EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Fundamentos de medicina basada en la evidencia

Correlación. Modelos de regresión

Molina Arias M[1], Ochoa Sangrador C[2], Ortega Páez E[3]
[1]Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.
[2]Servicio de Pediatría. Hospital Virgen de la Concha. Zamora. España.
[3]UGC de Maracena. Distrito Granada-Metropolitano. Granada. España.

Correspondencia: Manuel Molina Arias, mma1961@gmail.com

Palabras clave en español: estadística; correlación; regresión.
Palabras clave en inglés: statistics; correlation; regression.

Fecha de recepción: 12 de mayo de 2021 • **Fecha de aceptación:** 24 de mayo de 2021
Fecha de publicación del artículo: 10 de junio de 2021

Evid Pediatr. 2021;17:25.

CÓMO CITAR ESTE ARTÍCULO

Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Correlación. Modelos de regresión. Evid Pediatr. 2021;17:25.

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2021;17:25.
 • ISSN: 1885-7388

Correlación. Modelos de regresión

Molina Arias M[1], Ochoa Sangrador C[2], Ortega Páez E[3]
[1]Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.
[2]Servicio de Pediatría. Hospital Virgen de la Concha. Zamora. España.
[3]UGC de Maracena. Distrito Granada-Metropolitano. Granada. España.

Correspondencia: Manuel Molina Arias, mma1961@gmail.com

En el presente artículo de Fundamentos de Medicina Basada en la Evidencia trataremos dos técnicas que sirven para estudiar la asociación estadística entre variables cuantitativas: correlación y regresión.

Aunque las dos técnicas pueden parecer similares y comparten unos cálculos matemáticos parecidos, en realidad son conceptualmente diferentes. Como veremos más adelante, la correlación nos informa únicamente sobre si existe relación entre dos variables y cómo se modifica una de ellas con los cambios de la otra. Por su parte, el análisis de regresión va un paso más allá y trata de encontrar un modelo que nos permita predecir el valor de una de las variables (denominada dependiente) en función del valor que tome la otra variable (denominada independiente).

En el flujo de análisis de los resultados de un estudio, suele ser habitual estudiar la correlación entre dos variables como un paso previo para tratar de establecer un modelo de regresión entre ambas.

CORRELACIÓN

La correlación estudia si existe asociación entre dos variables cuantitativas y establece cuál es la dirección y la magnitud de esa asociación. De este modo, nos informa de cómo cambia de forma sistemática una de las variables con los cambios en la otra, sin suponer ningún tipo de dependencia entre las dos. Es importante comprender que correlación no es sinónimo de causalidad ni de dependencia.

La dirección de la correlación nos indicará el sentido de los cambios de una variable respecto a la otra. Cuando las dos variables cambien en el mismo sentido, hablaremos de correlación directa o positiva, mientras que cuando lo hagan en sentidos opuestos, diremos que existe una correlación inversa o negativa.

El segundo aspecto de esta asociación es su magnitud o intensidad, que se cuantificará mediante el uso de coeficientes de correlación, de los que hablaremos más adelante.

Un tercer aspecto de esta asociación sería su forma, que nos indica el tipo de línea que mejor se ajusta a la representación gráfica de los pares de valores de las dos variables. En este artículo nos centraremos en la forma más sencilla, en la que el ajuste se realiza con una línea recta, por lo que hablaremos de correlación lineal.

COEFICIENTES DE CORRELACIÓN

La forma más sencilla de medir la relación entre dos variables sería a través de su covarianza, que es el parámetro que indica la variabilidad conjunta de dos variables aleatorias.

Sin embargo, no podemos hacer un uso directo de la covarianza, ya que su valor depende de las escalas de medición de las variables, lo que nos impediría realizar comparaciones entre distintos pares de variables. Para solucionar este problema, podemos estandarizar la covarianza, obteniendo así los denominados coeficientes de correlación.

Todos los coeficientes de correlación tienen valores entre -1 y 1. Su valor nulo es el 0, que indica ausencia de correlación entre las dos variables.

El signo del coeficiente nos proporcionará la dirección de la asociación. Un coeficiente con signo positivo indica correlación directa: cuando una variable aumenta o disminuye, la otra lo hace en el mismo sentido. Por otra parte, un coeficiente de signo negativo nos indicará una correlación inversa: cuando una variable aumenta o disminuye, la otra varía en el sentido opuesto.

El valor del coeficiente nos informa de la intensidad de la asociación entre las dos variables. Cuanto más se aleje del 0, mayor será la fuerza de la asociación, de forma que los valores de -1 o 1 indican que existe una correlación perfecta, inversa o directa, respectivamente. Además, podemos hacer un paralelismo entre el valor y la intensidad de la asociación: valores de 0 indican asociación nula, de 0,1 asociación pequeña, de 0,3 mediana, 0,5 moderada, 0,7 alta y 0,9 asociación muy alta.

Como es lógico, la estimación del coeficiente debe ser lo suficientemente precisa para poder valorar su magnitud y solo cuando el intervalo de confianza se encuentre por encima o por debajo de cero podremos considerar que hay correlación. Para concretar la estimación podemos hacer un contraste de hipótesis con la hipótesis nula de ausencia de correlación, de manera que si la diferencia es estadísticamente significativa asumimos la existencia de correlación. La alternativa es calcular los intervalos de confianza que, además, nos permiten valorar mejor la estimación del valor poblacional.

Coeficiente de correlación de Pearson

El **coeficiente de correlación lineal producto-momento**, más conocido como **coeficiente de correlación de Pearson** (r), es el más utilizado y se obtiene al dividir la covarianza entre el producto de la varianza de las dos variables:

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

donde s_{xy} representa la covarianza y s_x y s_y las varianzas de las variables "x" e "y".

Para poder utilizar el coeficiente de correlación de Pearson entre dos variables cuantitativas deben cumplirse una serie de supuestos:

1. La relación entre las dos variables debe ser lineal. Esto puede comprobarse de forma sencilla dibujando un diagrama de puntos o de dispersión entre las dos variables, comprobando que la forma en que se modifica una variable en función de la otra sigue, de forma aproximada, una línea recta.
2. Ambas variables deben seguir una distribución normal en la población. Para comprobarlo podemos realizar una prueba numérica como la de Kolmogorov-Smirnov o la de Shapiro-Wilk. Recordemos que ambas pruebas son poco potentes y establecen la hipótesis nula de normalidad, por lo que siempre es conveniente, antes de asumir la normalidad de la variable, complementarlas con un método gráfico, como puede ser el histograma o el gráfico de cuantiles teóricos.
3. Debe existir homocedasticidad, lo que quiere decir que la varianza de la variable "y" debe ser constante a lo largo de los valores de la variable "x". Podemos comprobar si se cumple este supuesto de forma sencilla dibujando el diagrama de dispersión y comprobando que la nube de puntos se dispersa de forma similar a lo largo de los valores de la variable "x".

Por último, debemos saber que el valor de este coeficiente es sensible a la presencia de valores extremos en la distribución, que pueden sesgar la magnitud del efecto estimado. En estos casos, nos plantearemos si lo más idóneo es utilizar alguna alternativa al coeficiente de correlación de Pearson.

Coeficiente de correlación de Spearman

El **coeficiente de correlación por rangos**, más conocido como **coeficiente de correlación de Spearman** (ρ) es el equivalente no paramétrico del coeficiente de Pearson. Como ocurre con el resto de las técnicas no paramétricas, no se emplean los datos directos para el cálculo del coeficiente, sino su transformación en rangos.

Utilizaremos el coeficiente de Spearman cuando no se cumplan los supuestos necesarios para utilizar el coeficiente de Pearson o cuando tratemos con variables ordinales.

Aunque la potencia del coeficiente de Spearman es menor que la del coeficiente de Pearson, tiene una serie de ventajas sobre este último. En primer lugar, no exige supuesto de linealidad, por lo que puede utilizarse en casos de relación logística y exponencial. Solo debe cumplirse que la relación entre las dos variables sea monótona, lo cual quiere de decir que, cuando una de las variables cambia, la otra lo hace con una tendencia constante.

En segundo lugar, al ser una prueba no paramétrica, no precisa asumir el supuesto de normalidad de las variables. Por último, al calcularse con los rangos en lugar de con los datos directos, es mucho más robusto a la presencia de valores extremos que el coeficiente de Pearson.

Ejemplo de cálculo de los coeficientes de correlación

Tenemos una base de datos con 30 registros de niños asmáticos (Fundamentos_graficos.RData) y vamos a determinar si existe correlación entre los valores de peso y talla estandarizados (Z.Peso, Z.Talla). Emplearemos para ello el programa R utilizando su interfaz R-Commander.

Una vez cargados los datos, representamos el diagrama de dispersión (figura 1), con el que podemos asumir que ambas variables se relacionan de forma lineal.

Comprobamos ahora la asunción de normalidad, mediante una prueba de Shapiro-Wilk. Seleccionamos la opción del menú Estadísticos > Resúmenes > Test de normalidad… Marcamos la variable Z.Peso y la prueba elegida, en este caso, Shapiro-Wilk.

El programa nos ofrece el resultado, con un estadístico W = 0,948 y un valor de significación de p = 0,158. No podemos rechazar la hipótesis nula, por lo que asumimos que la variable peso estandarizado sigue una distribución normal. Podemos repetir el proceso para la variable Z.Talla, llegando a la misma conclusión (W = 0,982, p = 0,882).

Para mayor seguridad, complementamos las pruebas numéricas con un método gráfico, como el histograma. Seleccionamos, para las dos variables, la opción del menú Gráficos > Histograma… y, en la ventana emergente, seleccionamos la variable. En la figura 2 podemos ver el resultado.

Figura 1. Obtención del diagrama de dispersión de dos variables cuantitativas con la interfaz R-Commander

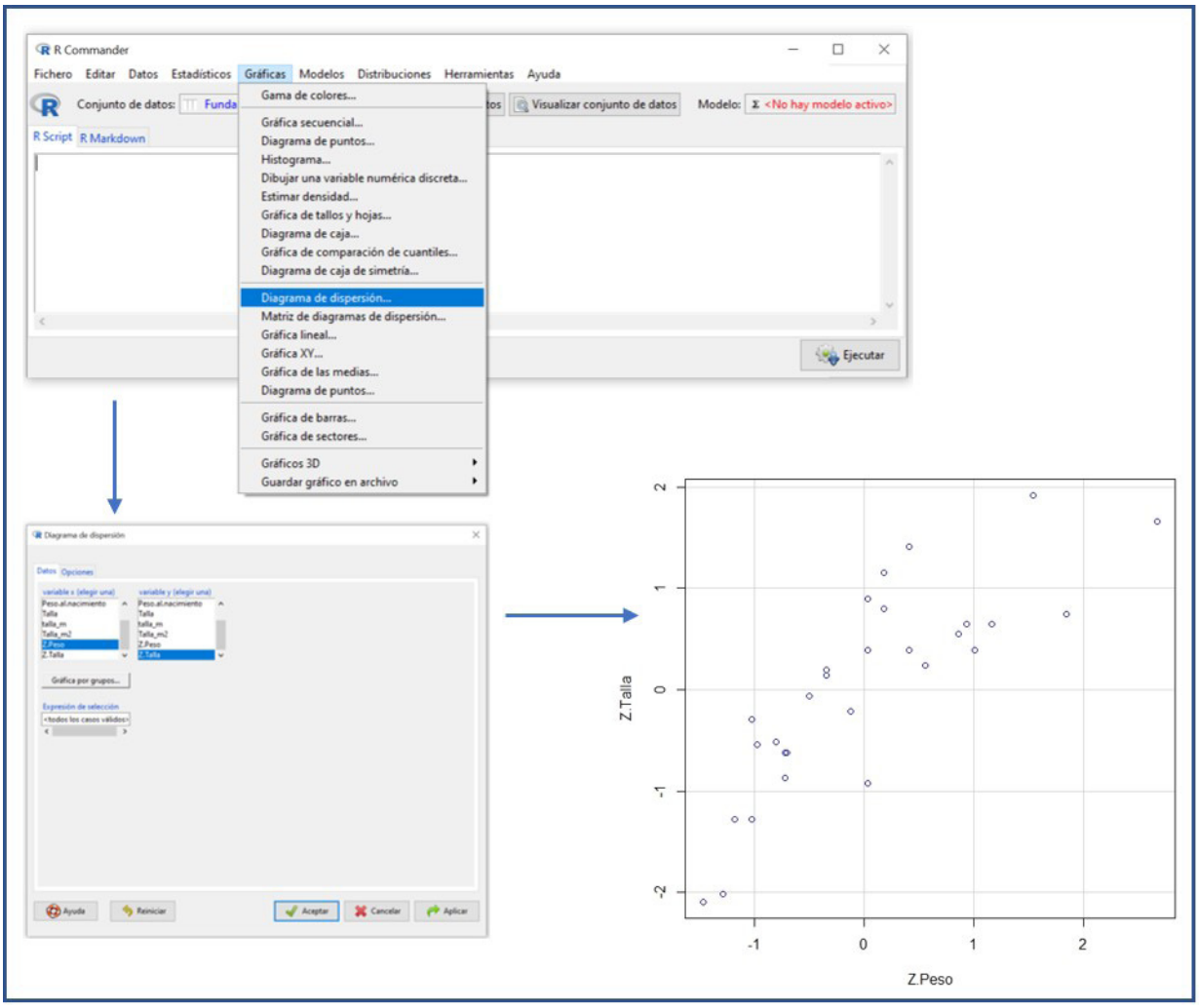
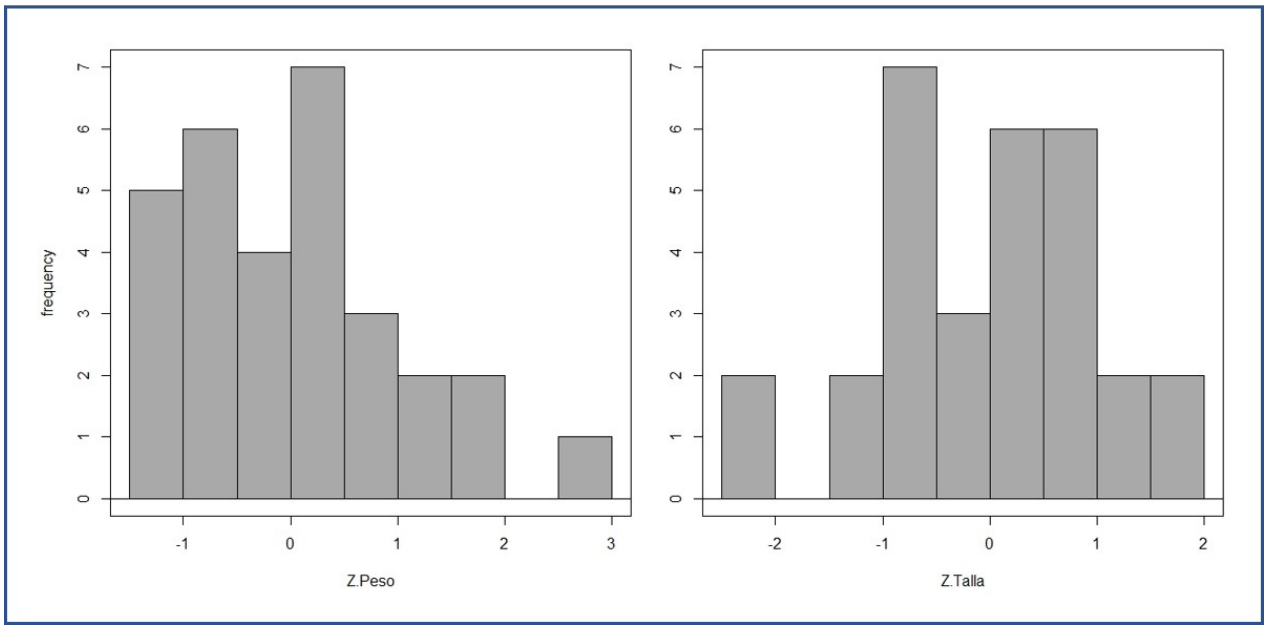


Figura 2. Histogramas de las variables peso estandarizado y talla estandarizadar



A la vista de los histogramas, ya no podemos estar tan seguros de poder asumir la condición de normalidad. Por razones didácticas, calcularemos el coeficiente de correlación de Pearson, aunque quizás lo más correcto sería calcular una alternativa no paramétrica, como el coeficiente de Spearman.

Por último, vamos a comprobar el supuesto de homocedasticidad. Si observamos el diagrama de dispersión (figura 1), podemos asumir que la nube se dispersa de forma similar en todo el rango de valores de la variable representada en el eje x.

Ya solo nos queda calcular el coeficiente de correlación. Seleccionamos la opción Estadísticos > Resúmenes > Test de correlación… y, en la ventana emergente, marcamos las dos variables y seleccionamos el coeficiente de correlación elegido (en este caso, el coeficiente de Pearson). Por último, marcamos la opción para un contraste bilateral (salvo que conozcamos el sentido de la asociación, en cuyo caso podríamos seleccionar una de las opciones de contraste unilateral).

El programa nos ofrece un valor de $r = 0,82$, con un valor de significación estadística $p < 0,05$. Por lo tanto, podemos concluir que existe una asociación alta entre las dos variables.
El programa R nos ofrece también el intervalo de confianza del 95% del coeficiente, que es de 0,66 a 0,91. El intervalo no incluye el valor nulo, por lo que ya sabemos que alcanza significación estadística sin necesidad de conocer el valor de *p*.

En el caso de no asumir la normalidad de las variables, hubiésemos seleccionado la opción del coeficiente de Spearman, obteniendo un valor $\rho = 0,85$, con un valor de $p < 0,05$. En el caso del coeficiente de Spearman, R-Commander no calcula de forma directa su intervalo de confianza, para lo cual habría que recurrir a paquetes adicionales.

Otros coeficientes de correlación

Además de los dos coeficientes descritos, existen numerosos coeficientes más, el más utilizado de los cuales es el **coeficiente tau de Kendall** (τ).

La tau de Kendall es otra alternativa no paramétrica al coeficiente de Pearson, cuyo uso puede preferirse al de Spearman en aquellos casos de muestras pequeñas y en las que exista una alta ligadura de rangos (al ordenar los datos por rangos, existen múltiples coincidencias en la misma posición).

Otros coeficientes menos utilizados son el **coeficiente de correlación parcial**, que estudia la relación entre dos variables pero teniendo en cuenta y eliminando la influencia de otras variables existentes; el **coeficiente de correlación semiparcial**, similar al anterior, pero que discrimina el efecto de terceras variables sobre las dos correlacionadas de forma independiente (no sobre las dos de forma simultánea, como el coeficiente parcial); y el **coeficiente de correlación múltiple**, que permite conocer la correlación entre una variable y un conjunto de variables, todas ellas cuantitativas.

REGRESIÓN

Una vez entendido el concepto de correlación, podemos comprender mejor el de regresión, que no solo explica la relación entre las variables, sino que trata de construir un modelo que nos permita predecir el valor de una de las variables (la dependiente o criterio) en función del valor que tome la otra variable (la independiente o explicativa).

Existen numerosos modelos de regresión que, a su vez, pueden clasificarse en simples o múltiples en función del número de variables independientes que lo componen. Los modelos de regresión simple son los que incluyen dos variables, una dependiente y otra independiente. Por su parte, los modelos de regresión múltiple incluyen una variable dependiente y más de una variable independiente.

A continuación, describiremos de forma somera los modelos de regresión simple más utilizados, que serán desarrollados con mayor profundidad en artículos sucesivos.

Modelos de regresión

Si tomamos una variable independiente "x" y una variable dependiente "y", todos los modelos de regresión simple se ajustan a la siguiente ecuación:

$$Función(y) = a + bx + e.$$

El componente "Función(y)" dependerá del tipo de variable dependiente del modelo, lo que nos condicionará el modelo de regresión concreto que tendremos que utilizar. En la figura 3 se muestra un ejemplo de diagrama de dispersión de dos variables con la línea de regresión del modelo, en este caso lineal, así como el significado de los diferentes coeficientes de la ecuación de regresión.

"a" y "b" son los denominados coeficientes de regresión. El componente "a" representa el valor de "y" cuando "x" vale 0. Suele denominarse interceptor, ya que es el punto donde la representación gráfica de la línea de regresión cruza el eje de ordenadas (eje y).

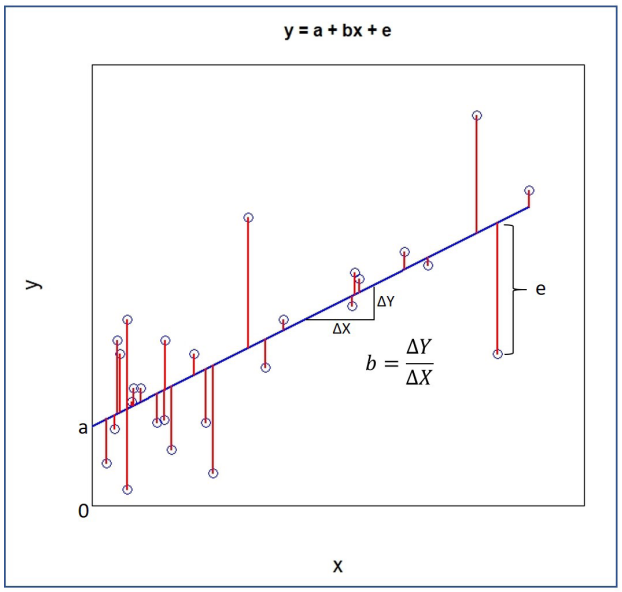
El componente "b" representa la pendiente de la línea y nos informa de en cuántas unidades aumenta la variable "y" por cada unidad que aumenta la variable "x".

Por último, el cuarto componente, "e", representa la variabilidad aleatoria del modelo. Esta variabilidad será la responsable de la diferencia que se produzca entre la predicción del modelo de regresión y el valor real observado en el estudio.

Según estos componentes, podemos definir los cuatro modelos de regresión simple utilizados con más frecuencia:

- **Regresión lineal simple**. Es el caso más sencillo y se aplica a dos variables cuantitativas. En este caso, la función

Figura 3. Ejemplo de diagrama de dispersión entre dos variables "x" e "y" con la representación de la recta del modelo de regresión. Se muestra el valor de los tres componentes de la ecuación: a (intercepto), b (la pendiente de la recta) y e (el error aleatorio o residuo)



del modelo es la media aritmética de la variable dependiente.

- **Regresión binaria logística**. Se utiliza cuando la variable dependiente es cualitativa dicotómica. En estos casos, la variable dependiente se codificará como 1 o 0 y la función del modelo será el logaritmo neperiano (natural) de la *odds* de que la variable tenga el valor 1. El coeficiente "b" representa el logaritmo neperiano de la *odds ratio* de que ocurra un fenómeno por unidad de cambio de la variable independiente, por lo que podremos estimar la *odds ratio* calculando su antilogaritmo (e^b).
- **Regresión de riesgos proporcionales de Cox**. Se utiliza en estudios de supervivencia, cuando la variable dependiente es de tipo tiempo a suceso. El modelo es similar al de la regresión logística, con la diferencia de que la función representa el logaritmo neperiano de la tasa de riesgos instantáneos (*hazard ratio*).
- **Regresión de Poisson**. Se utiliza cuando la variable dependiente se ajusta a una distribución de Poisson para cualquier combinación de valores de la variable independiente. La distribución de Poisson es una distribución discreta, por lo que los valores de la variable dependiente son enteros positivos, lo que la convierte en la técnica ideal para situaciones de recuento, como número de ingresos, número de hijos, etc.
  La función del modelo de regresión de Poisson es el logaritmo neperiano de lambda (λ), que representa la probabilidad de que ocurra un evento en un intervalo determinado, lo que suele corresponder a la densidad de incidencia en los estudios longitudinales.

La interpretación de todos estos modelos se verá de forma más clara cuando se desarrollen en próximos artículos, donde se describirán sus peculiaridades, sus requisitos de aplicación y su modo de llevar a cabo con ejemplos prácticos.

BIBLIOGRAFÍA

- Amat Rodrigo J. Correlación lineal y regresión lineal simple. En: Ciencia de datos [en línea] [consultado el 01/06/2021]. Disponible en: https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal
- Arriaza Gómez AJ, Fernández Palacín F, López Sánchez MA, Muñoz Márquez M, Pérez Plaza S, Sánchez Navas S. Estadística Básica con R y R-Commander. Cádiz: UCA; 2008.
- Estimating the difference between two population parameters. En: Bowers D (ed.). Medical statistics from scratch. An introduction for health professionals, 2nd ed. Reino Unido: John Wiley & Sons Ltd.; 2018. p. 119-31.
- Sánchez-Villegas A, Martín-Calvo N, Martínez-González MA. Correlación y regresión lineal simple. En: Martínez MA, Sánchez-Villegas A, Toledo EA, Faulin A (eds.). Bioestadística amigable (3.ª ed.). Barcelona: Elsevier España; 2014. p. 269-326.
- Solanas A, Guàrdia J. Modelos de regresión lineal. En: Peró M, Leiva D, Guàrdia J, Solanas A (eds.). Estadística aplicada a las ciencias sociales mediante R y R-Commander. Madrid: Ibergarceta Publicaciones; 2012. p. 434-97.