de regresión lineal (*SCR*) más la varianza que no queda explicada por el modelo, es decir, la varianza de los residuos (*SCE*).

Entonces podemos escribir la definición del **coeficiente de determinación** de esta manera:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

o también,

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Observando estas expresiones, es fácil apreciar las características de este coeficiente. Siempre será: $0 \leq R^2 \leq 1$, de manera que:

- $R^2 = 1$ cuando el ajuste es perfecto, es decir, cuando todos los puntos se encuentran sobre la recta de regresión. En este caso los residuos son cero y la suma de sus cuadrados también y, por tanto, *SCR* = *SCT*.
- $R^2 = 0$ denota la inexistencia de relación entre las variables *X* e *Y*. En este caso la suma de residuos es máxima y tenemos que *SCE* = *SCT*.
- 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.

Observación

Un coeficiente de determinación diferente de cero no significa que haya relación lineal entre las variables. Por ejemplo, $R^2 = 0,5$ sólo nos dice que el 50% de la varianza de las observaciones queda explicado por el modelo lineal.

Ejemplo de las alturas y los pesos

Consideremos las observaciones de los pesos (kg) y las alturas (cm) de un conjunto de diez personas: el individuo 1 tiene 161 cm de altura y 63 kg de peso, el individuo 2 tiene 152 cm de altura y 56 kg de peso, etc.

Individuos (y)	1	2	3	4	5	6	7	8	9	10
Altura (x_i)	161	152	167	153	161	168	167	153	159	173
Peso (y_i)	63	56	77	49	72	62	68	48	57	67

A partir de la recta de regresión:

$$\hat{y} = -96,1121 + 0,979009x$$

podemos calcular los valores estimados y los residuos. Es muy conveniente, por comodidad, disponer de los datos y los cálculos en forma de tabla; en concreto, construiremos una tabla de cálculos del coeficiente de determinación:

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	e_i	e_i^2
1	161	63	61,51	1,10	1,21	–0,39	0,15	1,49	2,23
2	152	56	52,70	–5,90	34,81	–9,20	84,69	3,30	10,91
3	167	77	67,38	15,10	228,01	5,48	30,06	9,62	92,50
4	153	49	53,68	–12,90	166,41	–8,22	67,63	–4,68	21,87
5	161	72	61,51	10,10	102,01	–0,39	0,15	10,49	110,07
6	168	62	68,36	0,10	0,01	6,46	41,75	–6,36	40,47
7	167	68	67,38	6,10	37,21	5,48	30,06	0,62	0,38
8	153	48	53,68	–13,90	193,21	–8,22	67,63	–5,68	32,22
9	159	57	59,55	–4,90	24,01	–2,35	5,52	–2,55	6,50
10	173	67	73,26	5,10	26,01	11,36	128,97	–6,26	39,14
Σ		**619**			**812,90**		**456,61**		**356,29**

Tenemos que:

$$SCR = 456{,}61$$

$$SCT = 812{,}90$$

Por tanto, tenemos un coeficiente de determinación:

$$R^2 = 456{,}61 \,/\, 812{,}90 = 0{,}5617$$

Con este ejemplo podemos comprobar la equivalencia entre las dos expresiones obtenidas antes por el coeficiente de determinación. A partir de la suma de los cuadrados de los residuos:

$$SCE = 356{,}29$$

tenemos para el coeficiente de determinación:

$$R^2 = 1 - (\, 356{,}29 \,/\, 812{,}90 \,) = 1 - 0{,}4383 = 0{,}5617$$

Evidentemente, coinciden los resultados.

Hemos obtenido un coeficiente de determinación $R^2 = 0{,}5617$ que nos informa de que el modelo de regresión lineal sólo nos explica el 56,17% de la varianza de las observaciones.

3. El coeficiente de correlación muestral, *r*

A partir del diagrama de dispersión podemos ver si hay algún tipo de relación entre dos variables X e Y.

Se suele decir que X e Y tienen una **relación positiva** si los valores grandes de X están aparejados con valores grandes de Y y valores pequeños de X, con valores pequeños de Y. De manera análoga, se dice que X e Y tienen una **relación negativa** si los valores grandes de X están aparejados con los valores pequeños de Y y los pequeños de X, con grandes de Y.

Ahora queremos medir estas relaciones de forma numérica. La covarianza muestral entre dos variables *X* e *Y*:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \text{ W}$$

nos puede servir para medir estas relaciones positivas y negativas entre las variables *X* e *Y*.

- Si tenemos una relación positiva, entonces la mayoría de los puntos de coordenadas $((x_i - \bar{x}), (y_i - \bar{y}))$ estarán en el primer y tercer cuadrante en que $(x_i - \bar{x})(y_i - \bar{y}) \geq 0$, de manera que contribuirán de forma positiva a la suma.

- Si tenemos una relación negativa, entonces la mayoría de los puntos de coordenadas $((x_i - \bar{x}), (y_i - \bar{y}))$ estarán en el segundo y cuarto cuadrante, en los que $(x_i - \bar{x})(y_i - \bar{y}) \leq 0$, de manera que contribuirán de forma negativa a la suma.

- Si, por el contrario, no existe ningún tipo de relación positiva o negativa, la covarianza será una cantidad pequeña al encontrarse todos los puntos aproximadamente igual repartidos por los cuatro cuadrantes, cosa que compensa de forma aproximada las cantidades positivas y negativas del sumatorio.

Observad la figura de los ejemplos de diagramas de dispersión en el apartado 3 de la sesión "El modelo de regresión simple" de este módulo.

Esquema de relaciones entre *X* e *Y*

La covarianza presenta el gran inconveniente de depender de las unidades de las variables que estudiamos.

Definimos el **coeficiente de correlación muestral** como:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Unidades del coeficiente de correlación muestral

Al dividir la covarianza por las desviaciones típicas de *X* y de *Y*, hemos conseguido una medida adimensional que no depende de las unidades de las variables.

El coeficiente de correlación se caracteriza por $-1 \leq r \leq 1$, de manera que:

- $r = 1$ o $r = -1$ cuando haya una asociación lineal exacta entre las variables (en el primer caso positiva y en el segundo, negativa).

- $-1 < r < 1$ cuando la relación entre las variables no sea lineal de forma exacta.

- Para los otros valores siempre se formula la misma pregunta: ¿a partir de qué valor de *r* podemos decir que la relación entre las variables es fuerte? Una regla razonable es decir que la relación es débil si $0 < |r| < 0{,}5$; fuerte si $0{,}8 < |r| < 1$, y moderada si tiene otro valor.

Para calcular el coeficiente de correlación muestral, podemos utilizar la misma tabla de cálculos que para obtener la recta de regresión. Lo ilustraremos con el ejemplo de las alturas y los pesos.

Ejemplo de las alturas y los pesos

Consideremos de nuevo el ejemplo de los pesos y las alturas. Buscaremos el coeficiente de correlación. Antes tendremos que calcular la covarianza y las varianzas muestrales.

i	x_i	y_i	$\bar{x}-x_i$	$\bar{y}-y_i$	$(\bar{x}-x_i)^2$	$(\bar{y}-y_i)^2$	$(\bar{x}-x_i)(\bar{y}-y_i)$
1	161	63	0,4	–1,1	0,16	1,21	–0,44
2	152	56	9,4	5,9	88,36	34,81	55,46
3	167	77	–5,6	–15,1	31,36	228,01	84,56
4	153	49	8,4	12,9	70,56	166,41	108,36
5	161	72	0,4	–10,1	0,16	102,01	–4,04
6	168	62	–6,6	–0,1	43,56	0,01	0,66
7	167	68	–5,6	–6,1	31,36	37,21	34,16
8	153	48	8,4	13,9	70,56	193,21	116,76
9	159	57	2,4	4,9	5,76	24,01	11,76
10	173	67	–11,6	–5,1	134,56	26,01	59,16
Σ	**1.614**	**619**			**476,40**	**812,90**	**466,40**

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{466,40}{10-1} = 51,822$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{476,40}{10-1} = 52,933 \text{ de manera que } s_x = 7,276$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{812,90}{10-1} = 90,322 \text{ de manera que } s_y = 9,504$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{51,822}{7,276 \cdot 9,504} = 0,749$$

El coeficiente de correlación lineal obtenido por nuestro ejemplo del peso y la altura es $r = 0,749$, que nos informa de la existencia de una moderada relación entre estas dos variables, así como de que, a medida que la altura crece, el peso también lo hace (ya que es positivo).

4. 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 $R^2 = r^2$, como fácilmente podemos comprobar.

Observación

En la regresión lineal múltiple ya no tendremos la igualdad $R^2 = r^2$.

Comprobación de que en regresión lineal simple $R^2 = r^2$

A partir de la ecuación del coeficiente de correlación:

$$r = \frac{s_{xy}}{s_x s_y}$$

y de la ecuación de la pendiente de la recta de regresión:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

tenemos la relación siguiente:

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

Por otra parte, tenemos el otro parámetro de la recta de regresión: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ y la ecuación de los valores estimados: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. De estas dos expresiones podemos escribir:

$$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

Aplicando todas estas relaciones a la ecuación del coeficiente de determinación, y a partir de la definición de varianza muestral, tenemos:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \hat{\beta}_1^2\frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = r^2 \underbrace{\frac{s_y^2\sum(x_i - \bar{x})^2}{s_x^2\sum(y_i - \bar{y})^2}}_{1} = r^2$$

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.

También 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 $r = -0{,}9$.

Ejemplo de las alturas y los pesos

Podemos comprobar la relación entre el coeficiente de determinación y el coeficiente de correlación con los resultados de nuestro ejemplo.

Hemos obtenido: $R^2 = 0{,}5617$ y $r = 0{,}749$.

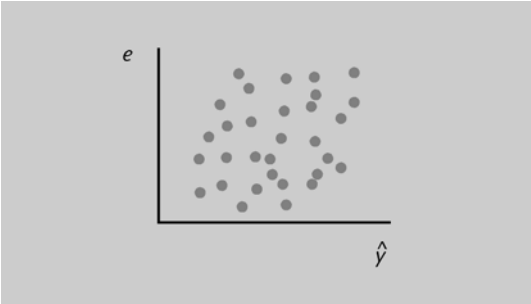
De manera que $r^2 = 0{,}749^2 = 0{,}561$.

5. Diagnóstico de la regresión: análisis de los residuos

Una vez hecho el ajuste de un modelo de regresión lineal a nuestros datos muestrales, hay que efectuar el análisis de los residuos.

Este análisis, que a continuación comentaremos de forma breve y muy intuitiva, nos servirá para hacer un diagnóstico de nuestro modelo de regresión.

El análisis de los residuos consiste en ver la distribución de los residuos. Esto lo haremos gráficamente representando un diagrama de dispersión de los puntos $(\hat{y}_i, e_i)$, es decir, sobre el eje de las abscisas representamos el valor estimado $\hat{y}_i$ y sobre el eje de ordenadas, el valor correspondiente del residuo, es decir, $e_i = y_i - \hat{y}_i$. Veamos un ejemplo:



Si el modelo lineal obtenido se ajusta bien a los datos muestrales, entonces la nube de puntos $(\hat{y}_i, e_i)$ no debe mostrar ningún tipo de estructura.

Lo ilustraremos con un ejemplo ya clásico en la bibliografía: el **ejemplo de Anscombe** (1973). A partir de las tablas de datos que se muestran a continuación discutiremos cuatro casos:

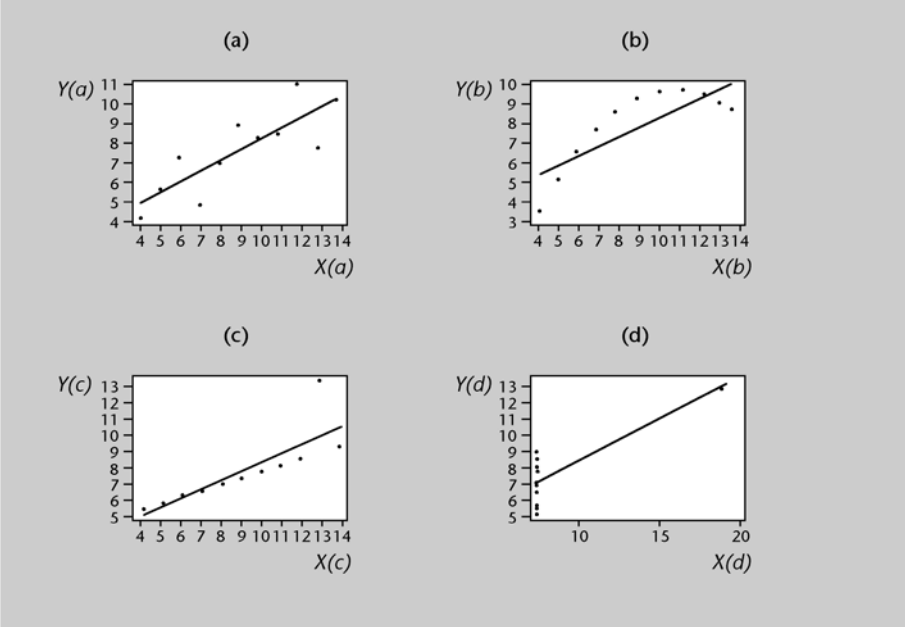
Lectura complementaria

Encontraréis el ejemplo de Anscombe en el artículo siguiente:

T.W. Anscombe (1973). "Graphs in Statistical Analysis". *The American Statistician* (núm. 27, pág. 17-21).

Caso (a)		Caso (b)		Caso (c)		Caso (d)	
X(a)	*Y(a)*	*X(b)*	*Y(b)*	*X(c)*	*Y(c)*	*X(d)*	*Y(d)*
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Dibujaremos a continuación el diagrama de dispersión y las rectas de regresión en el ejemplo de Anscombe.

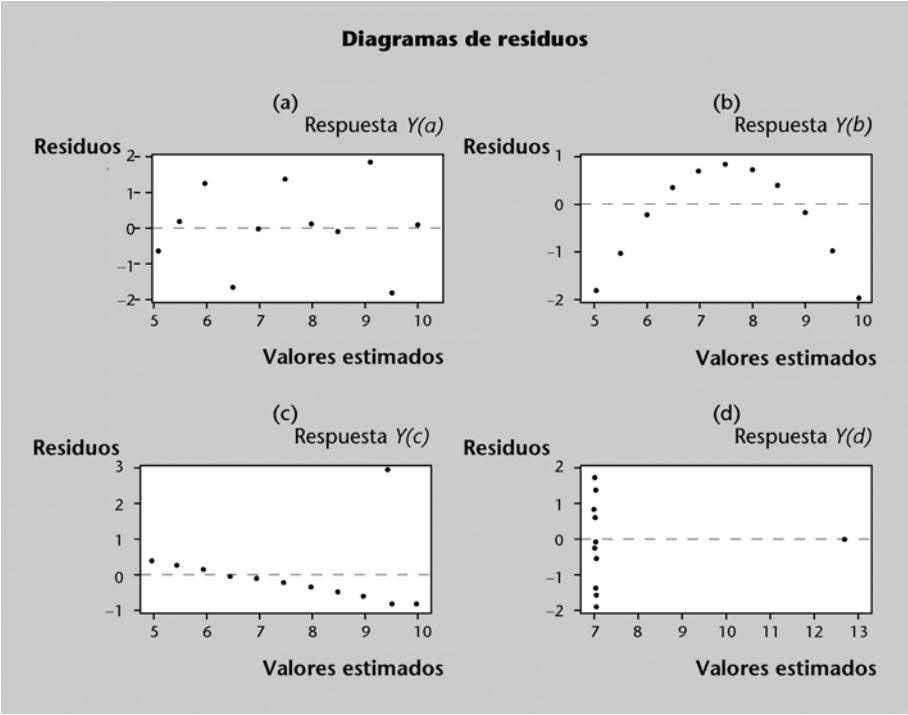


Si hacemos la regresión de Y sobre X, en los cuatro casos obtenemos la misma recta:

$$\hat{y} = 3 + 0{,}5x$$

El coeficiente de correlación es el mismo para las cuatro con valor $r = 0{,}82$.

Si ahora hacemos el estudio de los residuos tal como hemos indicado antes, tenemos la representación de los siguientes diagramas de residuos:



Podemos observar que de las cuatro, sólo la primera no presenta ningún tipo de estructura sobre la nube de puntos, de manera que sólo tendría sentido la regresión hecha sobre la muestra (a).

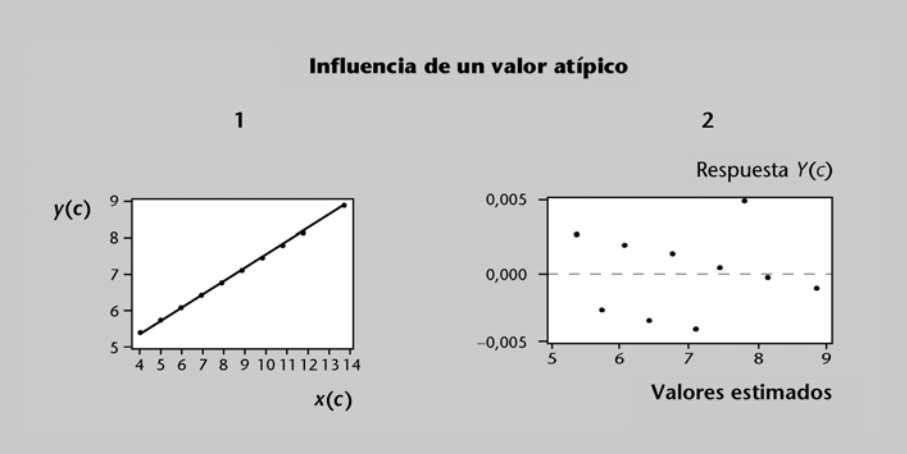
Consideremos a continuación el caso (b) del diagrama de dispersión. En éste se observa un comportamiento curvilíneo que nos hace pensar que un ajuste lineal no sería el más conveniente. Esto se manifiesta de forma mucho más evidente en el diagrama de residuos.

Si consideramos la muestra (c), en el diagrama de dispersión podemos observar la presencia del valor atípico (13, 12,74) que nos ha hecho ajustar un modelo erróneo al resto de las observaciones, ya que si lo eliminamos, entonces obtenemos una recta de regresión diferente:

$$\hat{y} = 4{,}01 + 0{,}345x$$

y un coeficiente de correlación $r = 1$. Podemos observar todos los puntos sobre la recta de regresión.

El diagrama de los residuos también nos sugiere un buen modelo de regresión para la muestra resultante de eliminar el valor atípico. A continuación representamos el diagrama de dispersión y el diagrama de residuos.



Influencia de un valor atípico

En la muestra (c) hemos eliminado el valor atípico y hemos representado de nuevo el diagrama de dispersión y la recta de regresión **1** y el diagrama de residuos **2**.

Finalmente, en la muestra (d) la pendiente está determinada por un único valor. Tampoco es un modelo demasiado fiable.

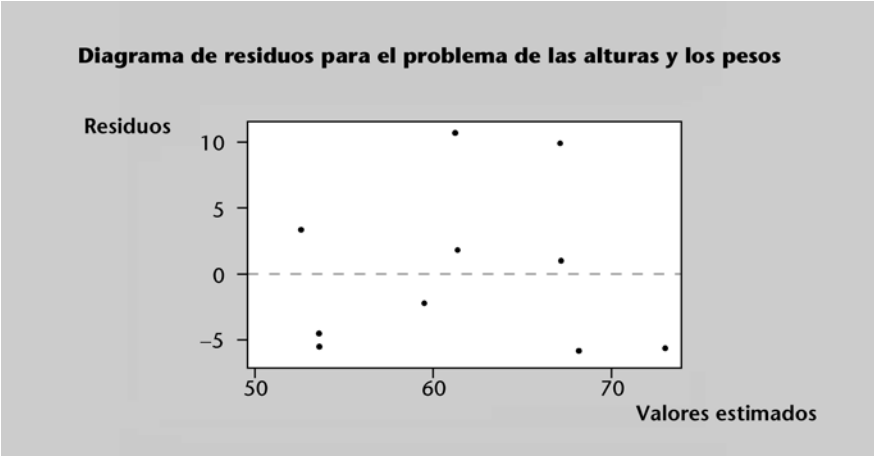
Ejemplo de las alturas y los pesos

Un último ejemplo que todavía podemos examinar es el de la relación de las alturas y pesos. A partir de los datos de la tabla ya vista:

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	e_i	e_i^2
1	161	63	61,51	1,10	1,21	–0,39	0,15	1,49	2,23
2	152	56	52,70	–5,90	34,81	–9,20	84,69	3,30	10,91
3	167	77	67,38	15,10	228,01	5,48	30,06	9,62	92,50
4	153	49	53,68	–12,90	166,41	–8,22	67,63	–4,68	21,87
5	161	72	61,51	10,10	102,01	–0,39	0,15	10,49	110,07

i	x_i	y_i	$\hat{y}_i$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$\hat{y}_i-\bar{y}$	$(\hat{y}_i-\bar{y})^2$	e_i	e_i^2
6	168	62	68,36	0,10	0,01	6,46	41,75	–6,36	40,47
7	167	68	67,38	6,10	37,21	5,48	30,06	0,62	0,38
8	153	48	53,68	–13,90	193,21	–8,22	67,63	–5,68	32,22
9	159	57	59,55	–4,90	24,01	–2,35	5,52	–2,55	6,50
10	173	67	73,26	5,10	26,01	11,36	128,97	–6,26	39,14
Σ		**61,9**			**812,90**		**456,61**		**356,29**

es fácil representar el diagrama de residuos:



No podemos observar ningún tipo de estructura en la representación; por tanto, podemos concluir que el modelo de regresión obtenido es un buen modelo para explicar la relación entre las dos variables.

6. Resumen

En esta segunda sesión hemos introducido una medida numérica de la bondad del ajuste de la recta de regresión en las observaciones. Esta medida se obtiene con el coeficiente de determinación R^2. Se ha discutido la interpretación de los valores que puede tomar. A continuación hemos visto el coeficiente de correlación muestral, r, que nos mide el grado de asociación entre dos variables. Hemos comprobado que en la regresión lineal simple R^2 y r coinciden. Finalmente, hemos comentado la importancia de analizar los residuos para hacer un diagnóstico del modelo lineal obtenido.

Ejercicios

1.

Una tienda de ordenadores llevó a cabo un estudio para determinar la relación entre los gastos de publicidad semanal y las ventas. Se obtuvieron los datos siguientes:

Gastos en publicidad (× 1.000 €)	Ventas (× 100.000 €)
40	380
25	410
20	390
22	370
31	475
52	450
40	500
20	390
55	575
42	520

Con estos datos se han obtenido las cantidades siguientes:

$$\sum_{i=1}^{10} x_i = 347 \qquad \sum_{i=1}^{10} y_i = 4.460 \qquad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 6.018$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1.522,1 \qquad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 43.590,0$$

$$\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 23.793,66$$

Y la recta de regresión: $\hat{y} = 308,88 + 3,95x$.

A partir de toda esta información, calculad el coeficiente de determinación y el coeficiente de correlación.

2.

El departamento de personal de una empresa informática dedicada a la introducción de datos ha llavado a cabo un programa de formación inicial del personal. La tabla siguiente indica el progreso obtenido en mecanografía de ocho estudiantes que siguieron el programa y el número de semanas que hace que lo siguen:

Número de semanas	Ganancia de velocidad (p.p.m.)
3	87
5	119
2	47
8	195
6	162

Número de semanas	Ganancia de velocidad (p.p.m.)
9	234
3	72
4	110

La recta de regresión calculada a partir de estos datos es:

$$\hat{y} = 1{,}659 + 25{,}318x$$

a) Calculad el coeficiente de determinación.

b) Haced un análisis de los residuos y comentadlo.

Solucionario

1.

Calculamos el coeficiente de determinación a partir de la expresión:

$$R^2 = \frac{SCR}{SCT}$$

El enunciado del problema nos proporciona estos datos, ya que:

- La suma de los cuadrados de la regresión es: $SCR = \sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 23.793{,}66$
- Y la suma de los cuadrados totales es: $SCT = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 43.590{,}0$

De manera que: $R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{10} (y_i - \bar{y})^2} = \frac{23.793{,}66}{43.590{,}0} = 0{,}5458$

Resultado que podemos interpretar como que el modelo de regresión lineal explica el 54,58% de la variabilidad de las ventas.

A partir de este valor podemos calcular el coeficiente de correlación teniendo en cuenta que:

$$R^2 = r^2$$

De manera que el coeficiente de correlación es la raíz cuadrada del coeficiente de determinación con el mismo signo que la pendiente de la recta de regresión.

La recta de regresión es: $\hat{y} = 308{,}8 + 3{,}95x$. La pendiente es positiva, de manera que tenemos una relación positiva entre los gastos en publicidad y ventas. Cuanto más se invierte en publicidad, más se vende.

Así pues, el coeficiente de correlación es:

$$r = +\sqrt{R^2} = +\sqrt{0{,}5458} = 0{,}7388$$

2.

a) Lo primero que haremos será construir la tabla de cálculos:

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
1	3	87	77,61	–41,25	1.701,56	–50,64	2.564,11
2	5	119	128,25	–9,25	85,56	0,00	0,00
3	2	47	52,30	–81,25	6.601,56	–75,96	5.769,16
4	8	195	204,20	66,75	4.455,56	75,95	5.768,86
5	6	162	153,57	33,75	1.139,06	25,32	640,95
6	9	234	229,52	105,75	11.183,06	101,27	10.255,82
7	3	72	77,61	–56,25	3.164,06	–50,64	2.564,11
8	4	110	102,93	–18,25	333,06	–25,32	641,05
Σ		**1.026**			**28.663,50**		**28.204,05**

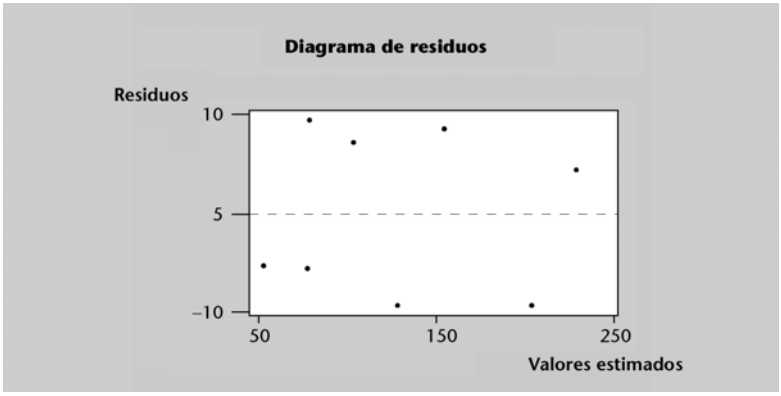
$$SCR = 28.204,05 \quad SCT = 28.663,50$$
$$R^2 = 28.204,05 \,/\, 28.663,50 = 0,9920$$

El modelo de regresión lineal explica el 99,20% de la varianza de la muestra. Tenemos bondad en el ajuste.

b) Para hacer el análisis de los residuos, en primer lugar calcularemos los residuos y después haremos la representación gráfica.

i	x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	3	87	77,61	9,39
2	5	119	128,25	–9,25
3	2	47	52,30	–5,30
4	8	195	204,20	–9,20
5	6	162	153,57	8,43
6	9	234	229,52	4,48
7	3	72	77,61	–5,61
8	4	110	102,93	7,07

Si representamos el valor del residuo frente al valor ajustado, tenemos el diagrama de residuos siguiente:



No observamos ningún tipo de forma determinada en los puntos de esta gráfica.

Este resultado, junto con el elevado coeficiente de determinación, nos hace llegar a la conclusión de que el modelo lineal es adecuado para tratar este problema.

Anexos

Anexo 1

Descomposición de la suma de cuadrados total

A continuación veremos que la suma de cuadrados total de las observaciones (*SCT*) se puede expresar de la manera siguiente:

$$SCT = SCR + SCE$$

donde:

- *SCR* es la suma de cuadrados de la regresión.
- *SCE* es la suma de cuadrados de los residuos.

A partir de la definición de residuos de la regresión como la diferencia entre los valores observados y los valores estimados por la recta de regresión:

$$e_i = y_i - \hat{y}_i$$

Podemos escribir:

$$y_i = \hat{y}_i + e_i$$

Y si ahora restamos a los dos miembros de esta igualdad la media de las observaciones y_i, obtenemos una expresión que nos relaciona las desviaciones con respecto a la media, las observaciones y los valores estimados:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

Elevando al cuadrado y sumando todos los valores:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(\hat{y}_i - \bar{y}) + e_i]^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i + \sum_{i=1}^{n} e_i^2$$

$$2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i = \sum_{i=1}^{n} \hat{y}_i e_i - \sum_{i=1}^{n} \bar{y} e_i = \sum_{i=1}^{n} \hat{y}_i e_i - \bar{y}\sum_{i=1}^{n} e_i = 0 + 0 = 0$$

$$0 \qquad 0$$

Por tanto, es suficiente con ver que $\sum_{i=1}^{n} \hat{y}_i e_i = 0$ y $\sum_{i=1}^{n} e_i = 0$

Observamos que a partir de las ecuaciones normales:

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i$$

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^{n} e_i x_i$$

Y, por tanto:

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 - \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i = 0$$

Hemos demostrado así que:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

Si denominamos:

$\sum_{i=1}^{n} (y_i - \bar{y})^2 = SCT$ **S**uma de **C**uadrados **T**otales.

$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SCR$ **S**uma de **C**uadrados de la **Regresión**.

$\sum_{i=1}^{n} e_i = SCE$ **S**uma de **C**uadrados de los **E**rrores.

Tenemos que: $SCT = SCR + SCE$.

Inferencia en la regresión

1. Introducción

En otras sesiones nos hemos preocupado de estudiar la relación lineal entre dos variables X e Y a partir de los valores observados en una muestra. Si en el diagrama de dispersión observábamos una relación lineal, entonces calculábamos la recta que mejor se ajustaba a nuestros datos haciendo que la suma de los cuadrados de los residuos fuese mínima. Es la llamada *recta de regresión*.

Ahora cambiaremos el punto de vista y pensaremos que esta muestra de observaciones proviene de una población. Nos preguntamos si esta relación lineal se puede extender de alguna manera a toda la población.

2. El modelo de regresión en la población

Modelo de regresión lineal

Es muy importante tener presente que, para un mismo valor de la variable X, se pueden observar diferentes valores de la variable Y, es decir, asociado a cada valor de X no hay un único valor de Y, sino una distribución de frecuencias de Y. Esto se debe al hecho de que Y no sólo depende de X, sino también de otros factores difícilmente cuantificables o simplemente desconocidos. La influencia de este conjunto de factores es la que determina que la relación entre X e Y sea estadística y no determinista. Todos estos factores son los responsables de los errores o residuos.

El peso depende de la altura y de otros factores

En el ejemplo de la relación entre el peso y la altura de las personas, es evidente que existen muchos factores, como pueden ser aspectos genéticos, la actividad física, la alimentación, etc., que hacen que una persona de una determinada altura tenga un peso u otro. Para una altura fija, de por ejemplo 170 cm, no todas las personas tienen el mismo peso.

Dada una muestra de observaciones (x_i, y_i), $y = 1, \ldots, n$ de individuos de una población, ya sabemos encontrar la recta de regresión lineal $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Si tenemos en cuenta que llamábamos *residuo* o *error* a la diferencia entre el valor observado y el valor estimado $e_i = y_i - \hat{y}_i$, para una observación y_i, podemos escribir: $y_i = \hat{y}_i + e_i$, es decir:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x + e_i$$

Podemos hacer lo mismo con varias muestras de esta misma población.

Ejemplo de las alturas y los pesos

Consideremos las observaciones de los pesos (kg) y alturas (cm) de tres muestras de alumnos de la UOC y las rectas de regresión correspondientes:

Muestra $j = 1$										
Individuos	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
Altura (x_{ij})	161	152	167	153	161	168	167	153	159	173
Peso (y_{ij})	63	56	77	49	72	62	68	48	57	67

La recta de regresión correspondiente es: $\hat{y} = -96{,}112 + 0{,}979x$.

Muestra $j = 2$								
Individuos	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
Altura (x_{ij})	161	152	167	153	161	168	167	153
Peso (y_{ij})	63	56	77	49	72	62	68	48

La recta de regresión correspondiente es: $\hat{y} = -82{,}614 + 1{,}029x$.

Muestra $j = 3$									
Individuos	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$
Altura (x_{ij})	161	152	167	153	161	168	167	153	159
Peso (y_{ij})	63	56	77	49	72	62	68	48	57

La recta de regresión correspondiente es: $\hat{y} = -98{,}582 + 0{,}94x$.

Observamos que los valores obtenidos para cada coeficiente son relativamente similares:

$\hat{\beta}_0$: –96,112; –82,614; –98,528

$\hat{\beta}_1$: 0,979; 1,029; 0,945

Podemos pensar que si recogemos más muestras de la misma población, iremos obteniendo coeficientes parecidos a éstos.

Ahora el objetivo es dar un modelo para todos los individuos de la población. Éste vendrá dado por una expresión análoga a las encontradas por las muestras.

Lllamamos **modelo de regresión lineal para la población** a:

$$y_i = \beta_0 + \beta_1 x + e_i$$

Notación

No ponemos los "sombreros" sobre los parámetros para indicar que ahora se trata de la recta de regresión para la población.

Para encontrar este modelo para la población, deberíamos estudiar a todos los individuos que la componen. Esto es prácticamente imposible, de manera que deberemos estimarla a partir de los resultados calculados para una muestra. Es decir, deberemos hacer inferencia estadística.

Antes de continuar, tenemos que hacer dos suposiciones muy importantes:

1) Los errores se distribuyen según una distribución normal de media cero y varianza σ^2.

2) Los errores son independientes.

Distribución de los errores en la realidad

La distribución de los errores es diferente para diferentes valores de *X*. Por ejemplo, las personas que miden cerca de 160 cm varían menos su peso que las personas que miden 185 cm. De todos modos, aceptaremos la suposición de que siempre son iguales.

Con estas suposiciones tenemos que:

1) Por cada valor fijo x de X obtenemos una distribución de valores y de la variable Y. Y podemos calcular la media o la esperanza matemática de cada una de estas distribuciones:

$$\mu_x = E(Y|x) = E(\beta_0 + \beta_1 x + e) = \beta_0 + \beta_1 x + E(e) = \beta_0 + \beta_1 x$$

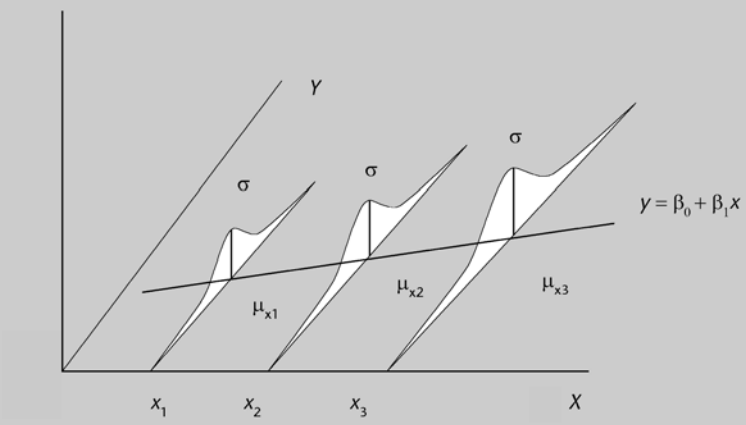
2) También podemos calcular su varianza:

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + e) = Var(\beta_0 + \beta_1 x) + Var(e) = 0 + \sigma^2 = \sigma^2$$

Cada distribución de valores de Y tiene la misma varianza σ^2, que es la varianza de los residuos.

En el gráfico vemos la recta de regresión lineal para la población.

Recta de regresión lineal para la población



Distribución de las medias

El primer resultado nos dice que estas medias se encuentran situadas sobre una recta.

Es importante tener presente que para tener bien determinado el modelo de regresión para la población, debemos conocer tres parámetros: β_0, β_1 y σ^2.

Estos parámetros desconocidos se tienen que estimar a partir de una muestra de la población.

Como se ve en la sesión "El modelo de regresión simple", los parámetros de la recta se estiman por el método de los mínimos cuadrados. Este método determina aquellos valores de los parámetros que hacen mínima la suma de los cuadrados de los residuos:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\,;\ \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

De manera que $\hat{\beta}_0$ y $\hat{\beta}_1$ son los valores estimados (o "estimadores") de los parámetros β_0 y β_1 de la población. Y la recta que mejor se ajusta a los datos es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Valor medio

Debemos interpretar:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

como la estimación del valor medio de la distribución *Y* para un valor fijo $X = x_i$.

Todavía nos falta estimar la varianza de los errores aleatorios, σ^2. Este término refleja la variación aleatoria en torno a la auténtica recta de regresión.

Si consideramos los residuos de la regresión como estimaciones de los valores de los errores aleatorios, entonces podemos estimar su varianza a partir de la varianza de los residuos:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Terminología

Habitualmente, s^2 se denomina *varianza residual*.

Hemos dividido la suma de las desviaciones al cuadrado por $n - 2$, no por $n - 1$. Esto se debe a que estimamos la media de *Y* para un valor dado de *X* con una fórmula que contiene dos parámetros estimados a partir de los datos de la muestra ($\hat{\beta}_0$ y $\hat{\beta}_1$). Diremos que "hemos perdido dos grados de libertad".

Pérdida de grados de libertad

El razonamiento es el mismo que el que hacemos al justificar la división por $(n - 1)$ en la fórmula de la varianza muestral:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$$

Lo hacemos porque hemos perdido un grado de libertad al estimar la media a partir de los datos de la muestra.

Ejemplo de las alturas y los pesos

Consideramos las observaciones de los pesos (kg) y alturas (cm) de un conjunto de diez personas:

Individuos (*i*)	1	2	3	4	5	6	7	8	9	10
Altura (*x*)	161	152	167	153	161	168	167	153	159	173
Peso (y_i)	63	56	77	49	72	62	68	48	57	67

La recta de regresión correspondiente es:

$$\hat{y} = -96{,}112 + 0{,}979x$$

Para hacer los cálculos más cómodos, es aconsejable construir la tabla de cálculos por la varianza de los residuos que se muestra a continuación.

En la sesión "El modelo de regresión simple" se deduce la recta de regresión correspondiente a este ejemplo.

i	x_i	y_i	$\hat{y}_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$e_i = y_i - \hat{y}_i$	e_i^2
1	161	63	61,51	–0,4	0,16	1,49	2,225
2	152	56	52,70	–9,4	88,36	3,30	10,908
3	167	77	67,38	5,6	31,36	9,62	92,498
4	153	49	53,68	–8,4	70,56	–4,68	21,868
5	161	72	61,51	–0,4	0,16	10,49	110,075
6	168	62	68,36	6,6	43,56	–6,36	40,468
7	167	68	67,38	5,6	31,36	0,62	0,381
8	153	48	53,68	–8,4	70,56	–5,68	32,220

i	x_i	y_i	$\hat{y}_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$e_i = y_i - \hat{y}_i$	e_i^2
9	159	57	59,55	–2,4	5,76	–2,55	6,504
10	173	67	73,26	11,6	134,56	–6,26	39,143
Σ	**1.614**	**619**			**476,4**		**356,290**

La octava columna contiene los cuadrados de los residuos. Sumando todos los datos y dividiendo por el número de observaciones menos 2, es decir, por 10 – 2 = 8, obtenemos la varianza de los residuos:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{356,290}{10-2} = 44,536$$

3. Distribución probabilística de la pendiente ($\hat{\beta}_1$)

La ordenada en el origen β_0 nos informa del valor medio de la variable *Y* para un valor de *X* igual a cero. No siempre tiene interpretación realista en el contexto del problema: por este motivo, únicamente consideraremos hacer inferencia estadística sobre la pendiente.

Para poder hacer inferencia estadística (hacer contrastes de hipótesis y buscar intervalos de confianza), será necesario conocer la distribución de probabilidad de $\hat{\beta}_1$.

Del modelo de regresión lineal tenemos que $\hat{\beta}_1$ es una combinación lineal de las observaciones y_i; y si éstas tienen una distribución normal y son independientes (tal como hemos supuesto al establecer el modelo de regresión), entonces $\hat{\beta}_1$ también tendrá una distribución normal. Tendremos bien determinada esta distribución cuando conozcamos la esperanza y la varianza.

A partir de la expresión de $\hat{\beta}_1$ podemos encontrar el valor esperado y la varianza.

Los desarrollos matemáticos se muestran en el anexo de esta sesión.

- Valor esperado de $\hat{\beta}_1$:

$$E(\hat{\beta}_1) = \beta_1$$

La pendiente estimada de la recta está distribuida según una distribución normal con una media igual al valor de este parámetro para la población. Aunque este valor es desconocido, este resultado nos será muy útil para tener información de la población haciendo inferencia estadística. Esto lo veremos un poco más adelante en esta sesión.

- Varianza de $\hat{\beta}_1$:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

A continuación veremos que necesitaremos la información de la muestra, ya que σ^2 es un valor desconocido que tendremos que estimar.

4. El intervalo de confianza para la pendiente

Acabamos de ver que las suposiciones del modelo de regresión lineal simple implican que el parámetro $\hat{\beta}_1$ es una variable aleatoria distribuida normalmente con:

- Media: β_1
- Varianza: $\sigma^2_{\hat{\beta}_1} = \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2}$

Dado que esta varianza σ^2 es desconocida, deberemos estimarla a partir de la varianza muestral que ya hemos calculado anteriormente:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Definimos el error estándar de la pendiente como:

$$s_{\hat{\beta}_1} = +\sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}$$

Dado que $\hat{\beta}_1$ sigue una distribución normal con varianza desconocida (ya que no se conoce σ^2), entonces la variable tipificada:

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

tiene una distribución *t* de Student con $n - 2$ grados de libertad.

Intervalo de confianza por la pendiente con un nivel significativo α.

Con todo esto, tenemos que un **intervalo de confianza** de 100 $(1 - \alpha)$% por la pendiente β_1 de la recta de regresión poblacional viene dado por:

$$[\hat{\beta}_1 - t_{\alpha/2,\,n-2}s_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,\,n-2}s_{\hat{\beta}_1}]$$

ya que:

$$P\left(-t_{\alpha/2,\,n-2} \le \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \le t_{\alpha/2,\,n-2}\right) = 1 - \alpha$$

Este intervalo está centrado en la estimación puntual del parámetro, es decir, en $\hat{\beta}_1$, y la cantidad en la que se alarga a cada lado de la estimación depende del nivel deseado de confianza, α (mediante el valor crítico $t_{\alpha/2,\,n-2}$) y de la variabilidad del estimador $\hat{\beta}_1$ (mediante $s_{\hat{\beta}_1}$).

Ejemplo de las alturas y los pesos

Consideremos una vez más el ejemplo de los pesos y las alturas de una muestra de diez personas. La recta de regresión correspondiente era: $\hat{y} = -96{,}112 + 0{,}979x$, de manera que $\hat{\beta}_1 = 0{,}979$.

Calcularemos un intervalo de confianza del 95% para la pendiente. Por tanto, $\alpha = 0{,}05$ y mirando la tabla de la *t* de Student tenemos un valor crítico de $t_{\alpha/2;\, n-2} = t_{0{,}025;8} = 2{,}3060$.

Para calcular el intervalo de confianza: $[\hat{\beta}_1 - t_{\alpha/2,\, n-2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\, n-2} s_{\hat{\beta}_1}]$, antes tenemos que calcular:

$$s_{\hat{\beta}_1}{}^2 = \frac{s^2}{\sum (x_i - \bar{x})^2}$$

donde:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Antes ya hemos calculado la varianza de los residuos:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{356{,}290}{10-2} = 44{,}536$$

De manera que:

$$s_{\hat{\beta}_1}{}^2 = \frac{s^2}{\sum (x_i - \bar{x})^2} = \frac{44{,}536}{476{,}4} = 0{,}093$$

Por tanto, el error estándar de la pendiente será: $s_{\hat{\beta}_1} = \sqrt{0{,}093} = 0{,}306$

Y el intervalo de confianza es: [0,979 – 2,3060 · 0,306; 0,979 + 2,3060 · 0,306].

Finalmente tenemos [0,274; 1,684]. Así pues, tenemos un 95% de probabilidad de que la pendiente de la recta de regresión para la población se encuentre en este intervalo.

5. El contraste de hipótesis sobre la pendiente

Observemos que si en el modelo de regresión lineal la pendiente es cero, entonces la variable *X* no tiene ningún efecto sobre la variable *Y*. En este caso diremos que *X* no es una **variable explicativa** del modelo.

En este apartado haremos un contraste de hipótesis sobre la pendiente de la recta de regresión para saber si podemos afirmar o no que éste es igual a cero.

Como en todos los contrastes de hipótesis, daremos los pasos siguientes:

1) Establecemos las hipótesis nula y alternativa:

- Hipótesis nula: H_0: $\beta_1 = 0$, es decir, la variable *X* no es explicativa
- Hipótesis alternativa: H_1: $\beta_1 \neq 0$, es decir, la variable *X* es explicativa

No rechazar la hipótesis nula significa que no se puede considerar el parámetro β_1 significativamente diferente de cero. Es decir, la variable X no tiene influencia sobre la variable Y y, por tanto, no existe una relación lineal entre las dos variables.

Interpretación geométrica

No rechazar H_0 significa que la recta estimada tiene una pendiente nula y, por tanto, para cualquier valor de X la variable Y toma un mismo valor.

2) Fijamos un nivel significativo α.

3) Bajo el supuesto de la hipótesis nula cierta ($\beta_1 = 0$) tenemos el **estadístico de contraste**:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

que corresponde a una observación de una distribución t de Student con $n - 2$ grados de libertad.

4) Finalmente, podemos actuar de dos maneras:

a) A partir del p-valor. Este valor es: $p = 2P(t_{n-2} > |t|)$.

Recordemos que...

... el p-valor es la probabilidad del resultado observado o de uno más alejado si la hipótesis nula es cierta.

- Si $p \leq \alpha$ se rechaza la hipótesis nula H_0
- Si $p > \alpha$ no se rechaza la hipótesis nula H_0

b) A partir de los valores críticos $\pm t_{\alpha/2,\, n-2}$, de manera que:

- Si $|t| > t_{\alpha/2,\, n-2}$, se rechaza la hipótesis nula H_0; por tanto, hay una relación lineal entre las variables X e Y.
- Si $|t| \leq t_{\alpha/2,\, n-2}$, no se rechaza la hipótesis nula H_0; por tanto, no hay una relación lineal entre X e Y. Decimos que la variable X es no explicativa.

Ejemplo de las alturas y los pesos

Continuando con el ejemplo de las alturas y los pesos, queremos contrastar la hipótesis nula de que la variable X no es explicativa de la variable Y, es decir, que la pendiente de la recta de regresión es cero.

1) Establecemos las hipótesis nula y alternativa:

Hipótesis nula: H_0: $\beta_1 = 0$
Hipótesis alternativa: H_1: $\beta_1 \neq 0$

2) Calculamos el estadístico de contraste: $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 3{,}202$

Sigue una distribución t de Student con $n - 2 = 10 - 2 = 8$ grados de libertad.

3) Establecemos un criterio de decisión a partir de un nivel significativo α fijado: si escogemos un nivel significativo de $\alpha = 0{,}05$:

a) A partir del p-valor: $P(|t| > 3{,}202) = 2P(t > 3{,}202) = 2 \cdot 0{,}0063 = 0{,}0126 < 0{,}05$; por tanto, rechazamos la hipótesis nula.

b) A partir del valor crítico que es $t_{0{,}025;8} = 2{,}3060$, dado que $3{,}202 > 2{,}306$, llegamos a la misma conclusión: rechazamos la hipótesis nula y podemos concluir que la variable altura es explicativa del peso de las personas con un 95% de confianza.

6. Resumen

En esta sesión dedicada a la regresión lineal simple hemos considerado que nuestras observaciones sobre dos variables X e Y son una muestra aleatoria de una población y que las utilizamos para extraer algunas conclusiones del comportamiento de las variables sobre la población. Hemos establecido el modelo de regresión lineal con sus hipótesis básicas más importantes y hemos visto cómo hacer inferencia sobre la pendiente de la recta obtenida a partir de la muestra y, en particular, cómo calcular un intervalo de confianza y cómo hacer un contraste de hipótesis para decidir si la variable X nos explica realmente el comportamiento de la variable Y.

Ejercicios

1.

El departamento de personal de una empresa informática dedicada a la introducción de datos ha llevado a cabo un programa de formación inicial del personal. La tabla siguiente indica el progreso obtenido en mecanografía de ocho estudiantes que siguieron el programa y el número de semanas que hace que lo siguen:

Número de semanas	Ganancia de velocidad (p.p.m.)
3	87
5	119
2	47
8	195
6	162
9	234
3	72
4	110

La recta de regresión calculada a partir de estos datos es:

$$\hat{y}_i = 1{,}659 + 25{,}318x_i$$

a) Calculad un intervalo de confianza del 95% para la pendiente de la recta de regresión.

b) Haced un contraste de hipótesis con un nivel de significación $\alpha = 0{,}05$, para saber si la variable "número de semanas" es explicativa de la variable "ganancia de velocidad".

2.

Una tienda de ordenadores llevó a cabo un estudio para determinar la relación entre los gastos de publicidad semanal y las ventas. Se obtuvieron los datos siguientes:

Gastos en publicidad (*x* 1.000 €)	Ventas (*x* 1.000 €)
40	380
25	410
20	390
22	370
31	475
52	450
40	500
20	390
55	575
42	520

Con estos datos se han obtenido las cantidades siguientes:

$$\sum_{i=1}^{10} x_i = 347 \qquad \sum_{i=1}^{10} y_i = 4.460 \qquad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 6.018$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1.522,1 \qquad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 43.590,0$$

$$\sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 23.793,66 \qquad \sum_{i=1}^{10} (y_i - \hat{y}_i)^2 = 19.796,34$$

Y la recta de regresión: $\hat{y} = 308,8 + 3,95\ x$.

A partir de toda esta información, calculad un intervalo de confianza del 95% para la pendiente.

Solucionario

1.

a) Intervalo de confianza:

Queremos un intervalo de confianza del 95%, por tanto, $\alpha = 0,05$ y observando la tabla de la *t* de Student para 6 grados de libertad, tenemos un valor crítico de $t_{\alpha/2;n-2} = = t_{0,025;6} = 2,4469$.

Como siempre, lo primero que haremos es una tabla de cálculos adecuada con lo que nos piden en este problema:

i	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	3	87	–2	4	77,61	9,39	88,116
2	5	119	0	0	128,25	–9,25	85,544
3	2	47	–3	9	52,30	–5,30	28,037
4	8	195	3	9	204,20	–9,20	84,695
5	6	162	1	1	153,57	8,43	71,115
6	9	234	4	16	229,52	4,48	20,061
7	3	72	–2	4	77,61	–5,61	31,506
8	4	110	–1	1	102,93	7,07	49,971
Σ	**40**	**1.026**	**35**	**44**			**459,045**

El intervalo de confianza viene dado por:

$$[\hat{\beta}_1 - t_{\alpha/2,\ n-2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\ n-2} s_{\hat{\beta}_1}]$$

Y ya estamos en condiciones de calcular cada uno de estos términos:

$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum (x_i - \bar{x})^2} = \frac{76,507}{44,0} = 1,739$$

donde $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{459{,}045}{10-2} = 76{,}507$

Por tanto, $s_{\hat{\beta}_1} = \sqrt{1{,}739} = 1{,}319$

Y el intervalo de confianza es:

$$[25{,}318 - 2{,}4469 \cdot 1{,}319;\ 25{,}318 + 2{,}4469 \cdot 1{,}319]$$

Es decir:

$$[22{,}092;\ 28{,}545]$$

b) Contraste de hipótesis para $\alpha = 0{,}05$:

1) Establecemos las hipótesis nula y alternativa:

Hipótesis nula: H_0: $\beta_1 = 0$
Hipótesis alternativa: H_1: $\beta_1 \neq 0$

2) Calculamos el estadístico de contraste:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 19{,}200$$

Sigue una distribución t de Student con $n - 2 = 6$ grados de libertad.

3) Conclusión: puesto que para $\alpha = 0{,}05$ tenemos un valor crítico $t_{0{,}025;6} = 2{,}4469$ menor que el estadístico de contraste $t = 19{,}200$, entonces rechazamos la hipótesis nula, de manera que la pendiente es diferente de cero y la variable "número de semanas" es explicativa de la "ganancia de velocidad".

2.

El intervalo de confianza viene dado por:

$$[\hat{\beta}_1 - t_{\alpha/2,\, n-2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,\, n-2} s_{\hat{\beta}_1}]$$

Necesitamos calcular el error estándar de la pendiente y encontrar los valores críticos.

1) Error estándar de la pendiente:

Primero calculamos:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{19.796{,}34}{10-2} = 2.474{,}54$$

de manera que:

$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum(x_i - \bar{x})^2} = \frac{2.474,54}{1.522,1} = 1,626$$

Por tanto, el error estándar de la pendiente vale: $s_{\hat{\beta}_1} = \sqrt{1,626} = 1,275$

2) Un intervalo de confianza del 95% con $n = 10$, tenenemos unos valores críticos:

$$t_{0,025;8} = \pm 2,3060$$

3) Por tanto, el intervalo de confianza es:

$$[3,953 - 2,3060 \cdot 1,275;\ 3,953 + 2,3060 \cdot 1,275]$$

Es decir:

$$[1,013;\ 6,894]$$

Este intervalo de confianza no contiene el valor cero; por tanto, este resultado nos indica que el gasto en publicidad es explicativo de las ventas con una confianza del 95%.

Anexos

Anexo 1

a) Valor esperado de $\hat{\beta}_1$:

$$E(\hat{\beta}_1) = \beta_1$$

Manipulando un poco la expresión que tenemos para $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}y_i$$

Si hacemos: $w_i = \frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}$, podemos escribir: $\hat{\beta}_1 = \sum_{i=1}^{n} w_i y_i$

Si ahora calculamos el valor esperado:

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} E(w_i y_i) = \sum_{i=1}^{n} w_i E(y_i) =$$

$$= \sum_{i=1}^{n} w_i(\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} w_i\beta_0 + \beta_1 w_i x_i =$$

$$= \sum_{i=1}^{n} w_i\beta_0 + \sum_{i=1}^{n} \beta_1 w_i x_i = \beta_0 \sum_{i=1}^{n} w_i + \beta_1 \sum_{i=1}^{n} w_i x_i$$

Propiedad de la linealidad

La propiedad de la linealidad de la esperanza de una variable es:

$E(kX) = kE(X)$.

Vemos que: $\sum_{i=1}^{n} w_i = 0$ y que $\sum_{i=1}^{n} w_i x_i = 1$

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

Observación

Puesto que:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

es fácil ver que:

$$\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

Para calcular el término $\sum_{i=1}^{n} w_i x_i$, utilizaremos la igualdad siguiente:

$$\sum_{i=1}^{n} w_i(x_i-\bar{x}) = \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i\bar{x} = \sum_{i=1}^{n} w_i x_i - \bar{x}\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i x_i$$

Ya que: $\sum_{i=1}^{n} w_i = 0$

De manera que:

$$\sum_{i=1}^{n} w_i(x_i-\bar{x}) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_i-\bar{x}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 1$$

Así pues: $\sum_{i=1}^{n} w_i x_i = 1$

Y, finalmente, tenemos que: $E(\hat{\beta}_1) = \beta_1$.

b) Varianza de $\hat{\beta}_1$:

$$\sigma_{\hat{\beta}_1} = Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} Var(w_i y_i) = \sum_{i=1}^{n} w_i^2 Var(y_i) =$$

$$= \sigma^2 \sum_{i=1}^{n} w_i^2 = \sigma^2 \left(\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)^2 = \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Propiedad de la varianza

$Var(kX) = k^2 Var(X)$.

Tenemos que la varianza de $\hat{\beta}_1$ es: $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$