Relación entre variables: causalidad, correlación y regresión

Correlación entre variables. Modelos de regresión simple (lineal, cuadrática, cúbica). Modelos de regresión múltiple

Blanca de la Fuente

PID_00161061

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

En este módulo se van a estudiar las relaciones que se pueden presentar entre diferentes variables. En concreto se estudiaran posibles relaciones de dependencia entre las variables para intentar encontrar una expresión que permita estimar una variable en función de otras. Para profundizar en el análisis es necesario determinar la *forma* concreta en que se relacionan y medir su *grado* de asociación.

Así, por ejemplo, el estudio de las relaciones entre variables se puede aplicar para dar respuestas a preguntas y casos como los siguientes:

- ¿Existe relación entre la edad de los lectores y el número de préstamos de libros?

- En otro caso, una editorial podría usar la relación entre el número de páginas de un trabajo y el tiempo de impresión para predecir el tiempo empleado en la impresión.

- Se quiere estudiar el "tiempo de respuesta" de unos ciertos programas de búsqueda bibliográfica en función del "número de instrucciones" en que están programados.

- En una determinada empresa de venta de libros en línea, ¿cómo representamos que el aumento de la cantidad gastada en publicidad provoca un incremento de las ventas?

Este módulo examina la relación entre dos variables, una variable independiente y otra dependiente, por medio de la regresión simple y la correlación. También se considera el modelo de regresión múltiple en el que aparecen dos o más variables independientes.

Objetivos

Los objetivos académicos del presente módulo se describen a continuación:

1. Comprender la relación entre correlación y regresión simple.

2. Usar gráficos para ayudar a comprender una relación de regresión.

3. Ajustar una recta de regresión e interpretar los coeficientes.

4. Obtener e interpretar las correlaciones y su significación estadística.

5. Utilizar los residuos de la regresión para comprobar la validez de las suposiciones necesarias para la inferencia estadística.

6. Aplicar contrastes de hipótesis.

7. Ajustar una ecuación de regresión múltiple e interpretar los resultados.

1. Relación entre variables

Cuando se estudian conjuntamente dos o más variables que no son independientes, la relación entre ellas puede ser **funcional** (relación matemática exacta entre dos variables, por ejemplo, espacio recorrido por un vehículo que circula a velocidad constante y el tiempo empleado en recorrerlo) o **estadística** (no existe una expresión matemática exacta que relacione ambas variables, existe una relación aproximada entre las dos variables, por ejemplo, incremento de las ventas de libros en función de la cantidad gastada en publicidad). En este último caso interesa estudiar el grado de dependencia existente entre ambas variables. Lo realizaremos mediante el **análisis de correlación** y, finalmente, desarrollaremos un modelo matemático para estimar el valor de una variable basándonos en el valor de otra, en lo que llamaremos **análisis de regresión**.

El análisis de regresión **no se puede** interpretar como un procedimiento para establecer una relación **causa-efecto o causalidad** entre variables. La regresión solo puede indicar cómo están **asociadas** las variables entre sí y nos permite construir un modelo para explicar la relación entre ellas. La correlación indica el grado de la relación entre dos variables sin suponer que una alteración en una cause un cambio en la otra variable.

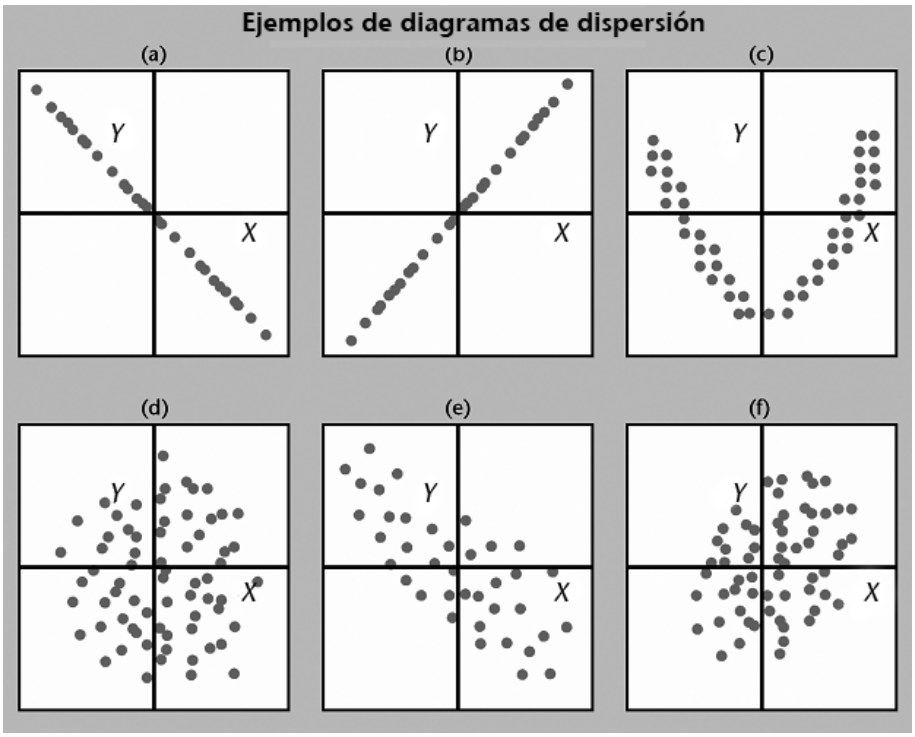
El objetivo principal del análisis de regresión es explicar el comportamiento de una **variable dependiente *Y*** (endógena o explicada) a partir de una o varias **variables independientes** (exógenas o explicativas). El tipo más sencillo de regresión es la **regresión simple**. La regresión lineal simple estima una ecuación lineal que describe la relación, mientras que la correlación mide la fuerza de la relación lineal. Aparte de los modelos lineales se pueden establecer otros modelos de regresión no lineales. El análisis de regresión donde intervienen dos o más variables independientes se llama análisis de regresión múltiple, donde una variable viene explicada por la acción simultánea de otras variables.

Diagrama de dispersión

Antes de abordar el problema, se puede intuir si existe relación entre las variables a través de la representación gráfica llamada **diagrama de dispersión** o **nube de puntos**.

A partir de un conjunto de observaciones (x_i, y_i) de dos variables X e Y sobre una muestra de individuos se representan estos datos sobre un eje de coordenadas x–y. En la figura 1 se incluyen varias gráficas de dispersión que ilustran diversos tipos de relación entre variables.

Figura 1. Diagramas de dispersión



En los casos (a) y (b) tenemos que las observaciones se encuentran sobre una recta. En el primer caso, con pendiente negativa, indica una relación inversa entre las variables (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, indica una relación directa entre las variables (a medida que aumenta X, la Y también aumenta). 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 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 relación lineal. En el caso (f) observamos una relación lineal con pendiente positiva, pero no tan fuerte como la anterior.

Después de estudiar el diagrama de dispersión, el siguiente paso es comprobar analíticamente la dependencia o independencia de ambas variables.

2. Análisis de la correlación

El análisis de correlación mide el grado de relación entre las variables. En este apartado veremos el análisis de correlación simple, que mide la relación entre sólo una variable independiente (X) y la variable dependiente (Y). En el apartado 4 de este módulo se describe el análisis de correlación múltiple que muestra el grado de asociación entre dos o más variables independientes y la variable dependiente.

La correlación simple determina la cantidad de variación conjunta que presentan dos variables aleatorias de una distribución bidimensional. En concreto, cuantifica la dependencia lineal, por lo que recibe el nombre de correlación lineal. El coeficiente de correlación lineal se llama coeficiente de correlación de Pearson designado $\boldsymbol{r}$, cuyo valor oscila entre –1 y +1. Su expresión es el cociente entre la covarianza muestral entre las variables y el producto de sus respectivas desviaciones típicas:

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

El valor de r se aproxima a +1 cuando la correlación tiende a ser lineal directa (mayores valores de X significan mayores valores de Y), y se aproxima a –1 cuando la correlación tiende a ser lineal inversa. Podemos formular la 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 $0 \leq |r| \leq 0{,}5$; fuerte si $0{,}8 \leq |r| \leq 1$, y moderada si tiene otro valor.

Dada una variable X con x_1, x_2... x_n valores muestrales y otra variable Y con y_1, y_2... y_n valores muestrales, siendo n el número total de observaciones y siendo

la media de X: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ y la media de Y: $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

La covarianza muestral entre dos variables X e Y nos permite medir estas relaciones positivas y negativas entre las variables X e Y:

$$Cov(X,Y) = S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_j - \overline{y})$$

La covarianza muestral podemos calcularla mediante otra expresión equivalente:

$$S_{XY} = \frac{\left[\sum_{i,j=1}^{n} x_i y_j\right] - n \cdot \overline{x} \cdot \overline{y}}{n-1}$$

Ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación".

Estamos realizando un proceso de evaluación de los servicios ofrecidos por un centro de documentación. Para conocer la opinión de los usuarios se les ha pedido que rellenen un cuestionario de evaluación del servicio. Hacemos dos preguntas, una para que valoren de 0 a 10 su impresión sobre el funcionamiento global del centro y otra pregunta que valora específicamente la atención a los usuarios, para determinar si las valoraciones respecto a la atención al usuario (representadas por la variable dependiente *Y*) están relacionadas con las valoraciones obtenidas respecto al funcionamiento global del centro (variable independiente *X*).

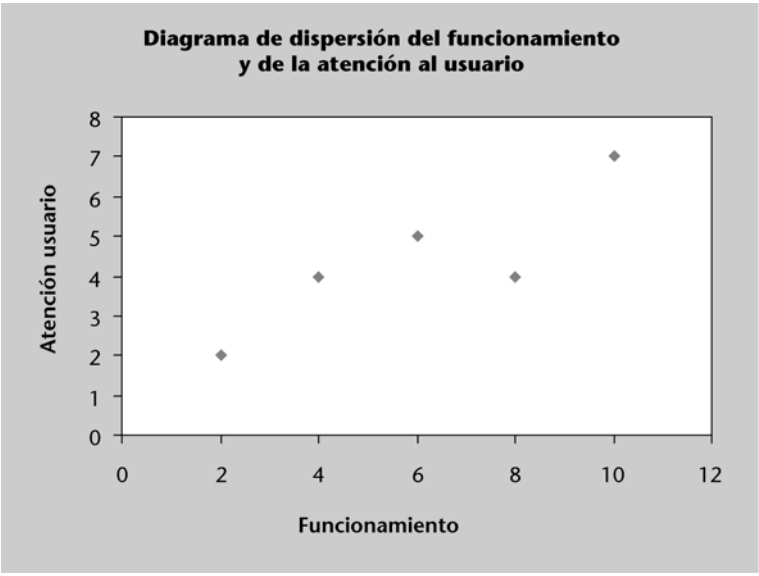
Para ello, un investigador ha seleccionado al azar cinco personas entrevistadas y dan las siguientes valoraciones:

Tabla 1. Datos obtenidos de respuestas a cinco entrevistas realizadas sobre valoraciones de funcionamiento y atención a usuarios de un centro de documentación

Entrevista (*i*)	Funcionamiento (*X*)	Atención (*Y*)
1	2	2
2	4	4
3	6	5
4	8	4
5	10	7

El diagrama de dispersión (figura 2) nos permite observar gráficamente los datos y sacar conclusiones. Parece que las valoraciones de atención al usuario son mejores para valoraciones elevadas del funcionamiento global del centro. Además, para esos datos la relación entre la atención al usuario y el funcionamiento parece poder aproximarse a una línea recta; realmente parece haber una relación lineal positiva entre *X* e *Y*.

Figura 2. Diagrama de dispersión del funcionamiento del centro y de la atención al usuario



Para determinar si existe correlación lineal entre las dos variables, calculamos el coeficiente de correlación ***r***.

En la tabla 2 se desarrollan los cálculos necesarios para determinar los valores de las varianzas, desviaciones típicas muestrales y covarianza muestral.

Tabla 2. Cálculo de las sumas de cuadrados para la ecuación estimada de regresión de mínimos cuadrados

Funcionamiento (*X*)	**Atención (*Y*)**	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
2	2	–4	–2,4	9,6	16	5,76
4	4	–2	–0,4	0,8	4	0,16
6	5	0	0,6	0	0	0,36
8	4	2	–0,4	–0,8	4	0,16
10	7	4	2,6	10,4	16	6,76

y_i representa las valoraciones observadas (reales) del funcionamiento global obtenidas en la entrevista i,

$$n = 5 \quad \sum_{i=1}^{5} x_i = 30 \quad \sum_{i=1}^{5} y_i = 4,4 \quad \sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = 20 \quad \sum_{i=1}^{5} (x_i - \bar{x})^2 = 40 \quad \sum_{i=1}^{5} (y_i - \bar{y})^2 = 13,2$$

realizando las siguientes operaciones obtendremos el coeficiente de correlación lineal.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{30}{5} = 6 \quad ; \quad S_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{40}{5-1}} = 3,16$$

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n} = \frac{22}{5} = 4,4 \; ; \quad S_Y = \sqrt{\frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n-1}} = \sqrt{\frac{13,2}{5-1}} = 1,82$$

$$Cov(X,Y) = \frac{1}{n-1} \sum_{i,j=1}^{n} (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{5-1} \cdot 20 = 5$$

El coeficiente de correlación lineal es:

$$r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{5}{3,16 \cdot 1,82} = 0,87$$

Como el valor del coeficiente de correlación lineal es próximo a 1, se puede afirmar que existe una correlación lineal positiva entre las valoraciones obte-

nidas de atención al usuario y las valoraciones del funcionamiento global del centro. Es decir el, funcionamiento global está asociado positivamente a la atención al usuario.

3. Modelos de regresión simple

3.1. Modelos de regresión lineal simple

Una vez que hemos obtenido el diagrama de dispersión y después de observar una posible relación lineal entre las dos variables, el paso siguiente sería encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos. Esta recta se denomina **recta de regresión.** Una recta queda bien determinada si el valor de su pendiente (b) y de la ordenada en el origen (a) son conocidas. De esta manera la ecuación de la recta viene dada por:

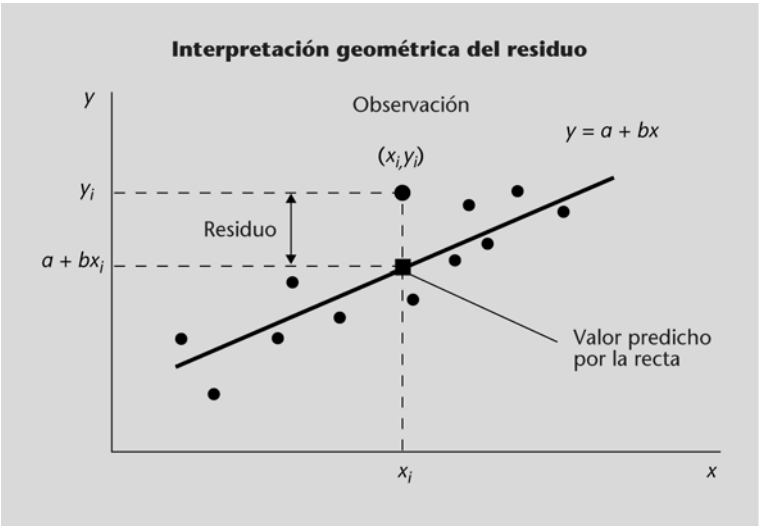
$$Y = a + bx$$

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 minimice la suma de los cuadrados de los residuos. Este es el **método de los mínimos cuadrados**, un procedimiento para encontrar la ecuación de regresión que consiste en buscar los valores de los coeficientes a y b de manera que la suma de los cuadrados de los residuos sea mínima, obteniéndose la **recta de regresión por mínimos cuadrados** (figura 3).

Figura 3. Recta de regresión por mínimos cuadrados



Nota

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

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 obtenida al determinar a y b.

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

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

El modelo de regresión lineal permite hallar el valor esperado de la variable aleatoria *Y* cuando *X* toma un valor específico.

La **recta de regresión *Y*/*X*** permite predecir un valor de *y* para un determinado valor de *x*.

Para cada observación (x_i, y_i) definimos:

- El valor estimado o predicho para la recta de regresión:

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

- Los parámetros o coeficientes de la recta y vienen dados por:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{y} \quad \hat{\beta}_1 = \frac{Cov(XY)}{S_X^2} = \frac{S_{XY}}{S_X^2}$$

Siendo:

$\hat{\beta}_0$ es la ordenada en el origen de la ecuación estimada de regresión.

$\hat{\beta}_1$ es la pendiente de la ecuación estimada de regresión.

S_{XY} la covarianza muestral, S_X^2 la varianza muestral de *X*, $\overline{x}$ *e* $\overline{y}$ son las medias aritméticas de las variables *X* e *Y* respectivamente.

- El residuo o error es la diferencia entre el valor observado y_i y el valor estimado $\hat{y}_i$:

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

Ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación".

Hemos comprobado en el ejemplo anterior que existe correlación lineal entre ambas variables, ahora calcularemos la **recta de regresión por mínimos cuadrados *Y*/*X*.**

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

en la que,

x_i = valor de funcionamiento para la i-ésima entrevista

$\hat{\beta}_0$ = ordenada en el origen de la línea estimada de regresión

$\hat{\beta}_1$ = pendiente de la línea estimada de regresión

$\hat{y}_i$ = valor estimado de la atención al usuario para la i-ésima entrevista

Para que la línea estimada de regresión ajuste bien con los datos, las diferencias entre los valores observados y los valores estimados de atención al usuario deben ser pequeñas.

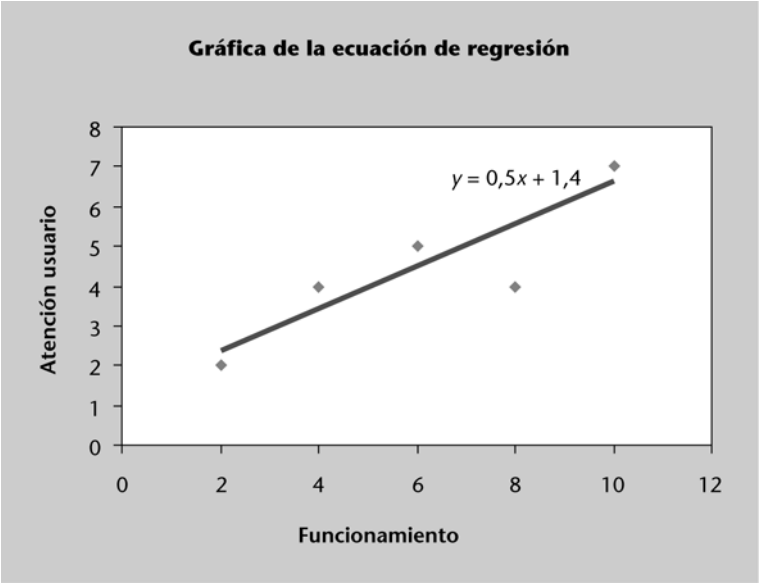
Utilizando los valores obtenidos en la tabla 2 podemos determinar la pendiente y la ordenada en el origen de la ecuación estimada de regresión en este ejemplo. Los cálculos son los siguientes:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{5}(x_i - \overline{x})^2} = 0,5\ ; \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 1,4$$

Por lo anterior, la ecuación estimada de regresión deducida con el método de mínimos cuadrados, será:

$$\hat{y} = 1,4 + 0,5x$$

Figura 4. Gráfica de la ecuación de regresión ejemplo 1



Interpretación de los parámetros de la recta de regresión

Es importante interpretar los coeficientes de la ecuación en el contexto del fenómeno que se está estudiando.

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

  Este coeficiente representa la estimación del valor de Y cuando X es igual a cero. No siempre tiene una interpretación práctica. Para que sea posible, es preciso que:

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

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

  Este coeficiente representa la estimación del incremento que experimenta la variable Y cuando X aumenta en una unidad. Este coeficiente nos informa de cómo están relacionadas las dos variables en qué cantidad varían los valores de Y cuando varían los valores de la X en una unidad.

La calidad o bondad del ajuste

Una vez acumulada la recta de regresión por mínimos cuadrados debemos analizar si este ajuste al modelo 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. La medida de bondad de ajuste para una ecuación de regresión es el **coeficiente de determinación R^2.** 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 en Y explicada por la recta de regresión. La expresión de R^2 es la siguiente:

$$R^2 = \frac{\text{Varianza en } Y \text{ explicada por la recta de regresión}}{\text{Varianza total de los datos } Y}$$

La varianza explicada por la recta de regresión es la varianza de los valores estimados y la varianza total de los datos es la varianza de los valores observados. Por tanto, podemos establecer que:

Varianza total de Y = varianza explicada por la regresión +
+ varianza no explicada (residual o de los errores)

Es decir, podemos descomponer la variabilidad total (*SSTotal*) de las observaciones de la forma:

$$SSTotal = SSRegresión + SSError$$

en la que,

SSTotal, es la suma de cuadrados totales $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$

SSRegresión, mide cuánto se desvían los valores de $\hat{y}_i$ medidos en la línea de regresión, de los valores de $\overline{y}_i$, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$

SSError, representa el error que se comete al usar $\hat{y}_i$ para estimar y_i, es la suma de cuadrados de estos errores, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

Ahora vemos cómo se pueden utilizar las tres sumas de cuadrados, *SST, SSR* y *SSE* para obtener la medida de bondad de ajuste para la ecuación de regresión, que es el coeficiente de determinación R^2. Vendrá dado por la expresión:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Los valores del coeficiente de determinación están comprendidos entre cero y uno: $0 \leq R^2 \leq 1$
- $R^2 = 1$ cuando el ajuste es perfecto, es decir, todos los puntos están sobre la recta de regresión.
- $R^2 = 0$ muestra la inexistencia de relación entre las variables X e Y.
- Como R^2 explica la proporción de variabilidad de los datos explicada por el modelo de regresión, cuanto más próximo a la unidad, será mejor el ajuste.

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^2 es el coeficiente de correlación, mide el grado de asociación lineal entre las dos variables.
- No obstante, en la regresión lineal simple tenemos que $R^2 = r^2$.

Observaciones

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

La relación entre R^2 y r ayuda a comprender lo expuesto en el análisis de la correlación: que un valor de $r^2 = 0,5$ indica una correlación débil. Este valor representará un $R^2 = 0,25$; es decir, el modelo de regresión sólo explica un 25% de la variabilidad total de las observaciones.

El signo de r da información de si la relación es positiva o negativa. Así pues, con el valor de r siempre se puede calcular el valor de R^2, pero al revés quedará indeterminado el valor del signo a menos que conozcamos la pendiente de la recta. Por ejemplo, dado un $R^2 = 0,81$, si se sabe que la pendiente de la recta de regresión es negativa, entonces se puede afirmar que el coeficiente de correlación r será igual a 0,9.

Predicción

La predicción constituye una de las aplicaciones más interesantes de la técnica de regresión. La predicción consiste en determinar a partir del modelo estimado el valor que toma la variable endógena para un valor determinado de la exógena. La fiabilidad de esta predicción será tanto mayor, en principio, cuanto mejor sea el ajuste (es decir, cuanto mayor sea R^2), en el supuesto de que exista relación causal entre la variable endógena y la variable exógena.

Nota

Variable endógena es la variable dependiente. Es la variable que se predice o se explica. Se representa por *Y*.

Variable exógena es la variable independiente. Es la variable que sirve para predecir o explicar. Se representa por *X*.

Ejemplo 1. Estudio de los servicios ofrecidos por un centro de documentación.

Una vez obtenida la ecuación estimada de regresión $\hat{y} = 1,4 + 0,5x$ del ejemplo anterior, interpretamos los resultados:

En este caso la ordenada en el origen ($\hat{\beta}_0 = 1,4$) si puede tener interpretación con sentido, ya que correspondería a la estimación de la puntuación obtenida para la atención al usuario cuando la puntuación del funcionamiento global es cero. La pendiente ($\hat{\beta}_1 = 0,5$) es positiva, lo que indica que el aumento en una unidad de la valoración del funcionamiento global del centro está asociado con un aumento de 0,5 unidades en la puntuación de atención al usuario.

Si quisiéramos predecir la valoración de la atención para una persona que ha valorado 7 el funcionamiento global, el resultado sería:

$$\hat{y} = 1,4 + 0,5 \cdot 7 = 4,9$$

En el ejemplo hemos obtenido la ecuación de regresión y debemos analizar la bondad de dicho ajuste que daría respuesta a la siguiente pregunta: ¿se ajustan bien los datos a esta ecuación de regresión?

Calcularemos el coeficiente de determinación que es una medida de la corrección del ajuste. Para ello tenemos que descomponer la variabilidad total de las observaciones de la forma:

$$SST = SSR + SSE$$

Utilizando los valores de la tabla 2 (cálculo de las sumas de cuadrados para la ecuación estimada de regresión con mínimos cuadrados), calculamos *SST* = suma de cuadrados total, es la suma de la última columna de la tabla 2.

$$SST = \sum_{i=1}^{5} (y_i - \overline{y})^2 = 13,2$$

En la tabla 3 vemos los cálculos necesarios para determinar la *SSE* = suma de cuadrados debida al error

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = 3,2$$

Tabla 3. Cálculo de las sumas de cuadrados debidas al error SCE

Funcionamiento (*X*)	**Atención (*Y*)**	$\hat{y} = 1,4 + 0,5x_i$	$e = y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
2	2	2,4	–0,4	0,16
4	4	3,4	0,6	0,36
6	5	4,4	0,6	0,36
8	4	5,4	–1,4	1,96
10	7	6,4	0,6	0,36

$$SSE = \sum_{i=1}^{5} (y_i - \hat{y}_i)^2 = 3,2$$

La *SSR* = suma de cuadrados debida a la regresión se puede calcular con facilidad usando esta expresión:

$$SSR = \sum_{i=1}^{5} (\hat{y}_i - \overline{y})^2$$

o bien si se conocen *SST* y *SSE* se puede obtener fácilmente.

$$SSR = SST - SSE = 13,2 - 3,2 = 10$$

El valor del coeficiente de determinación será:

$$R^2 = \frac{SSR}{SST} = \frac{10}{13,2} = 0,7576$$

Si lo expresamos en porcentaje, R^2= 75,76 %. Podemos concluir que el 75,76% de la variación de la puntuación en la atención al usuario se puede explicar con la relación lineal entre las valoraciones del funcionamiento global del centro y la atención al usuario. El ajuste al modelo lineal es bueno. Se considera un buen ajuste cuando R^2 es mayor o igual que 0,5.

El coeficiente de correlación lineal "*r*" será $\sqrt{0,75760} = |0,87|$, resultado acorde con la estimación obtenida usando la covarianza.

Solución de problemas de regresión lineal simple con programas informáticos

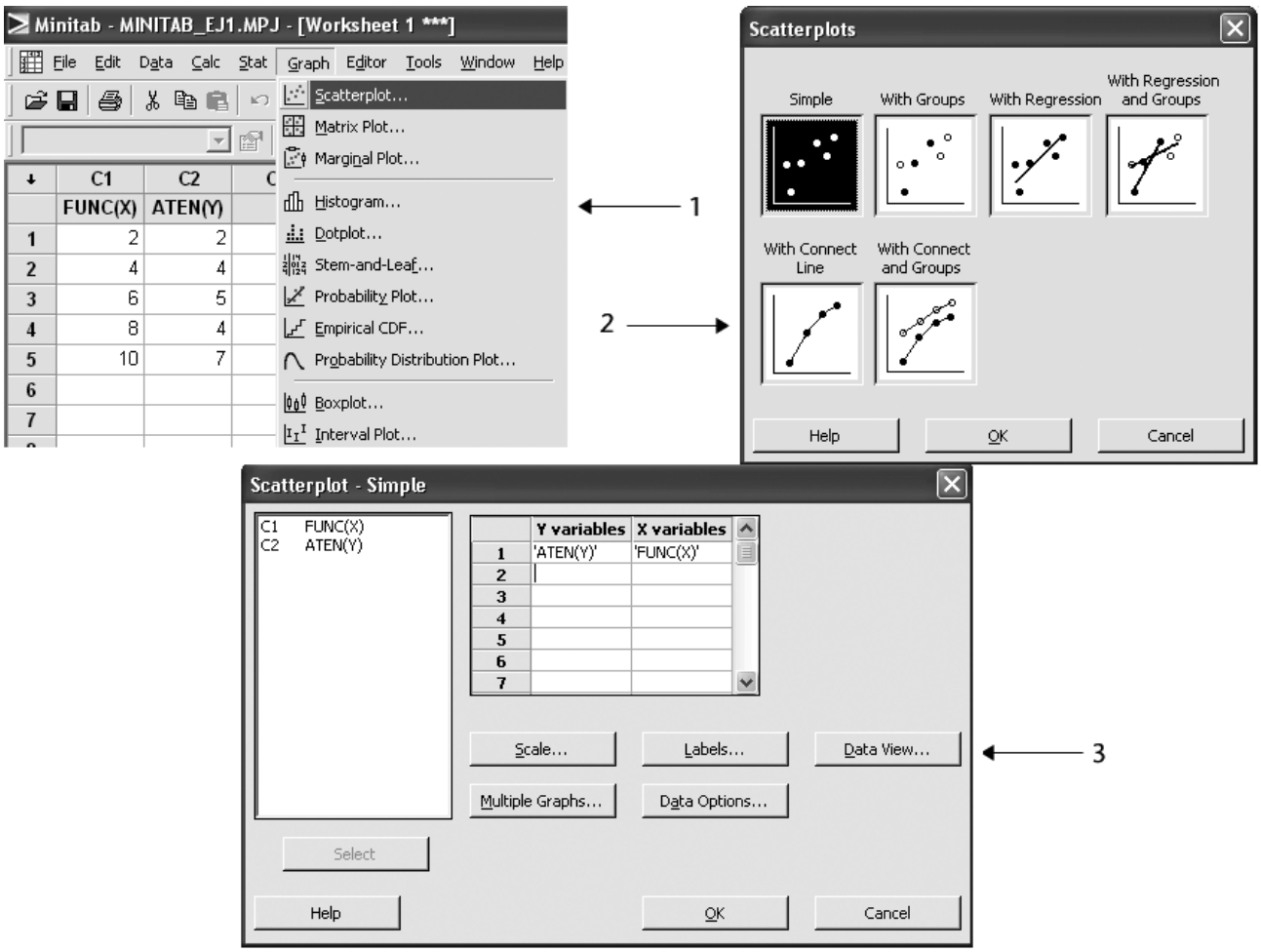
Para resolver el ejercicio empleamos el programa Minitab.

Insertamos los datos del ejemplo 1: "Estudio de los servicios ofrecidos por un centro de documentación". A la variable independiente (*Y*) la llamamos **ATEN** (de atención al usuario) y a la variable dependiente (*X*) la llamamos **FUNC** (de funcionamiento global) para facilitar la interpretación de los resultados. Insertamos los datos **FUNC** en la columna C1 y los datos de **ATEN** en la columna C2, con encabezados para obtener el diagrama de dispersión.

Pasos a seguir

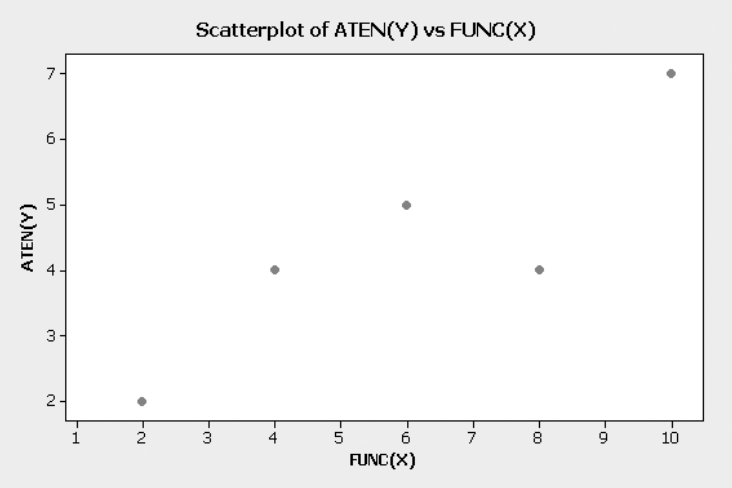
Para crear el gráfico una vez introducidos los datos en el programa ***(1)***, se sigue la ruta ***Graph*** > ***Scatterplot*** > ***Simple*** ***(2)*** y se rellenan los campos en la ventana correspondiente seleccionando las variables ***(3)***. Seleccionad ***OK*** para obtener el diagrama de dispersión.

Figura 5. Pasos a seguir para obtener el diagrama de dispersión



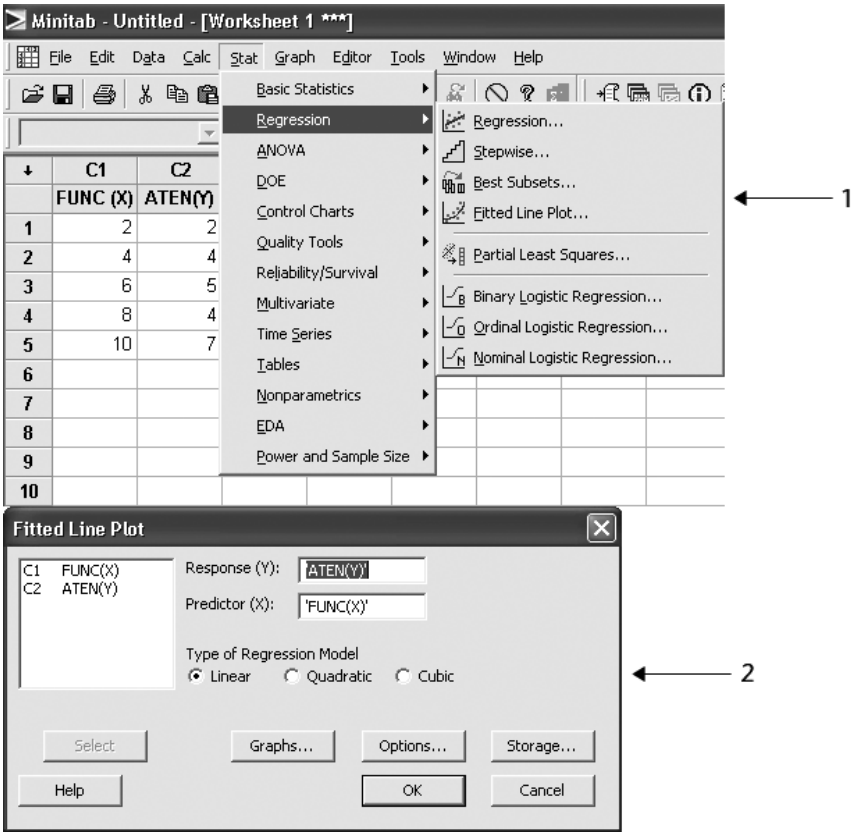
Obtuvimos el diagrama de la figura 6.

Figura 6. Diagrama de dispersión. Minitab



La figura 7 muestra los pasos a seguir para representar la recta de de regresión de mínimos cuadrados:

Figura 7. Pasos a seguir para representar la recta de regresión de mínimos cuadrados



Pasos a seguir

Usamos la opción ***Stat***, se sigue la ruta ***Regression*** > ***Regression*** > ***Fitted Line Plot (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. Seleccionad ***OK*** para obtener el gráfico.

Obtuvimos los resultados que aparecen en la figura 8.

A continuación interpretaremos los resultados:

La figura 8 muestra la gráfica de la ecuación de regresión sobre el diagrama de dispersión. La pendiente de la ecuación de regresión ($\hat{\beta}_1 = 0{,}50$) es positiva, lo

que implica que al aumentar las valoraciones del funcionamiento global, las puntuaciones de atención al usuario también aumentan.

Figura 8. Gráfica de la ecuación de regresión de mínimos cuadrados

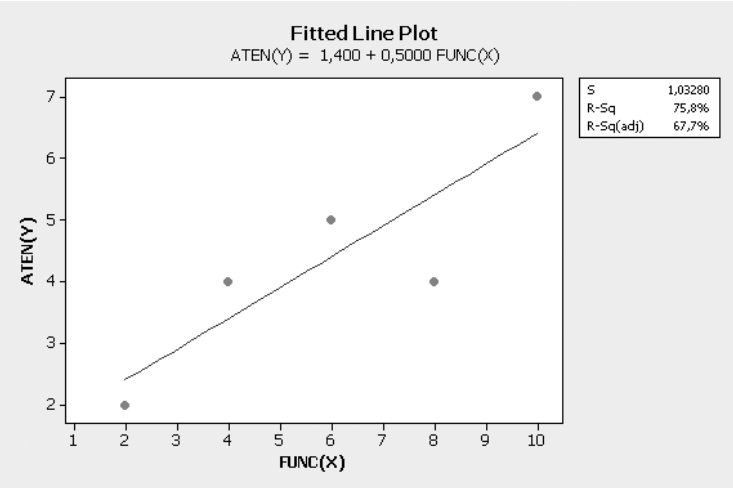
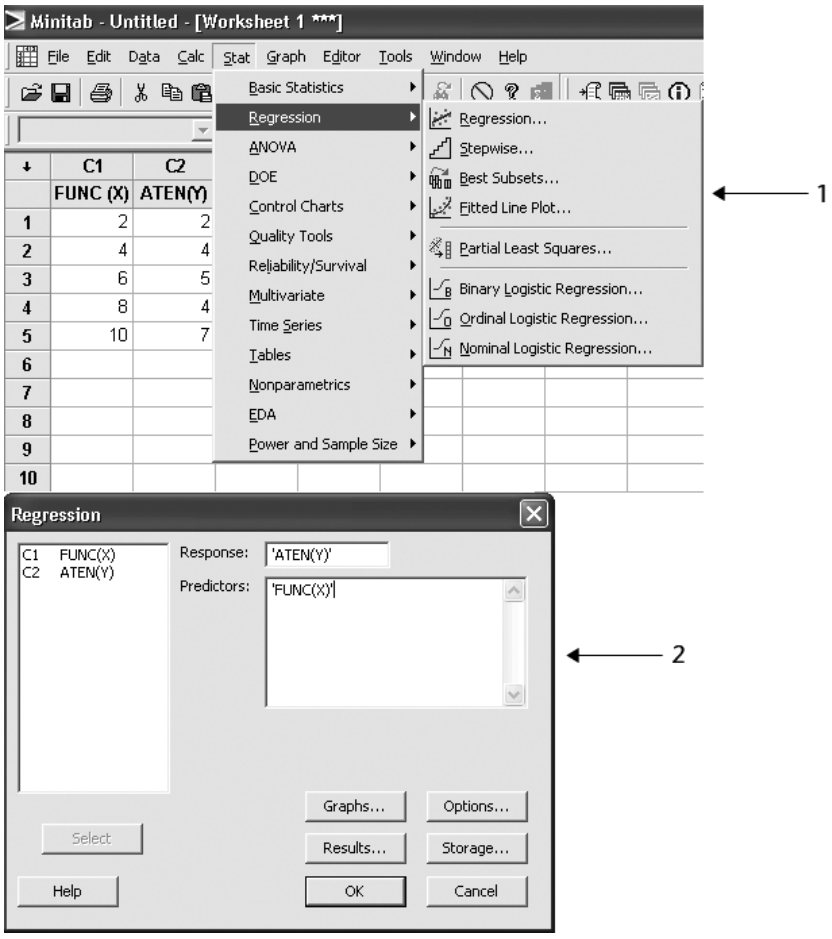


Figura 9. Pasos a seguir para realizar el análisis de regresión



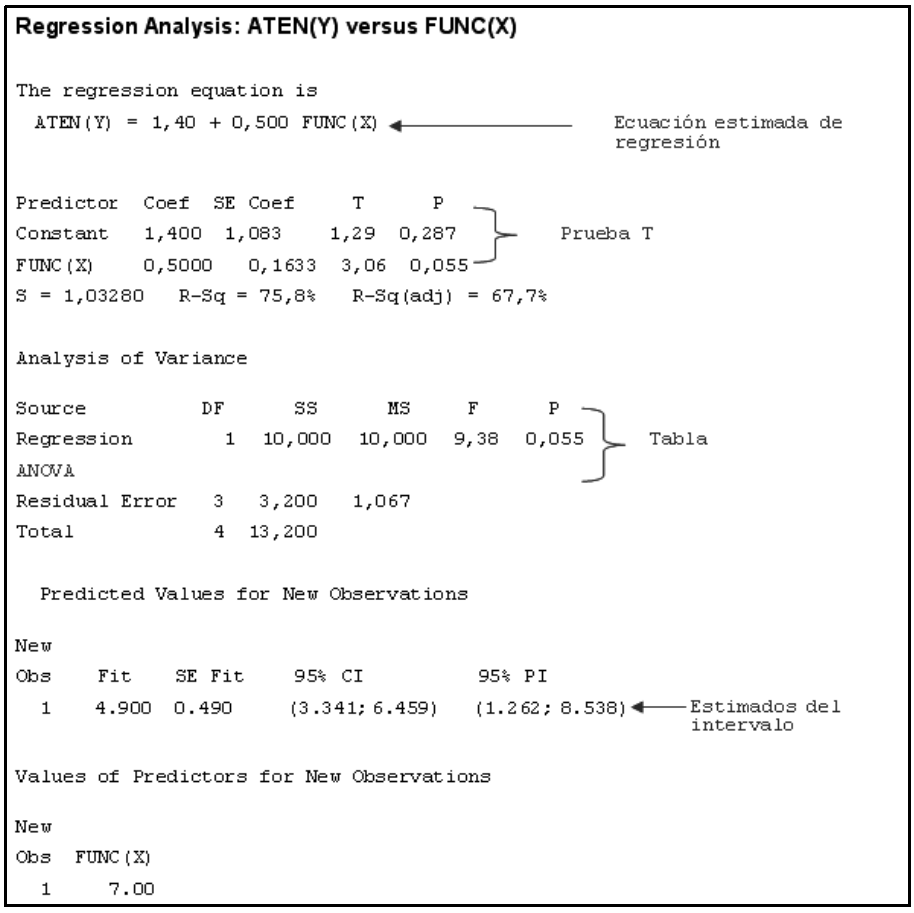
Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regression*** > ***Regression (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. Seleccionad ***OK*** para obtener el análisis de regresión.

En el cuadro de diálogo de Minitab puede obtenerse más información sobre resultados seleccionando las opciones deseadas. Por ejemplo, con este cuadro de diálogo se pueden obtener los residuos, los residuales estandarizados, los puntos de alta influencia y la matriz de correlación (estos resultados los comentaremos más adelante).

Obtenemos los resultados que aparecen en la figura 10.

Figura 10. Resultados del análisis de regresión. Minitab



- Interpretación de las estadísticas de regresión:

  Minitab imprime la ecuación de regresión en la forma:

  $$\text{ATEN}(Y) = 1{,}40 + 0{,}500\ \text{FUNC}(X).$$

  Se imprime una tabla que muestra los valores de los coeficientes *a* y *b*. El coeficiente *Constant* (ordenada en el origen) es 1,4, y la pendiente con base en la variable FUNC es 0,50. *SE Coef* son las desviaciones estándar de cada coeficiente. Los valores de las columnas *T* y *P* los analizaremos más adelante al estudiar la inferencia en la regresión.

  El programa imprime el error estándar del valor estimado, $S = 1{,}03280$ mide el tamaño de una desviación típica de un valor observado (x,y) a partir de la recta de regresión. También proporciona la información sobre la bondad de ajuste. Observad que *R-Sq* = 75,8% ($R^2 = 0{,}758$) es el coeficiente de determinación expresado en porcentaje. Como hemos comentado en la solución manual del ejercicio, un valor del 75,8% significa que el 75,8% de la variación en la puntuación de atención al usuario puede explicarse por medio de la valoración obtenida en el funcionamiento global del centro. Se supone que el 24,2 % restante de la variación se debe a la variabilidad aleatoria. El resultado *R-Sq*(adj) = 67,7% (R^2 ajustado) es un valor corregido de

acuerdo con la cantidad de variables independientes. Se tiene en cuenta al realizar una regresión con varias variables independientes y se estudiará más adelante al tratar la regresión múltiple.

- Interpretación del análisis de la varianza:

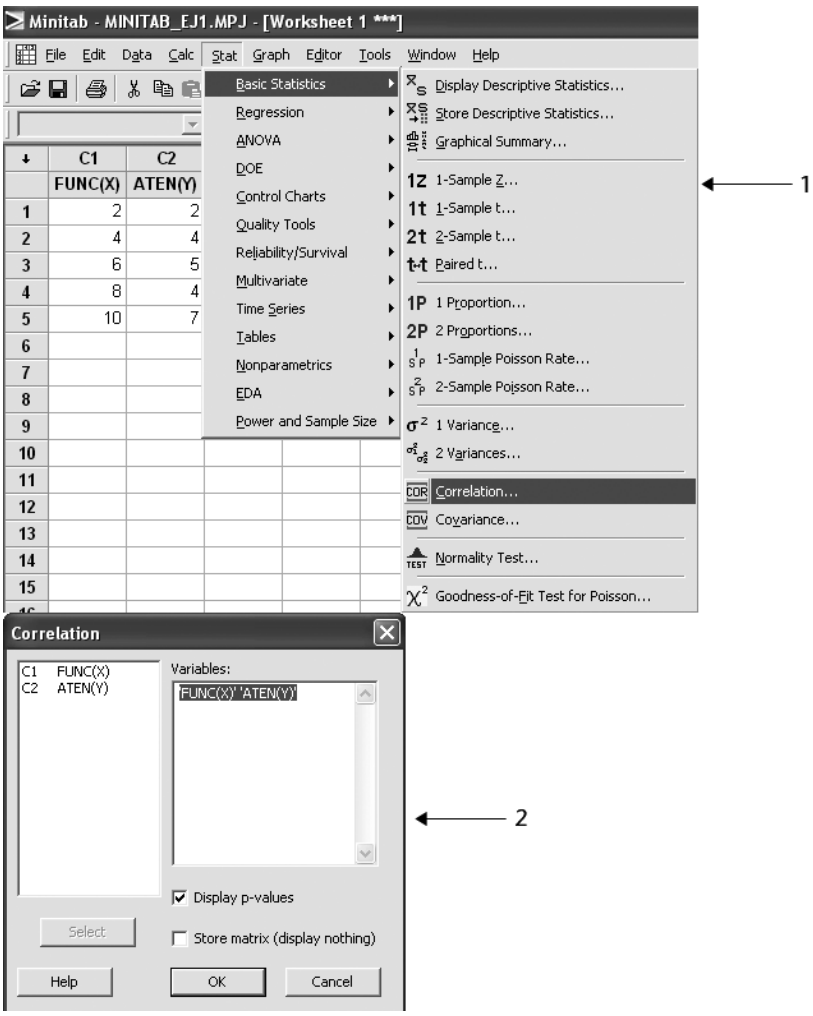
  La salida de Minitab analiza la variabilidad de las puntuaciones de atención al usuario. La variabilidad, como hemos explicado anteriormente, se divide en dos partes: *SST* = *SSR* + *SSE*.

  SS Regresión (*SSR*) es la variabilidad debida a la regresión, ***SS Error*** (*SSE*) es la variabilidad debida al error o variabilidad aleatoria, ***SS Total*** (*SST*) es la variabilidad total. El resto de la información se irá viendo mas adelante al tratar la regresión lineal múltiple.

- Interpretación del valor estimado de predicción y del intervalo de confianza de 95% (95% C.I.) y el estimado del intervalo de predicción (95% P.I.) de la atención al usuario para el valor 7 de funcionamiento global. El valor estimado para Atención al usuario es 4,9.

  A continuación calcularemos el coeficiente de correlación lineal como se indica en la figura 11.

Figura 11. Pasos a seguir para calcular el coeficiente de correlación



Pasos a seguir

Para crear el gráfico se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Correlation (1)*** y se rellenan los campos en la ventana correspondiente ***(2)***. Seleccione ***OK*** para obtener el coeficiente de correlación lineal.

Obtuvimos los resultados que aparecen en la figura 12.

Figura 12. Resultados del análisis de correlación

Correlations: FUNC(X); ATEN(Y)

```
Pearson correlation of FUNC(X) and ATEN(Y) = 0,870
P-Value = 0,055
```

- Interpretación del análisis de correlación:

  Como r =0,870, podemos decir que existe correlación lineal positiva entre las valoraciones obtenidas de atención al usuario y las valoraciones del funcionamiento global del centro. El funcionamiento está asociado positivamente con la atención al usuario.

Obsérvese que $R^2 = 0{,}758$, por lo que $\sqrt{R^2} = \sqrt{0{,}758} = 0{,}87 = r$

Para resolver el ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación" se emplea **Microsoft Excel**.

La figura 13 muestra el correspondiente *output* que ofrece **Microsoft Excel**.

Se observa que las estadísticas de regresión coinciden con las obtenidas con Minitab.

Atención

Para poder hacer la regresión con **MS Excel** es necesario instalar previamente un complemento llamado "Análisis de datos". Para instalar las herramientas de análisis de datos, haced clic en ***Herramientas*** > ***Complementos,*** y en el cuadro de diálogo activar: ***Herramientas para análisis***.

Figura 13. Resultados del análisis de regresión del ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación". Excel

	A	B	C	D	E	F	G	H	I
1	Resumen								
2									
3	*Estadísticas de la regresión*								
4	Coeficiente de correlación múltiple	0,87038828							
5	Coeficiente de determinación R^2	0,757575758							
6	R^2 ajustado	0,676767677							
7	Error típico	1,032795559							
8	Observaciones	5							
9									
10	ANÁLISIS DE VARIANZA								
11		*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
12	Regresión	1	10	10	9,375	0,054912524			
13	Residuos	3	3,2	1,066666667					
14	Total	4	13,2						
15									
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
17	Intercepción	1,4	1,083205121	1,292460655	0,28674468	-2,047242134	4,847242134	-2,047242134	4,847242134
18	Funcionamiento (X)	0,5	0,163299316	3,061862178	0,05491252	-0,019691305	1,019691305	-0,019691305	1,019691305
19									
20									
21									
22	Análisis de los residuales				Resultados de datos de probabilidad				
23									
24	*Observación*	*Pronóstico Atención (Y)*	*Residuos*		*Percentil*	*Atención (Y)*			
25	1	2,4	-0,4		10	2			
26	2	3,4	0,6		30	4			
27	3	4,4	0,6		50	4			
28	4	5,4	-1,4		70	5			
29	5	6,4	0,6		90	7			
30									

Diagnóstico de la regresión

Al igual que en cualquier procedimiento estadístico, cuando se efectúa una regresión en un conjunto de datos se hacen algunas suposiciones importantes, y en este caso son cuatro:

1) El modelo de línea recta es correcto.

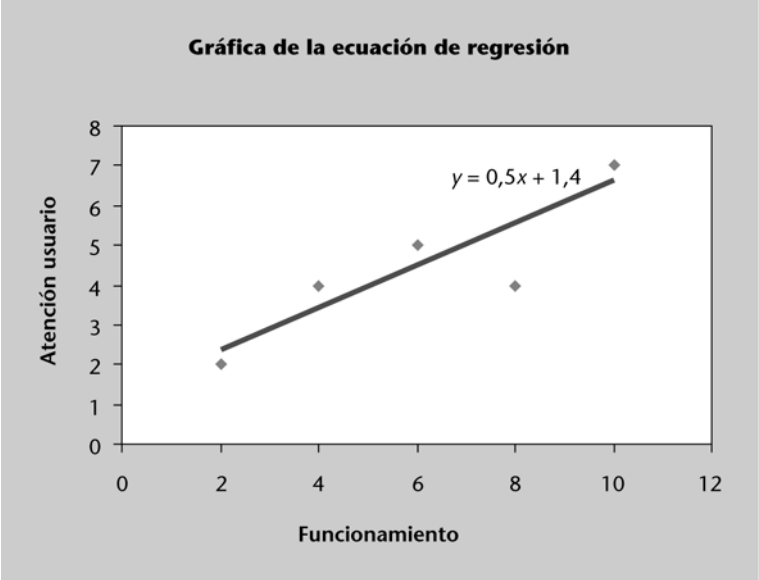
2) Los errores o residuos siguen una distribución aproximadamente normal de media cero.
3) Los errores o residuos tienen una varianza constante σ^2.
4) Los errores o residuos son independientes.

Siempre que usen regresiones para ajustar una recta a los datos, deben considerarse estas suposiciones. Comprobar que los datos cumplen estas suposiciones supone pasar por una serie de pruebas llamadas **diagnosis** que se describen a continuación.

Prueba de suposición de línea recta.

Para comprobar si es correcto el modelo de línea recta se usa el gráfico de dispersión con el ajuste a la recta de mínimos cuadrados (ejemplo 1, figura 14).

Figura 14. Gráfica de la ecuación de regresión del ejemplo 1



Análisis de residuos

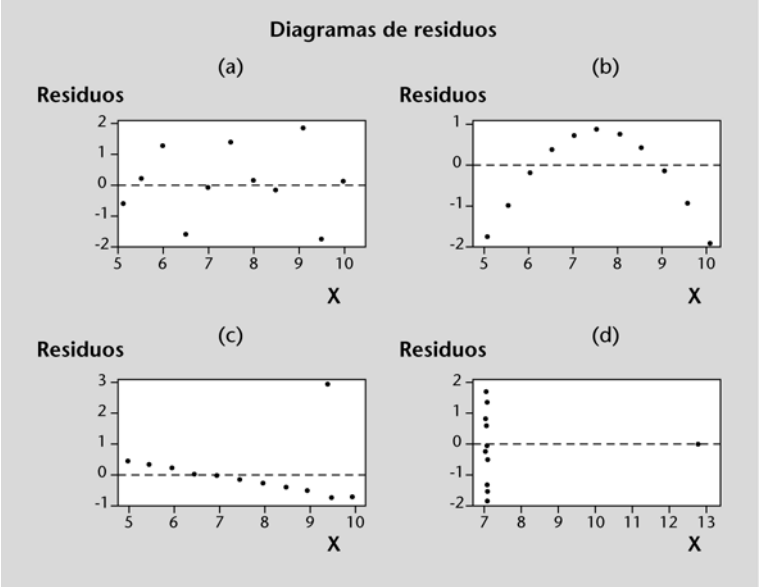
Una vez hecho el ajuste de un modelo de regresión lineal a los datos muestrales, hay que efectuar el análisis de los residuos o errores. Este análisis, que a continuación comentaremos de forma breve e intuitiva, nos servirá para hacer un diagnóstico del modelo de regresión.

Otra forma de ver si los datos se ajustan a una recta es realizando un gráfico de los residuos ($e_i = y_i - \hat{y}_i$) en función de la variable predictora (X). En el eje horizontal se representa el valor de la variable independiente (X) y en el vertical los valores de los residuos (e_i).

Podemos calcular los residuos manualmente según habíamos indicado en la tabla 3.

En la figura 15 presentamos 4 ejemplos de gráficos de residuos o errores.

Figura 15. Diagrama de residuos



Podemos observar que de los cuatro, sólo el primero no presenta ningún tipo de estructura, los residuos se distribuyen aleatoriamente, de manera que sólo tendría sentido la regresión hecha sobre la muestra (a). Si los puntos se orientasen en forma de "U" (o "U" invertida), habría problemas con este supuesto, como es el caso de la muestra (b). Los residuos del diagrama (c) y (d) no se distribuyen aleatoriamente, por lo que no se cumple el supuesto de linealidad.

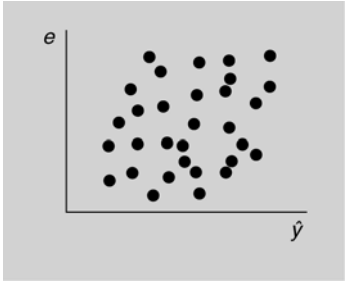
En el mismo gráfico también podemos observar si los residuos tienen varianza constante (supuesto 3). Si la varianza de los errores es constante para todos los valores de X, la gráfica de residuales debe mostrar un patrón similar a una banda horizontal de los puntos, como en (a). Si forman una flecha (en un extremo se agrupan mucho más que en el otro), caso (d), entonces este supuesto falla. También es conveniente estar atentos ante la posible existencia de valores atípicos o valores extremos (*outliers*), pues éstos podrían afectar.

Valor atípico

Por *valor atípico* entendemos un valor muy diferente de los otros y que muy posiblemente es erróneo.

También podemos usar un gráfico de residuos en función del valor estimado o predicho $\hat{y}$. Esto lo representaremos gráficamente mediante un diagrama de dispersión de los puntos $\left(\hat{y}_i, e_i\right)$, es decir, sobre el eje de las abscisas representamos el valor estimado $\hat{y}$, y sobre el eje de ordenadas, el valor correspondiente del residuo, es decir, $e_i = y_i - \hat{y}_i$.

Figura 16. Gráfico de residuos en función de valor estimado o predicho $\hat{y}$



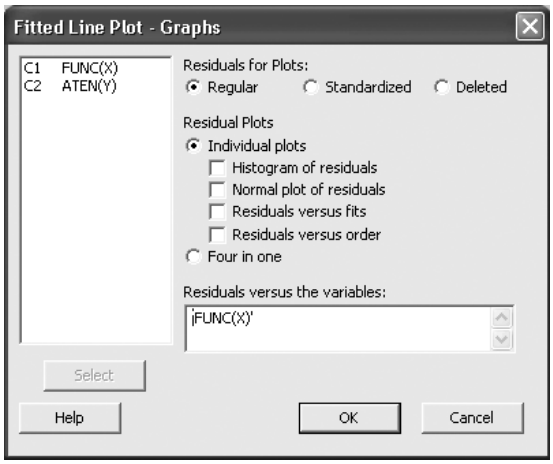
Si el modelo lineal obtenido se ajusta bien a los datos muestrales, entonces la nube de puntos $\left(\hat{y}_i, e_i\right)$ no debe mostrar ningún tipo de estructura. Para la regresión lineal simple, la gráfica de residuos en función de X y los de residuos en función de $\hat{y}$ dan la misma información. Para la regresión múltiple, la gráfica de residuos en función de $\hat{y}$ se usa con más frecuencia porque se maneja más de una variable independiente.

Para comprobar el segundo supuesto de que los errores o residuos siguen una distribución aproximadamente normal usaremos la gráfica de probabilidad normal.

Consideramos de nuevo el ejemplo 1. "Estudio de los servicios ofrecidos por un centro de documentación" y realizamos la diagnosis con Minitab a fin de comprobar si se cumplen las condiciones del modelo.

En la figura 17 se indican los pasos a seguir para crear un gráfico de los residuos en función de la variable de predicción con Minitab:

Figura 17. Pasos a seguir para crear un gráfico de los residuos en función de la predicción

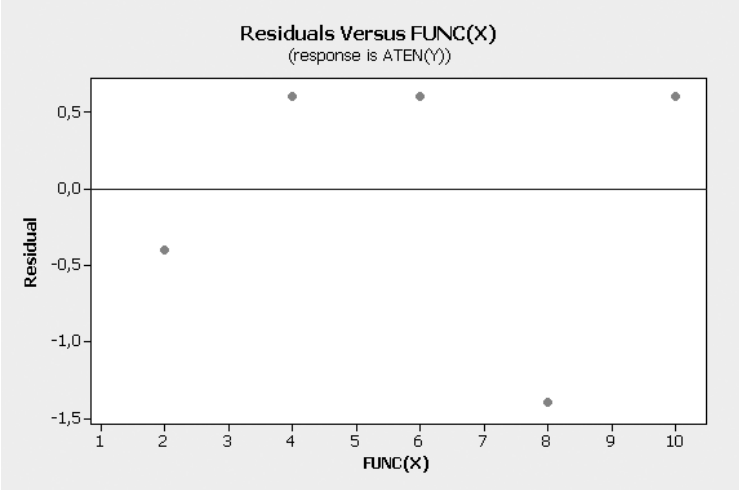


Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regression*** > ***Fitted Line Plot*** > ***Linear*** > ***Graph*** y se rellenan los campos correspondientes. Seleccione ***OK*** para obtener el gráfico de residuos.

Obtenemos la gráfica que aparece en la figura 18.

Figura 18. Gráfica de los residuos en función de la variable independiente

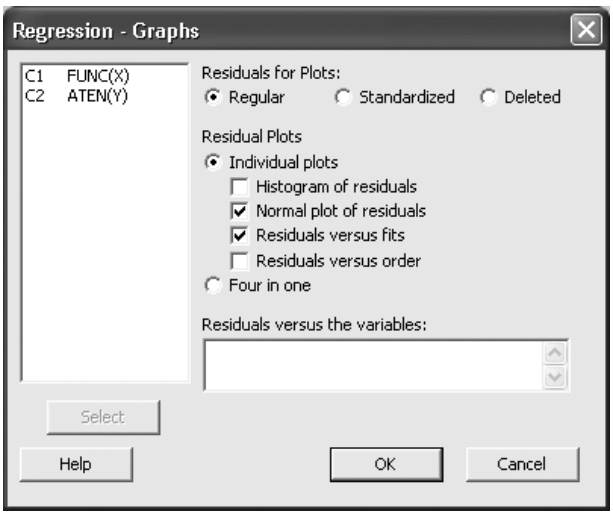


Los valores residuales se distribuyen aleatoriamente y no presenta ningún tipo de estructura, por consiguiente concluimos que la gráfica de los residuos no muestra evidencia de incumplir el supuesto de linealidad y podemos por ahora concluir que el modelo lineal simple es válido para el ejemplo "Estudio de los servicios ofrecidos por un centro de documentación".

En el mismo gráfico podemos observar que los residuos tienen varianza constante ya que parecen estar en la banda horizontal.

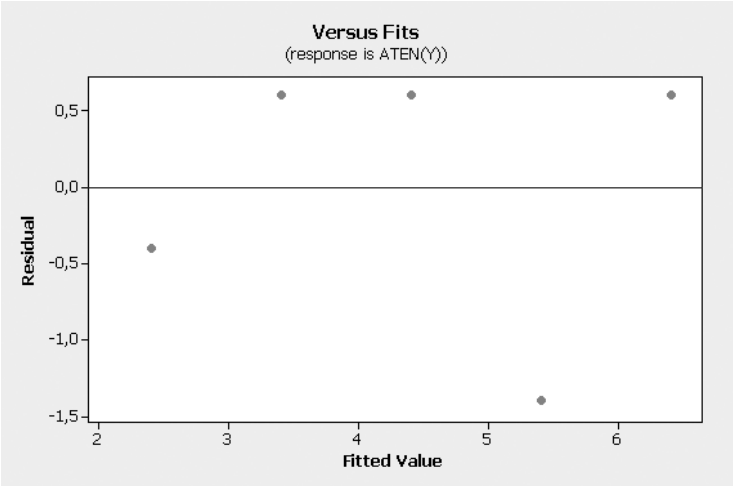
A fin de comprobar si se cumplen el resto de las condiciones del modelo, seleccionamos la opción ***Graphs*** y completamos los campos según se indica en la figura 19:

Figura 19. Pasos a seguir para crear un gráfico de los residuos en función de los valores estimados (fits)



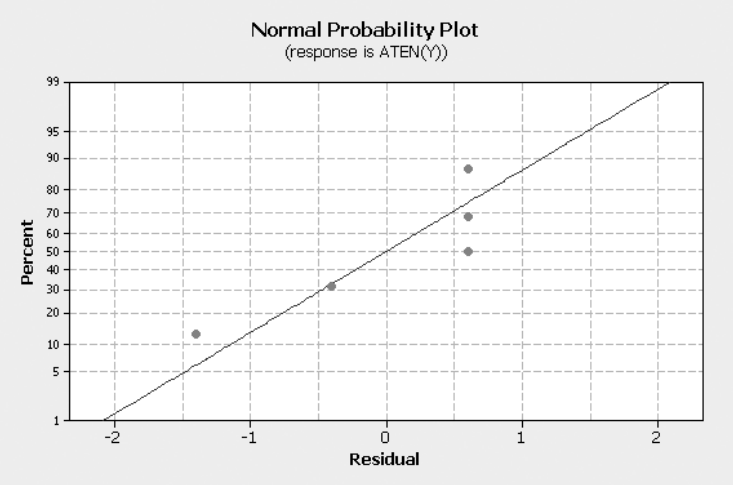
La figura 20 presenta el gráfico de los valores residuales frente a los valores estimados y el significado es análogo al de la figura 18. Los residuos se distribuyen aleatoriamente, no presenta ningún tipo de estructura, y podemos concluir que es válido el modelo lineal simple.

Figura 20. Gráfica de los residuos en función de los valores estimados



En la gráfica de la figura 21 podemos comprobar que los residuos siguen una distribución aproximadamente normal, ya que los puntos se acercan bastante a una recta (esta hipótesis sólo plantearía dificultades si estos puntos se alejasen de la forma lineal):

Figura 21. Gráfica de probabilidad normal



Inferencia en la regresión: constrastes de hipótesis e intervalos de confianza

Al hacer un análisis de regresión se comienza proponiendo una hipótesis acerca del modelo adecuado de la relación entre las variables dependiente e independiente. Para el caso de regresión lineal simple, el modelo de regresión supuesto es:

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A continuación aplicamos el método de mínimos cuadrados para determinar los valores de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros del modelo. La ecuación estimada de regresión que resulta es:

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

Ya hemos visto que el valor del coeficiente de determinación (R^2) es una medida de bondad de ajuste de esta ecuación. Sin embargo, aun con un valor grande de R^2 no se debería usar la ecuación de regresión sin antes efectuar un análisis de la adecuación del modelo propuesto. Para ello se debe determinar el significado (o importancia estadística) de la relación. Las pruebas de significación en el análisis de regresión se basan en los siguientes supuestos acerca del término del error ε:

1) El término del error ε es una variable aleatoria con distribución normal con media, o valor esperado, igual a cero.

2) La varianza del error, representada por σ^2, es igual para todos los valores de x.

3) Los valores de los errores son independientes.

Base para la inferencia sobre la pendiente de la regresión poblacional

Sea β_1 la pendiente del modelo de regresión y $\hat{\beta}_1$ su estimación por mínimos cuadrados (basada en observaciones muestrales). Si se cumplen los supuestos acerca del término del error expuestos anteriormente, la pendiente del modelo de regresión β_1 se distribuye como una *t* de Student con $(n - 2)$ grados de libertad.

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

Para obtener el estadístico de contraste, calcularemos:

$S_{\hat{\beta}_1}$ es la desviación estándar estimada de β_1,

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}$$

s es el error estándar de los estimados. Para calcularlo, se divide la suma de las desviaciones al cuadrado por $n - 2$, que son los grados de libertad.

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

En el análisis de regresión aplicado, primero se desea conocer si existe una relación entre las variables *X* e *Y*. En el modelo se ve que si β_1 es cero, entonces no existe relación lineal: *Y* no aumentaría o disminuiría cuando aumenta *X*. Para averiguar si existe una relación lineal, se puede contrastar la hipótesis

$$H_0\colon\ \beta_1 = 0$$

frente a

$$H_1\colon\ \beta_1 \neq 0$$

Se puede contrastar esta hipótesis utilizando el estadístico *t* de Student

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}},$$

que se distribuye como una *t* de Student con *n* – 2 grados de libertad. La mayoría de los programas que se emplean para estimar regresiones la desviación estándar de los coeficientes y el estadístico *t* de Student para $\beta_1 = 0$. Las figuras 10 y 13 muestran respectivamente las salidas de Minitab y Excel correspondientes al ejemplo del estudio de los servicios ofrecidos por un centro de documentación.

En el caso del modelo de ejemplo, el coeficiente de la pendiente es $\hat{\beta}_1 = 0,50$ con una desviación estándar $S_{\hat{\beta}_1} = 0,1633$. Para saber si existe relación entre la atención al usuario, Y, y el funcionamiento global, X, se puede contrastar la hipótesis $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$. Este resultado se obtiene en el caso de un contraste de dos colas con un nivel de significación $\alpha = 0,05$ y 3 grados de libertad.

El estadístico *t* calculado es:

$$t = \frac{0,50 - 0}{0,1633} = 3,06$$

Recordad

El *p*-valor es la probabilidad de que una variable aleatoria supere el valor observado para el estadístico de contraste.

- Si *p*-valor $< \alpha$, se rechaza H_0.
- Si *p*-valor $\geq \alpha$, no se rechaza H_0.

El estadístico *t* resultante, *t* = 3,06, mostrado en la salida de regresión de la figura 22, es la prueba definitiva para rechazar o aceptar la hipótesis nula. En este caso el *p*-valor es 0,055; como *p*-valor > 0,05 (no podemos rechazar la H_0: $\beta_1 = 0$ al nivel de significación de $\alpha = 0,05$), se acepta que $\hat{\beta}_1 = 0$. Por lo tanto, no se puede afirmar que exista una relación lineal entre las valoraciones del funcionamiento global y la atención al usuario a un nivel de confianza del 95% (nivel de significación del 0,05).

Figura 22. Resumen de la figura 10. Resultados del análisis de regresión. Minitab

Regression Analysis: ATEN(Y) versus FUNC(X)

The regression equation is

ATEN(Y) = 1,40 + 0,500 FUNC(X)

Predictor	Coef	SE Coef	T	P
Constant	1,400	1,083	1,29	0,287
FUNC(X)	0,5000	0,1633	3,06	0,055

S = 1,03280 R-Sq = 75,8% R-Sq(adj) = 67,7%

Si el nivel de significación se hubiera fijado del 10% ($\alpha = 0,10$), se podría rechazar H_0, ya que el *p*-valor < 0,10, los resultados indicarían que $\beta_1 \neq 0$ y en este caso se podría decir que a un nivel de confianza del 90% existe relación lineal entre ambas variables.

Intervalo de confianza para la pendiente

Se puede obtener intervalos de confianza para la pendiente β_1 del modelo de regresión utilizando los estimadores de los coeficientes y de las varianzas que se han desarrollado y el razonamiento utilizado en el módulo 2.

Si los errores de la regresión ε_i siguen una distribución normal y se cumplen los supuestos de la regresión, se obtiene un intervalo de confianza al $(1 - \alpha)\%$ de la pendiente del modelo de regresión simple β_1 de la siguiente forma:

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

donde $t_{n-2,\alpha/2}$ es el número para el que

$$P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$$

el estadístico t_{n-2} sigue una distribución t de Student con $(n - 2)$ grados de libertad.

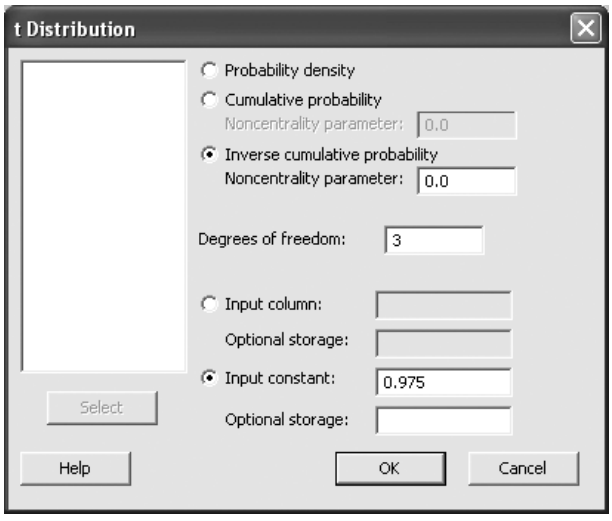
En la salida del análisis de regresión de la atención al usuario respecto al funcionamiento global del centro de documentación de la figura 22, se observa que

$$n = 5 \quad \hat{\beta}_1 = 0{,}50 \quad S_{\hat{\beta}_1} = 0{,}1633$$

Para obtener el intervalo de confianza al 95% de β_1, $(1 - \alpha) = 0{,}95$ y $n - 2 = 3$ grados de libertad, es necesario calcular el valor crítico de la t-Student. En este caso con $n - 2 = 5 - 2 = 3$ grados de libertad y $\alpha/2 = 0{,}05/2 = 0{,}025$. Se puede obtener utilizando las tablas de la distribución t de Student o con el ordenador.

Si se utiliza Minitab, los pasos a seguir se muestran en la figura 23.

Figura 23. Pasos a seguir para calcular el valor crítico *t*



Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***t*** y se rellenan los campos en la ventana correspondiente. Seleccionad ***OK*** para obtener el *output* de la figura 24.

Figura 24. Resultados de cálculo del valor crítico *t*. Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 3 DF

P( X <= x )        x
      0.975  3.18245
```

el valor de $t_{n-2,\alpha/2} = t_{3;0,025} = 3,18$

Por lo tanto, el intervalo de confianza al 95% será

$$0,50 - (0,1633)\ (3,18) < \beta_1 < 0,50 + (0,1633)\ (3,18)$$

O sea

$$-0,019 < \beta_1 < 1,0193$$

Por tanto, el intervalo de confianza buscado es: $0,50 \pm 3,18245 \cdot 0,1633$, *i. e.*, se puede afirmar con una probabilidad del 95% que β_1 se encuentra en el intervalo de extremos –0,0197 y 1,0197.

En la tabla 4 se presentase el intervalo de confianza calculado con Excel. El resumen muestra en las ultimas columnas los valores estimados de intervalo de confianza del 95% para los parámetros de regresión β_0 y β_1, también las desviaciones estándar estimadas (columna *Error típico*), el valor estadístico *t* (columna *Estadístico t*) y los *p*-valores (columna *Probabilidad*).

Tabla 4. Resumen de la figura 13 (Resultados del análisis de regresión. Excel)

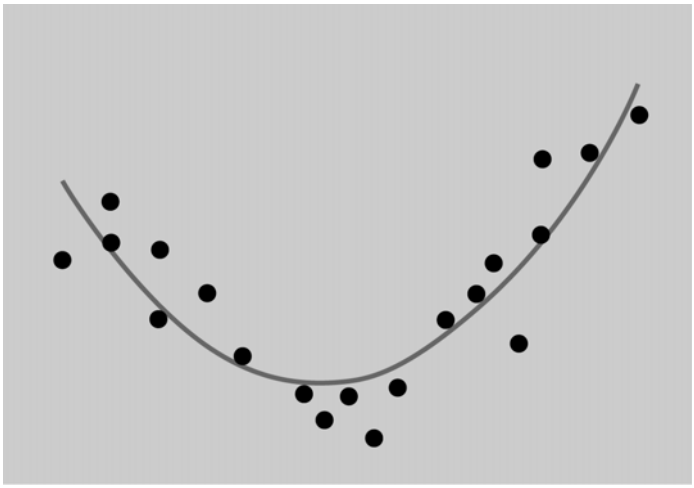
	Coeficientes	**Error típico**	**Estadístico *t***	**Probabilidad**	**Inferior 95%**	**Superior 95%**
Intercepción	1,4	1,08320512	1,29246066	0,286745	–2,047242	4,847242134
Funcionamiento (*X*)	0,5	0,16329932	3,06186218	0,054913	–0,019691	1,019691305

3.2. Modelos de regresión simple no lineales: modelo cuadrático y cúbico

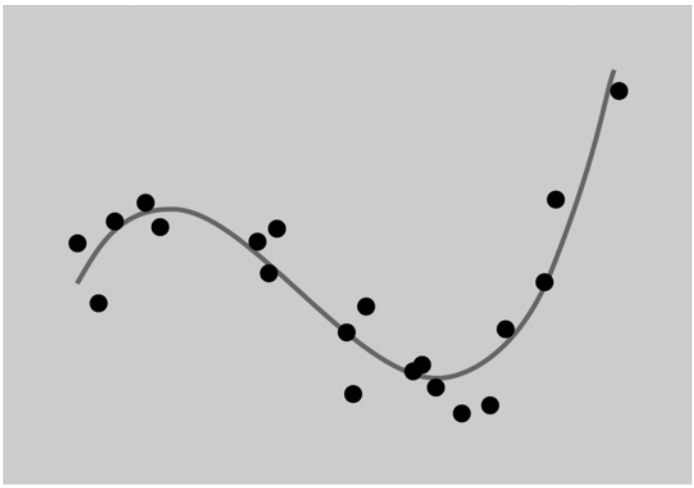
Existen algunas relaciones que no son estrictamente lineales, y se pueden desarrollar métodos con el fin de poder utilizar los métodos de regresión para estimar los coeficientes del modelo.

Aparte de los modelos de regresión lineales, se pueden establecer otros que no son lineales, entre los cuales destacamos: el modelo cuadrático y el cúbico, que son modelos curvilíneos. Cada modelo corresponde con el grado de la ecuación, siendo *Y* la respuesta y *X* la variable predictora, β_0 la ordenada en el origen, y β_1, β_2, y β_3 los coeficientes. Es importante escoger el modelo apropiado cuando se modelizan datos usando regresión y análisis de tendencia.

Modelo cuadrático: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$



Modelo cúbico: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$



Para determinar qué modelo utilizar, se representan previamente los datos (diagrama de dispersión) y se calcula el coeficiente de correlación lineal de Pearson. Conviene recordar que dicho coeficiente "*r*" mide el grado de asociación que existe entre las variables *X* e *Y* cuando se ajusta a su nube de puntos una *línea recta,* pero no mide el grado de ajuste de una curva a la nube de puntos. Podría darse el caso de que la relación entre las variables fuera grande, sólo que distribuida a lo largo de una curva, en cuyo caso, al ajustar a una recta se obtendría un coeficiente de correlación lineal "*r*" y un coeficiente de determinación "R^2" bajo. Calcularíamos el ajuste simultáneo a los modelos no lineales (cuadrático y cúbico) y se calcularían los coeficientes de determinación para ambos modelos para determinar la bondad del ajuste. El mejor modelo será el que presente el valor más elevado de R^2.

Los métodos de inferencia para los modelos no lineales transformados son los mismos que se han desarrollado para los modelos lineales. Así, si se tiene un modelo cuadrático, el efecto de una variable *X* está indicado por los coeficientes tanto de los términos lineales como de los términos cuadráticos.

Ejemplo. Número de visitantes a un museo (estimación de un modelo cuadrático utilizando Minitab)

Se desea estudiar la variación entre el número de visitantes a un museo en función del número de obras visitadas. La tabla 5 muestra el número de visitantes y el número de obras visitadas. Se han seleccionado aleatoriamente los datos correspondientes a 6 días.

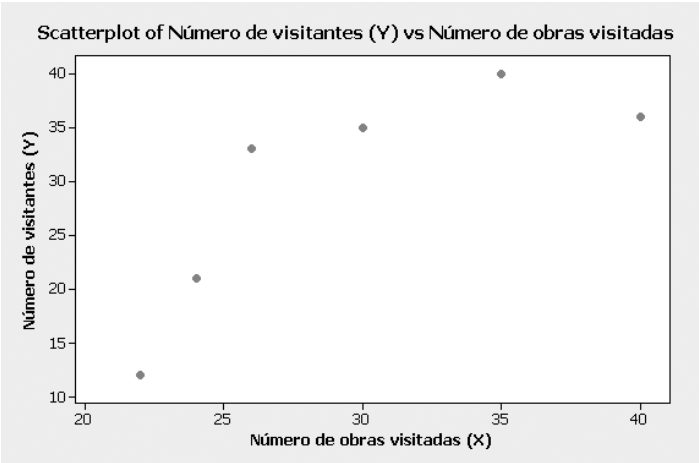
Tabla 5. Número de visitantes a un museo

Número de visitantes (*Y*)	22	24	26	30	35	40
Número de obras visitadas (*X*)	12	21	33	35	40	36

Con estos datos podemos deducir si existe relación entre ambas variables y si las variables están relacionadas establecer el mejor modelo.

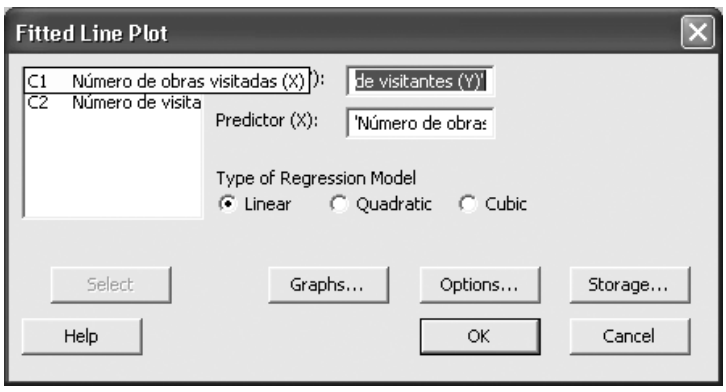
La figura 25 representa el diagrama de dispersión para estos datos. El diagrama de dispersión indica que posiblemente hay una relación curvilínea entre el número de de obras visitadas y el número de visitantes.

Figura 25. Diagrama de dispersión para ejemplo 2. Minitab



Antes de deducir la ecuación curvilínea entre número de obras visitadas y número de visitantes, se realiza el ajuste a un modelo de regresión lineal simple (de primer orden) siguiendo los pasos que muestra la figura 26.

Figura 26. Pasos a seguir para comprobar el modelo lineal



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regresión*** > ***Fitted Line Plot*** > ***Linear*** y se rellenan los campos en la ventana correspondiente. Seleccionad ***OK*** para obtener el *output* de la figura 27 y 28.

Figura 27. Gráfica de la ecuación de regresión de mínimos cuadrados

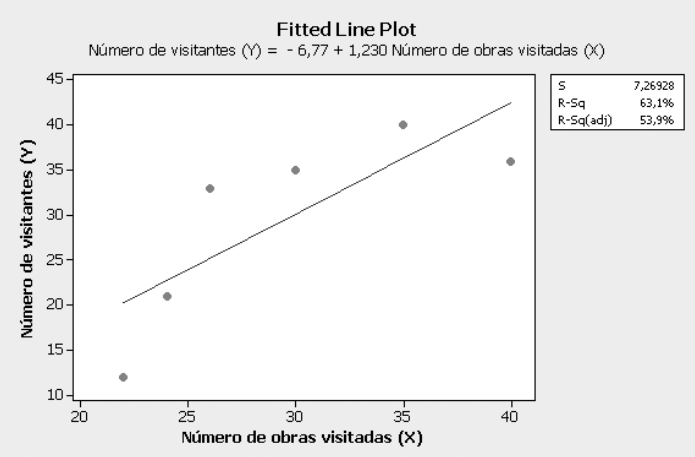


Figura 28. Resultados del análisis de regresión. Modelo lineal simple

Regression Analysis: Número de visitantes (Y) versus Número de obras visitada

The regression equation is

Número de visitantes (Y) = - 6,77 + 1,230 Número de obras visitadas (X)

S = 7,26928 R-Sq = 63,1% R-Sq(adj) = 53,9%

Analysis of Variance

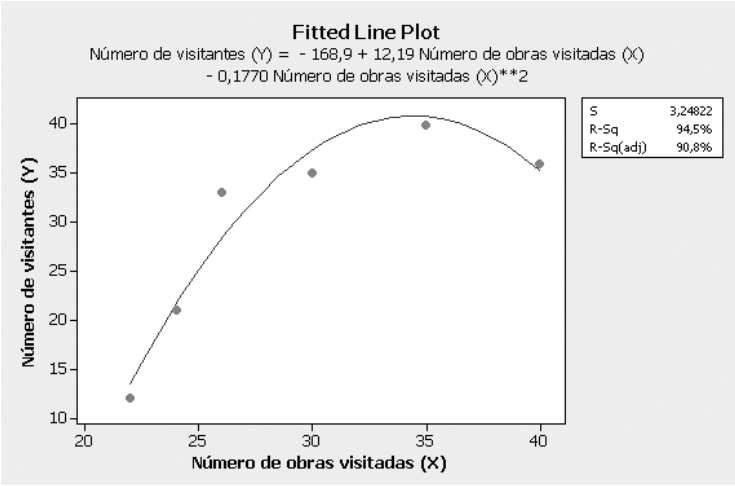
Source	DF	SS	MS	F	P
Regression	1	362,130	362,130	6,85	0,059
Error	4	211,370	52,842		
Total	5	573,500			

Observamos que con el modelo lineal se explica un 63,1% de la variabilidad del número de visitantes (R^2 = 63,1%). La ecuación de ajuste es:

Número de visitantes (Y) = –6,77 + 1,230; número de obras visitadas (X)

A continuación se presenta el ajuste del modelo cuadrático y, como se puede ver en la gráfica de la figura 29, los puntos se ajustan mejor a una función no lineal.

Figura 29. Gráfica del ajuste cuadrático



Observamos que el ajuste cuadrático es muy bueno con un valor de R^2= 94,5% que mejora el ajuste lineal. La ecuación de ajuste es:

Número de visitantes (Y) = –168,9 + 12,19; número de obras visitadas –0,1770; número de obras visitadas^2

Figura 30. Resultados del análisis de regresión. Modelo cuadrático

Polynomial Regression Analysis: Número de visita versus Número de obras

The regression equation is
Número de visitantes (Y) = - 168,9 + 12,19 Número de obras visitadas (X) - 0,1770 Número de obras visitadas (X)**2

S = 3,24822 R-Sq = 94,5% R-Sq(adj) = 90,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	541,847	270,924	25,68	0,013
Error	3	31,653	10,551		
Total	5	573,500			

Sequential Analysis of Variance

Source	DF	SS	F	P
Linear	1	362,130	6,85	0,059
Quadratic	1	179,717	17,03	0,026

A continuación se presenta el ajuste del modelo cúbico:

Figura 31. Gráfica del ajuste cúbico

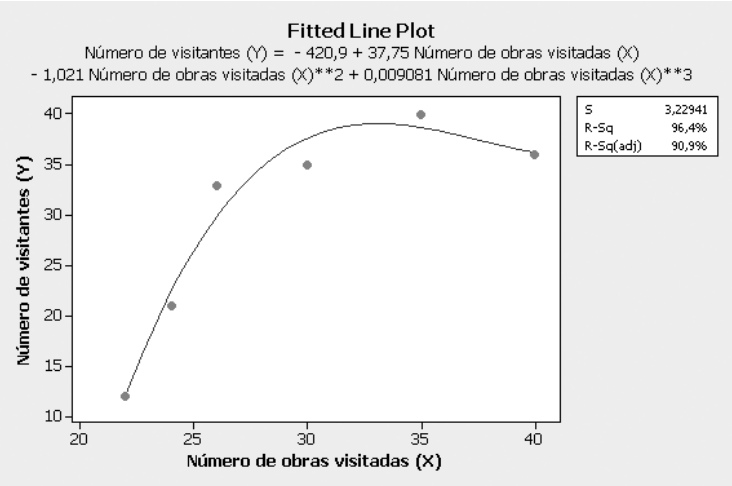


Figura 32. Resultados del análisis de regresión. Modelo cúbico

Polynomial Regression Analysis:
Número de visita versus Número de obras

The regression equation is

Número de visitantes (Y) =
- 420,9 + 37,75 Número de obras visitadas (X)
- 1,021 Número de obras visitadas (X)**2
+ 0,009081 Número de obras visitadas (X)**3

S = 3,22941 R-Sq = 96,4% R-Sq(adj) = 90,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	552,642	184,214	17,66	0,054
Error	2	20,858	10,429		
Total	5	573,500			

Sequential Analysis of Variance

Source	DF	SS	F	P
Linear	1	362,130	6,85	0,059
Quadratic	1	179,717	17,03	0,026
Cubic	1	10,795	1,04	0,416

El ajuste al modelo cúbico también es bueno con un valor alto de $R^2 = 96,4$ % que mejora el ajuste lineal e iguala al cuadrático.

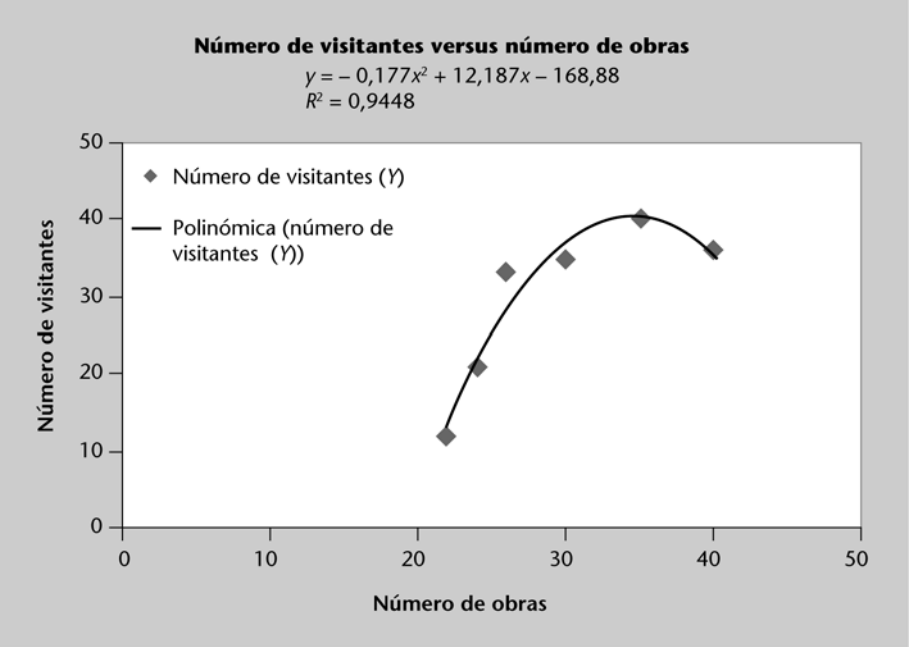
La ecuación de ajuste es:

$$\text{Número de visitantes } (Y) = -420,9 + 37,75 \text{ Número de obras visitadas} - 1,021 \text{ Número de obras visitadas}^2 + 0,009081 \text{ Número de obras visitadas}^3$$

Analizando la significatividad de los modelos mediante el *p*-valor, el modelo cuadrático por tener el menor *p*-valor (*p*-valor = 0,026) es el más significativo, por lo que se elegiría como mejor ajuste el cuadrático.

La figura 33 muestra el correspondiente *output* que ofrece **Microsoft Excel** del ejemplo 2. "**Número de visitantes a un museo**". Seleccionando la opción Tipo de tendencia poligonal de segundo orden, que coincide con el ajuste cuadrático elegido con Minitab (figuras 29 y 30). La ecuación de ajuste y el valor de R^2 coinciden con las obtenidas con Minitab.

Figura 33. Gráfica del ajuste cuadrático. Excel



3.3. Transformaciones de modelos de regresión no lineales: modelos exponenciales

Algunas relaciones entre variables pueden analizarse mediante modelos exponenciales. Por ejemplo las relaciones entre la variable tiempo (X) y otras variables (Y) como la población, los precios de algunos productos, el número de ordenadores infectados son exponenciales. Los modelos exponenciales de demanda se utilizan mucho en el análisis de conducta del mercado.

Curva en un modelo exponencial

En el modelo lineal se ajusta la nube de puntos a una recta de ecuación:

$$y = a + bx$$

En el modelo exponencial se ajusta a una curva de ecuación:

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

El modelo exponencial es del tipo:

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

donde k y a son valores constantes.

Para tratar este modelo se realizará una transformación de las variables de manera que el modelo 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:

Propiedades de los logaritmos

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

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

Esta ecuación muestra un modelo lineal entre las variables X y $\ln Y$.

Si representamos el diagrama de dispersión de los puntos $(x_i, \ln y_i)$ y la nube de puntos presenta una estructura lineal, se puede pensar que entre las variables X e Y hay una relación exponencial.

4. Modelos de regresión múltiple

En el apartado 3.1 hemos presentado el método de regresión simple para obtener una ecuación lineal que predice una variable dependiente o endógena en función de una única variable independiente o exógena: número total de libros vendidos en función del precio. Sin embargo, en muchas situaciones, varias variables independientes influyen conjuntamente en una variable dependiente. La regresión múltiple permite averiguar el efecto simultáneo de varias variables independientes en una variable dependiente utilizando el principio de los mínimos cuadrados.

Existen muchas aplicaciones de la regresión múltiple para dar respuesta a preguntas como las siguientes:

¿En qué medida el precio de un ordenador depende de la velocidad del procesador, de la capacidad del disco duro y de la cantidad de memoria RAM?

¿Cómo relacionar el índice de impacto de una revista científica con el número total de documentos publicados y el número de citas por documento?

¿El sueldo de un titulado depende de la edad, de los años que hace que acabó los estudios, de los años de experiencia en la empresa, etc.?

¿El precio de alquiler de un piso depende de los metros cuadrados de superficie, de la edad de la finca, de la proximidad al centro de la ciudad, etc.?

¿El precio de un coche depende de la potencia del motor, del número de puertas y de multitud de accesorios que puede llevar: airbag, ordenador de viaje, equipo de alta fidelidad volante deportivo, llantas especiales, etc.?

Los métodos para ajustar modelos de regresión múltiple se basan en el mismo principio de mínimos cuadrados explicado en el apartado 3.1.

Nuestro objetivo es aprender a utilizar la regresión múltiple para crear y analizar modelos. Por lo tanto se aprenderá cómo funciona la regresión múltiple y algunas directrices para interpretarla. Comprendiendo perfectamente la regresión múltiple, es posible resolver una amplia variedad de problemas aplicados. Este estudio de los métodos de regresión múltiple es paralelo al de regresión simple. El primer paso para desarrollar un modelo consiste en la selección de las variables y de la forma del modelo. A continuación, estudiamos el método de mínimos cuadrados y analizamos la variabilidad para identificar los efectos de cada una de las variables de predicción.

Después estudiamos la estimación, los intervalos de confianza y el contraste de hipótesis. Utilizamos aplicaciones informáticas para indicar cómo se aplica la teoría a problemas reales.

Desarrollo del modelo

Cuando se aplica la regresión múltiple, se construye un modelo para explicar la variabilidad de la variable dependiente. Para ello hay que incluir las influencias simultáneas e individuales de varias variables independientes. Se supone, por ejemplo, que se quiere desarrollar un modelo que prediga el precio de las impresoras láser que desea liquidar una empresa. Un estudio inicial indicaba que el precio estaba relacionado con el número de páginas por minuto que la impresora es capaz de imprimir y los años de antigüedad de la impresora en cuestión. Eso llevaría a especificar el siguiente modelo de regresión múltiple con dos variables independientes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

donde:

Y = precio en euros
X_1 = número de páginas impresas por minuto
X_2 = años de antigüedad de la impresora

La tabla 6 contiene 12 observaciones de estas variables. Utilizaremos estos datos para desarrollar el modelo lineal que prediga el precio de las impresoras en función del número de páginas impresas por minuto y de los años de antigüedad de la impresora.

Tabla 6. Datos del ejemplo "Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo".

X_1	6	6	6	6	8	8	8	8	12	12	12	12
X_2	6	4	2	0	6	4	2	0	6	4	2	0
Y	466	418	434	487	516	462	475	501	594	553	551	589

Nota

En el caso general emplearemos k para representar el número de variables independientes.

Pero antes de poder estimar el modelo es necesario desarrollar y comprender el método de regresión múltiple.

El modelo de regresión múltiple es

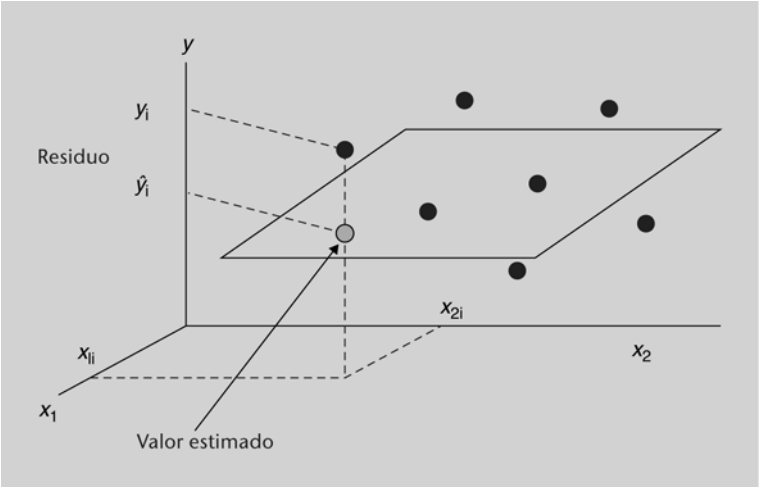
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

Donde β_0, β_1, β_2,... , β_k son los coeficientes de las variables independientes o exógenas y ε (letra griega épsilon) es el error o residuo y es una variable alea-

toria. Más adelante describiremos todos los supuestos del modelo para el modelo de regresión múltiple y para ε.

Los coeficientes en general no se conocen y se deben determinar a partir de los datos de una muestra y empleándose el **método de mínimos cuadrados** para llegar a la ecuación estimada de regresión que más se aproxima a la relación lineal entre las variables independientes y dependiente. El procedimiento es similar al utilizado en la regresión simple. En la regresión múltiple el mejor ajuste es un hiperplano en espacio *n*-dimensional (espacio tridimensional en el caso de dos variables independientes, figura 34).

Figura 34. Gráfica de la ecuación de regresión, para el análisis de regresión múltiple con dos variables independientes



Los valores estimados de la variable dependiente se calculan con la ecuación estimada de regresión múltiple:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

Donde $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ son los valores de los estimadores de los parámetros o coeficientes de la ecuación de regresión múltiple, la deducción de estos coeficientes requiere el empleo del álgebra de matrices y se sale del propósito de este texto. Así, al describir la regresión múltiple lo enfocaremos hacia cómo se pueden emplear los programas informáticos de cálculo para obtener la ecuación estimada de regresión y otros resultados y su interpretación, y no hacia cómo hacer los cálculos de la regresión múltiple.

Criterio de mínimos cuadrados

$$\min \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Donde:

y_i = valor observado de la variable dependiente en la *i*-ésima observación.

$\hat{y}_i$ = valor estimado de la variable dependiente en la *i*-ésima observación.

Considerando de nuevo el modelo de regresión con dos variables independientes del ejemplo 3. "Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo". Utilizando los datos de la tabla 6 se ha estimado un modelo de regresión múltiple, que se observa en la salida Minitab de la figura 35.

Figura 35. Resultados del ejemplo 3 del análisis de regresión múltiple para dos variables independientes

```
Regression Analysis: Y versus X1; X2

The regression equation is
Y = 330 + 20,2 X1 - 0,50 X2

Predictor     Coef  SE Coef      T      P
Constant   330,38    29,40  11,24  0,000
X1         20,187    3,056   6,61  0,000
X2         -0,500    3,410  -0,15  0,887

S = 26,4100   R-Sq = 82,9%   R-Sq(adj) = 79,1%
```

Pasos a seguir

Para estimar el modelo de regresión múltiple introducimos los datos en Minitab para calcular el modelo.

Se sigue la ruta ***Stat*** > ***Regression*** > ***Regression*** y se rellenan los campos en la ventana correspondiente. Se selecciona ***OK*** para obtener el análisis de regresión.

Los coeficientes estimados se identifican en la salida de los programas informáticos

La ecuación de regresión múltiple es: $Y = 330 + 20{,}2\ X1 - 0{,}50\ X2$

La interpretación de los coeficientes es la siguiente:

- Coeficiente de $X1$ (20,2 euros): sería el aumento del precio de la impresora cuando aumenta en una unidad el número de páginas por minuto que imprime, cuando las demás variables independientes se mantienen constantes (en este caso $X2$, la antigüedad no varía).
- Coeficiente $X2$ (–0,50 euros): sería la disminución del precio por cada año más de antigüedad de la impresora, cuando $X1$ permanece constante (el número de páginas por minuto no varía).
- Término independiente (330): no tiene mucho sentido interpretarlo en este caso ya que representaría el precio de una impresora que no puede imprimir ninguna página.

El coeficiente de determinación múltiple

Coeficiente de determinación R^2

El coeficiente de determinación R^2 en Minitab se designa como ***R-sq***.

En la regresión lineal simple vimos que la suma total de cuadrados se puede descomponer en dos componentes: la suma de cuadrados debida a la regresión y la suma de cuadrados debida al error. Este mismo procedimiento se aplica a la suma de cuadrados de la regresión múltiple. El coeficiente de determinación múltiple mide la bondad de ajuste para la ecuación de regresión múltiple. Este coeficiente se calcula como sigue:

$$R^2 = \frac{SSR}{SST}$$

Se puede interpretar como la proporción de variabilidad de la variable dependiente que se puede explicar con la ecuación de regresión múltiple. Cuando se

multiplica por cien, se interpreta como la variación porcentual de y que se explica con la ecuación de regresión.

En general, R^2 aumenta cuando se añaden variables independientes (variables explicativas o predictoras) al modelo. Si se añade una variable al modelo, R^2 se hace mayor (o permanece igual), aun cuando esa variable no sea estadísticamente significativa. El **coeficiente de determinación corregido** o **adjusted *R-sq*** elimina el efecto que se produce sobre el *R-sq* cuando se aumenta el número de variables independientes.

El **coeficiente de correlación múltiple** se define como la raíz cuadrada positiva del *R-sq*. Este coeficiente nos proporciona la correlación existente entre la variable dependiente (respuesta) y una nueva variable formada por la combinación lineal de los predictores.

Continuando con el **ejemplo 3. "Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo"**, interpretaremos el resultado del coeficiente de determinación *R-Sq* = 82,9% (figura 35). Significa que el 82,9% de la variabilidad en el precio de impresoras láser se explica con la ecuación de regresión múltiple, con el número de páginas que imprime por minuto y los años de antigüedad. La figura 35 muestra que el valor *R-Sq* (adj) = 79,1%, significa que si se agregase una variable independiente (predictora) el valor de R^2 no aumentaría.

Supuestos del modelo

Los supuestos acerca del término del error ε, en el modelo de regresión múltiple, son similares a los del modelo de regresión lineal simple.

Por simplicidad, consideraremos un modelo de regresión con sólo dos variables explicativas ($X1$ y $X2$). La ecuación de regresión múltiple, con dos variables independientes será:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

donde los β_i representan coeficientes reales y ε representa el error aleatorio.

1) El error es una variable aleatoria cuyo valor medio u esperado es cero; esto es $E(\varepsilon) = 0$.

2) Para todos los valores de $X1$ y $X2$, los valores de Y (o, alternativamente, los valores de (ε) muestran varianza constante σ^2.

3) Para cada valor de $X1$ y $X2$, la distribución de Y (o, alternativamente, la de ε) es aproximadamente normal.

4) Los valores de *Y* obtenidos (o, alternativamente, los de ε) son independientes.

Hay toda una serie de gráficos que nos pueden ayudar a analizar los resultados de una regresión lineal múltiple y a comprobar si se cumplen o no los supuestos anteriores:

1) Un gráfico de la variable dependiente frente a los valores estimados por el modelo nos ayudará a comprobar visualmente la bondad del ajuste.

2) Representando los residuos frente a los valores estimados podremos comprobar la variabilidad vertical en los datos. Ello nos permitirá saber si se cumple el supuesto de varianza constante.

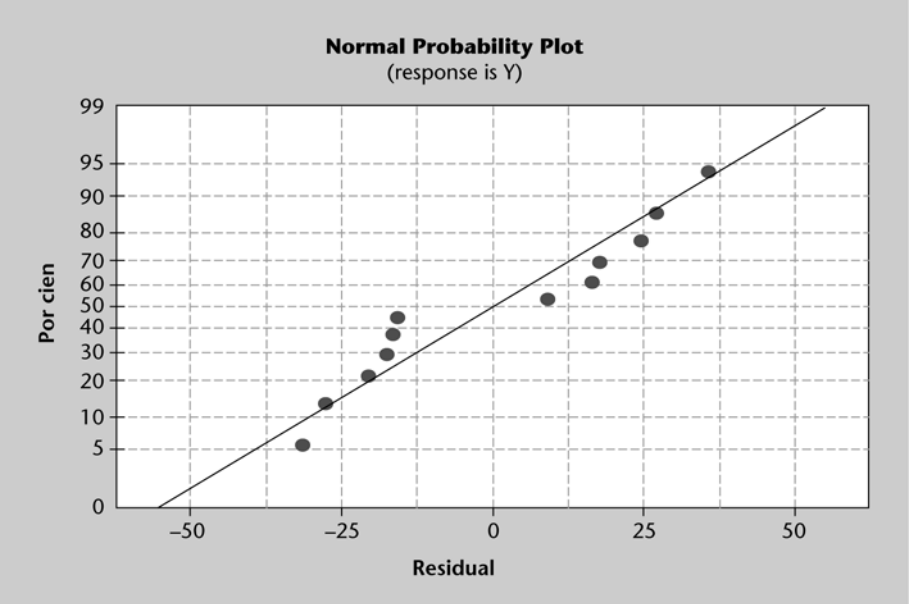
3) Un gráfico de residuos frente a cada una de las variables explicativas puede revelar problemas adicionales que no se hayan detectado en el gráfico anterior.

4) Para comprobar la hipótesis de normalidad suele ser conveniente realizar un test y un gráfico de normalidad para los residuos.

En el ejemplo se comprueba si se cumplen los supuestos del modelo utilizado.

En la gráfica de la figura 36 podemos comprobar que los residuos siguen una distribución aproximadamente normal, ya que los puntos se acercan bastante a una recta.

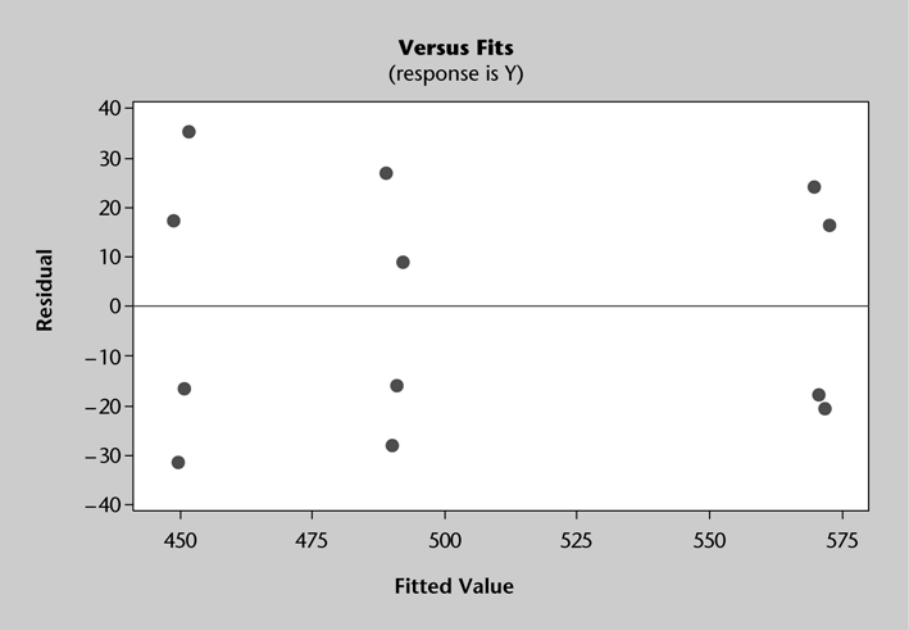
Figura 36. Gráfica de probabilidad normal



La figura 37 presenta el gráfico de los valores residuales frente a los valores estimados. Los residuos se distribuyen aleatoriamente, no presenta ningún tipo de estructura y podemos concluir que es válido el modelo lineal múltiple. También observamos en este gráfico que las varianzas de los residuos son constantes. El procedimiento y la interpretación de los supuestos se explica-

ron en el apartado 3.1. (modelos de regresión lineal simple) y son iguales a los correspondientes de regresión múltiple.

Figura 37. Gráfica de los residuos en función de los valores estimados



Pruebas de significación

Las pruebas de significación que empleamos en la regresión lineal fueron una prueba *t* y una prueba *F*. En ese caso, ambas pruebas dan como resultado la misma conclusión: si se rechaza la hipótesis nula, la conclusión es que $\beta_1 \neq 0$. En la regresión múltiple la prueba *t* y *F* tienen distintas finalidades.

La prueba *F* se usa para determinar si hay una relación significativa entre la variable dependiente y el conjunto de todas las variables independientes. En estas condiciones se le llama **prueba de significación global.**

La prueba *t* se aplica para determinar si cada una de las variables independientes tiene significado. Se hace una prueba *t* por separado para cada variable independiente en el modelo y a cada una de estas pruebas se le llama **prueba de significación individual.**

Prueba *F* o análisis de la varianza en regresión lineal

Las hipótesis para la prueba *F* implican los parámetros del modelo de regresión múltiple:

Hipótesis nula: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Hipótesis alternativa: H_1: uno o más de los parámetros no es igual a cero (al menos un parámetro es $\neq 0$). Debemos fijar el nivel se significación α.

Si se rechaza H_0 tendremos suficiente evidencia estadística para concluir que uno o más de los parámetros no es igual a cero y que la relación general entre y y el conjunto de variables independientes x_1, x_2,.....x_k es significativa. Sin embargo, si no podemos rechazar H_0, no tenemos la evidencia suficiente para llegar a la conclusión de que la relación es significativa.

Para realizar el contraste debemos calcular el estadístico de contraste F. **El estadístico *F* es una variable aleatoria que se comporta según una distribución *F*-Snedecor con *k* grados de libertad en el numerador (*DF-Regresión*) y *n-k*-1 grados de libertad en el denominador (*DF- Error*).** Donde k son los grados de libertad de la regresión iguales a la cantidad de variables independientes y n es el número de observaciones. Así pues, el estadístico de contraste es:

$$F^* = \frac{SSR/k}{SSE/n-k-1}$$

También podemos definir el estadístico de contraste como el cociente de cuadrados medio (*mean squares*).

Cuadrado medio

Es la suma de cuadrados dividida por los grados de libertad (DF) correspondientes. Esta cantidad se usa en la prueba *F* para determinar si hay diferencias significativas entre medias.

El *cuadrado medio debido a la regresión* o simplemente *regresión del cuadrado medio* se representa por **MSR** (*mean square regression*):

$$MSR = \frac{SSR}{\text{grados de libertad de la regresión}} = \frac{SSR}{k}$$

El cuadrado medio debido a los errores o residuos se llama *cuadrado medio residual* o *cuadrado medio del error* se representa por **MSE** (*mean square residual error*):

$$MSE = \frac{SSE}{\text{grados de libertad del error}} = \frac{SSR}{n-k-1}$$

El valor del estadístico de contraste F podemos definirlo como: $F^* = \frac{MSR}{MSE}$

Regla de decisión del contraste de hipótesis

Podemos actuar de dos maneras:

a) A partir del p-valor. Este valor es: p-valor = $P(F_{\alpha;\, k,n-k-1} > F^*)$, donde F_α es un valor de la distribución F con k grados de libertad en el numerador y n–k–1 grados de libertad en el denominador.

- Si p-valor < α se rechaza la hipótesis nula H_0; por tanto, el modelo en conjunto explica de forma significativa la variable Y. Es decir, el modelo sí contribuye con información a explicar la variable Y.

- Si *p*-valor ≥ α no se rechaza la hipótesis nula H_0; por tanto, no hay una relación significativa. El modelo en conjunto no explica de forma significativa la variable *Y*.

b) A partir de los valores críticos

- Si $F^* > F_{\alpha;\, k,,n-k-1}$, se rechaza la hipótesis nula H_0
- Si $F^* < F_{\alpha;\, k,,n-k-1}$, no se rechaza la hipótesis nula H_0

Los cálculos necesarios se pueden resumir en la tabla 7, conocida como **tabla de análisis de la varianza:**

Tabla de análisis de varianza

En la primera columna se pone la **fuente de variación**, los elementos del modelo responsables de la variación.

En la segunda columna ponemos la **suma de cuadrados** correspondientes.

En la tercera columna ponemos los grados de libertad correspondientes a las **sumas de cuadrados.**

En la cuarta columna con el nombre de **cuadrados medios** se ponen las sumas de cuadrados divididas por los grados de libertad correspondientes. Sólo para *SSR* y *SSE*.

En la quinta columna ponemos el estadístico de contraste ***F***.

Tabla 7. Análisis de varianza para un modelo de regresión múltiple con *k* variables independientes

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*F*
Regresión	SSR	*k*	$MSR = SSR/k$	$F = \frac{MSR}{MSE}$
Error	SSE	$n–k–1$	$MSE = SSE/n–k–1$	
Total	SST	n-1		

Aplicaremos la prueba *F* al ejemplo 3. Con dos variables independientes "número de páginas por minuto (*X*1)" y "antigüedad de la impresora (*X*2)".

Las hipótesis se formulan como sigue:

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_1 \text{ y/o } \beta_2 \text{ no es igual a cero}$$

Fijamos un nivel de significación del 5% (α = 0,05).

La figura 38 muestra los resultados del modelo de regresión múltiple, en la parte de resultados correspondiente al análisis de varianza.

Figura 38. Resultados obtenidos con Minitab.Tabla de análisis de varianza

```
Analysis of Variance

Source          DF     SS     MS      F      P
Regression       2  30444  15222  21.82  0.000
Residual Error   9   6277    697
Total           11  36722
```

El valor del estadístico de contraste es $F^* = 21{,}82$, el *p*-valor = 0,000

Como *p*-valor < α, rechazamos la hipótesis nula, por tanto, el modelo **en conjunto** explica de forma significativa la variable *Y*. Es decir, llegamos a la con-

clusión de que hay una relación significativa entre el precio de la impresora y las dos variables independientes que son número de páginas impresas por minuto ($X1$) y la antigüedad de la impresora ($X2$).

Prueba *t*

Se utiliza para determinar el significado de cada uno de los parámetros individuales. Las hipótesis para la prueba t implican los parámetros del modelo de regresión múltiple, se hace un contraste para cada parámetro β:

Hipótesis nula: H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Hipótesis alternativa: H_1: uno o más de los parámetros no es igual a cero (al menos un parámetro es $\neq 0$). Debemos fijar el nivel se significación α.

El estadístico de contraste es:

$$t^* = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Sigue una distribución t de Student con n-k-1 grados de libertad

Regla de decisión del contraste de hipótesis

Podemos actuar de dos maneras:

a) A partir del p-valor. Este valor es: $p = 2P(t_{n-k-1} > |t^*|)$.

- Si $p < \alpha$ se rechaza la hipótesis nula H_0; se rechaza la hipótesis nula H_0; por tanto, hay una relación lineal entre la variable X_i e Y. Por consiguiente, dicha variable debe permanecer en el modelo.

- Si $p \geq \alpha$ no se rechaza la hipótesis nula H_0; por tanto, no hay una relación lineal entre la correspondiente variable X_i e Y. Decimos que la variable implicada X_i es no explicativa y podemos eliminarla del modelo.

b) A partir de los valores críticos $\pm t\alpha/2, n-k-1$, de manera que:

- Si $|t^*| > t\alpha/2, n-k-1$, se rechaza la hipótesis nula H_0; por tanto, la variable es significativa.

- Si $|t^*| > t\alpha/2, n-k-1$, no se rechaza la hipótesis nula H_0; por tanto, la variable no es significativa. Decimos que la variable implicada X_i no es explicativa.

Si la prueba *F* del ejemplo (figura 38) ha mostrado que la relación múltiple tiene significado, se puede hacer una prueba *t* para determinar el significado de cada uno de los parámetros individuales. El nivel de significación es $\alpha = 0{,}05$. Obsérvese que los valores de los estadísticos *t* aparecen en la figura 39. Los *p*-valores de los contrastes individuales son para el contraste de β_1 el *p*-valor = 0,000 y para β_2, *p*-valor = 0,887.

Figura 39. Resultados obtenidos con Minitab

Predictor	Coef	SE Coef	T	P	VIF
Constant	330.38	29.40	11.24	0.000	
X1	20.187	3.056	6.61	0.000	1.000
X2	-0.500	3.410	-0.15	0.887	1.000

Interpretamos el contraste para el parámetro β_1, la H_0: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$. Como $0{,}000 < 0{,}05$ se rechaza H_0, y, por tanto, la variable $X1$ (número de páginas impresas por minuto) es significativa.

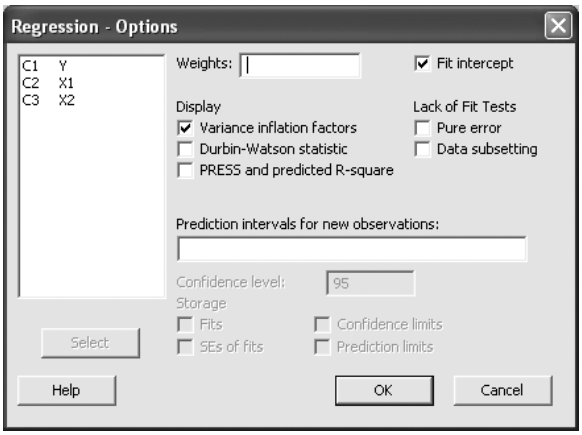
El contraste para el parámetro β_2, la H_0: $\beta_2 = 0$, H_1: $\beta_2 \neq 0$. Como $0{,}887 > 0{,}05$ no podemos rechazar H_0, por lo que la variable $X2$ (antigüedad) no es significativa y podríamos eliminarla del modelo porque no influye significativamente en el precio.

El problema de la multicolinealidad

En los problemas de regresión lineal múltiple esperamos encontrar dependencia entre la variable *Y* y las variables explicativas $X1$, $X2$, ..., Xk, pero en algunos problemas de regresión podemos tener también algún tipo de dependencia entre algunas de las variables Xj. En este caso tenemos información redundante en el modelo. Este fenómeno se llama **multicolinealidad** y suele ser bastante frecuente en los modelos de regresión lineal múltiple.

El término **multicolinealidad** en análisis de regresión múltiple indica la correlación entre variables independientes. La multicolinealidad puede tener efectos muy importantes en las estimaciones de los coeficientes de la regresión y, por tanto, sobre las posteriores aplicaciones del modelo estimado. Cuando las variables independientes están muy correlacionadas no es posible determinar el efecto por separado de una de ellas sobre la variable dependiente. Cuando existe multicolinealidad, los resultados de los contrastes de hipótesis sobre el modelo conjunto y los resultados de los contrastes individuales son aparentemente contradictorios, pero realmente no lo son. Este efecto lo veremos en el ejemplo propuesto (figura 40). Minitab dispone de una opción, llamada **Variance Inflation Factors** (VIF), que nos permite identificar la multicolinealidad entre los predictores del modelo. La figura 40 indica los pasos a seguir.

Figura 40. Pasos a seguir para identificar la multicolinealidad



Pasos a seguir

Se sigue la ruta ***Stat*** > ***Regression*** > ***Regression*** > ***Options*** y se rellenan los campos en la ventana correspondiente. Seleccionad ***OK***.

Ahora la figura 41 de los resultados del análisis de regresión múltiple contiene los valores VIF. Cada coeficiente VIF es de 1,000. Estos valores son bajos, lo que indica que las variables independientes no están correlacionadas. Dado que estos valores indican que el grado de colinearidad es bajo. No existe multicolienalidad en el modelo propuesto.

Figura 41. Resultados del ejemplo 3 del análisis de regresión múltiple, que incluye los *Variance Inflation Factors* (VIF) o factores de inflación de la varianza

Regression Analysis: Y versus X1; X2

The regression equation is
Y = 330 + 20.2 X1 - 0.50 X2

Predictor	Coef	SE Coef	T	P	VIF
Constant	330.38	29.40	11.24	0.000	
X1	20.187	3.056	6.61	0.000	1.000
X2	-0.500	3.410	-0.15	0.887	1.000

S = 26.4100 R-Sq = 82.9% R-Sq(adj) = 79.1%

Usando **Microsoft Excel para obtener el análisis de regresión del ejemplo 3. "Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo".**

Pasos a seguir

Para efectuar la regresión múltiple con **MS Excel**, una vez introducidos los datos en la hoja de cálculo se sigue la siguiente ruta: clic en ***Herramientas*** > ***Análisis de datos*** > ***Regresión*** > ***OK***.

A continuación se seleccionan los rangos de datos de las variables.

La tabla 8 muestra el correspondiente *output* que ofrece **Microsoft Excel**.

Tabla 8. Resultados del análisis de regresión del ejemplo 3. Estudio sobre el precio de impresoras láser en función de su velocidad de impresión y la antigüedad del modelo. Excel

	B	C	D	E	F	G	H
1	Resumen						
2							
3	*Estadísticas de la regresión*						
4	Coeficiente de correlación múltiple	0,910524728228339					
5	Coeficiente de determinación R^2	0,829055280715291					
6	R^2 ajustado	0,791067565318689					
7	Error típico	26.40996235					
8	Observaciones	12					
9							
10	ANÁLISIS DE VARIANZA						
11		*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
12	Regresión	2	30444.29167	15222.14583	21.82429957	0.000353062	
13	Residuos	9	6277.375	697.4861111			
14	Total	11	36721.66667				
15							
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
17	Intercepción	330.375	29.40041791	11.23708517	1.34464E-06	263.8666342	396.8833658
18	X1	20.1875	3.056359247	6.605080872	9.86968E-05	13.27353505	27.10146495
19	X2	-0.5	3.409511478	-0.146648575	0.886641778	-8.212850796	7.212850796

Resumen

En este módulo hemos introducido conceptos de relaciones funcionales y estadísticas, así como el de variables dependientes y el de variables independientes. Hemos comentado la construcción de un diagrama de dispersión como paso inicial a la hora de buscar algún tipo de relación entre dos variables. Si el diagrama muestra una estructura lineal, entonces se buscará la recta que mejor se ajusta a las observaciones. Hemos puesto de manifiesto la importancia de interpretar correctamente los coeficientes de la recta. También hemos visto cómo se debe utilizar la recta de regresión para realizar predicciones. Hemos introducido una medida numérica de la bondad de ajuste. Esta medida se obtiene con el coeficiente de determinación, discutiendo los valores que puede tomar. Finalmente, hemos comentado la importancia de analizar los residuos para hacer un diagnóstico del modelo lineal obtenido.

En este módulo de regresión lineal simple hemos considerado que las observaciones sobre dos variables X e Y son una muestra aleatoria de una población y que se utilizan para extraer algunas conclusiones del comportamiento de las variables sobre la población, y para ello hemos visto cómo hacer inferencia sobre la pendiente de la recta obtenida a partir de la muestra y cómo hacer un contraste de hipótesis para decidir si la variable X explica realmente el comportamiento de la variable Y. También hemos comentado algunas las relaciones no lineales y la manera en que se puede transformar en una lineal.

Hemos tratado la regresión lineal múltiple como una generalización del modelo de regresión lineal simple en aquellos casos en los que se tiene más de una variable explicativa. Finalmente, hemos visto cómo hacer inferencia sobre los coeficientes de regresión obtenidos a partir de la muestra, cómo hacer un contraste de hipótesis para cada uno de los coeficientes obtenidos para decidir si las variables independientes explican realmente el comportamiento de la variable dependiente o se puede prescindir de alguna de ellas. También hemos realizado un contraste conjunto del modelo. Finalmente, hemos presentado el posible problema de multicolinealidad que puede aparecer y que es debido a la relación entre algunas de las variables explicativas que supuestamente son independientes.

Ejercicios de autoevaluación

1) Los precios de una pantalla TFT de una conocida marca son los siguientes:

Tamaño (pulgadas)	15	17	19	24
Precio (euros)	251	301	357	556

Calculad la recta de regresión para explicar el precio a partir del tamaño.

2) Con los datos de la cuestión anterior queremos decidir si se trata de un buen modelo. ¿Qué método proponéis para determinar si se ajusta bien? ¿Qué podemos decir del caso concreto del ejemplo anterior?

3) Consideramos un modelo lineal para explicar el rendimiento de un sistema informático (variable Y) en relación con el número de *buffers* y el número de procesadores (variables $X1$ y $X2$ respectivamente). Se obtiene el modelo $Y = -3{,}20 + 2X1 + 0{,}0845X2$ con un coeficiente de determinación de 0,99. ¿Se trata de un buen modelo? ¿Cuál será el rendimiento esperado si tenemos 1 *buffer* y 1 procesador? Comentad si este valor os parece lógico y si puede relacionarse con la bondad del modelo.

4) La empresa Ibérica editores tiene que decidir si firma o no un contrato de mantenimiento para su nuevo sistema de procesamiento de palabras. Los directivos creen que el gasto de mantenimiento debe estar relacionado con el uso y han reunido la información que vemos en la tabla siguiente sobre el uso semanal, en horas, y el gasto anual de mantenimiento (cientos de euros).

Uso semanal (horas)	**Gastos anuales de mantenimiento**
13	17,0
10	22,0
20	30,0
28	37,0
32	47,0
17	30,5
24	32,5
31	39,0
40	51,5
38	40,0

a) Determinad la ecuación de regresión que relaciona el costo anual de mantenimiento con el uso semanal.

b) Probad el significado de la relación obtenida en el apartado *a* al nivel de significación 0,05.

c) Ibérica editores espera usar el procesador de palabras 30 horas semanales. Determinad un intervalo de predicción del 95% para el gasto de la empresa en mantenimiento anual.

d) Si el contrato de mantenimiento cuesta 3.000 euros anuales, ¿recomendaríais firmarlo? ¿Por qué?

5) Una biblioteca pública de una ciudad española ofrece un servicio vía Internet de préstamo de libros a los usuarios. Se quiere estudiar la correlación entre el número de usuarios de esta biblioteca virtual y cuántos de ellos acaban realizando los préstamos.

Los datos de los últimos doce meses son:

Usuarios	296	459	602	798	915	521	362	658	741	892	936	747
Préstamos	155	275	322	582	761	324	221	415	562	628	753	569

a) Determina el coeficiente de correlación entre las dos variables. Calcula y representa la recta de regresión.

b) ¿Qué número de préstamos se esperaría si el número de usuarios aumentase a 1.000?

6) Un experto documentalista necesita saber si la eficiencia de un nuevo programa de búsqueda bibliográfica depende del volumen de los datos entrantes. La eficiencia se mide con el número de peticiones por hora procesadas. Aplicando el programa a distintos volúmenes de datos, obtenemos los resultados siguientes:

Volumen (gigabytes), *X*	6	7	7	8	10	10	15
Peticiones procesadas, *Y*	40	55	50	41	17	26	16

a) Calculad la recta de regresión para explicar las peticiones procesadas por hora a partir del volumen de datos e interpretad los parámetros obtenidos.

b) Cread el gráfico de ajuste a la recta de mínimos cuadrados.

c) Determinad el coeficiente de correlación lineal entre las dos variables e interpretad su significado.

d) Determinad el coeficiente de determinación R^2 e interpretad su significado.

e) Calculad, a partir de la recta anterior, cuántas peticiones podemos esperar para un volumen de datos de 12 gigabytes.

f) Realizad el contraste de hipótesis sobre la pendiente. ¿Podemos afirmar a un nivel de significación de 0,05 que la pendiente de la recta es cero?

Solucionario

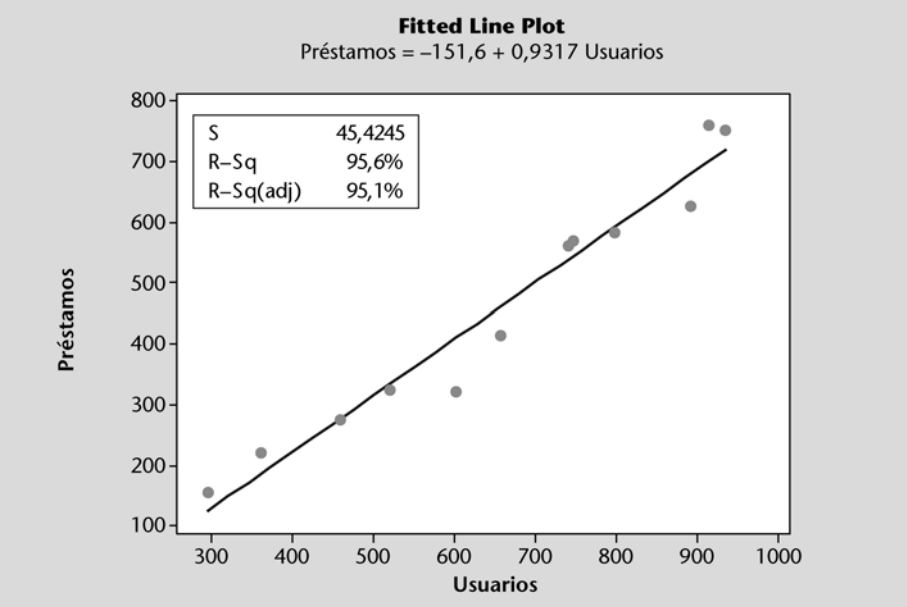
1) Precio = –279,11 + 34,42 · tamaño.

2) Para estudiar la calidad del ajuste, se calcula el coeficiente de correlación muestral $r = 0,994$

3) Es un buen modelo ya que el coeficiente de determinación es muy cercano a 1. El rendimiento, si tenemos un *buffer* y un procesador sería: $Y = -3,20 + 2 \cdot 1 + 0,0845 \cdot 1 = -1,1155$. Este valor no tiene sentido, ya que el rendimiento no puede ser negativo. De todas las maneras, este hecho no es contradictorio con tener un buen modelo ya que estamos fuera del intervalo donde la regresión funciona.

4)
a) $\hat{y} = 10,5 + 0,953x \cdot$
b) Relación significativa; p-valor = 0,000.
c) $[2.874; 54.952]$ euros.
d) Sí, la probabilidad de encontrar el gasto de mantenimiento dentro del intervalo de confianza es del 95%.

5)
a) $r = 0,978$.



b) $-151,6 + 0,9317 \times 1.000 \simeq 780$ préstamos

6)
a)

Regression Analysis: Peticiones procesadas versus Volumen (gigabytes)

```
The regression equation is
Peticiones procesadas = 72,29 - 4,143 Volumen (gigabytes)
S = 9,90815   R-Sq = 66,2%   R-Sq(adj) = 59,4%

Analysis of Variance
Source       DF       SS        MS     F      P
Regression    1   961,14  961,143  9,79  0,026
Error         5   490,86   98,171
Total         6  1452,00
```

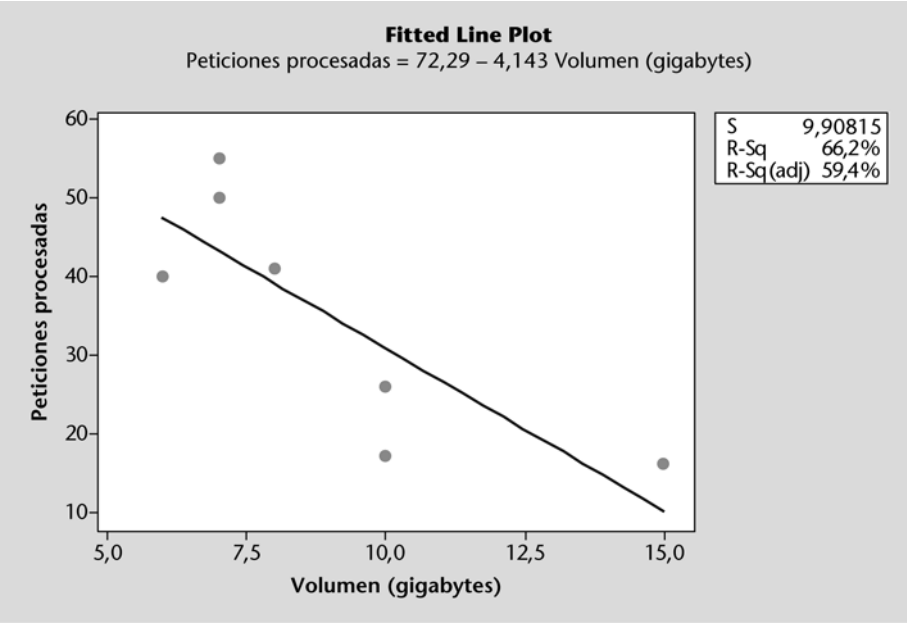
La recta de regresión será:

Peticiones procesadas = 72,29 – 4,143 volumen (gigabytes).

La ordenada en el origen: 72,29; en este caso su significado no tiene ningún sentido.

La pendiente de la recta: –4,143; es negativa: indica que, por cada unidad de volumen de datos (gigabytes) que aumenten los datos entrantes, el número de peticiones procesadas disminuye en 4,143 unidades.

b) El gráfico de ajuste a la recta de mínimos cuadrados es:



c)

Correlations: Volumen (gigabytes); Peticiones procesadas

```
Pearson correlation of Volumen (gigabytes) and Peticiones
procesadas = -0,814
 P-Value = 0,026
```

El coeficiente de correlación $r = -0{,}814$ nos indica que hay una correlación alta negativa entre volumen de datos entrantes y el número de peticiones procesadas.

d) El coeficiente de determinación *R-Sq* es el 66,2%. Esto quiere decir que nuestro modelo lineal explica el 66,2% del comportamiento de la variable *Y* (en este caso, número de peticiones procesadas).

e) Con 12 gigabytes, habrá $72{,}3 - 4{,}14 \cdot 12 = 22{,}57$ peticiones.

f) En el *output* anterior podemos ver que el *p*-valor asociado al contraste de hipótesis anterior es 0,026. Como este valor es menor que $\alpha = 0{,}05$, debemos rechazar la hipótesis nula; es decir, podemos concluir que la pendiente de la recta es distinta de cero o, lo que es lo mismo, que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo estudiado tiene sentido).