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

Fundamentos de medicina basada en la evidencia

Regresión lineal simple

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; inferencia estadística; regresión lineal.
Palabras clave en inglés: statistics; statistical inference; linear regression.

Fecha de recepción: 14 de diciembre de 2021 • **Fecha de aceptación:** 20 de diciembre de 2021
Fecha de publicación del artículo: 22 de diciembre de 202

Evid Pediatr. 2021;17:46.

CÓMO CITAR ESTE ARTÍCULO

Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Regresión lineal simple. Evid Pediatr. 2021;17:46.

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2021;17:46.
 • ISSN: 1885-7388

Regresión lineal simple

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

Vimos en un artículo anterior de esta serie cómo la correlación estudia si existe asociación entre dos variables cuantitativas y establece cuál es la dirección y la magnitud de esa asociación. Por su parte, la regresión va un paso más allá y 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).

La regresión puede ser simple o múltiple, según el número de variables independientes o explicativas que se introduzcan en el modelo. Además, existen una serie de modelos de regresión, que pueden expresarse mediante la siguiente ecuación:

$$\text{Función}(y) = a + bx_1 + cx_2 + \ldots + nx_n + e.$$

Según la función que apliquemos a la variable dependiente "y" y el tipo de esta variable, definiremos los distintos modelos de regresión. En este artículo nos centraremos en el caso más sencillo, el de la regresión lineal simple, aplicado a dos variables cuantitativas, una de ellas la variable independiente (x) y la otra la dependiente (y).

REGRESIÓN LINEAL SIMPLE

En el caso de la regresión lineal, la función del modelo que aplicamos a la variable dependiente es la media aritmética, por lo que la recta de regresión pasará por el punto que definen las coordenadas de las medias de x e y.

Podemos representar la recta de regresión lineal simple mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x + e.$$

En la ecuación, β_0 y β_1 son los denominados coeficientes de regresión. El componente β_0 ("a" en la figura) 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 (figura 1).

Por su parte, β_1 ("b" en la figura) representa la pendiente (inclinación) de la recta de regresión. Este coeficiente nos dice el incremento de unidades de la variable "y" que se produce por cada incremento de una unidad de la variable "x".

Por último, el 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.

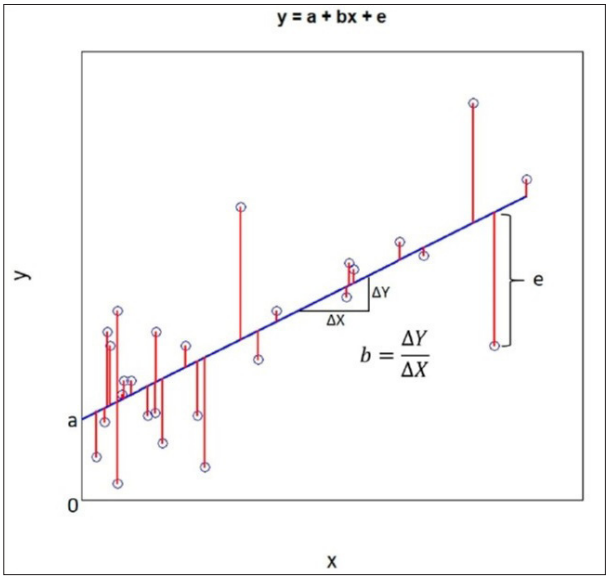
CÁLCULO DE LOS COEFICIENTES DE REGRESIÓN

El método más empleado para calcular el valor de los coeficientes de regresión es el denominado método de los mínimos cuadrados. Veamos someramente el razonamiento matemático que subyace a este método.

Hemos visto la ecuación de la recta del modelo de regresión. El problema es que, una vez que tenemos representado el diagrama de dispersión, ninguna recta se va a ajustar de manera perfecta a la nube de puntos. Sabemos que la recta pasará por el punto que marcan las coordenadas de las medias de "x" e "y", pero el problema es que pueden trazarse infinitas rectas que pasen por un punto dado. ¿Cuál será la recta que nos interesa?

Imaginemos cualquiera de estas posibles rectas de regresión. Si intentamos calcular un valor de "y" determinado (y_i) a par-

Figura 1. Representación gráfica de una recta de regresión. Significado de sus componentes



tir de un valor de "x" (x_i) habrá una diferencia entre el valor real de y_i y el que obtengamos (el valor estimado, representado como $\hat{y}_i$) con la fórmula de la recta:

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

Fijémonos en el valor de ε_i. Representa esta diferencia entre el valor real de y_i en nuestra nube de puntos y el que nos proporcionaría la ecuación de la recta. Podemos representarlo matemáticamente de la siguiente forma:

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

Este valor se conoce con el nombre de residuo y su valor depende del azar. Por tanto, nos interesará calcular los coeficientes de la recta de regresión que minimice estas diferencias.

Los residuos siguen una distribución normal con una media de cero, por lo que no podemos sumarlos directamente, ya que los positivos se cancelarían con los negativos. Por este motivo recurrimos a la suma de los cuadrados de las diferencias y buscamos la recta con el valor más bajo de esta suma de los cuadrados de los residuos. De ahí el nombre del método de los mínimos cuadrados.

Obviando la demostración matemática, a partir de este razonamiento calcularemos del valor del coeficiente, según la siguiente fórmula:

$$\beta_1 = \frac{s_{xy}}{s_x^2}$$

Donde tenemos, en el numerador, la covarianza de las dos variables y, en el denominador, la varianza de la variable independiente. A partir de aquí, el cálculo de $_0$ es sencillo, despejándolo de la recta de regresión:

$$\beta_0 + \hat{y} - b_1\bar{x}$$

VALIDACIÓN DEL MODELO DE REGRESIÓN

Una vez calculados los coeficientes, debemos calcular sus intervalos de confianza o su nivel de significación estadística, ya que solo si ambos coeficientes son estadísticamente significativos podremos aplicar el modelo a nuestra población.

Para ello, plantearemos un contraste de hipótesis para los dos coeficientes de la recta con la hipótesis nula de que su valor en la población es cero. Este contraste puede hacerse de dos formas:

- Si dividimos cada coeficiente por su error estándar, obtendremos un estadístico que sigue una distribución de la *t* de Student con n - 2 grados de libertad. Podemos calcular el valor de *p* asociado a ese valor y resolver el contraste de hipótesis rechazando la hipótesis nula si el valor de $p < 0{,}05$.
- Una forma un poco más compleja es fundamentar este contraste de hipótesis sobre un análisis de la varianza (ANOVA), considerando que la variabilidad de la variable dependiente se descompone en dos términos: uno explicado por la variable independiente y otro no asignado a ninguna fuente y que se considera no explicada (aleatoria).

En cualquier caso, no debemos preocuparnos por los detalles del contraste, ya que habitualmente utilizaremos un programa informático para realizarlo.

DIAGNÓSTICO DEL MODELO DE REGRESIÓN

Una vez comprobado que los coeficientes son significativos, tendremos que comprobar que se cumplen una serie de supuestos necesarios para que el modelo sea válido. Es lo que se conoce como diagnóstico del modelo de regresión.

Estos supuestos son cuatro: linealidad, homocedasticidad, normalidad e independencia. Una vez más, utilizaremos un programa de estadística para realizar un correcto diagnóstico del modelo regresión.

Supuesto de linealidad

La relación entre la variable de predicción (independiente) y de criterio (dependiente) debe ser lineal en el rango de valores observados de la variable de predicción. Una manera sencilla de comprobar este supuesto es la de representar un diagrama de dispersión y ver si la distribución de los puntos tiene lugar, de forma aproximada, a lo largo de una línea recta.

Un método numérico que permite comprobar el supuesto de linealidad es la prueba RESET de Ramsey. Para que el modelo sea correcto, la mediana de los residuos debe estar próxima a cero y los valores absolutos de los residuos deben distribuirse de manera uniforme entre los cuartiles (similar entre máximo y mínimo y entre primer y tercer cuartil). Si esto se cumple significará que los residuos siguen una distribución normal cuya media es cero, condición necesaria para la validez del modelo.

Supuesto de homocedasticidad

Esto significa que los residuos deben distribuirse de forma homogénea para todos los valores de la variable de predicción. Podemos comprobarlo de forma sencilla con un diagrama de dispersión que represente, en el eje de abscisas, las estimaciones de la variable dependiente para los distintos valores de la variable independiente y, en el eje de coordenadas, los residuos correspondientes. Se aceptará el supuesto de homocedasticidad si los residuos se distribuyen de forma aleatoria, en cuyo caso veremos una nube de puntos de forma similar en todo el rango de las observaciones de la variable independiente.

También existen métodos numéricos para comprobar el supuesto de homocedasticidad, como la prueba de Breusch-Pagan-Godfrey, cuya hipótesis nula supone que se cumple este supuesto.

Supuesto de normalidad

Como ya hemos mencionado, los residuos deben distribuirse de forma normal.

Una forma sencilla de comprobarlo sería representar el histograma o el gráfico de cuantiles teóricos de los residuos, en el que deberíamos ver su distribución a lo largo de la diagonal del gráfico.

También podremos aplicar un método numérico, como la prueba de Kolmogorov-Smirnov o la de Shapiro-Wilk.

Supuesto de independencia

Para comprobar este supuesto, es necesario comprobar que los residuos sean independientes entre sí y que no haya ningún tipo de correlación entre ellos.

Esto puede contrastarse realizando la prueba de Durbin-Watson, cuya hipótesis nula supone, precisamente, que los residuos son independientes.

EJEMPLO PRÁCTICO DE REGRESIÓN LINEAL SIMPLE

Veamos un ejemplo utilizando un programa de acceso libre, el software estadístico R (https://www.r-project.org/) con el *plugin* RCommander (https://www.rcommander.com/) y la base de datos IndCT_IMC.Rdata, disponible en la web de *Evidencias en Pediatría*. Si necesita saber cómo instalar RCommander, puede consultar el siguiente tutorial en línea (http://sct.uab.cat/estadistica/sites/sct.uab.cat.estadistica/files/instalacion_r_commander_0.pdf).

Esta base de datos recoge algunos datos antropométricos de una serie de 58 niños, que incluyen dos variables que se asocian a riesgo cardiovascular: el índice de cintura-talla (IndCT) y el índice de masa corporal estandarizado por edad y sexo (IMC_DS). Estamos interesados en estimar el IMC_DS (que tiene valores entre -2,2 y +2,7) a partir de la medición del IndCT (que tiene valores entre 0,36 y 0,58, por lo que nos proponemos calcular un modelo de regresión lineal simple entre las dos variables, siendo IndCT la variable independien-te o explicativa e IMC_DS la variable dependiente o criterio.

En primer lugar, representaremos los datos de las dos variables en un gráfico de dispersión para comprobar la forma de la nube de puntos y la tendencia que siguen las variables. Una vez cargada la base de datos, seleccionamos la opción Gráficas->Diagrama de dispersión y, en la ventana emergente que nos aparece, marcamos IndCT como variable "x" e IMC_DS como variable "y" (figura 2).

Si observamos el gráfico, vemos que los puntos tienden a distribuirse, aproximadamente, a lo largo de una recta en sentido ascendente hacia la derecha. Viendo la forma de la nube de puntos, parece razonable suponer que exista una relación lineal entre las dos variables.

Calculemos ahora la recta de regresión lineal. Seleccionamos la opción Estadísticos/Ajuste de modelos/Regresión lineal. En la ventana emergente, marcamos IMC_DS como variable explicada e IndCT como variable explicativa. Pulsamos aceptar y obtenemos el resumen del modelo en la ventana de resultados (figura 3).

Vemos que R nos ofrece, en primer lugar, la distribución de los residuos. Podemos ver como su mediana está próxima a cero (0,08) y como sus valores se distribuyen de forma uniforme entre los cuantiles primero y tercero. Esto es indicativo de que los residuos siguen una distribución normal.

Seguidamente, podemos ver el valor de los coeficientes de regresión, su error estándar (con el que podríamos calcular sus intervalos de confianza añadiendo o restando del coeficiente 1,96 veces el error estándar) y su significación estadística (por ambos métodos descritos previamente, el de la t de Student y el del análisis de la varianza con la F de Snedecor). Asimismo, el programa proporciona una estimación del porcentaje de variabilidad de la variable IMC_DS que explica la variable IndCT, el "Adjusted R squared" o R cuadrado ajustado de valor 0,3682; se interpreta como que el IndCT explica el 36,82% de la variabilidad del IMC_DS.

Vemos que ambos coeficientes son significativos, con lo que validamos el siguiente modelo de regresión lineal simple:

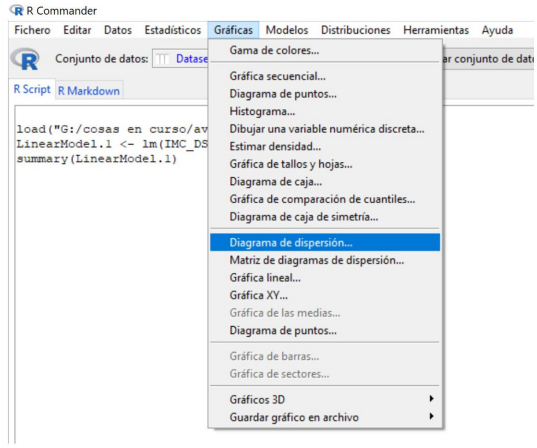
$$IMC\_DS = -7{,}4 + 16{,}2IndCT.$$

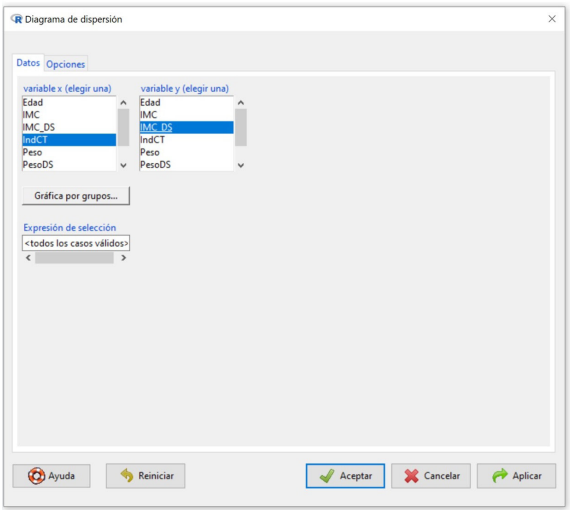
El modelo nos dice que por cada unidad de IndCT el IMC_DS aumenta 16,2 unidades. Considerando que el IndCT adopta valores entre 0,36 y 0,58, resulta más intuitivo interpretar que por cada 0,1 unidades de IndCT, el IMC_DS aumenta 1,62 unidades.

Procederemos, a continuación, con el diagnóstico del modelo, comprobando los supuestos de linealidad, homocedasticidad, normalidad e independencia.

Ya vimos con el diagrama de dispersión (figura 2) que, gráficamente, era razonable asumir que existe una relación lineal entre las dos variables. Podemos completar la valoración con un método numérico. Seleccionamos la opción del menú Modelos->Diagnósticos numéricos->Test RESET de no linealidad (figura 4). R nos da un valor RESET = 0,64, con una p = 0,52. Como p >0,05, no podemos rechazar la hipótesis nula de que el modelo es lineal, con lo que corroboramos la im-

Figura 2. Obtención del diagrama de dispersión





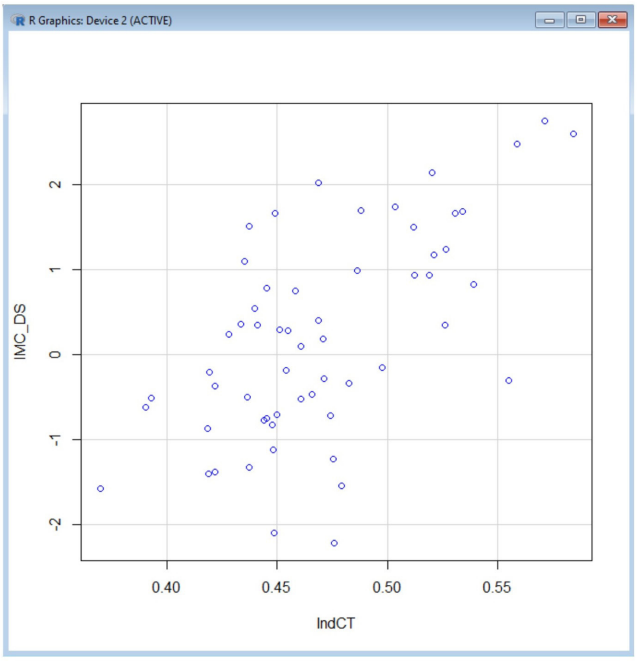
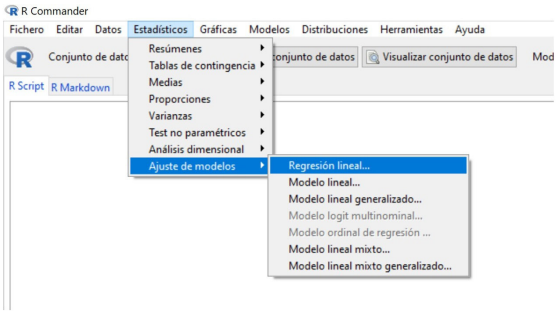
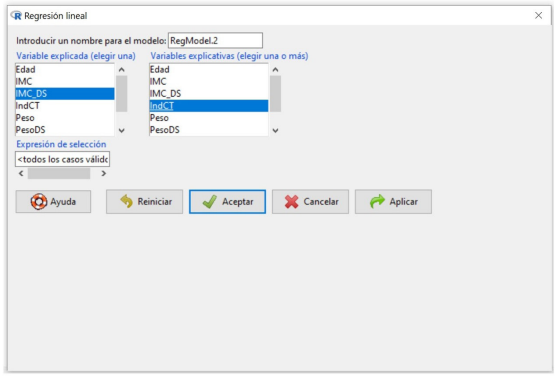


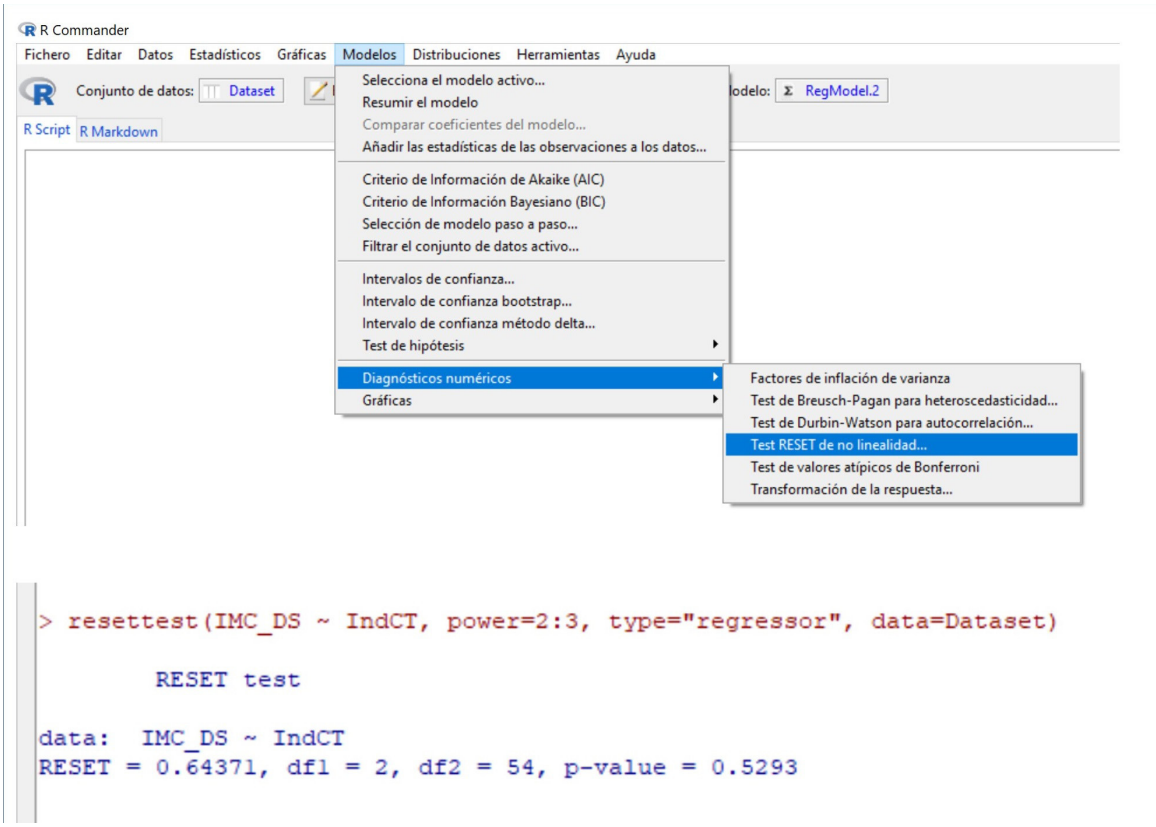
Figura 3. Cálculo del modelo de regresión lineal simple






```
> summary(RegModel.2)

Call:
lm(formula = IMC_DS ~ IndCT, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.52522 -0.60399  0.08176  0.61355  1.83523

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   -7.406      1.309  -5.659 0.000000541 ***
IndCT         16.198      2.769   5.849 0.000000267 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9632 on 56 degrees of freedom
Multiple R-squared:  0.3792, Adjusted R-squared:  0.3682
F-statistic: 34.21 on 1 and 56 DF,  p-value: 0.0000002672
```

Figura 4. Comprobación del supuesto de linealidad mediante el test RESET de no linealidad



presión que obtuvimos con el método gráfico.
Pasemos a comprobar el supuesto de homocedasticidad. Para el método gráfico recurrimos a la opción del menú Modelos/Gráficas/Gráficas básicas de diagnóstico. El programa nos proporciona 4 gráficas (figura 5), aunque, en este momento, solo nos interesa la primera de ellas, que representa los valores de la variable dependiente predichos por el modelo frente a los residuos.

Podemos ver que la distribución de los residuos es bastante homogénea, aunque quizás haya una mayor dispersión en los valores centrales de la variable dependiente. Para tratar de aclararnos, realizaremos un método numérico. Seleccionamos la opción del menú Modelos/Diagnósticos numéricos/Test de Breusch-Pagan para heterocedasticidad (figura 6). El valor del estadístico BP que nos proporciona R es de 0,10, con un valor de p = 0,74. Como p >0,05, no podemos rechazar la hipótesis nula, así que asumimos que se cumple el supuesto de homocedasticidad.

Comprobemos ahora el supuesto de normalidad. Podemos obtener los gráficos de comparación de cuantiles de las dos variables (Gráficas/Gráfica de comparación de cuantiles), que podemos ver en la figura 7. Comprobamos que los puntos se distribuyen a lo largo de la diagonal en la representación de las dos variables, por lo que asumimos que siguen una distribución normal.

Para terminar, comprobemos el supuesto de independencia. Realizamos un método numérico para comprobar el supuesto de independencia. Seleccionamos la opción Modelos/Diagnósticos numéricos/Test de Durbin-Watson para autocorrelación.

En la ventana emergente se nos pide que especifiquemos el valor de rho, que representa el coeficiente de autocorrelación de los residuos. Lo habitual es realizar un contraste bilateral y seleccionar un valor de rho distinto de cero, ya que es infrecuente conocer el sentido de la autocorrelación de los residuos, si esta existe. Lo hacemos así (figura 8) y R nos da un valor del estadístico DW = 2,37, con valor de p = 0,14.

En consecuencia, al ser p >0,05, no podemos rechazar la hipótesis nula de que los residuos son independientes, cumpliéndose así la última condición para considerar el modelo como válido.

Figura 5. Método gráfico para el estudio de la homocedasticidad

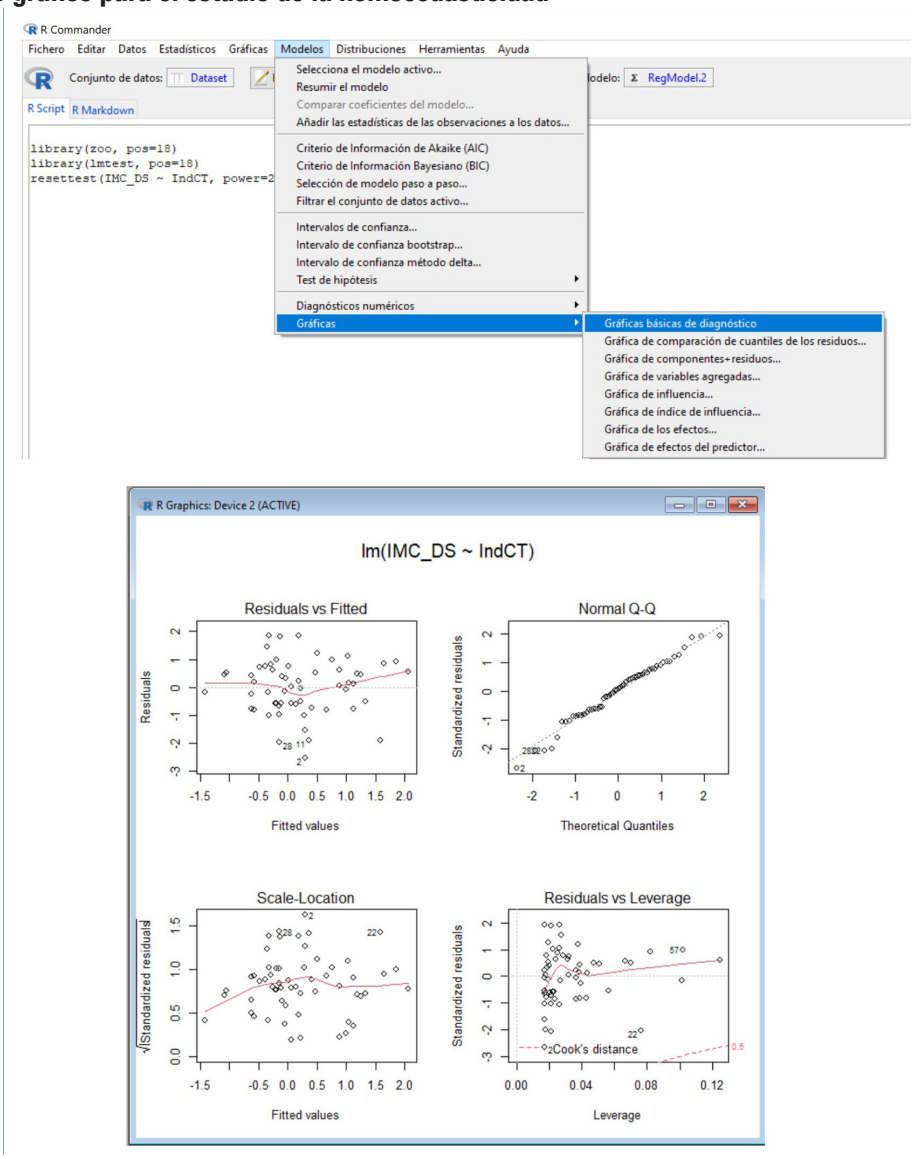


Figura 6. Comprobación del supuesto de homocedasticidad mediante el test de Breusch-Pagan para heterocedasticidad

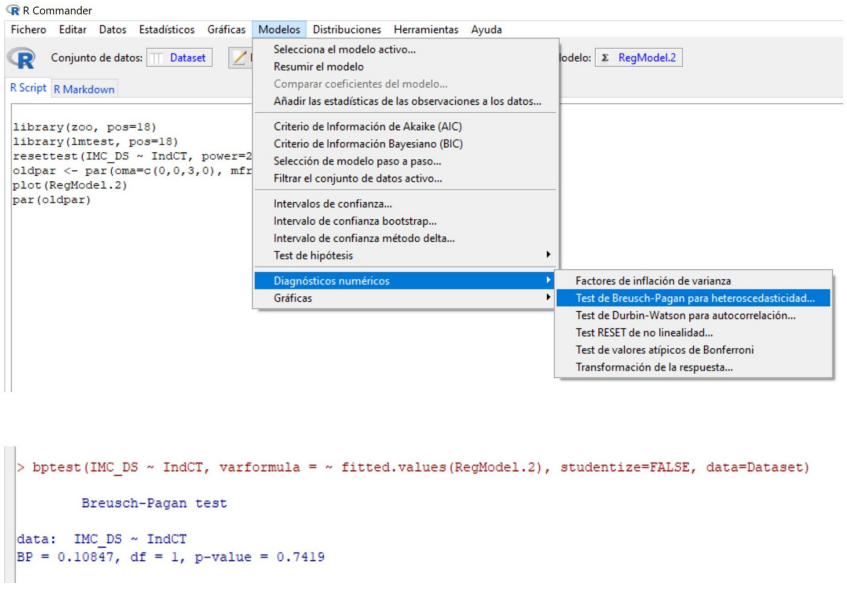


Figura 7. Gráficos de comparación de cuantiles. A: índice de masa corporal estandarizada (IMC_DS). B: Índice cintura-talla (IndCT)

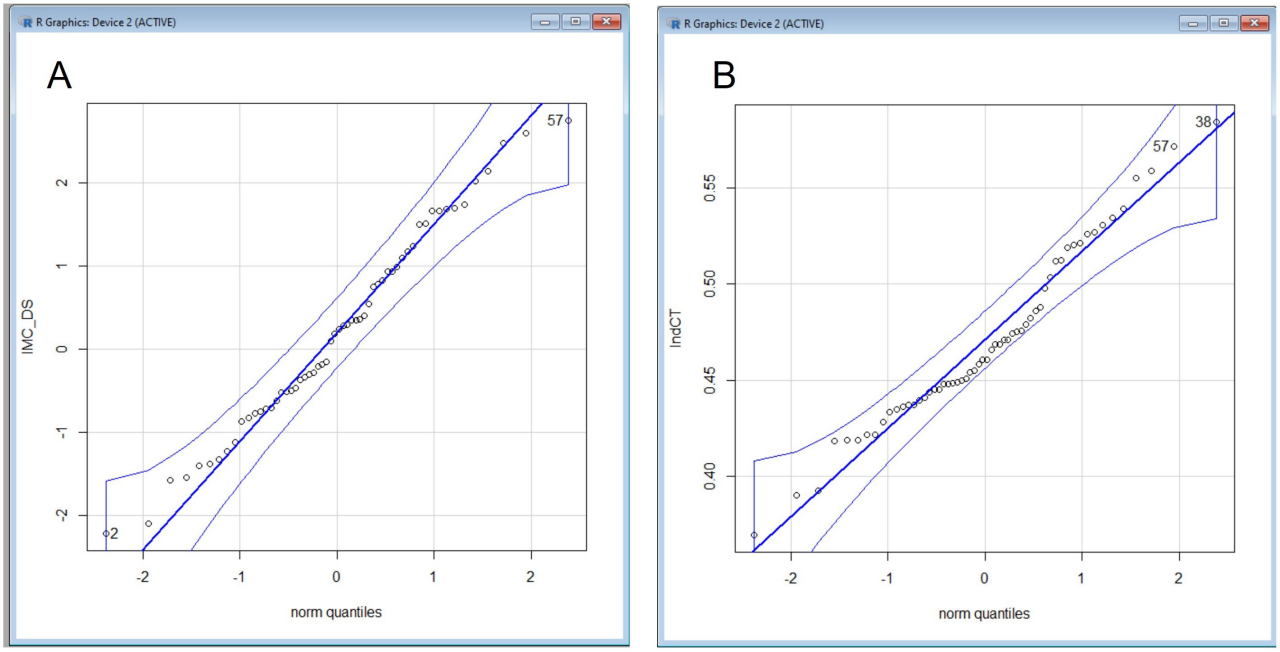
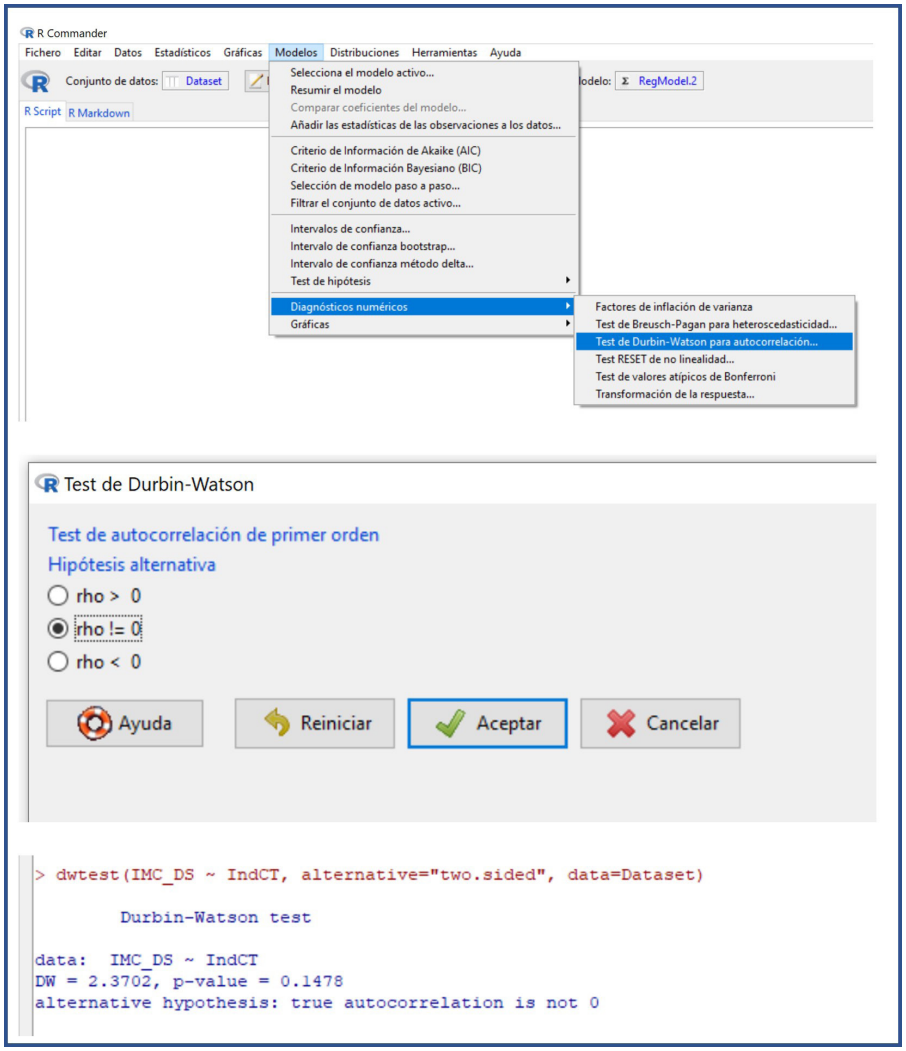


Figura 8. Comprobación del supuesto de independencia mediante el test de Durbin-Watson para autocorrelación



BIBLIOGRAFÍA

- Amat Rodrigo J. Correlación lineal y regresión lineal simple. En: Estadística con R [en línea] [consultado el 20/12/2021]. Disponible en https://github.com/JoaquinAmatRodrigo/Estadistica-con-R/blob/master/PDF_format/24_Correlaci%C3%B3n_y_Regresi%C3%B3n_lineal.pdf
- Molina Arias M, Ochoa Sangrador C, Ortega Páez E. Correlación. Modelos de regresión. Evid Pediatr. 2021;17:25.
- 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; 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.