Bioestadística – Grado en Biología

Curso 2022-2023

Tema 6. El modelo de regresión lineal simple

Profesores: Mª Ángeles Casares de Cal, Fernando Castro Prado, Laura Davila Pena y Pedro Faraldo Roca

Índice

1 Introducción

El objetivo del análisis de regresión es construir modelos matemáticos que describan o expliquen las relaciones que pueden existir entre las variables. El caso más simple es cuando solo hay dos variables, como altura y peso, longitud y anchura de las hojas, temperatura y presión de un determinado volumen de gas, etc.

Por ejemplo, podemos estar interesados en una variable Y (*variable respuesta o variable dependiente*), y queremos estudiar cómo depende de un conjunto de variables llamadas *variables explicativas o variables independientes*. Por ejemplo, nuestra variable respuesta podría ser el riesgo de ataque cardíaco y las variables explicativas podrían incluir presión arterial, edad, sexo, nivel de colesterol, etc. Sabemos que las relaciones estadísticas no implican necesariamente relaciones causales, pero la presencia de cualquier relación estadística nos da un punto de partida para futuras investigaciones. Una vez que estamos seguros de que existe una relación estadística, podemos intentar modelizar esta relación matemáticamente y luego usar el modelo para la predicción. Para una persona determinada, podemos usar sus valores de las variables explicativas para predecir su riesgo de un ataque cardíaco.

Aunque los modelos de regresión fueron utilizados con anterioridad en Astronomía y Física por Laplace y Gauss, su nombre genérico, *modelos de regresión*, proviene de los trabajos de Galton en Biología a finales del siglo XIX. Galton estudió la dependencia de la estatura de los hijos (Y) respecto a la de sus padres (X), encontrando lo que denominó una regresión a la media: los padres altos tienen en general hijos altos, pero en promedio no tan altos como sus padres; los padres bajos tienen hijos bajos, pero en promedio más altos que sus padres. Desde entonces, los modelos estadísticos que explican la dependencia de una variable Y respecto de una o varias variables X se denominan modelos de regresión.

Los modelos de regresión se diseñan con dos objetivos:

- Conocer de qué modo la variable Y depende de X. En este sentido, el modelo de regresión permite describir la forma de dependencia.
- Una vez construido el modelo de regresión, podemos utilizarlo para realizar predicciones del valor de Y cuando se conoce el valor de X.

Por ejemplo, podemos pensar en un modelo de regresión que represente la longitud (cm) de los pétalos de los lirios (*Iris Versicolor*), en función de su anchura (cm). En este caso, la variable Y sería la longitud de los pétalos, mientras que la anchura de los pétalos sería la variable X.

Resulta muy interesante disponer de un modelo de regresión que represente cómo cambia la longitud de los pétalos según sea su anchura. En principio, parece que anchuras grandes darán lugar a una mayor longitud de los pétalos. Pero además, el modelo de regresión servirá para predecir la longitud cuando se conoce la anchura, y esta predicción será mucho más precisa que la que podríamos obtener sin tener en cuenta la anchura.

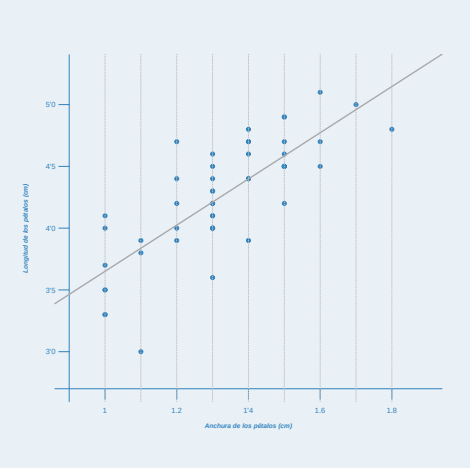


Diagrama de dispersión de la longitud frente a la anchura de los pétalos de lirios *Iris Versicolor*, junto con la recta de regresión ajustada.

Los modelos de regresión se pueden comparar con otros modelos de las ciencias experimentales, como las leyes de los gases ideales o las leyes de la gravitación, que se suelen plantear como *modelos deterministas*: conocidas las variables explicativas la variable respuesta se puede predecir con total exactitud. En el ejemplo de los gases, conocida la temperatura, podemos predecir la presión que ejercerá el gas.

Sin embargo, en la vida real a menudo la predicción con exactitud es imposible, y en su lugar necesitamos modelos que permitan aprovechar el conocimiento de variables explicativas, pero que además incorporen una componente de error impredecible, que vendría ocasionado por errores de medida, por la influencia de otras variables no controlables, o por una aleatoriedad intrínseca a la variable respuesta. Cuando un modelo matemático incorpora una componente aleatoria se dice que es un *modelo estocástico*, a diferencia de los modelos deterministas, que carecen de ella. Los modelos de regresión que vamos a estudiar en este tema son modelos estocásticos.

En este tema nos apoyaremos en los siguientes ejemplos.

Ejemplo 1
Se ha registrado el rendimiento del cultivo de alfalfa (*Medicago Sativa*) para distintas cantidades de agua de riego utilizadas. Los datos siguientes corresponden a doce ciclos anuales diferentes en una comarca productora de alfalfa:

Riego (en cm)	80	80	80	90	90	100	100	110	110	110	120	120
Rendimiento (en t/ha)	54'2	55'9	56'8	58'3	56'7	61'7	59'8	63'2	61'9	62'8	68'7	63'5

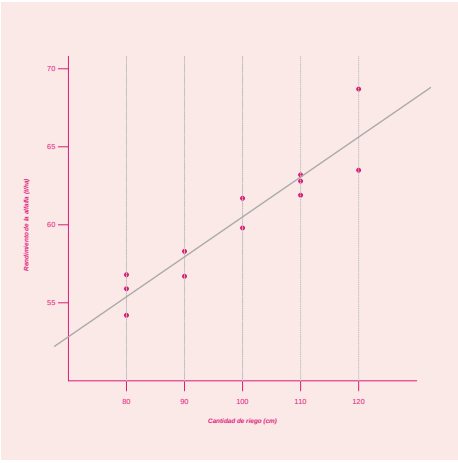
Estamos ante un problema de regresión, pues se quiere estudiar cómo influye la cantidad de agua empleada en el rendimiento de la alfalfa, e interesa predecir el rendimiento para ciertas cantidades de agua recibida.

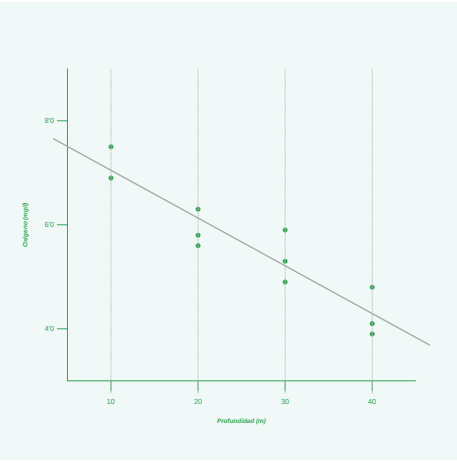
Ejemplo 2
Se ha medido el contenido de oxígeno del agua de un lago a ciertas profundidades. Los datos figuran a continuación.

Profundidad (en metros)	10	10	20	20	20	30	30	30	40	40	40
Contenido de oxígeno (en mg/l)	7'5	6'9	5'6	6'3	5'8	5'3	5'9	4'9	4'8	3'9	4'1

Estamos ante un problema de regresión, pues interesa conocer la dependencia del contenido de oxígeno, en función de la profundidad. También interesa predecir el contenido de oxígeno que se va encontrar a determinada profundidad.

A continuación representamos el diagrama de dispersión de Y frente a X para los dos ejemplos, junto con las rectas de regresión ajustadas.





2 Elementos de un modelo de regresión: el modelo lineal

En términos generales, un modelo de regresión es un modelo estadístico formulado mediante una expresión matemática, que nos permite conocer la variable respuesta Y a partir de la variable explicativa X, de modo que dado un valor específico de X, podamos predecir "aproximadamente" el valor de Y. Tenemos, entonces:

$$Y = f(X) + \varepsilon$$

es decir, podemos descomponer la variable respuesta Y en función del resultado de X más una cantidad que llamaremos **error**[*], y que es variable.[†]

Formalmente, esta relación se expresa mediante el valor medio de la variable respuesta Y cuando la variable explicativa X toma el valor x:

$$f(x) = E(Y|X = x), \quad \text{para cada posible valor } x \text{ de } X,$$

y verificándose, además, que:

$$E(\varepsilon|X = x) = 0$$

El modelo de regresión lineal simple asume una *relación lineal* entre el valor medio de la variable respuesta Y y la variable explicativa X, es decir,

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

donde la pendiente (β_1) y la ordenada en el origen (β_0) de la recta reciben el nombre de **coeficientes de regresión**.

El modelo, en este caso, tiene la forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Para construir el modelo de regresión específico en cada caso, se tiene en cuenta si hay una sola o varias variables explicativas, o variables respuesta, si éstas son discretas o continuas, la forma de la función de regresión (lineal, polinómica, u otras), el tipo de distribución del error, la forma de obtener los datos muestrales, y otros aspectos que al final permiten configurar el modelo adecuado.

En este tema se considerará un modelo muy sencillo de regresión: el modelo de regresión lineal simple. En este modelo tanto la variable respuesta Y, como la variable explicativa X, se suponen univariantes, esto es, cada una de ellas refleja el valor de una sola característica.

2.1 Hipótesis del modelo

Las hipótesis básicas de este modelo son las siguientes:

- **Linealidad.** La función de regresión es una línea recta. En consecuencia, el modelo se suele escribir así:

  $$Y = \beta_0 + \beta_1 X + \varepsilon$$

  donde β_0 y β_1 son parámetros, en principio desconocidos, y ε es lo que hemos definido como error, que es una variable aleatoria no observable que contiene la variabilidad no achacable a la variable explicativa sino debida a errores de medición u otros factores no controlables.

- **Homocedasticidad.** La varianza del error es la misma cualquiera que sea el valor de la variable explicativa:

  $$\text{Var}(\varepsilon|X = x) = \sigma^2 \qquad \text{para todo} \quad x.$$

[*] El término "error" indica ausencia de una relación "exacta" entre X e Y.

[†] Como ejemplo explicativo, consideremos las edades (X) y alturas (Y) de niños menores de 14 años: no todos los niños tienen la misma altura, para la misma edad, por lo tanto habrá un error a la hora de conocer la altura a partir de la edad, error que es impredecible y que varía de unos niños a otros.

- **Normalidad.** El error tiene distribución normal

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

- **Independencia.** Las variables aleatorias que representan los errores $\varepsilon_1, \ldots, \varepsilon_n$ son mutuamente independientes. Se entiende que vamos a obtener una muestra de n observaciones bajo el modelo de regresión. Pues bien, esta suposición dice que los n errores serían mutuamente independientes.

Observación

La hipótesis de linealidad consiste en suponer que la media de la variable respuesta toma un valor inicial β_0 cuando la variable explicativa x vale cero, y además dicha media crece en una cantidad fija β_1 cada vez que x se incrementa en una unidad:

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

La hipótesis de linealidad hace que estemos ante un **modelo paramétrico**, porque supone que la función de regresión es una recta pero deja libertad al valor concreto de la pendiente β_1 y la ordenada en el origen β_0, que son parámetros que debemos estimar a partir de una muestra $(x_1, y_1), \ldots, (x_n, y_n)$.

Por supuesto, en la Estadística se estudian otros modelos paramétricos (por ejemplo, modelos polinómicos o exponenciales), e incluso existen métodos de regresión que no requieren suposición paramétrica alguna, a través de métodos no paramétricos. En esta tema, nos centraremos únicamente en el modelo paramétrico lineal.

Las hipótesis de homocedasticidad y normalidad constituyen simplificaciones muy útiles para poder llevar a cabo las tareas de inferencia bajo un modelo de regresión cualquiera, y también en el caso del modelo lineal.

Finalmente, la suposición de independencia de los errores es conveniente para poder desarrollar inferencia, pero además es razonable suponerla cierta, por ejemplo, en los casos en que la muestra está constituida por experimentos sobre individuos diferentes.

2.2 Tipos de diseño

Para poder estimar los parámetros del modelo (β_0 y β_1), como ya hemos adelantado, necesitamos datos experimentales (una muestra). Distinguiremos dos tipos de diseño experimental.

- **Diseño fijo.** Los valores de la variable explicativa están fijados por el experimentador, de acuerdo a un diseño conveniente de cara a la viabilidad del experimento o a su eficiencia estadística.

  Por ejemplo, podemos fijar distintas concentraciones de nutrientes y medir el crecimiento bacteriano que se obtiene en cada una de ellas.

  En este caso los valores de la variable explicativa no son aleatorios, y sólo es aleatorio el error y, en consecuencia, la variable respuesta. Por tanto, la muestra resultante de un diseño fijo sería del tipo:

$$(x_1, Y_1), \ldots, (x_n, Y_n)$$

- **Diseño aleatorio.** En este caso tanto la variable explicativa como la variable respuesta son aleatorias.

  Por ejemplo, nos interesa un modelo de regresión donde la variable explicativa sea el tamaño de los peces de cierta especie (medido mediante la longitud) y la variable respuesta sea la concentración de cierto ácido graso. Si el experimento consiste en tomar peces al azar en un río y medir su longitud y su concentración del ácido graso, entonces ambas variables son aleatorias y por tanto se trata de un diseño aleatorio.

  En definitiva, la muestra resultante de un diseño aleatorio sería del tipo:

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

En adelante supondremos diseño fijo, aunque los procedimientos estadísticos que veremos también son aplicables bajo diseño aleatorio.

En resumen, un modelo de regresión lineal simple, homocedástico, con errores normales e independientes, del que extraemos una muestra bajo diseño fijo nos proporciona datos del tipo $(x_1, Y_1), \ldots, (x_n, Y_n)$, donde $x_1, \ldots, x_n$ son valores fijados por el experimentador, mientras que

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{para} \quad i \in \{1, \ldots, n\}$$

siendo $\varepsilon_1, \ldots, \varepsilon_n \sim N(0, \sigma^2)$, independientes.

En los ejemplos 1 y 2, *supondremos que son ciertas las hipótesis de linealidad, homocedasticidad, normalidad de los errores e independencia. De momento no vamos a cuestionar la veracidad de estas hipótesis, porque además en ambos casos los datos disponibles no son suficientes para verificar su cumplimiento. Lo que sí constatamos es que se trata de situaciones de diseño fijo, pues tanto las cantidades de agua utilizadas para el riego como las profundidades no son fruto del azar, sino que se han fijado de antemano.*

3 Estimación de los parámetros β_0, β_1, y σ^2

Supondremos las hipótesis de linealidad, homocedasticidad, normalidad, diseño fijo e independencia de los errores.

- Vamos a obtener en primer lugar los estimadores para los parámetros β_0 y β_1 a partir de una muestra $(x_1, Y_1), \ldots, (x_n, Y_n)$.

  Si llamamos $\widehat{\beta}_0$ y $\widehat{\beta}_1$ a los estimadores de los parámetros β_0 y β_1, la recta "ajustada" ("estimada") será:
  $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

  Entonces, para un nuevo valor x_0 de la variable explicativa daríamos "la predicción" de la variable respuesta:
  $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.

  Ahora bien, esto también se puede aplicar a los datos muestrales, y así, para el valor observado x_i tendríamos "el valor ajustado":
  $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$,

  pero, además, para el valor observado x_i tenemos que el valor observado ha sido Y_i.

  Entonces, tendríamos las siguientes diferencias entre el valor ajustado y el valor observado:

  $$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad \text{para} \quad i \in \{1, \ldots, n\},$$

  y que se denominan **residuos** de la regresión.

  El método que se utiliza para obtener $\widehat{\beta}_0$ y $\widehat{\beta}_1$ (los estimadores de los parámetros parámetros β_0 y β_1) es conocido como el **método de los mínimos cuadrados**.
  La idea consiste en escoger los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que den lugar a los residuos más pequeños. Para ello, para evitar que se compensen los residuos positivos con los negativos, se usa la suma de los cuadrados de los residuos como criterio a minimizar.
  Así, los estimadores por mínimos cuadrados son $\widehat{\beta}_0$ y $\widehat{\beta}_1$ tales que:

  $$\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

La minimización se realiza calculando las derivadas parciales respecto de β_0 y β_1, igualándolas a cero y despejando de ambas ecuaciones los valores de β_0 y β_1, candidatos a mínimo. El cálculo de las segundas derivadas prueba que en efecto constituyen un mínimo absoluto de la suma de cuadrados de los residuos. Como resultado se obtienen los estimadores de β_0 y β_1, respectivamente:

$$\boxed{\widehat{\beta}_0 = \overline{Y} - \frac{S_{xY}}{S_x^2}\,\overline{x} \qquad\qquad \widehat{\beta}_1 = \frac{S_{xY}}{S_x^2}}$$

donde

$\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$,

$\overline{Y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i$

son las medias respectivas de la variable explicativa y la variable respuesta,

$S_{xY} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})$

es la covarianza [‡]

y $S_x^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$

es la varianza de la variable explicativa.

La recta de regresión estimada por mínimos cuadrados es la que pasa por el vector de medias o centro de gravedad, $(\overline{x}, \overline{Y})$, y tiene pendiente $\widehat{\beta}_1 = \dfrac{S_{xY}}{S_x^2}$.

El requisito de pasar por el vector de medias es muy natural, pues impone que la recta se sitúe allí donde se encuentran los puntos muestrales.

Respecto de la pendiente, su signo es el de la covarianza S_{xY}, pues el denominador S_x^2 siempre es positivo.

- Por otro lado, estimaremos la varianza del error σ^2 mediante

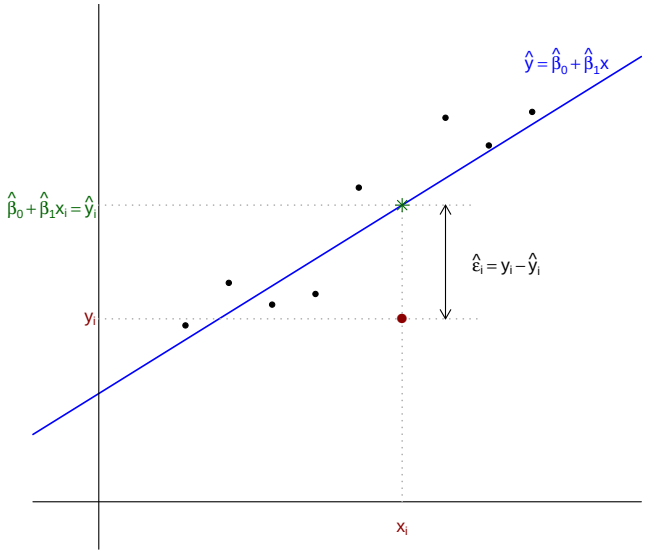
$$\boxed{\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2}$$

Empleamos la suma de cuadrados de los residuos, pero dividimos por $(n-2)$ en lugar de hacerlo por n, para que el estimador sea insesgado.

Como ilustración, en la figura siguiente hemos representado un diagrama de dispersión, la recta de regresión ajustada, los puntos (x_i, y_i) y $(x_i, \widehat{y}_i)$, siendo $\widehat{y}_i$ el valor ajustado $\widehat{y}_i = \beta_0 + \beta_1 \cdot x_i$, y el residuo

[‡] Más adelante veremos con más detalle una descripción de la covarianza.

correspondiente $\widehat{\epsilon}_i = y_i - \widehat{y}_i$.



Ejemplo 1

Estimaremos los coeficientes de la recta de regresión, β_0 y β_1, y la varianza del error, σ^2, con los datos de rendimiento de la alfalfa y cantidad de riego.

En la tabla siguiente presentamos todos los cálculos necesarios del ejemplo 1:

	Riego (mm) x_i	*Rendimiento (t/ha)* Y_i	$(x_i-\overline{x})^2$	$(Y_i-\overline{Y})^2$	$(x_i-\overline{x})\cdot(Y_i-\overline{Y})$	*Valores ajustados* $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$	*Residuos* $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$	$\widehat{\epsilon}_i^2$
	80	54′2	367′4	37′11	116′76	55′38	−1′18	1′40
	80	55′9	367′4	19′29	84′17	55′38	0′52	0′27
	80	56′8	367′4	12′19	66′92	55′38	1′42	2′00
	90	58′3	84′0	3′97	18′26	57′94	0′36	0′13
	90	56′7	84′0	12′90	32′92	57′94	−1′24	1′55
	100	61′7	0′7	1′98	1′17	60′51	1′19	1′43
	100	59′8	0′7	0′24	−0′41	60′51	−0′71	0′50
	110	63′2	117′4	8′46	31′51	63′07	0′13	0′02
	110	61′9	117′4	2′59	17′42	63′07	−1′17	1′36
	110	62′8	117′4	6′29	27′17	63′07	−0′27	0′07
	120	68′7	434′0	70′70	175′17	65′63	3′07	9′45
	120	63′5	434′0	10′29	66′84	65′63	−2′13	4′52
SUMAS	1190	723′5	2491′67	186′01	637′92		**0′00**	22′69

Por lo tanto:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1190}{12} = 99'17 \ cm$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{723'5}{12} = 60'3 \ t/ha$$

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{2491'67}{12} = 207'64$$

$$S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \frac{186'01}{12} = 15'50$$

$$S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(Y_i-\overline{Y}) = \frac{637'92}{12} = 53'16$$

Estimación de los parámetros

- *La ordenada en el origen*
  $\widehat{\beta}_0 = \overline{Y} - \frac{S_{xY}}{S_x^2}\,\overline{x} = 34'90$
- *La pendiente*
  $\widehat{\beta}_1 = \frac{S_{xY}}{S_x^2} = 0'26$
- *La varianza del error*
  $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n-2}\widehat{\epsilon}_i^2 = 2'27$
- *La desviación típica del error:*
  $\widehat{\sigma} = \sqrt{9'28} = 1'51$

4 Propiedades de los estimadores

En esta sección estudiaremos las propiedades de los estimadores que acabamos de obtener, en términos de sesgo y varianza. Omitiremos las demostraciones, y nos centraremos en el análisis de las expresiones para la varianza.

4.1 Propiedades de $\widehat{\beta}_1$

Se puede demostrar que el estimador de la pendiente es insesgado, esto es:

$$\mathrm{E}\left(\widehat{\beta}_1\right) = \beta_1$$

Para la varianza se tiene:

$$\mathrm{Var}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{nS_x^2}$$

Siendo su error típico la raíz cuadrada de esta cantidad.

Interpretación de la varianza del estimador $\widehat{\beta}_1$.

De esta expresión deducimos que la varianza del estimador de la pendiente será:

– Tanto mayor cuanto mayor sea la varianza del error, σ^2. Esto es lógico pues al aumentar la varianza del error, los datos aparecerán más alejados de la recta de regresión, y será más imprecisa la estimación de los parámetros de la recta a partir de ellos.

– Será más pequeña si los valores $x_1, \ldots, x_n$ tienen mucha dispersión. Esto es muy interesante. Dice que para anclar bien la pendiente de la recta de regresión conviene que los valores de la variable explicativa estén suficientemente espaciados.

– Será más pequeña si disponemos de muchos datos, o lo que es lo mismo, si el tamaño muestral n es grande.

Además,

usando la independencia y normalidad de los errores, se puede demostrar que el estimador $\widehat{\beta}_1$ tiene distribución normal, y se tiene en este caso:

$$\boxed{\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right)}$$

4.2 Propiedades de $\widehat{\beta}_0$

La ordenada en el origen, β_0, es el valor que toma la recta de regresión cuando $x = 0$.
Salvo en las ocasiones en que nos interese la media de la variable respuesta cuando la variable explicativa tome el valor cero, la ordenada en el origen tiene menos interés que la pendiente.

Empezamos diciendo que $\widehat{\beta}_0$ es un estimador insesgado, esto es:

$$\mathrm{E}\left(\widehat{\beta}_0\right) = \beta_0$$

La varianza se puede expresar así:

$$\mathrm{Var}\left(\widehat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}\right)$$

Y su error típico será la raíz cuadrada de esta cantidad.

Interpretación de la varianza del estimador $\widehat{\beta}_0$.

Podemos descomponer esta expresión de la varianza en dos términos: σ^2/n y $(\sigma^2\overline{x}^2)/(nS_x^2)$, que asociamos respectivamente con $\overline{Y}$ y $\widehat{\beta}_1\overline{x}$, de cuya diferencia se obtiene $\widehat{\beta}_0$. Así, σ^2/n es la parte de la varianza de $\widehat{\beta}_0$ que se debe a la estimación de la media $\overline{Y}$, mientras que $(\sigma^2\overline{x}^2)/(nS_x^2)$ es la parte asociada a la estimación de la pendiente. Aquí el factor $\overline{x}^2$ indica que cuanto más lejos esté $\overline{x}$ del origen, más varianza tendrá el estimador de la ordenada en el origen, siendo por tanto más impreciso.

Igual que antes, bajo suposición de independencia y normalidad de los errores, tendríamos que $\widehat{\beta}_0$ tiene distribución normal, y se tiene en este caso:

$$\boxed{\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}\right)\right)}$$

4.3 Propiedades de $\widehat{\sigma}^2$

Para el estimador de la varianza del error, una demostración algo más compleja que las anteriores y que vamos a omitir, nos conduciría a la siguiente distribución Ji-cuadrado:

$$\boxed{\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}}$$

De aquí se deduce que $\widehat{\sigma}^2$ es un estimador insesgado de σ^2. De hecho, la aparición de $(n-2)$ grados de libertad es el motivo por el que hemos dividido la suma de cuadrados de los residuos por $(n-2)$, en lugar de por n, para calcular el estimador de la varianza.

5 Inferencia sobre los parámetros

Hasta aquí hemos visto cómo se estiman los parámetros β_0, β_1 y σ^2 involucrados en el modelo lineal simple, y hemos analizado las propiedades de los estimadores: esperanza, varianza y distribución. En esta sección realizaremos las otras dos tareas de la Inferencia: intervalos de confianza y contraste de hipótesis; para cada uno de ellos.

5.1 Inferencia sobre β_1

Para construir intervalos de confianza o realizar contrastes de hipótesis para la pendiente, β_1, podríamos usar como pivote la estandarización de $\widehat{\beta}_1$, esto es:

$$\frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{S_x\sqrt{n}}} \sim N(0,1)$$

aunque para ello habría que conocer la varianza del error, σ^2.
Como lo más habitual es que σ^2 sea desconocida, se suele estimar el error típico mediante:

$$\widehat{\text{Error Típico}}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}}{S_x\sqrt{n}}$$

modificándose, entonces el pivote.
Por lo tanto,
cuando la varianza del error es desconocida, usamos el pivote:

$$\boxed{\frac{\widehat{\beta}_1 - \beta_1}{\frac{\widehat{\sigma}}{\left(S_x\sqrt{n}\right)}} \sim T_{n-2}}$$

donde la distribución normal estándar ha sido sustituida por la T de Student.

- Basándonos en este pivote (cuando σ es desconocida), se construye **el intervalo de confianza para β_1 con nivel de confianza $(1-\alpha)$**.

  Este intervalo de confianza estará centrado en $\widehat{\beta}_1$, y su radio será el producto del cuantil de la T de Student por el error típico estimado:

$$\boxed{\left(\widehat{\beta}_1 - t_{n-2,\alpha/2}\frac{\widehat{\sigma}}{S_x\sqrt{n}}\ ,\ \widehat{\beta}_1 + t_{n-2,\alpha/2}\frac{\widehat{\sigma}}{S_x\sqrt{n}}\right)}$$

- En caso de contraste de hipótesis, tiene especial interés **el contraste de la hipótesis nula $H_0 : \beta_1 = 0$**, pues, de ser cierta esta hipótesis, la función de regresión sería una recta horizontal, y nos encontraríamos con que la variable explicativa no influye en la variable respuesta.
  En este caso, el estadístico del contraste, bajo $H_0 : \beta_1 = 0$, es:

$$\frac{\widehat{\beta}_1}{\frac{\widehat{\sigma}}{\left(S_x\sqrt{n}\right)}} \sim T_{n-2}$$

  y rechazaremos $H_0 : \beta_1 = 0$ si

$$\frac{\widehat{\beta}_1}{\frac{\widehat{\sigma}}{\left(S_x\sqrt{n}\right)}} > t_{n-2,\alpha/2}$$

  En tal caso diremos que $\widehat{\beta}_1$ ha tomado un valor significativamente distinto de cero.

5.2 Inferencia sobre β_0

En el caso de β_0, la ordenada en el origen, para construir intervalos de confianza o realizar contrastes de hipótesis, usaríamos como pivote:

$$\frac{\widehat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}} \sim N(0,1) \qquad \text{si } \sigma \text{ es conocida}$$

Cuando σ es desconocida, se estima el error típico mediante

$$\widehat{\text{Error Típico}}\left(\widehat{\beta}_0\right) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}$$

Por lo tanto,

cuando σ es desconocida, usamos el pivote:

$$\boxed{\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}} \sim T_{n-2}}$$

Basándonos en este pivote, (cuando σ es desconocida, la situación más habitual) el intervalo de confianza para β_0 con nivel de confianza $(1-\alpha)$, estará centrado en $\widehat{\beta}_0$, y su radio será el producto del cuantil de la T de Student por el error típico estimado:

$$\boxed{\left(\widehat{\beta}_0 - t_{n-2,\alpha/2}\ \widehat{\sigma}\ \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}\ ,\ \widehat{\beta}_0 + t_{n-2,\alpha/2}\ \widehat{\sigma}\ \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}\ \right)}$$

Para el contraste de hipótesis relativas a β_0 también podemos utilizar este pivote. Así, por ejemplo, rechazaremos la hipótesis nula $H_0 : \beta_0 = 0$ en favor de $H_a : \beta_0 \neq 0$ si

$$\frac{\left|\widehat{\beta}_0\right|}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}} > t_{n-2,\alpha/2}$$

5.3 Inferencia sobre σ^2

Para la varianza del error, σ^2, el pivote sería

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

En función de este pivote, el intervalo de confianza para σ^2 con nivel de confianza $(1-\alpha)$, se puede construir así

$$\left(\frac{(n-2)\widehat{\sigma}^2}{\chi^2_{n-2,\alpha/2}}\ ,\ \frac{(n-2)\widehat{\sigma}^2}{\chi^2_{n-2,1-\alpha/2}}\right)$$

El intervalo de confianza para la desviación típica del error, σ, se obtendría efectuando la raíz cuadrada de los extremos del intervalo anterior.

En el caso de contraste de hipótesis en relación a σ^2, a modo de ejemplo, se rechazaría ría la hipótesis nula $H_0 : \sigma^2 \geq 1$ en favor de la alternativa $H_a : \sigma^2 < 1$ si

$$\frac{(n-2)\widehat{\sigma}^2}{1} < \chi^2_{n-2,1-\alpha}$$

Ejemplo 2

Vamos a realizar las siguientes tareas:

(a) Obtener las estimaciones de los parámetros del modelo de regresión lineal simple del contenido de oxígeno sobre la profundidad.

(b) Contrastar si la pendiente es cero.

(c) Calcular los intervalos de confianza para cada uno de los parámetros: ordenada en el origen, pendiente, varianza del error y desviación típica del error; al nivel de confianza del 95%.

Tal como hicimos en el ejemplo 1, presentamos a continuación una tabla con todos los cálculos necesarios.

	Profundidad (m) x_i	*Oxígeno (mg/l)* Y_i	$(x_i-\overline{x})^2$	$(Y_i-\overline{Y})^2$	$(x_i-\overline{x})\cdot(Y_i-\overline{Y})$	*Valores ajustados* $\widehat{Y}_i=\widehat{\beta}_0+\widehat{\beta}_1 x_i$	*Residuos* $\widehat{\epsilon}_i=Y_i-\widehat{Y}_i$	$\widehat{\epsilon}_i^2$	$(\widehat{Y}_i-\overline{Y}_i)^2$
	10	7'5	267'77	3'82	−31'98	7'05	0'45	0'20	2'26
	10	6'9	267'77	1'83	−22'17	7'05	−0'15	0'02	2'26
	20	5'6	40'50	0'00	− 0'35	6'13	−0'53	0'28	0'34
	20	6'3	40'50	0'57	− 4'80	6'13	0'17	0'03	0'34
	20	5'8	40'50	0'06	− 1'62	6'13	−0'33	0'11	0'34
	30	5'3	13'22	0'06	− 0'89	5'21	0'09	0'01	0'11
	30	5'9	13'22	0'13	1'29	5'21	0'69	0'47	0'11
	30	4'9	13'22	0'42	− 2'35	5'21	−0'31	0'10	0'11
	40	4'8	185'95	0'56	−10'17	4'29	0'51	0'26	1'57
	40	3'9	185'95	2'71	−22'44	4'29	−0'39	0'15	1'57
	40	4'1	185'95	2'09	−19'71	4'29	−0'19	0'04	1'57
SUMAS	290	61	1254'5455	12'2473	−115'1818		**0'00**	1'67225	10'575

(a) Datos muestrales:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{290}{11}=26'36\ m$$

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{61}{11}=5'55\ mg/l$$

$$S_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1254'5455}{11}=114'05$$

$$S_Y^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\frac{12'2473}{11}=1'113$$

$$S_{xY}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(Y_i-\overline{Y})=\frac{-115'1818}{11}=-10'471$$

Estimación de los parámetros

- *La ordenada en el origen*
  $\widehat{\beta}_0=\overline{Y}-\frac{S_{xY}}{S_x^2}\overline{x}=7'97$
- *La pendiente*
  $\widehat{\beta}_1=\frac{S_{xY}}{S_x^2}=-0'0918$
- *La varianza del error*
  $\widehat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=\frac{1}{n-2}\widehat{\epsilon}_i^2=\frac{1}{9}1'67225=0'1858$
- *La desviación típica del error:*
  $\widehat{\sigma}=\sqrt{0'1858}=0'431$

(b) Contraste de hipótesis sobre el parámetro β_1 (la pendiente de la recta de regresión).

- *Hipótesis nula:* $H_0: \beta_1 = 0$
  Hipótesis alternativa: $H_a: \beta_1 \neq 0$ *(lo que queremos demostrar)*
- *Criterio de decisión:*
  Se probará H_a (la pendiente es distinta de cero) si la pendiente estimada es "mucho menor" o "mucho mayor" de 0.
- *Estadístico del contraste:*
  $\dfrac{\widehat{\beta}_1}{\dfrac{\widehat{\sigma}}{S_x\sqrt{n}}} \sim T_6$ *bajo* H_0.
- *Valor del estadístico del contraste para la muestra dada:*
  $$\frac{\widehat{\beta}_1}{\dfrac{\widehat{\sigma}}{S_x\sqrt{n}}} = \frac{-0'0918}{\dfrac{0'431}{\sqrt{114'05 \cdot 11}}} = -7'544$$
- *Cálculo del nivel crítico:*
  $$P\left[\frac{|\widehat{\beta}_1|}{\dfrac{\widehat{\sigma}}{Sx\sqrt{n}}} > |-7'544|\right] = P\left(|T_9| > 7'544\right) = P\left(T_9 > 7'544\right) + P\left(T_9 < -7'544\right) = 2 \cdot P\left(T_9 > 7'544\right)$$

  De las tablas de la distribución T de Student (de las que presentamos a continuación un fragmento), y por la simetría de esta distribución, se obtiene que esta probabilidad es menor que $2 \cdot 0'005 = 0'01$.

α / **m**	**0'45**	**0'4**	**0'3**	**0'25**	**0'2**	**0'1**	**0'05**	**0'025**	**0'01**	**0'005**
6	0'131	0'265	0'553	0'718	0'906	1'44	1'94	2'45	3'14	3'71
7	0'130	0'263	0'549	0'711	0'896	1'41	1'89	2'36	3'00	3'50
8	0'130	0'262	0'546	0'706	0'889	1'40	1'86	2'31	2'90	3'36
9	0'129	0'261	0'543	0'703	0'883	1'38	1'83	2'26	2'82	3'25
10	0'129	0'260	0'542	0'700	0'879	1'37	1'81	2'23	2'76	3'17

- *A la vista de los resultados, podemos afirmar que la pendiente de la recta de regresión es significativamente distinta de cero, dicho de manera más detallada, la pendiente estimada es distinta de cero y constituye una prueba significativa de que la "verdadera pendiente" es distinta de cero.*
  En este ejemplo, si la pendiente fuera cero, estaríamos en un modelo que predice el mismo contenido en oxígeno para cualquier profundidad, lo cual es contradictorio con los datos.

(c) Intervalos de confianza

- *Intervalo de confianza para β_1 al nivel de confianza del 95%*
  Basándonos en el pivote
  $$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/\left(S_x\sqrt{n}\right)} \sim T_{n-2}$$
  se construye el intervalo de confianza para β_1
  $$\left(\widehat{\beta}_1 - t_{n-2,\alpha/2}\frac{\widehat{\sigma}}{S_x\sqrt{n}}\ ,\ \widehat{\beta}_1 + t_{n-2,\alpha/2}\frac{\widehat{\sigma}}{S_x\sqrt{n}}\right)$$
  que en este caso es
  $$\begin{aligned}&\left(-0'09 - t_{9,0'025} \cdot 0'012, -0'09 + t_{9,0'025} \cdot 0'012\right) = \\ &= \left(-0'09 - 2'26 \cdot 0'012, -0'09 + 2'26 \cdot 0'012\right) = \\ &= (-0'12, -0'06)\end{aligned}$$
- *Intervalo de confianza para β_0 al nivel de confianza del 95%*
  Análogamente, basándonos en el pivote
  $$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}} \sim T_{n-2}$$
  se construye el intervalo de confianza para β_0
  $$\left(\widehat{\beta}_0 - t_{n-2,\alpha/2}\ \widehat{\sigma}\ \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}\ ,\ \widehat{\beta}_0 + t_{n-2,\alpha/2}\ \widehat{\sigma}\ \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}\ \right)$$

que en este caso es

$$(7'97 - t_{9,0'025} \cdot 0'346\ ,\ 7'97 + t_{9,0'025}\ \cdot 0'346\ ) =$$
$$= (7'97 - 2'26 \cdot 0'346\ ,\ 7'97 + 2'26\ \cdot 0'346\ ) =$$
$$= (7'18\ ,\ 8'75)$$

- *Intervalo de confianza para σ^2 al nivel de confianza del 95%*

  Basándonos en el pivote

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

  se construye el intervalo de confianza para la varianza del error

$$\left(\frac{(n-2)\widehat{\sigma}^2}{\chi^2_{n-2,\alpha/2}}\ ,\ \frac{(n-2)\widehat{\sigma}^2}{\chi^2_{n-2,1-\alpha/2}} \right)$$

  que en este caso es

$$\left(\frac{9 \cdot 0'19}{\chi_{9,0'025}}\ ,\ \frac{9 \cdot 0'19}{\chi_{9,0'975}} \right) = \left(\frac{9 \cdot 0'19}{19'0}\ ,\ \frac{9 \cdot 0'19}{2'70} \right) =$$
$$= (0'09\ ,\ 0'63)$$

- *Intervalo de confianza para σ al nivel de confianza del 95%*

  El intervalo de confianza para la desviación típica del error, σ, se obtiene efectuando la raíz cuadrada de los extremos del intervalo anterior.

$$= \left(\sqrt{0'09}\ ,\ \sqrt{0'63}\right) = (0'3\ ,\ 0'8)$$

Tabla resumen

					Intervalos de confianza	
	Estimaciones	*Error típico*	*Valor T*	*Niv. Crít.*	*Extr inferior*	*Extr superior*
Ordenada en el origen	*7'97*	*0'346*	–	–	*7'18*	*8'75*
Pendiente	*– 0'09*	*0'012*	*– 7'544*	$< 0'005$	*– 0'12*	*– 0'06*
Varianza	*0'1858*	–	–	–	–	–
Desviación típica	*0'431*	–	–	–	–	–

6 Covarianza, coeficiente de correlación y coeficiente de determinación

La covarianza es la forma más común de medir la relación lineal (creciente o decreciente) entre dos variables. Se obtiene así:

$$S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) \tag{1}$$

En la figura siguiente mostramos un diagrama de dispersión de valores de x e Y, con el cual ilustraremos el cálculo de la covarianza. Así, en los puntos (x_i, Y_i) situados en los cuadrantes primero y tercero, que son los señalados con el signo +, presentan valores positivos en el producto $(x_i - \overline{x})(Y_i - \overline{Y})$, pues ambos factores tienen el mismo signo.

Por ello, aportarán sumandos positivos en la expresión de la covarianza (véase la ecuación de la covarianza).

Por el contrario, los puntos (x_i, Y_i) situados en los cuadrantes segundo y cuarto, regiones señaladas con el signo −, aportan sumandos negativos a la expresión de la covarianza, pues las diferencias respecto de la media tienen distinto signo.

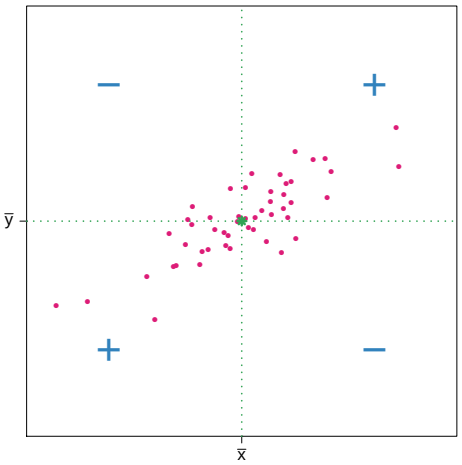


Diagrama de dispersión con relación creciente entre x e Y.

De este modo, si hay muchos puntos en los regiones con signo + y pocos en las regiones con signo −, la covarianza será positiva. En ese caso, la nube de puntos tendrá orientación creciente, y podemos interpretar que al aumentar la variable x, también aumentará (en términos generales) la variable Y.

Por el contrario, si abundan más los puntos de las regiones con signo − , la covarianza será negativa, y nos estará indicando la orientación decreciente de la nube de puntos. Interpretaremos que al aumentar el valor de x, disminuye el valor de Y.

Es decir, si las dos variables x e Y presentan una relación lineal creciente o relación directa (al aumentar una variable aumenta la otra) entonces la covarianza será positiva. Por el contrario, si la relación es decreciente o inversa (al aumentar una variable disminuye la otra) entonces la covarianza será negativa.

Propiedad de la covarianza

Respecto del cálculo de la covarianza, observamos que la covarianza no se ve afectada por cambios de localización, pero sí por cambios de escala en cualquiera de las dos variables. Esto lo podemos resumir así:

$$S_{a+bX,c+dY} = bdS_{XY}$$

siendo a, b, c y d constantes.

Así, por ejemplo, si la variable X es una longitud y se mide en metros, y la variable Y es un peso y se mide en kilogramos, entonces ya sabíamos que la media y la desviación típica de X se miden en metros (m), su varianza en m^2, y que la media y desviación típica de Y se miden en kg. Pues bien, la covarianza

entre X e Y, S_{XY} se mide en $m \cdot kg$. De este modo, si pasamos las mediciones de X a centímetros, todos los valores quedarán multiplicados por 100, y también quedarán multiplicadas por 100 su media, su desviación típica y la covarianza, S_{XY}. Este fenómeno de cambiar de metros a centímetros, con la consiguiente multiplicación por 100, es lo que conocemos como cambio de escala.

Definición 1
Para obtener una medida de la relación lineal que no se vea afectada por cambios de escala, se define el ***coeficiente de correlación*** *(*lineal o de Pearson*), que se obtiene dividiendo la covarianza por las desviaciones típicas de las dos variables, esto es:*

$$R = \frac{S_{xY}}{S_x S_Y}$$

siendo $S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ *y* $S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$.

El coeficiente de correlación carece de unidades, y de hecho su valor siempre se encuentra entre -1 y +1, esto es: $R \in [-1, +1]$.

Su signo goza de la misma interpretación que la covarianza.
Si vale cero no hay relación lineal, si es positivo hay relación lineal creciente, y si es negativo hay relación lineal decreciente.

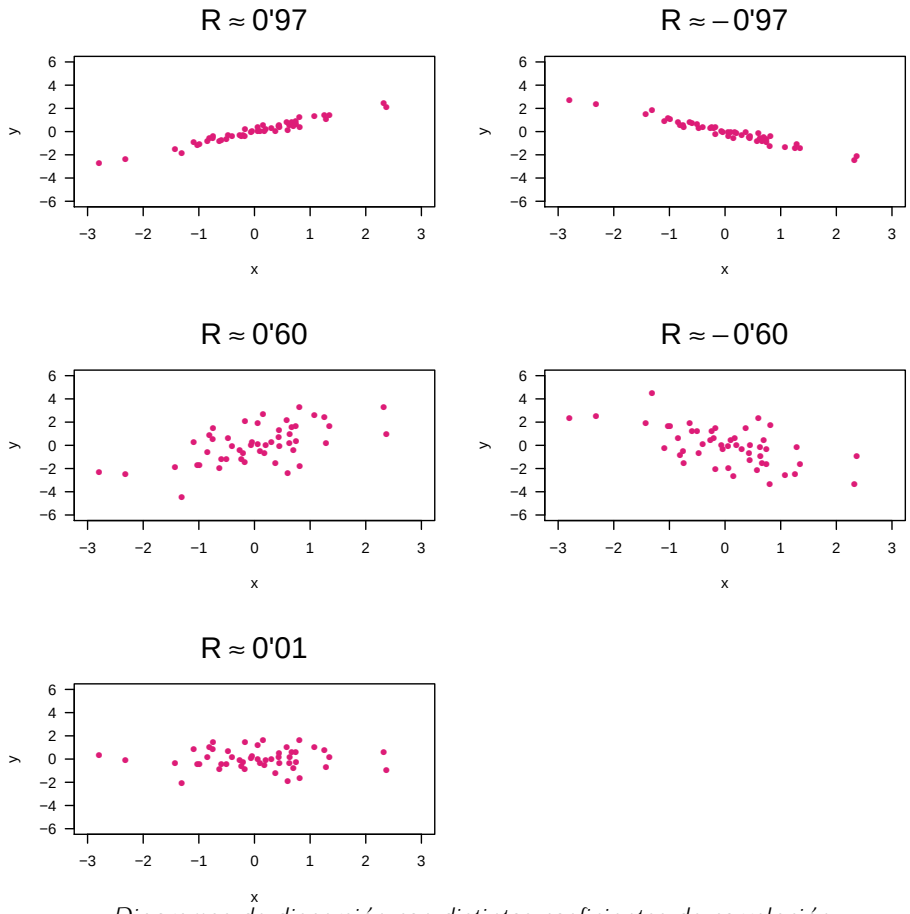
Pero ahora, al estar estandarizado entre -1 y +1, se puede interpretar su magnitud. Así, si los datos se aproximan mucho a una recta creciente, el coeficiente de correlación estará próximo a +1, mientras que si se aproximan a una recta decreciente, el coeficiente de correlación estará próximo a -1. Por el contrario, si pierden el alineamiento, el coeficiente de correlación va haciéndose más pequeño (en valor absoluto), hasta llegar al cero, cuando ya no se aprecia una recta creciente o decreciente.

Observación 1
Las medidas de la correlación son medidas de la asociación entre variables. Es importante señalar que la asociación es un concepto que no tiene repercusión sobre las causas. Dos variables pueden tener correlación alta y sin embargo que no haya ninguna relación de causa–efecto.
Por ejemplo, en un clima cálido tanto el consumo de helado como la población de mosquitos aumentan. Las variables que miden cada una de estas cantidades tendrían una alta correlación. Esta relación, sin embargo, no es más que una relación de coincidencia y no tiene mayor interés estadístico. No obstante mirar las correlaciones proporciona un análisis preliminar útil para examinar las relaciones entre las variables.

En la figura siguiente se muestran diagramas de dispersión de cinco situaciones diferentes, con distintos coeficientes de correlación. En la primera fila se encuentran los datos más alineados, con coeficientes de correlación de 0'97 y -0'97, con orientación creciente en el gráfico de la izquierda y decreciente en el de la derecha, en coherencia con el signo de la correlación.

En la segunda fila las correlaciones son de 0'60 y -0'60 a izquierda y derecha, respectivamente. Vemos que, en efecto, los datos se alejan más de la recta.

Por último, el quinto gráfico, situado en la tercera fila, muestra una ausencia casi total de orientación creciente o decreciente.



Diagramas de dispersión con distintos coeficientes de correlación.

Definición 2

Para medir la proximidad de los datos a la recta, sin atender a si ésta es creciente o decreciente, es frecuente calcular el ***coeficiente de determinación****, que es el cuadrado del coeficiente de correlación, y se suele denotar por* R^2.

Al efectuar el cuadrado del coeficiente de correlación, obtendremos que el coeficiente de determinación siempre es positivo o cero, y será tanto mayor cuanto más alineadas se encuentren las observaciones, bien en sentido creciente o decreciente.

Además, se puede demostrar (omitiremos los detalles) que

$$R^2 = 1 - \frac{RSS}{TSS}$$

siendo

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

la suma de cuadrados total de la variable respuesta y

$$RSS = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$$

la suma de cuadrados residual correspondiente a los residuos o defectos de ajuste del modelo.

La suma de cuadrados total siempre es mayor que la suma de cuadrados residual, pues la total correspondería al ajuste basándose en una recta horizontal a altura $\overline{Y}$.

Cuanto más próximos se encuentren los puntos a una recta (no horizontal), menor será la suma de cuadrados residual, y mayor será el valor del coeficiente de determinación, que se interpreta como una

proporción de variabilidad de la respuesta, Y, explicada por el modelo de regresión (lo veremos en la siguiente sección).

Ejemplo 1

Sobre los datos del ejemplo 1 calcúlese la covarianza, el coeficiente de correlación y el coeficiente de determinación e interprétese los resultados.

Para el cálculo de estas medidas nos basamos en la tabla del ejemplo 1 y en los cálculos ya realizados antes.

- *La covarianza:*
  $S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (Y_i - \overline{Y}) = 53'16$
- *El coeficiente de correlación:*
  $R = \frac{S_{xY}}{S_x \cdot S_Y} = \frac{53'16}{\sqrt{207'64} \cdot \sqrt{15'50}} = \frac{53'16}{14'40 \cdot 3'94} = 0'937$
  El coeficiente de correlación es positivo y muy alto, esto indica una relación lineal creciente entre las variables "Rendimiento de la alfalfa" y "Cantidad de agua de riego utilizada".
- *El coeficiente de determinación:*
  $R^2 = 0'937^2 = 0'878$
  El coeficiente de determinación es alto. Indica un buen ajuste del modelo.
  $R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{22'69}{186'01} = 0'878$
  También nos indica que el 87'8% de la variabilidad de Y está explicada por el modelo de regresión.

Ejemplo 2

Calcúlese la covarianza, el coeficiente de correlación y el coeficiente de determinación sobre los datos del ejemplo 2.

Utilizamos la tabla del ejemplo 2.

- *La covarianza:*
  $S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (Y_i - \overline{Y}) = -10'471$
- *El coeficiente de correlación:*
  $R = \frac{S_{xY}}{S_x \cdot S_Y} = \frac{-10'471}{\sqrt{114'05} \cdot 1'113} = \frac{-10'471}{11'27} = -0'9291$
  El coeficiente de correlación es negativo y alto, esto indica una relación lineal decreciente entre las variables "Contenido de oxígeno" y "Profundidad".
- *El coeficiente de determinación:*
  $R^2 = (-0'9291)^2 = 0'863$
  El coeficiente de determinación es alto. Indica un buen ajuste del modelo.
  $R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{1'6722}{12'2473} = 0'863$
  También nos indica que el 86'35% de la variabilidad de Y está explicada por modelo de regresión.

7 Descomposición de la variabilidad. El test F

Los métodos de regresión explican cómo la variable respuesta, Y, se comporta de distinta manera en función del valor que tome la variable explicativa, X.
En consecuencia, parte de la variabilidad de Y quedaría justificada por la influencia de la variable X, mientras que otra parte sería debida al error del modelo.
Además, gracias al modelo de regresión podemos obtener predicciones más precisas de Y a partir del valor conocido de $X = x_i$, que si no conociéramos dicho valor.

Así:

– sin tener en cuenta la variable explicativa X, la mejor predicción que podemos hacer de Y es su media, $\overline{Y}$, mientras que si sabemos que $X = x_i$ la predicción será $\widehat{Y}_i$, el valor correspondiente en la recta de regresión.

– sin usar la recta de regresión, los residuos obtenidos serían $Y_i - \overline{Y}$ mientras que, usando la recta de regresión, los residuos vendrían dados por $Y_i - \widehat{Y}_i$,

Por otro lado,
sabemos que una medida de la variabilidad de Y es su cuasivarianza, que viene dada por:

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

además:

$$Y_i - \overline{Y} = \left(Y_i - \widehat{Y}_i\right) + \left(\widehat{Y}_i - \overline{Y}\right),$$

por lo que, haciendo algunas cálculos, podemos descomponer la "variabilidad de Y" (sin tener en cuenta los grados de libertad –el denominador–) en dos sumandos:

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2$$

Es habitual representar esta situación mediante la llamada **tabla de análisis de la varianza**.

Fuente de variación	Suma de cuadrados	Grados de libertad
Debida a la regresión	$\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2$	1
Debida al error	$\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$	$n-2$
Total	$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$	$n-1$

La variabilidad de toda la muestra la denominamos suma total de cuadrados y la denotamos por sus siglas en inglés de la siguiente manera:

$$TSS = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

Según indica la tabla de análisis de la varianza, la variabilidad de toda la muestra se descompone en dos sumandos:

– El primero de ellos representa las desviaciones de las predicciones respecto a la media global. Por tanto, sirve como medición de la "variabilidad que se puede explicar basándose en el modelo de regresión".
– El segundo representa las desviaciones de los valores observados Y_i respecto de las predicciones, y en consecuencia refleja la *variabilidad no explicada por la regresión, sino debida al error*. Por ello

se interpreta como "variabilidad residual", se calcula mediante la suma de los residuos al cuadrado, denominada más brevemente como suma residual de cuadrados, y denotada por sus siglas en inglés como:

$$RSS = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

Definición 3
En este punto surge de nuevo un concepto muy interesante que ya hemos introducido anteriormente: el ***coeficiente de determinación****, e indica la proporción de variabilidad (de Y) explicada (por el modelo de regresión):*

$$R^2 = 1 - \frac{RSS}{TSS}$$

Recordemos que el coeficiente de determinación es un número entre cero y uno, y cuanto más próximo a uno, más cerca estarán las observaciones de la recta ajustada. Además, coincide con el cuadrado del coeficiente de correlación entre la variable explicativa y la variable respuesta.

El test F

La descomposición de la variabilidad, aparte del interés en sí misma, se suele emplear para efectuar lo que se conoce como test F. Consiste en contrastar:

$H_0 : Y = \beta_0 + \varepsilon$	para algún β_0	*(no hay relación lineal)*
$H_a : Y = \beta_0 + \beta_1 X + \varepsilon$	para algún β_0 y algún β_1	*(sí hay relación lineal)*

La hipótesis nula nos dice que no sería necesaria la regresión, pues la media de Y es la misma, β_0, cualquiera que sea el valor de X.
La hipótesis alternativa nos dice que la recta no es horizontal, $\beta_1 \neq 0$, por lo que la regresión aportaría información relevante sobre Y.
Por este motivo, se conoce como un **contraste de la regresión**.

Observamos también que coincide con el contraste de la hipótesis nula $H_0 : \beta_1 = 0$ frente a la alternativa $H_0 : \beta_1 \neq 0$, que hemos visto anteriormente.

Para efectuar el contraste de la regresión, usamos lo que se conoce como test F, que está basado en el estadístico F, que se construye así

$$F = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2 / 1}{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 / (n-2)} \sim F_{1,n-2}$$

que tiene distribución conocida como *distribución F o distribución de Fisher-Snedecor con* (1,n–2) *grados de libertad* –la distribución da nombre al test–.

Rechazaremos la hipótesis nula si la variabilidad explicada es grande en comparación con la variabilidad residual, pues esto constituiría una prueba de que la regresión es relevante ya que a ella se debe una parte sustancial de la variabilidad.
Así, si se hubiera fijado un nivel de significación α, se rechazaría la hipótesis nula cuando $F > f_{1,n-2,\alpha}$, porque se habrían encontrado pruebas significativas a ese nivel de que la función de regresión no es una recta horizontal.

Ejemplo 2

En el ejemplo 2 obtener la tabla de análisis de la varianza y efectuar el test F para contrastar si el contenido de oxígeno cambia con la profundidad, bajo el modelo lineal.

Tabla de análisis de la varianza

Fuente de variación	Suma de cuadrados	Grados de libertad
Debida a la regresión	$\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}_i\right)^2=10'575$	1
Debida al error	$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=1'672$	$n-2=9$
Total	$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=12'247$	$n-1=10$

El test F

- H_0: No hay relación lineal, la pendiente vale cero
  H_a: Sí hay relación lineal, la pendiente es distinta de cero
  o, lo que es lo mismo:
  $H_0: Y=\beta_0+\varepsilon$ para algún β_0
  $H_a: Y=\beta_0+\beta_1X+\varepsilon$ para algún β_0 y algún β_1
- El estadístico F es:
  $$F=\frac{\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2/1}{\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2/(n-2)}\sim F_{1,n-2} \text{ bajo } H_0$$
- Valor del estadístico para la muestra dada:
  $$F=\frac{\frac{10'575}{1}}{\frac{1'67225}{9}}=\frac{10'575}{0'1858}=56'91$$
- Nivel crítico
  $P[F>56'91]<0'01$
  Aunque no podemos calcular el valor exacto de este nivel crítico (solo tenemos tablas de la distribución F para los niveles de significación $\alpha=0'05, 0'025$ y $0'01$), al ser un valor tan alto podemos deducir que la probabilidad que nos deja a su derecha es muy pequeña.
  Por otra parte, aunque no podamos comprobarlo ahora, este nivel crítico coincide con el nivel crítico obtenido en el contraste para el coeficiente β_1 (la pendiente de la recta), pues, como ya indicamos, el test F es equivalente al contraste de hipótesis sobre la pendiente, en el caso de la regresión lineal simple.
- Conclusiones
  En este ejemplo, el valor del estadístico de contraste es $F=56'916$, con un nivel crítico asociado muy pequeño ($<0'01$). Por lo tanto, se rechaza la hipótesis nula de que la recta de regresión es horizontal, con las significaciones usuales (5% y 1%).

Tabla resumen del análisis de la varianza y el test F

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico F	Nivel crítico
Debida a la regresión	10'575	1	10'575	56'91	$<0'01$
Debida al error	1'672	9	0'1858		
Total	12'247	10	1'2247		

8 Predicción

Un modelo de regresión permite estimar la media de Y, así como prever valores individuales de la variable respuesta, para distintos valores x de la variable explicativa X.

Tanto la estimación de la media, como la predicción del valor de Y se obtienen sustituyendo en la recta de regresión el valor de X, x_0, y calculando el valor $\widehat{Y}$:

$$\boxed{\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_0}$$

por tanto, sus valores numéricos son idénticos.
Sin embargo, la precisión (dada por los errores típicos) de estas estimaciones es distinta, como se puede ver observando los respectivos intervalos de confianza.

8.1 Estimación de la media condicionada

Supongamos que se desea estimar la media de la variable respuesta Y condicionada por $X = x_0$, es decir, la estimación del valor esperado de Y cuando $X = x_0$.

Basándonos en el modelo de regresión obtenemos que la estimación de la media de Y cuando $X = x_0$ es:

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

Además, basándonos en este estimador $\widehat{Y}_0$, conocida su distribución (T de Student con n-2 grados de libertad), así como su media y su error típico, podemos obtener el "intervalo de confianza" del 100(1-α)% para la media de Y condicionada por $X = x_0$:

$$\left(\widehat{Y}_0 - t_{n-2,\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n_0}}\ ,\ \widehat{Y}_0 + t_{n-2,\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n_0}}\right)$$

donde:
$t_{n-2,\alpha/2}$ es el cuantil de la distribución T_{n-2} que verifica $P[T_{n-2} > t_{n-2,\alpha/2}] = \alpha/2$,
y
$n_0 = \dfrac{n}{1 + \frac{(x_0-\overline{x})^2}{S_x^2}}$

8.2 Predicción de una nueva observación

Para predecir el valor concreto que tomará la variable Y cuando $X = x_0$ usaremos el mismo valor $\widehat{Y}_0$.

Para obtener un intervalo de confianza para la predicción de un nuevo valor de Y cuando $X = x_0$, debemos conocer su distribución (T de Student con n-2 grados de libertad), su media y su error típico, que no coincide con la anterior.

En este caso, el intervalo de confianza 100(1-α)% para la predicción de la nueva observación Y_0 cuando $X = x_0$ –conocido como "intervalo de predicción"– será:

$$\left(\widehat{Y}_0 - t_{n-2,\alpha/2}\,\widehat{\sigma}\sqrt{1+\frac{1}{n_0}}\ ,\ \widehat{Y}_0 + t_{n-2,\alpha/2}\,\widehat{\sigma}\sqrt{1+\frac{1}{n_0}}\right)$$

Observación 2
En el "intervalo de predicción" para Y dada $X = x_0$, la variabilidad es mayor que la del "intervalo de confianza" para la media de Y cuando $X = x_0$, pues hay un uno añadido en la fórmula de la medida de su dispersión.

El *intervalo de predicción* es similar al *intervalo de confianza*, excepto que el *intervalo de predicción* está diseñado para cubrir un "objetivo móvil", el nuevo valor "aleatorio" de Y, mientras que el *intervalo*

de confianza está diseñado para cubrir el "objetivo fijo", el valor –esperado– de Y para un x_0 dado. De ahí que tenga sentido que el *intervalo de predicción* sea más amplio (más dispersión), para un mismo coeficiente de confianza.

Ejemplo 2

Sobre los datos del contenido de oxígeno a ciertas profundidades, la estimación del contenido medio de oxígeno a una profundidad de 18 metros (o la predicción del valor del oxígeno a una profundidad de 18 metros) es:

$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = 7'97 - 0'09 \cdot 18 = 6'3$ *mg/l*

El intervalo de confianza del 95% para la media condicionada es:
$(5'94, 6'69)$

La predicción del contenido de oxígeno a una profundidad de 18 metros es la misma:

$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = 7'97 - 0'09 \cdot 18 = 6'3$ *mg/l*

Pero el intervalo de confianza del 95% para la predicción es más amplio, como podemos ver:
$(5'27, 7'36)$

Observación 3

Tenemos que ser muy cautelosos al usar la línea de regresión para la predicción fuera del rango de valores observados de las variables explicativas. Para el ejemplo anterior, los valores observados de X en nuestra muestra están entre 10 m y 40 m. El uso de la línea de regresión para la predicción fuera de este rango se llama *extrapolación*. A medida que nos alejamos de este rango, la relación general entre la variable respuesta y la variable explicativa puede cambiar. En general, no se debe hacer extrapolación, no debemos hacer estimaciones más allá del rango de valores observados para la variable explicativa.

9 Ejercicios

Las tablas con los cálculos necesarios para resolver estos dos ejercicios están en el anexo, al final de este documento.

1. Con la finalidad de buscar el mayor rendimiento de la tierra, un agricultor, preocupado por su cosecha de naranjas, está interesado en estudiar el grado de relación entre la cantidad de fruta recogida (Tm) y la lluvia caída en los últimos diez años (m^3). La información de la que dispone es la siguiente:

Naranjas (Tm)	10'1	8'2	7'2	11'4	8'5	9'6	5'9	6'8	7'9	6'9
Lluvia (m^3)	1'3	0'9	0'8	1'7	1'0	1'4	0'7	0'6	1'2	0'7

   (a) Ajusta un modelo de regresión lineal simple de la cantidad de fruta recogida sobre la cantidad de lluvia caída. Para ello se debe calcular las estimaciones de la ordenada en el origen, de la pendiente y de la desviación típica del error.

   (b) Calcula el coeficiente de correlación y el coeficiente de determinación.

   (c) ¿Qué cosecha se espera recoger si la cantidad de lluvia caída es 1'45 m^3?

2. Para saber si existe una asociación entre la cantidad de algas en el agua de los estanques de una zona y la claridad del agua se recogieron muestras de agua de siete estanques locales. Para medir la claridad del agua, se corta un pequeño disco de cartulina blanca, y se divide en cuatro partes iguales, pintando de negro dos partes opuestas; a continuación se coloca el disco en el fondo de un vaso cilíndrico graduado. Para cada muestra, se vierte lentamente el agua del estanque en el cilindro hasta que el disco ya no es visible desde arriba, apuntando en ese momento el volumen de agua contenido en el vaso. Como indicador de la concentración de algas se ha utilizado la concentración de clorofila en el agua (se extrae la clorofila de las muestras de agua y se utiliza un espectrofotómetro para determinar su concentración). En la tabla siguiente figuran los resultados obtenidos:

Concentración de clorofila (X) (μg/l)	14	5	10	7	17	16	3
Claridad del agua (Y) (ml)	28	68	32	54	18	25	77

   Realiza un análisis de regresión lineal simple de la claridad del agua en función de la concentración de clorofila:

   (a) Identifica la variable explicativa y la variable respuesta.

   (b) Representa el diagrama de dispersión de la variable respuesta sobre la variable explicativa. Explica alguna característica visible en el diagrama: tendencia creciente o decreciente, posible linealidad, dispersión en torno a la tendencia.

   (c) Ajusta el modelo: estimaciones de la ordenada en el origen y de la pendiente, varianza y desviación típica del error.

   (d) Calcula los intervalos de confianza del 95% para la ordenada en el origen y la pendiente, y efectúa el contraste de que la pendiente vale cero.

Anexo

Cálculos del ejercicio 1

	Lluvia x_i	Cosecha Y_i	$(x_i-\overline{x})^2$	$(Y_i-\overline{Y})^2$	$(x_i-\overline{x})\cdot(Y_i-\overline{Y})$	Valores ajustados $\widehat{Y}_i=\widehat{\beta}_0+\widehat{\beta}_1 x_i$	Residuos $\widehat{\epsilon}_i=Y_i-\widehat{Y}_i$	$\widehat{\epsilon}_i^2$	$(\widehat{Y}_i-\overline{Y}_i)^2$
	1′3	10′1	0′073	3′423	0′500	9′444	0′66	0′4301	1′43
	0′9	8′2	0′017	0′003	0′007	7′675	0′52	0′2756	0′33
	0′8	7′2	0′053	1′103	0′242	7′233	−0′03	0′0011	1′03
	1′7	11′4	0′449	9′923	2′111	11′213	0′19	0′0348	8′78
	1′0	8′5	0′001	0′063	−0′008	8′117	0′38	0′1464	0′02
	1′4	9′6	0′137	1′823	0′499	9′886	−0′29	0′0821	2′68
	0′7	5′9	0′109	5′523	0′776	6′790	−0′89	0′7929	2′13
	0′6	6′8	0′185	2′103	0′624	6′348	0′45	0′2042	3′62
	1′2	7′9	0′029	0′123	−0′059	9′002	−1′10	1′2142	0′57
	0′7	6′9	0′109	1′823	0′446	6′790	0′11	0′0120	2′13
Sumas	10′3	82′5	1′161	25′905	5′135	–	**0′00**	3′1934	22′7

Cálculos del ejercicio 2

	Clorofila x_i	Caridad Y_i	$(x_i-\overline{x})^2$	$(Y_i-\overline{Y})^2$	$(x_i-\overline{x})\cdot(Y_i-\overline{Y})$	Valores ajustados $\widehat{Y}_i=\widehat{\beta}_0+\widehat{\beta}_1 x_i$	Residuos $\widehat{\epsilon}_i=Y_i-\widehat{Y}_i$	$\widehat{\epsilon}_i^2$	$(\widehat{Y}_i-\overline{Y}_i)^2$
	14	28	13′80	229′31	− 56′24	28′17	− 0′17	0′03	224′10
	5	68	27′94	617′88	−131′39	64′45	3′55	12′63	453′84
	10	32	0′08	124′16	3′18	44′29	−12′29	151′15	1′33
	7	54	10′80	117′88	− 35′67	56′39	− 2′39	5′69	175′37
	17	18	45′08	632′16	−168′82	16′08	1′92	3′68	732′30
	16	25	32′65	329′16	−103′67	20′11	4′89	23′89	530′41
	3	77	53′08	1146′31	−246′67	72′51	4′49	20′19	862′25
Sumas	72	302	183′43	3196′86	−739′29	–	**0′00**	217′26	2979′60