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

Fundamentos de medicina basada en la evidencia

Análisis multivariante. Regresión lineal múltiple

Carlos Ochoa Sangrador[1], Manuel Molina Arias[2], Eduardo Ortega Páez[3]
[1]Servicio de Pediatría. Hospital Virgen de la Concha. Complejo Asistencial de Zamora. Zamora. España.
[2]Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.
[3]Pediatra. Unidad de Gestión Clínica Góngora. Distrito Granada-Metropolitano. Granada España.

Correspondencia: Manuel Molina Arias: mma1961@gmail.com

Palabras clave en español: análisis multivariante; estadística; inferencia estadística.
Palabras clave en inglés: multivariate analysis; statistics; statistical inference.

Fecha de recepción: 24 de mayo de 2023 • **Fecha de aceptación:** 31 de mayo de 2023
Fecha de publicación del artículo: 7 de junio de 2023

Evid Pediatr. 2023;19:22.

CÓMO CITAR ESTE ARTÍCULO

Ochoa Sangrador C, Molina Arias M, Ortega Páez E. Análisis multivariante. Regresión lineal múltiple. Evid Pediatr. 2023;19:22.

Para recibir Evidencias en Pediatría en su correo electrónico debe darse de alta en nuestro boletín de novedades en
http://www.evidenciasenpediatria.es

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2023;19:22.
©2005-23 • ISSN: 1885-7388

Análisis multivariante. Regresión lineal múltiple

Carlos Ochoa Sangrador[1], Manuel Molina Arias[2], Eduardo Ortega Páez[3]
[1]*Servicio de Pediatría. Hospital Virgen de la Concha. Complejo Asistencial de Zamora. Zamora. España.*
[2]*Servicio de Gastroenterología. Hospital Universitario La Paz. Madrid. España.*
[3]*Pediatra. Unidad de Gestión Clínica Góngora. Distrito Granada-Metropolitano. Granada España.*

Correspondencia: Manuel Molina Arias: mma1961@gmail.com

INTRODUCCIÓN

En un documento previo de Fundamentos hemos expuesto la necesidad de realizar ajustes de covariables para controlar sesgos de análisis como la confusión e interacción y hemos revisado el método básico de ajuste: el análisis estratificado. Mencionamos que este método tiene dos limitaciones importantes: (1) si hay más de un factor de confusión, la aplicación es laboriosa debido al mayor número de estratos y además obliga a tener un tamaño de muestra relativamente grande y (2) requiere que los factores de confusión continuos se recodifiquen en un número limitado de categorías, lo que podría generar confusión residual. En este documento expondremos las técnicas de análisis multivariante o multivariado, que permiten superar algunas de estas limitaciones.

Las distintas técnicas de análisis multivariante se diseñaron como respuesta a la necesidad de estudiar fenómenos biológicos en los que estaban implicadas más de dos variables, algo habitual en epidemiología. Su desarrollo ha ido de la mano del aumento de la capacidad de procesamiento de los ordenadores, pudiéndose analizar modelos cada vez más complejos. La mayor sofisticación del análisis multivariante no debe llevarnos a prescindir del análisis descriptivo bivariante y tampoco del análisis estratificado, porque a menudo será la única forma de identificar la relación entre las variables implicadas.

Hay numerosos métodos de análisis multivariante, que varían en función del tipo de datos involucrados en el análisis (número de variables dependientes e independientes y escalas de medida de las variables, etc.) y del objetivo del mismo (de clasificación, estimativo o predictivo). Habitualmente estos análisis se emplean para identificar si una serie de variables independientes están asociadas o no a una variable dependiente. Algunas técnicas permiten analizar la relación de varias variables dependientes con una o varias independientes.

El análisis multivariante es más potente y preciso que el análisis estratificado y, al integrar más información, ofrece una estimación más válida de la realidad. La complejidad de los análisis, superadas las limitaciones de capacidad de cálculo con los procesadores modernos, radica principalmente en la interpretación de los numerosos resultados que nos van a ofrecer. Por ello, aunque recurramos a las técnicas multivariantes, no podemos renunciar a la exploración descriptiva de las variables.

La selección de la técnica multivariante adecuada depende de:

a) Si existen variables dependientes e independientes definidas en el análisis.
b) Cuántas variables dependientes están involucradas en cada análisis.
c) Qué escalas de medida tienen las variables dependientes e independientes.

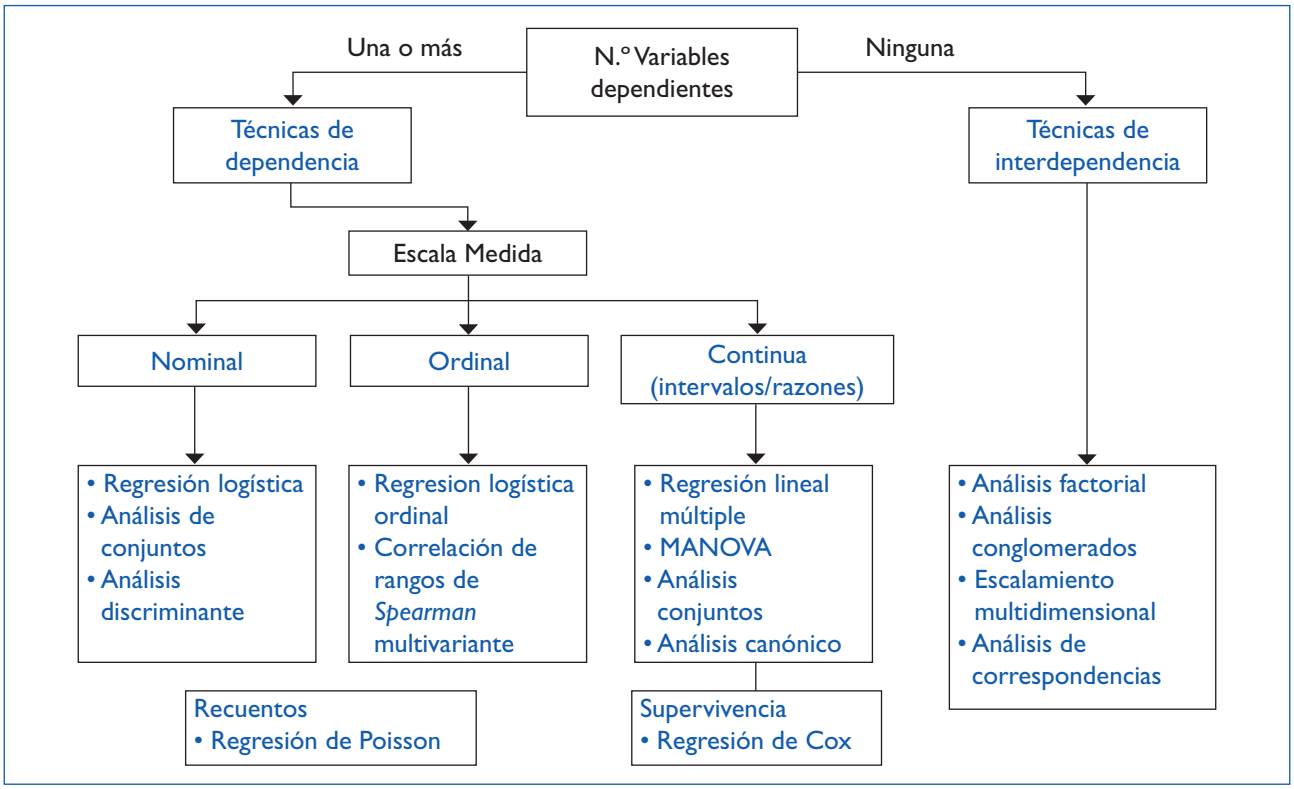
Si las variables involucradas tienen una relación de dependencia (podemos identificar una o más variables como dependientes y las variables restantes como independientes), usaremos técnicas de dependencia, como la regresión logística o la regresión lineal múltiple. En el caso contrario, usaremos técnicas de interdependencia, como el análisis factorial o el análisis de correspondencia. En la **Figura 1** se presenta un esquema de las principales técnicas, que incorpora una diferenciación en función de las escalas de medida de las variables dependientes.

OBJETIVOS DEL ANÁLISIS MULTIVARIANTE

Las técnicas de análisis multivariante pueden tener objetivos diferentes:

1. Reducir o simplificar datos. Ayuda a que los datos se sinteticen al máximo sin sacrificar información valiosa para facilitar la interpretación.
2. Ordenar y agrupar. Cuando tenemos múltiples variables, se crean grupos de objetos o variables "similares", basados en características medidas.
3. Estimar la dependencia entre variables. Permite identificar si las variables son mutuamente independientes o una o más variables dependen de las demás y, en este caso, estimar el grado de dependencia.
4. Predecir variables. Busca predecir los valores de una o más variables en función de los valores de otras variables.

Figura 1. Cuadro de clasificación de técnicas multivariantes



5. Contraste o validación de hipótesis. Permite probar hipótesis estadísticas específicas, formuladas en términos de parámetros multivariantes, para validar o reforzar suposiciones o convicciones previas.

PROCESO DE MODELIZACIÓN MULTIVARIANTE

El diseño y optimización de modelos multivariantes, tanto predictivos como estimativos, es un proceso estadístico relativamente complejo, que sigue una serie de pasos:

1. Elección del modelo máximo

En cualquier modelo multivariante debe definirse *a priori* qué variables independientes van a ser analizadas. El número de variables se ve limitado por el número de observaciones disponibles. En un modelo predictivo es habitual que se seleccionen todas las variables que presentan asociación con la variable dependiente en el análisis bivariable. En un modelo estimativo de riesgos ajustados, además de las variables significativas tendremos que incluir aquellas covariables que puedan comportarse como factores de confusión o interacción. Además de elegir las variables implicadas habrá que valorar si interesa introducir términos de interacción que exploren si existe modificación del efecto de una variable en función de diferentes valores de otra.

2. Codificación de variables y elección de los valores de referencia

Para variables métricas elegiremos si el valor de referencia es el valor más bajo o el más alto. En el caso de variables codificadas como 0 y 1, el valor de referencia será habitualmente el 0. Para variables nominales politómicas tendremos que crear variables indicadoras (*dummies*), que condicionarán la interpretación de los resultados en función del método de codificación elegido. El tipo de codificación condicionará el sentido de los coeficientes y su interpretación. Más adelante veremos un ejemplo de ellas.

3. Elección de la estrategia de modelización

La estrategia de modelización recoge el proceso de introducción y exclusión de variables en el modelo hasta conseguir el modelo más simple que explique la relación entre variables (principio de parsimonia). Los resultados pueden cambiar si elaboramos primero un modelo con todas las variables y vamos extrayendo las que no muestran asociación (Backward), si empezamos con una variable y vamos añadiendo el resto (Forward) o bien si usamos estrategias de comprobación de todas las variables ya incluidas con la estrategia anterior al introducir una nueva (Stepwise). Asimismo, debemos definir qué criterio estadístico, vinculado a cada variable (cambios en los coeficientes o significación individual) o a la bondad del modelo global (ej.: cambios en la F parcial para regresión

lineal, cambios en la razón de verosimilitudes para regresión logística), se emplea para ir comparando los modelos a cada paso.

4. Evaluación de la fiabilidad del modelo

Consiste en, una vez seleccionado el mejor modelo, comprobar si los datos predichos por él concuerdan con los datos observados en una nueva muestra de sujetos, obtenida de la misma población de la que se obtuvo la muestra original. Con frecuencia, esto resulta poco factible, por lo que se recurre a fraccionar la muestra original, estimar el modelo en una de las submuestras y contrastar su fiabilidad con los datos de la submuestra restante. Existen diferentes medidas que se pueden aplicar para contrastar la fiabilidad del modelo, adaptadas a cada tipo de análisis.

A continuación, mostraremos los pasos del análisis multivariante para la regresión lineal múltiple. En próximos artículos repasaremos otras técnicas.

REGRESIÓN LINEAL MÚLTIPLE

La regresión lineal múltiple es una extensión de la regresión lineal simple, previamente descrita en un documento anterior de Fundamentos. Se utiliza cuando queremos predecir el valor de una variable medida en una escala continua, de razones o intervalos, en función del valor de otras dos o más variables. La variable que queremos predecir es la variable dependiente.

La regresión múltiple tiene una serie de requisitos: linealidad (la relación entre las variables debe ser lineal), normalidad y homocedasticidad de los residuos (las diferencias entre los valores observados y los predichos por el modelo deben seguir una distribución normal con varianzas iguales a lo largo de todos los valores de las variables), equilibrio entre número de casos y variables (se requieren al menos 20 observaciones por cada variable incluida en el modelo), ausencia de colinealidad (debemos evitar introducir en el modelo variables independientes interrelacionadas) y control de valores fuera de rango (las observaciones con valores atípicos distorsionan los modelos).

Como se mencionó en el documento de regresión lineal simple, la regresión lineal se basa en la correlación entre variables y se concreta en la fórmula de regresión que permite estimar el valor de la variable dependiente a partir del valor de la variable independiente. La regresión lineal múltiple es una extensión de la simple, en la que la estimación se realiza a partir de los valores de dos o más variables independientes.

La ecuación general de la regresión lineal múltiple es de la forma:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

En la que Y es el valor de la variable dependiente, a es la constante del modelo (equivale al valor de la variable independiente cuando todas las dependientes valen cero o cuando no hay efecto de las variables independientes sobre la dependiente), b_i los coeficientes de cada variable independiente (equivale al cambio de valor de Y por cada unidad de cambio de X_i) y X_i sus valores.

Para variables nominales dicotómicas, el coeficiente se interpreta como la diferencia de medias de la variable dependiente para cada categoría de la variable independiente (en igualdad del resto de variables independientes). Los modelos calculan los errores estándar de cada coeficiente, a partir de los cuales obtenemos los intervalos de confianza. La estimación de los coeficientes resulta más compleja que la empleada con la regresión lineal simple, pero se sustenta en los mismos principios.

Para ilustrar el procedimiento de modelización emplearemos los datos de un estudio sobre factores predictivos de gastroenteritis bacteriana. En la **Figura 2** se presenta el diagrama de dispersión, del peso estandarizado (Zscore_Peso) en función de la edad de los pacientes, diferenciando los puntos en función del sexo. A la izquierda se presenta con la línea de regresión global y a la derecha por subgrupos de sexo.

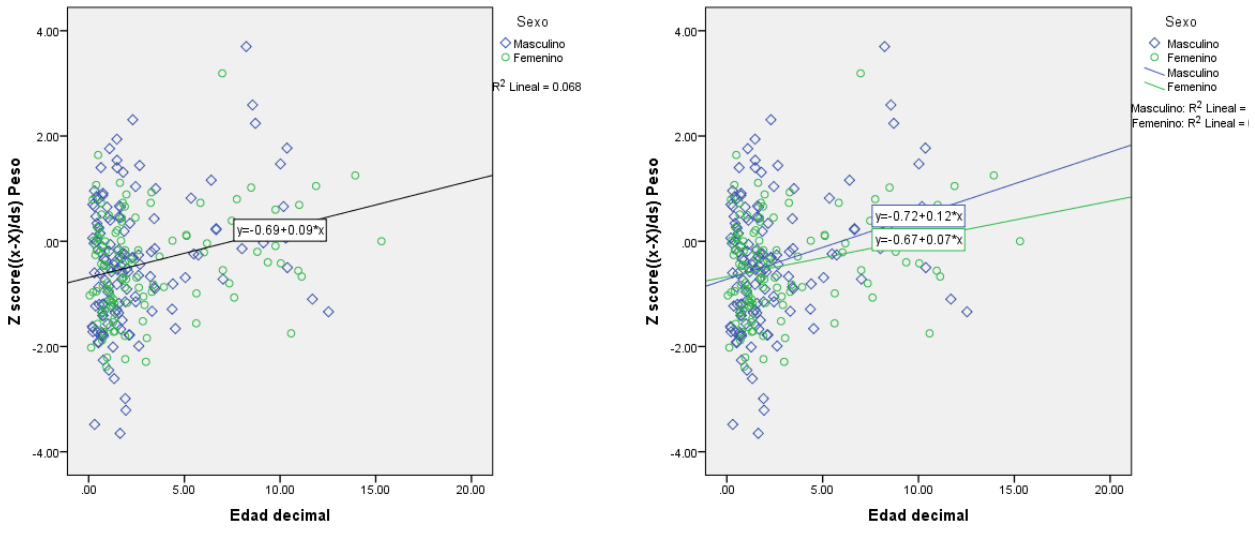
La fórmula insertada en cada gráfica nos permite estimar para cada edad el peso estandarizado. Vemos que por cada año aumenta el peso 0,09 desviaciones estándar; la interpretación natural es que los niños más pequeños tienen más afectación ponderal que los mayores. Si queremos mejorar la predicción incorporando al modelo el sexo, vemos que tanto las constantes (a: 0,72 para niños y 0,67 para niñas), como los coeficientes de la regresión (b: 0,12 para niños y 0,07 para niñas) son diferentes. Sumando las diferencias de constantes y coeficientes podemos estimar la diferencia de peso estandarizado entre sexos, ajustada por edad; en nuestro ejercicio, aproximadamente 0,10 desviaciones estándar (0,72 - 0,67 + 0,12 - 0,07). Además del sexo, nos puede interesar estimar el efecto de otras variables en el peso, por ejemplo, la frecuencia de deposiciones. Si añadimos nuevas variables independientes la representación gráfica, necesariamente multidimensional, es más compleja, por lo que prescindiremos de ella.

Veamos a continuación los pasos de la modelización.

1. Elección del modelo máximo

El primer paso será elegir las variables independientes que entrarán en el análisis. En función de nuestro conocimiento previo de la enfermedad a estudiar y del comportamiento de los datos en análisis previos, hemos decidido incluir en el modelo tres variables independientes: la edad, el sexo y la frecuencia de deposiciones.

Figura 2. Diagramas de dispersión del peso estandarizado y la edad, con las rectas de regresión global y separadas por sexos



2. Codificación de variables y elección de los valores de referencia

La variable edad se procesará con sus valores originales, considerando la unidad el año. Para poder operar con la variable sexo hemos recodificado los valores originales de la variable (1 masculino, 2 femenino) en nuevos valores que van a facilitar la interpretación de las estimaciones: 1 sexo masculino, 0 sexo femenino; de esta manera los coeficientes estimados para esta variable mostrarán la diferencia de peso estandarizado de los niños respecto a las niñas. Para la tercera variable, número de deposiciones, como es una variable ordinal, que no garantiza el supuesto de relación lineal, la simplificamos recodificándola como dicotómica: más de 3 deposiciones al día (1) o menos (0); de esta manera los coeficientes expresarán la diferencia de peso de los que tienen 3 o más deposiciones respecto al resto.

3. Elección de la estrategia de modelización

De las posibles estrategias disponibles, optamos por una estrategia "hacia atrás", introducimos todas las variables del modelo máximo y vamos eliminándolas de una en una, en función de los cambios que se producen en la significación estadística de las variables, indicada en el estadístico t correspondiente, y del modelo en su conjunto, representada en el estadístico F. Aunque podríamos usar un nivel de significación estándar (0,05) a la hora de eliminar variables, es habitual aumentar el umbral hasta 0,10, para evitar prescindir de variables potencialmente predictoras.

En la **Figura 3** presentamos una salida de resultados de la regresión lineal múltiple, realizada en RCommander, para tres modelos, el primero con las tres variables del modelo máximo y los siguientes, con la eliminación de las variables con menor significación. Podemos ver las variables de cada modelo, los coeficientes (Estimate), los errores estándares (Std. Error), el estadístico de contraste (t value) y su significación (Pr(>|t|)); junto a cada modelo se representa su estadístico F, su nivel de significación y los coeficientes de determinación múltiple, R^2 ("Multiple R-squared"), y sus estimadores ajustados al número de variables ("Adjusted R-squared"). Los coeficientes de determinación son los cuadrados de los coeficientes de correlación entre los valores observados de la variable dependiente y los predichos a partir de los coeficientes y valores de las variables independientes. R^2 equivale al porcentaje de varianza que el modelo explica, cuanto mayor sea este porcentaje mejor será el modelo.

Comprobamos que la variable edad tiene un coeficiente 0,08969, aproximadamente 0,09, que es similar al anteriormente observado en la **Figura 2**. Se interpreta como que por cada año de edad el peso aumenta 0,09 desviaciones estándar. Además, vemos que lleva asociado un contraste t significativo (0,00004). También vemos el coeficiente de sexo, que se asemeja a la estimación de la diferencia de medias realizada anteriormente (0,093); sin embargo, su contraste t no es significativo (p = 0,484), con un valor superior al de las otras variables. Por ello, es la primera variable a eliminar, lo que se observa en el modelo 2. Su eliminación no afecta a la validez del modelo, sin afectar a la significación del contraste F y mejorando discretamente el R^2 ajustado, a pesar de haber prescindido de una variable (de 0,07079 a 0,07266). En el modelo 2 permanece la variable número de deposiciones (más de 3), que presenta un coeficiente negativo -0,23812

Figura 3. Regresión lineal múltiple para peso estandarizado

```
> LinearModel.1 <- lm(dspeso ~ edaddec + Sexo1_0 + Heces3mas, data=GeaPed)
> summary(LinearModel.1)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.62661 0.13745 -4.559 <0.00001
edaddec 0.08969 0.02147 4.178 0.00004
Sexo1_0 0.09396 0.13430 0.700 0.4848
Heces3mas -0.23888 0.13413 -1.781 0.0761

Residual standard error: 1.062 on 252 degrees of freedom
Multiple R-squared: 0.08172,        Adjusted R-squared: 0.07079
F-statistic: 7.475 on 3 and 252 DF, p-value: 0.00008185

> LinearModel.2 <- lm(dspeso ~ edaddec + Heces3mas, data=GeaPed)
> summary(LinearModel.2)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.57032 0.11133 -5.123 <0.00001
edaddec 0.08809 0.02132 4.131 0.00004
Heces3mas -0.23812 0.13399 -1.777 0.0767

Residual standard error: 1.061 on 253 degrees of freedom
Multiple R-squared: 0.07994,        Adjusted R-squared: 0.07266
F-statistic: 10.99 on 2 and 253 DF, p-value: 0.00002648

> Confint(LinearModel.2, level=0.95)
 Estimate 2.5 % 97.5 %
(Intercept) -0.57031646 -0.78956320 -0.35106972
edaddec 0.08808917 0.04609393 0.13008441
Heces3mas -0.23812387 -0.50200110 0.02575336

> LinearModel.3 <- lm(dspeso ~ edaddec, data=GeaPed)
> summary(LinearModel.3)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.68838 0.08971 -7.673 <0.00001
edaddec 0.09202 0.02130 4.320 <0.00001

Residual standard error: 1.065 on 254 degrees of freedom
Multiple R-squared: 0.06845,        Adjusted R-squared: 0.06478
F-statistic: 18.66 on 1 and 254 DF, p-value: 0.00002238
```

(a mayor número de deposiciones menor peso) y un nivel de significación 0,0767; como es <0,10 la variable permanece en el modelo. Aunque podríamos finalizar la modelización aquí, hemos procedido a explorar el potencial efecto de su eliminación, lo que se observa en el modelo 3. Podemos ver que el contraste del estadístico F apenas cambia y que la R^2 ajustada desciende (de 0,07266 a 0,06478), por lo que nos quedaremos con el modelo 2. En la parte inferior del mismo se muestran los intervalos de confianza de los coeficientes de las variables (2,5% y 97,5%).

4. Evaluación de la fiabilidad del modelo

Nuestro modelo final, que incluye las variables edad y número de deposiciones, presenta un R^2 ajustado de 0,07266, lo que implica que el 7,2% de la varianza de peso se explica por el modelo; es el mejor porcentaje obtenido, aunque es bajo, por lo que podría haber otras variables que expliquen mejor la variable dependiente. El último paso sería comprobar la fiabilidad del modelo en otra muestra de pacientes. Se harían estimaciones de la variable dependiente a partir de los coeficientes estimados en la muestra original y los valores de los nuevos pacientes, lo que permitiría estimar nuevos R^2 ajustados. Si el nuevo R^2 fuera menor que el original, la fiabilidad del modelo quedaría cuestionada.

BIBLIOGRAFÍA

- Abraira V, Pérez de Vargas Luque A. Métodos multivariantes en bioestadística. Ed. Centro de Estudios Ramón Areces. Madrid 1996.
- Arezina R, Duolao W. Source and control of bias. En: Duolao W, Bakhai A (eds.). Clinical trials. A practical guide to design, analysis and reporting. Londres: Remédica; 2006. p. 55-64.
- Argimón Pallás JM, Jimenez Villa J. Métodos de investigación clínica y epidemiológica. Barcelona: Elsevier; 2006.
- Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Correlación. Modelos de regresión. Evid Pediatr. 2021;17:25.
- Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Regresión lineal simple. Evid Pediatr. 2021;17:46.
- Ochoa Sangrador C, Molina Arias M, Ortega Páez E. Métodos de ajuste de sesgos. Análisis estratificado. Evid Pediatr. 2022;18:31.
- Ortega Páez E, Ochoa Sangrador C, Molina Arias M. Regresión logística binaria simple. Evid Pediatr. 2022;18:11.
- Rosner B. Fundamentals of Biostatistics. 7th Edition. Boston: Brooks/Cole, Cengage Learning; 2011.
- Rothman K. J. Epidemiología Moderna. Madrid: Díaz de Santos; 1987.
- Woodwark M. Epidemiology study design and data analysis. London: Chapman & Hall/CRC; 1999.