Ecología Austral 30:344-353 Diciembre 2020
Asociación Argentina de Ecología

Variables explicativas que no pueden controlarse ni fijarse: ¿Funciona la regresión?

Teresa Boca[1,2,✉]; Adriana Pérez[1,3] & Susana Perelman[1,4]

[1] Facultad de Agronomía, Universidad de Buenos Aires. [2] Instituto Nacional de Tecnología Agropecuaria. [3] Grupo de Bioestadística Aplicada, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. [4] IFEVA, FAUBA-CONICET.

Resumen. El análisis de regresión lineal es una de las técnicas estadísticas que más se usan en los experimentos planificados para estudiar el funcionamiento de los sistemas naturales, en especial en estudios mensurativos. Muchas veces, el investigador no tiene capacidad de controlar la porción explicativa del modelo de regresión, por lo que las variables explicativas pueden resultar tan aleatorias o más que la variable respuesta. Esto podría generar sesgos en las estimaciones de las pendientes asociadas y conducir a conclusiones equivocadas. Una alternativa al método de regresión clásico es la regresión tipo II, diseñada para cuando no se pueden fijar los valores de la variable explicativa. En este trabajo se presentan distintas situaciones basadas en investigaciones publicadas en ecología y agronomía con diferentes objetivos: predicción, estimación de la pendiente y comparación de pendientes entre dos grupos, en las que el problema de variación aleatoria en las variables explicativas está presente con distinto grado de relevancia. En cada caso se identifica cuál es el camino más adecuado para el análisis. También se realizó una simulación que consideró distintas combinaciones para los errores aleatorios en las variables regresora y respuesta con el objetivo de visualizar el sesgo de los estimadores en cada situación para los diferentes métodos de regresión. De lo presentado se desprende que se debe poner énfasis en dos cuestiones muy importantes para poder decidir el método de regresión tipo II más adecuado: tener en claro cuál es el objetivo del trabajo y si se cumplen las condiciones de aplicación requeridos por cada método. Esta revisión pretende ser una sencilla guía de cuándo y qué método aplicar en cada situación.

[Palabras clave: variables independientes con error, simulaciones, sesgo].

Abstract. Explanatory variables that cannot be controlled or fixed: Does the regression work? Linear regression analysis is one of the most used statistical techniques in experiments planned to study the functioning of natural systems, especially in measurable studies. Many times, the researcher does not have the ability to control the explanatory portion of the regression model, so the explanatory variables can be as random or more than the response variable. This could generate biases in the estimates of the associated slopes and lead to wrong conclusions. An alternative to the classical regression method is type II regression when the values of the explanatory variable cannot be controlled. This paper presents different situations based on published research in ecology and agronomy for different purposes: prediction, estimate of the slope and comparison of slopes between two groups, in which the problem of random variation in the explanatory variables is present with different degrees of relevance. In each case, the most appropriate path for the analysis will be identified. A simulation was also carried out that considered different combinations for the random errors in the regressor and response variables in order to visualize the bias of the estimators in each situation for the different regression methods. It is clear from the foregoing that it is necessary to emphasize two very important issues in order to decide the most appropriate type II regression method: be clear about the objective of the work and if the application conditions required by each method are met. This review aims to be a simple guide to when and what method to apply in each situation.

[Keywords: independent variables, explanations with error, simulations, bias]

Introducción

El análisis de regresión lineal es una de las técnicas estadísticas más utilizadas para cuantificar relaciones entre variables en las investigaciones agronómicas, biológicas y ecológicas, entre otras. Con frecuencia, en la ejecución de experimentos planificados para contestar preguntas acerca del funcionamiento de los sistemas naturales —pero sobre todo cuando se trata de estudios mensurativos (*sensu* Hurlbert 1984)—, el investigador tiene poca o ninguna capacidad de controlar la porción explicativa del modelo de regresión. En consecuencia, las variables candidatas a predecir la respuesta pueden resultar tan aleatorias como la respuesta misma.

Si los valores de una variable surgen a partir de un muestreo aleatorio —por ejemplo, de la selección aleatoria de un conjunto de arbustos a los cuales se les mide la cobertura basal—, dicha variable es aleatoria y susceptible de poder modelarse mediante alguna distribución

Editor asociado: Diego Vazquez
✉ rtboca@agro.uba.ar
Recibido: 8 de Noviebre de 2019
Aceptado: 12 de Mayo de 2020

de probabilidad. Si los valores están fijados por el investigador, como podría ser el caso de la dosis de un fertilizante aplicado a los mismos arbustos, dicha variable no es considerada aleatoria sino fija y no es susceptible de ser modelada. Ambos tipos de variables pueden participar de un modelo estadístico como variables explicativas (también denominadas regresoras) para predecir, por ejemplo, la biomasa de los arbustos. Más allá del tipo de relación funcional que se proponga entre la variable respuesta (biomasa del arbusto en el ejemplo) y las explicativas, los modelos serán intrínsecamente diferentes, ya que en el primer caso la variable explicativa posee un error aleatorio —entendido como variaciones aleatorias entre cada observación y la esperanza de dicha variable—, mientras que el segundo, no.

Desconocer la presencia de errores aleatorios en las variables explicativas podría generar sesgo en las estimaciones de los coeficientes del modelo y, eventualmente, conducir a conclusiones equivocadas sobre el funcionamiento del sistema bajo estudio. Sin embargo, al momento de elegir el método de regresión para el análisis de la información, pocas veces se considera la violación del supuesto referido a variables explicativas fijas y medidas sin error. En los últimos años se observa un renovado ímpetu entre los revisores de las revistas de ecología y de agronomía por reclamar la aplicación de las técnicas de regresión especialmente diseñadas para este tipo de problemas, conocidas en conjunto como regresión tipo II (Ludbrook 2012). Algunas veces, se debería responder al reclamo con una negativa bien fundamentada en conceptos que surgen de la literatura especializada (Sokal and Rohlf 1995; Draper and Smith 2004; Legendre and Legendre 2012; Taskinena and Wartona 2013); otras veces, por lo contrario, con fundamento en la misma literatura, se deberían reanalizar los datos atendiendo al reclamo.

En este trabajo nos proponemos discutir algunos ejemplos de la investigación publicada en ecología y agronomía en los que el problema de variación aleatoria en las variables explicativas está presente y adquiere diferente grado de relevancia. Se identificarán las claves principales que marcan esas diferencias y definen el camino más adecuado para el análisis. Finalmente, se presentarán las técnicas de análisis utilizadas para el abordaje de este problema, algunos criterios de selección entre estas alternativas y una breve guía hacia paquetes desarrollados en R (R Core Team 2019) para ejecutar estos análisis: *smatr* (Warton et al. 2012) y *lmodel2* (Legendre 2015).

El modelo de regresión lineal

El modelo de regresión lineal simple o múltiple (con una o varias variables explicativas, respectivamente) más difundido en los cursos universitarios y en los textos de estadística, se fundamenta en los supuestos de linealidad, homogeneidad de varianzas, distribución normal de los errores aleatorios asociados a la variable respuesta e independencia entre las observaciones, además del supuesto de variables explicativas medidas sin error, en el que enfocamos nuestra atención en este trabajo. El cumplimiento de este conjunto de supuestos, derivado tanto del proceso de obtención de los datos como de las características del fenómeno bajo estudio, le confiere propiedades muy beneficiosas a los estimadores de los coeficientes de la regresión, tanto que éstos se denominan *BLUEs: best linear unbiased estimators*, es decir, los mejores estimadores lineales insesgados de mínima varianza; además, resultan perfectamente coincidentes, aunque sean derivados mediante diferentes métodos de estimación, el de Mínimos Cuadrados Ordinarios (del inglés, *ordinary least squares* u OLS), el de máxima verosimilitud y el de los momentos. En particular, aplicando el método OLS se obtienen estimadores que minimizan la suma de cuadrados de las distancias verticales entre los valores observados de la respuesta y los predichos por el modelo (residuales). Independientemente del método de estimación, a los modelos de regresión que cumplen con el supuesto de variables explicativas fijas se los conoce como de tipo I. En cambio, para situaciones en las que no se cumple el supuesto de valores de las variables explicativas fijados por el investigador se recomienda aplicar el modelo de regresión de tipo II.

El universo al que se aplican los modelos de regresión se fue ampliando con el desarrollo de métodos de análisis apropiados a situaciones en las que algunos supuestos no se cumplen. Por ejemplo, para el caso de heterogeneidad de varianzas, se pueden obtener los estimadores mediante el método de mínimos cuadrados ponderados, o ante la falta de normalidad se pueden utilizar distintas formas de regresión

robusta, no paramétrica o modelos lineales generalizados (Faraway 2016).

Las pendientes estimadas por el método OLS son insesgadas cuando el investigador fija los valores de la variable regresora. En cambio, si la variable regresora es aleatoria, el error asociado a la variación no controlada en las x produce un sesgo en la estimación de la pendiente.

La ecuación para una regresión OLS de y en x es:

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

donde ε_i representa toda la variación aleatoria en y no explicada por la variable explicativa x. Toda la variación no explicada por la recta $y_{i=} b_{0+} b_1 X_i$, está representada por e, el error en el ajuste x o residual. El estimador OLS del coeficiente β_1 es:

$$b_{1\,OLS} = \frac{S_{xy}}{S_x^2}$$

donde S_{xy} es el estimador de la covarianza entre x e y, σ_{xy} , y S^2_x es el estimador de la varianza poblacional de x, σ^2_x

En el caso en que exista variación aleatoria en la variable explicativa, se suma una componente de variación aleatoria δ_i a cada x_i; entonces, el estimador del coeficiente β_1 sería:

$$b_1 = \frac{S_{xy}}{S_x^2 + S_\delta^2} \qquad (1)$$

La S^2_δ modifica la estimación del coeficiente de regresión, ya que el denominador resulta mayor que si el investigador hubiera fijado los x_i; por lo tanto, se subestima el valor absoluto de la pendiente. Este sesgo hacia cero en el coeficiente se conoce como atenuación de la regresión y la magnitud del efecto depende de la magnitud de S^2_δ con respecto a S^2_x (Smith 2009).

Métodos de regresión tipo II

Los métodos y estimadores alternativos a OLS propuestos en la bibliografía (Gillard 2006; Legendre 1998; Warton et al. 2006) son: *Major Axis* (MA) o eje principal, *Standarized Major Axis* (SMA) o eje principal estandarizado y *Ranged Major Axis* (RMA) o eje principal relativizado a la amplitud, que difieren entre sí en la dirección en la que se miden los errores a minimizar (Figura 1) y, en consecuencia, resultan más adecuados para diferentes condiciones de aplicación.

En el método del eje principal (MA) los parámetros estimados minimizan la suma de cuadrados de las distancias ortogonales o perpendiculares de los datos a la recta de ajuste (Figura 1b), siendo el estimador de la pendiente:

$$b_{MA} = \frac{s_y^2 - s_x^2 + \sqrt{\left(s_y^2 - s_x^2\right)^2 + 4s_{xy}^2}}{2s_{xy}}$$

(Sokal and Rohlf 1995)

donde S^2_y y S^2_x son los estimadores de las varianzas de x e y respectivamente y S^2_{xy} es el estimador de la covarianza entre e x e y.

Este método es también conocido como regresión por análisis de componentes principales, ya que la línea de regresión estimada es la primera componente principal de la dispersión de puntos. La pendiente de la regresión aplicando MA también se puede calcular utilizando estimaciones de la pendiente de la recta de regresión OLS, y del coeficiente de correlación r_{xy}:

$$b_{MA} = \frac{d + \sqrt{d+4}}{2}, \text{ donde } d = \frac{{b_{OLS}}^2 - r_{xy}^2}{r_{xy}^2 \times b_{OLS}}$$

Este método es válido cuando las dos varianzas de los errores (en y y en x) son iguales, de modo que su cociente, denominado lamda, es S^2_y / S^2_x= 1. Esta suposición requiere que ambas variables posean las mismas escalas. El método MA se aplica entonces cuando ambas variables se expresan en las mismas escalas y cuando se puede suponer razonablemente que las varianzas de los errores de ambas variables son aproximadamente iguales.

La significancia de b_{MA} se puede poner a prueba por el método de permutaciones (Legendre and Legendre 2012) o, de forma alternativa, estimando el intervalo de confianza para verificar si incluye el valor cero u otro valor de interés (Sokal and Rohlf 1995; Jolicoeur 1990). En el caso de que se desee comparar pendientes, la regresión MA se puede utilizar, aunque los pares de variables estudiadas difieran entre sí en las unidades de medida.

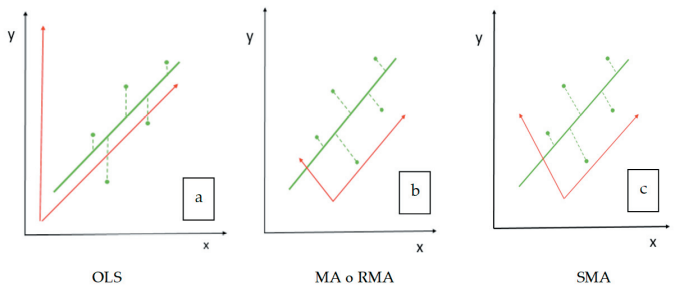

Figura 1. a) OLS; b) MA; c) SMA (Warton et al. 2012). OLS minimiza la suma de cuadrados de las distancias sólo en el sentido de la variabilidad en *y*, MA minimiza la suma de cuadrados de las distancias en ambas direcciones y con igual peso (siendo la dirección más corta desde las observaciones en sentido perpendicular a la línea), y SMA minimiza las distancias tomando en cuenta distinto peso en cada dirección. Las líneas punteadas indican residuales, las flechas representan los ejes donde fueron ajustados los residuales.

Figure 1. a) OLS; b) MA; c) SMA (Warton et al. 2012). OLS minimizes the sum of squares of the distances only in the sense of the variability in *y*, MA minimizes the sum of squares of the distances in both directions and with equal weight (being the shortest direction from the observations perpendicular to the line), and SMA minimizes the distances taking into account different weight in each direction. Dotted lines indicate residuals, the arrows represent the axes where the residuals were adjusted.

El método del eje principal estandarizado (SMA) consiste en aplicar el método MA sobre las variables *x* e *y* estandarizadas. La pendiente resultante valdrá siempre 1 o -1, según el signo del coeficiente de correlación. Luego se vuelve a las unidades originales multiplicando por (S_y/S_x). Por lo tanto, la pendiente es computada como:

$$b_{SMA} = signo\,(r_{xy})\,\frac{s_y}{s_x}$$

Este método minimiza el producto de las distancias verticales y horizontales a la recta, por lo que también es conocido como el método de los mínimos rectángulos (Francq 2014). Es útil cuando las dos variables no son medidas en la misma escala, en cuyo caso no parecería razonable dar el mismo peso a las distancias a la recta en *x* e *y*.

El estimador de la pendiente b_{SMA} se relaciona con el estimador b_{OLS} a través del coeficiente de correlación (Legendre and Legendre 2012) según:

$$b_{SMA} = \frac{b_{OLS}}{r_{xy}} \quad \text{para } r_{xy} \neq 0 \qquad (2)$$

Esta ecuación muestra que cuando las variables están altamente correlacionadas y por lo tanto r_{xy} tiende a 1, b_{SMA} tiende a b_{OLS}. A medida que la correlación entre ambas variables se hace más débil (es decir que el módulo de tiende a cero), b_{SMA} tiende, en valor absoluto, a ser mayor que b_{OLS}. Si r_{xy} no difiere significativamente de cero, la estimación de b_{SMA} obtenida de la ecuación (2) no es válida y carece de sentido aplicar el método SMA. Por dicha razón, se recomienda poner a prueba la significancia de la correlación entre ambas variables previo a aplicar este método. Según Legendre (1998) no es posible aplicar la prueba de permutaciones o el intervalo de confianza utilizado para MA, aunque algunos autores reportan aproximaciones (Warton et al. 2012; Jolicoeur 1990).

El método RMA constituye otra alternativa para variables con distintas unidades de medida donde los parámetros sí pueden ser sometidos a pruebas de hipótesis (Legendre 1998). Consiste en re-escalar a las observaciones de *x* e *y* de manera de llevarlas al rango 0 -1, sin unidades, mediante la siguiente ecuación:

$$y_i' = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Sobre las variables re-escaladas x' e y' se aplica la regresión MA y luego se vuelve la pendiente estimada a las unidades originales multiplicándola por la razón de los rangos, siendo entonces...

$$b_{SMA} = b'_{SMA}\,\frac{y_{max} - y_{min}}{x_{max} - x_{min}}$$

Como se mencionó antes, en el método RMA es posible someter a prueba la hipótesis de pendiente igual a un valor establecido, en particular 0 ó 1. Al igual que en MA, esto puede hacerse ya sea por el método de permutaciones o comparando el intervalo de confianza de la pendiente con el hipotético valor de interés. RMA es sensible a observaciones extremas (*outliers*) que tienen mucho impacto sobre

los datos relativizados, por lo que no se recomienda utilizarlo sin revisar previamente la presencia de valores extremos.

A diferencia de OLS, en los tres métodos descriptos (MA, SMA, RMA), la pendiente de la regresión de y vs x es la inversa de la pendiente de x vs y. Esta propiedad de simetría es deseable ya que relaja la discusión acerca de cuál de las dos variables se asocia al eje x y cuál al y. Asimismo, en los tres métodos la recta estimada pasa por el centroide $(\bar{y}, \bar{x})$ y por lo tanto $b_0 = \bar{y} - b_1\bar{x}$

En consecuencia, no existe una recomendación única, la decisión en cada circunstancia dependerá de las unidades de medidas de ambas variables, de las magnitudes relativas de los errores, y, como veremos más adelante, del objetivo central del análisis.

Paquetes de R para el ajuste de regresiones tipo II

Los dos paquetes más importantes de R para aplicar el método de regresión tipo II son *lmodel2* (Legendre 2015) y *smatr* (Warton et al. 2012). Estos paquetes estiman la regresión lineal simple del modelo II utilizando OLS, MA, SMA y RMA. El paquete *smatr* incluye también pruebas para la comparación entre dos pendientes con ordenada al origen común o no.

Presentación de ejemplos con diversidad de objetivos

Caso A) Objetivo: Predicción. Relación altura del árbol y diámetro a la altura del pecho. En este caso, el objetivo es obtener un modelo para predecir o pronosticar valores medios de la variable respuesta y: altura del árbol, medida en metros, para valores dados de la variable explicativa x: diámetro a la altura del pecho, medido en cm. Se analizaron datos obtenidos de mediciones anuales tomadas sobre 70 árboles durante un período de 15 años, provenientes de un ensayo de *Pinus taeda* instalado en la Región Mesopotámica de la Argentina por el INTA, EEA Montecarlo (Fassola et al. 2002). La variable explicativa diámetro a la altura del pecho no toma valores fijados de antemano por el investigador, sino que es una variable aleatoria y, por lo tanto, los paráme tros de la regresión podrían estar sesgados si se aplica OLS. Sin embargo, Legendre (1998) recomienda que cuando el propósito del estudio no es estimar los parámetros de la relación funcional, sino simplemente predecir valores de y dado x, se debe usar OLS, ya que es el único método que minimiza los residuales al cuadrado en y. Se obtuvieron los estimadores de los parámetros del modelo con los distintos métodos descriptos y se compararon mediante el error cuadrático medio de predicción (ECMP).

Los estimadores de la pendiente obtenidos con los distintos métodos de regresión y sus ECMP fueron: b_{OLS} = 0.52-ECMP$_{OLS}$ = 24.35%; b_{MA}= 0.61-ECMP$_{MA}$ = 23.35%; b_{SMA}= 0.68-ECMP$_{SMA}$ = 27.65% y b_{RMA} = 0.70-ECMP$_{MA}$= 28.80%. El valor menor de ECMP se obtuvo con el método OLS y el mayor con el de RMA, es decir que —en coincidencia con la recomendación de Legendre (1998)— OLS resultó el método para situaciones en las que el objetivo es la predicción (ver comandos R en el Anexo).

Caso B) Objetivo: Estimación de los coeficientes. Relación entre el rendimiento de trigo pan y el de trigo duro. En este caso, el objetivo es resumir la relación entre dos variables: rendimiento de trigo pan *vs.* rendimiento de trigo duro. El trabajo fue realizado por Marti y Slafer (2014), quienes se replantearon el hecho aceptado comúnmente que asume que el trigo pan es menos tolerante al estrés que el trigo duro. Para aportar evidencia al respecto, compararon los rendimientos de trigo pan y trigo duro obtenidos en diversos ensayos en un rango amplio de ambientes en una región agrícola donde se sembraron ambas variedades (Figura 2).

A diferencia del caso A, aquí no se pretende predecir la variable y a partir de la variable x, sino, por el contrario, se trata de un problema simétrico: las dos variables se podrían intercambiar ya que no cumplen roles de explicativa y de respuesta. Claramente aquí sólo se busca estimar la pendiente para comparar los incrementos relativos entre las dos variedades. El coeficiente de correlación, por el contrario, no captaría eso, sino la variabilidad conjunta. Podríamos, por ejemplo, tener distintos valores de pendiente, que reflejarían eficiencias relativas muy distintas, para un mismo valor de coeficiente de correlación. En este caso en particular se trató al rendimiento del trigo duro como la variable x y al rendimiento del trigo pan como la variable y, ambas variables están medidas en las mismas unidades (rendimiento, en Mg/ha) y abarcan igual rango de valores. Si fuese cierta la hipótesis de menor tolerancia de los trigos blandos, se esperaría que la pendiente, que relaciona el rendimiento del trigo pan con el del duro, sea menor a 1.

Claramente, utilizar un método que subestima la pendiente, como OLS, podría llevar a conclusiones incorrectas. Dado que ambas variables están en las mismas unidades, los métodos recomendados son MA y SMA. La elección de MA a favor de SMA dependerá de que se pueda sostener el supuesto mencionado anteriormente referido a varianzas iguales para ambas variables. A partir de la prueba F se rechazó la hipótesis nula de igualdad de varianzas (valor P=0.02), por lo que se recomienda SMA. A fines comparativos se muestran los estimadores de pendiente y los intervalos de confianza al 95% (IC95%) resultantes para cada método: b_{OLS}= 0.853-$IC95\%_{OLS}$= (0.830; 0.876); b_{MA}= 0.911-$IC95\%_{MA}$ = (0.887; 0.936); b_{SMA}= 0.917-$IC95\%_{SMA}$ = (0.894; 0.941) y b_{RMA}= 0.915, $IC95\%_{RMA}$= (0.891; 0.941). En la Figura 2 se muestran las rectas ajustadas por OLS y SMA.

Todos los estimadores de la pendiente de la regresión y sus intervalos de confianza arrojaron valores menores que 1, con amplitud de intervalo muy semejante. Aún cuando los resultados son similares en este caso particular, los métodos de regresión más adecuados son los de tipo II, dado que no se cumplen los supuestos de regresión tipo I.

Otro contexto de aplicación es aquel en el cual la teoría o investigaciones previas predicen cierto valor para los coeficientes y por lo tanto interesa que su estimación no esté sesgada. Por ejemplo, en poblaciones vegetales compuestas por individuos de una misma cohorte y, como consecuencia del autorraleo, se ha observado para un rango relativamente amplio de densidades que la relación entre el tamaño individual y la densidad es negativa, y se ha hipotetizado que la pendiente tiene un valor de -3/2 cuando ambas variables son expresadas en escala logarítmica (Yoda et al. 1963; Harper 1977). La utilización de OLS puede llevar a concluir que no se cumple dicha relación. En algunas disciplinas, como la alometría, donde el valor de la pendiente es clave para identificar el tipo de crecimiento, la regresión tipo II es el estándar de análisis (Warton et al. 2006).

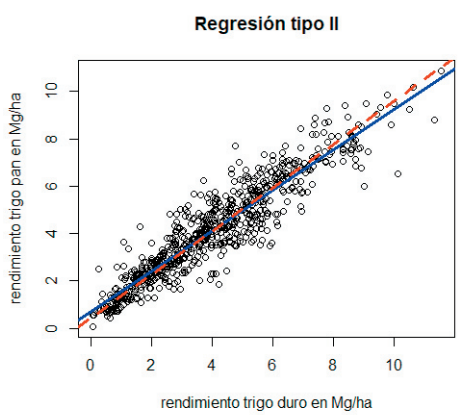

Figura 2. Gráfico de dispersión y recta de regresión del rendimiento de trigo pan (Mg/ha) y el rendimiento de trigo duro (Mg/ha). Se muestran las rectas de regresión usando OLS (línea llena) y aplicando el método de regresión tipo II SMA (línea punteada). Promedio y DE del rendimiento de trigo pan: 4.21 y 2.10 Mg/ha; promedio y DE del rendimiento de trigo duro: 4.11 y 2.29 Mg/ha.

Figure 2. Scatter plot and regression line of wheat bread yield (Mg/ha) and hard wheat yield (Mg/ha). The regression lines are shown using OLS (full line) and applying the type II regression method SMA (dotted line). Mean and SD of wheat bread yield: 4.21 and 2.10 Mg/ha; Mean and SD of hard wheat yield: 4.11 and 2.29 Mg/ha.

Finalmente, otra situación de aplicación es aquella en la cual se desea comparar la concordancia entre dos métodos de medición (Francq and Govaerts 2014). Por ejemplo, supongamos que existe un método de referencia para la medición del contenido de materia orgánica en muestras de agua y se lo desea reemplazar por otro más preciso, más rápido, menos contaminante, más económico o más fácil de medir. Para estudiar la equivalencia entre ambos métodos, con frecuencia se usa la regresión lineal entre ambas mediciones. En esos casos, una ordenada al origen significativamente distinta de cero sugiere un sesgo sistemático entre ambos métodos, mientras que una pendiente significativamente distinta de uno indica un sesgo proporcional (Bland and Altman 1999). Como ya se discutió, usar OLS proveerá estimaciones sesgadas de los coeficientes (pendientes menores y ordenadas al origen mayores, en valor absoluto), por lo que se aconseja utilizar alguno de los métodos aquí expuestos, tal como se observa en algunos estudios de calibración para decisiones racionales de fertilización en cultivos (Correndo et al. 2017) (ver comandos R en el Anexo).

Caso C) Objetivo: Comparar pendientes entre dos grupos. Relación entre el número total de granos por planta y la tasa de crecimiento en período crítico para dos híbridos. Es una técnica que se usa para poner a prueba la igualdad de pendientes o intersecciones en regresiones es el análisis de regresión con variables categóricas. Sin embargo, este método sólo se puede aplicar correctamente para el modelo I, OLS, y no es apropiado cuando los valores de x están sujetos a error (Ludbrook 2012).

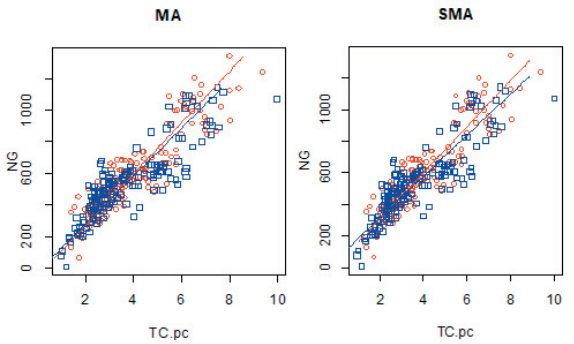

Figura 3. Gráfico de dispersión y recta de regresión del número total de granos por planta (NG) y la tasa de crecimiento en período crítico (TC.pc g/día) para dos híbridos, correspondiente a los datos del ejemplo de Ciano et al (2016). DK7210 (cuadrado) y DK747 (círculo). A la izquierda, los modelos estimados sin estandarización (MA); a la derecha, con estandarización (SMA). Promedio y DE de NG: 563 y 239 granos por planta; promedio y DE de TC.pc: 3.8 y 1.7 g/día.

Figure 3. Scatter plot and regression line of the total number of grains per plant (NG) and the critical period growth rate (TC.pc g/day) for two hybrids. DK7210 (square) and DK747 (circle). On the left, the estimated models without standardization (MA); on the right, with standardization (SMA). Mean and SD of NG: 563 and 239 grains per plant; mean and SD of TC.pc: 3.8 and 1.7 g/day.

El caso de estudio propuesto para este objetivo es el de Ciano et al (2016), que evalúa, como parte de un estudio más amplio, el impacto de la variación de la tasa de crecimiento en período crítico (TCpc, medida en g/día) sobre el número total de granos por planta (NG) en dos híbridos de maíz (DK7210 y DK747). Interesa conocer si esta relación difiere entre ambos híbridos, lo que implicaría distintas pendientes en la relación NG *vs.* TCpc de los híbridos. Los híbridos se cultivaron en distintas fechas de siembra y densidades para generar variabilidad en la TCpc. Las variables involucradas presentan diferentes magnitudes (Figura 3); esto nos indicaría que es necesario estandarizar las variables o sea que en este caso el método SMA es más apropiado que el método MA.

Los modelos estimados para ambos híbridos mediante SMA presentaron mayores diferencias entre sí que los estimados por MA (Figura 3). El análisis de comparación de pendientes por el método MA no muestra evidencias para el rechazo de la hipótesis nula de pendientes iguales (valor *P*=0.247), en cambio por el método SMA, que es el recomendado para este conjunto de datos, se concluye que sí existen evidencias para rechazar la hipótesis de igualdad de pendientes entre ambos híbridos (valor *P*=0.027) (ver comandos R en el Anexo).

Teoría de los modelos de regresión tipo II

Esta sección describe los fundamentos teóricos de los métodos propuestos para resolver problemas de regresión Tipo II. A su vez, también permite comprender el procedimiento de las simulaciones que se presentan en la siguiente sección.

Consideremos, según Faraday (2009) y Gillard (2006), que lo que el investigador registra al efectuar un estudio son las variables observadas x_i^o, y_i^o para i=1,…n, que están relacionadas a los verdaderos valores de las variables explicativa y respuesta x^A_i e y^A_i, respectivamente, según:

$$x_i^o = x_i^A + \delta_i$$
$$y_i^o = y_i^A + \varepsilon_i \qquad (3)$$

donde δ_i y ε_i son los componentes de error aleatorio o ruido de las variables x e y, respectivamente, y se asumen independientes entre sí. La verdadera relación subyacente entre las variables no observables y^A_i, x^A_i es:

$$y_i^A = \beta_0 + \beta_1 x_i^A \qquad (4)$$

Combinando (3) y (4) resulta:

$$y_i^o = \beta_0 + \beta_1 x_i^o + (\varepsilon_i - \beta_1 \delta_i) \qquad (5)$$

Es razonable asumir que los errores δ_i y ε_i son insesgados y por lo tanto su esperanza es $E[\delta_i]=E[\varepsilon_i]=0$. Además siendo la $var[\delta_i]=\sigma^2_\delta$, $var[\varepsilon_i]=\sigma^2_\varepsilon$ y asumiendo $Cov[x_i, \delta_i]=0$, es posible demostrar que:

$$E(\hat{\beta}_1) = \beta_1 \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\delta^2} \qquad (6)$$

En la ecuación (6) puede verse claramente la diferencia respecto al modelo de regresión estándar con x no aleatoria. Cuando $\sigma^2_\delta = 0$, $E(\beta_1)=\beta_1$ y el modelo de la ecuación (6) será equivalente al de la regresión tipo I.

Si $\sigma^2_\delta > 0$, la pendiente estará sesgada hacia cero, independientemente del tamaño muestral (comparar esta expresión con la ecuación [1]). La magnitud de este sesgo dependerá de la relación entre σ^2_δ y σ^2_x. Si σ^2_δ

es pequeña en relación con σ^2_x el modelo de regresión tipo I constituye una buena aproximación. En otras palabras, si la variabilidad en los errores de observación de *x* es pequeña en relación con el rango de x en el modelo, entonces el problema puede ser ignorado.

En cualquier otro caso, el estimador de β_1 será según la ecuación (7) (Gillard 2006):

$$b1 = \frac{s_y^2 - \lambda\, s_x^2 + \sqrt{\left(s_y^2 - \lambda\, s_x^2\right)^2 + 4}}{2s_{xy}} \quad (7)$$

Donde S^2_y y S^2_x son los estimadores de las varianzas de *x* e *y* respectivamente, S^2_{xy} es su covarianza y λ es el cociente $\sigma^2_e / \sigma^2_\delta$ de las varianzas de los términos aleatorios.

Simulación

Para visualizar el sesgo de los estimadores obtenidos mediante regresión tipo II bajo distintas condiciones se realizó una simulación siguiendo el modelo descripto en la ecuación (5) y aplicando el estimador de la ecuación (7). Se definió un escenario que considerara distintas combinaciones para los errores de medición en las variables regresora y respuesta. Los valores considerados fueron valores de σ_δ =0, 5, 7, 10; σ^2_e =5, 7, 10; cov(δ_i, ε_j)=0, mientras que β_0 se fijó en 10 y β_1 en 1, siendo *x* comprendida entre 0 y 200. Se realizaron 10000 simulaciones de pares de valores (x_i, y_i). La visualización de los valores resultantes de la simulación (Figura 4) muestra cómo los estimadores obtenidos bajo el método OLS solo son insesgados cuando la variable regresora se observa sin error, es decir σ^2_δ =0. A medida que x presenta errores crecientes, el sesgo en la estimación de la pendiente se incrementa, haciéndose más cercana a cero (atenuación). En cambio, si se aplica regresión tipo II, las estimaciones de la pendiente son insesgadas, independientemente de la magnitud de σ^2_e (ver comandos R en el Anexo).

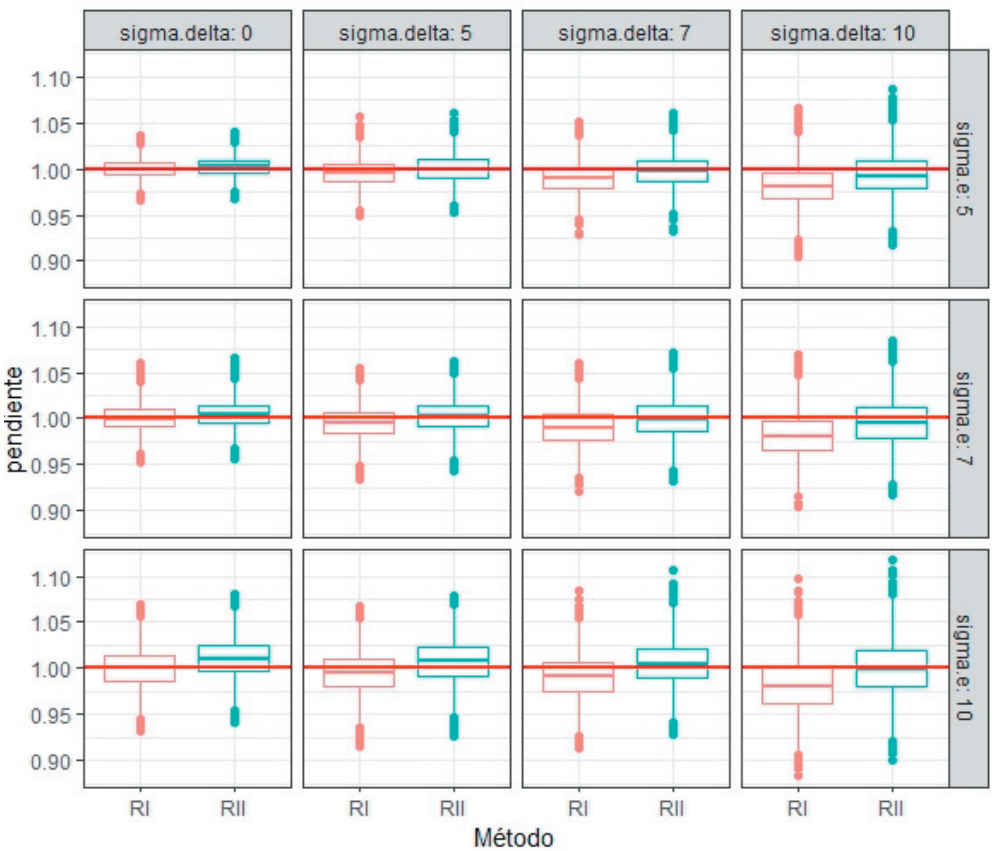

Figura 4. Resultados de las simulaciones. Se muestran las estimaciones de la pendiente utilizando regresión I (OLS) y regresión II (MA) bajo distintos escenarios (σ_δ=0,5,7,10; σ^2_e =5,7,10), siendo β_1=1 y el rango de x e y (0-200). La línea horizontal muestra el verdadero valor de la pendiente.

Figure 4. Simulation results. Slope estimates using regression I (OLS) and regression II (MA) are shown, under different scenarios (σ_δ=0,5,7,10; σ^2_e =5,7,10), with β_1=1 and x and y ranged between (0-200). The horizontal line shows the true value of the slope.

Consideraciones finales

De lo presentado se desprende que se debe poner énfasis en dos cuestiones muy importantes para poder decidir el método adecuado: tener en claro cuál es el objetivo del trabajo y, por otro lado, verificar si se cumplen las condiciones de aplicación requeridas por cada método. Frente a la pregunta ¿Deberíamos 'siempre' aplicar regresión tipo II cuando la variable explicativa no está fijada por el investigador?, respondemos que, si el objetivo es efectuar predicciones, el uso de regresión tipo II está decididamente desaconsejado. Si el objetivo es determinar si existe una relación funcional entre x e y, se podría utilizar OLS en vez de regresión tipo II, aunque sabiendo que la magnitud del efecto (la pendiente) puede estar atenuada. Si lo que se desea es estimar la fuerza de la asociación entre las dos variables, entonces debería aplicarse un análisis de correlación. Si el objetivo es poner a prueba si la pendiente es igual a cierto valor particular determinado por la teoría o por investigaciones previas (1, 3/2, etc.), se recomienda usar regresión tipo II. La misma recomendación surge cuando se desea comparar pendientes entre estudios que pueden tener distinta magnitud de variación en x y por lo tanto distintos sesgos en la estimación de la pendiente.

En cuanto a la elección del método en el caso que se decida aplicar regresión tipo II, la consideración más obvia tiene que ver con la escala de las variables. Si están medidas en distintas escalas se debería utilizar algún método que soslaye esta situación, como SMA o RMA. En modelos múltiples, el escalado tiene la ventaja de poner a las variables explicativas y a la respuesta en una escala comparable, lo que simplifica las comparaciones. También evita algunos problemas numéricos que pueden surgir cuando las variables son de escalas muy diferentes (Faraday 2009). La misma recomendación aplica cuando la escala es la misma pero las varianzas son muy distintas.

Por otro lado, SMA tiene la ventaja de generar estimaciones más precisas (y, por lo tanto, intervalos de confianza más angostos) que MA (Jolicoeur 1990). En la práctica, los tres métodos dan resultados similares si las varianzas de x e y son similares (por ejemplo, dentro de un factor de 1.2) o si la correlación entre ambas variables es alta, en cuyo caso es indistinto qué método se utilice (Warton et al. 2006). En otros casos, los resultados pueden ser muy diferentes. En la Tabla 1 se resumen algunas recomendaciones para elegir el método. Los supuestos habituales de independencia, linealidad, normalidad y varianzas homogéneas de los modelos lineales se aplican en regresión tipo II y para su exploración se pueden utilizar las mismas herramientas que para regresión tipo I. Sin embargo, sólo los supuestos de independencia y linealidad serían relevantes, ya que estas técnicas demostraron ser robustas al incumplimiento de los otros dos (Warton et al. 2006).

En este trabajo se presentaron distintas situaciones basadas sobre datos reales y escenarios simulados en los que la variable explicativa se mide con error y no cumple los supuestos de los modelos de regresión clásicos tipo I. Para cada caso se mostró de forma numérica cuándo aplicar los distintos métodos de regresión tipo II desarrollados en la literatura, cuándo no es recomendable y cuáles serían las posibles consecuencias de no emplear el método adecuado en cada situación. Los casos en los que resulta necesario aplicar los métodos de regresión tipo II aquí presentados se encuentran frecuentemente en el área de la ecología y en general en las ciencias biológicas y esta revisión pretende ser una sencilla guía de cuándo y qué método aplicar en cada situación.

Agradecimientos. Agradecemos a los autores de Ciancio et al. (2016), de Fassola et al. (2002) y de Marti y Slafer (2014) por facilitar los resultados de sus investigaciones para utilizar como casos de estudio en el presente trabajo.

Tabla 1. Comparación de los distintos métodos de regresión cuando las variables explicativas no son fijadas por el investigador (adaptada de Legendre [2015] y Warton et al. [2006]).

Table 1. Comparison of the different regression methods when the explanatory variables are not fixed by the researcher (adapted from Legendre [2015] and Warton et al. [2006]).

Método	Objetivos	Condiciones específicas de aplicación
OLS	Predecir y en función de x Determinar si existe una relación funcional entre x e y	-
MA	Poner a prueba si la pendiente es igual a cierto valor particular	Variables x e y en la misma escala Varianzas de x e y similares
SMA, RMA	Poner a prueba si la pendiente es igual a cierto valor particular	Sin observaciones extremas (*outliers*)

Referencias

Bland, J. M., and D. G. Altman. 1999. Measuring agreement in method comparison studies. Statistical Methods in Medical Research **8**(2):135-160. https://doi.org/10.1177/096228029900800204.

Ciancio, N., M. Parco, S. J. P. Incognito, and G. A. Maddonni. 2016. Kernel setting at the apical and sub-apical ear of older and newer Argentinean maize hybrids. Field Crops Research **191**:101-110. https://doi.org/10.1016/j.fcr.2016.02.021.

Correndo, A. A., F. Salvagiotti, F. O. García, and F. H. Gutiérrez-Boem. 2017. A modification of the arcsine-log calibration curve for analysing soil test value–relative yield relationships. Crop and Pasture Science **68**(3):297-304. https://doi.org/10.1071/CP16444.

Draper, N. R., and H. Smith. 2014. Applied regression analysis (Vol. 326). John Wiley and Sons.

Faraway, J. J. 2016. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Vol. **124**. CRC press. https://doi.org/10.1201/9781315382722.

Fassola, H. E., F. A Moscovich, P. Ferrere, and F. Rodríguez. 2002. Evolución de las principales variables de árboles de *Pinus taeda* L. sometidos a diferentes tratamientos silviculturales en el nordeste de la provincia de Corrientes, Argentina. Ciência Florestal **12**(2). https://doi.org/10.5902/198050981680.

Francq, B. G., and B. B. Govaerts. 2014. Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives. Chemometrics and Intelligent Laboratory Systems **134**:123-139. https://doi.org/10.1016/j.chemolab.2014.03.006.

Gillard, J. W. 2006. An historical overview of linear regression with errors in both variables. Math. School, Cardiff Univ., Wales, UK, Tech. Rep.

Harper, J. L. 1977. Population biology of plants. Academic Press, New York.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**(2): 187-211. https://doi.org/10.2307/1942661.

Jolicoeur, P. 1990. Bivariate allometry: interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. Journal of Theoretical Biology **144**(2):275-285. https://doi.org/10.1016/S0022-5193(05)80326-1.

Legendre, P., and L. F. Legendre. 2012. Numerical ecology. Vol. **24**. Elsevier.

Legendre, P. 1998. Model II regression user's guide, R edition. R Vignette, 14.

Legendre, P. 2015. lmodel2: Model II Regression. R package version 1.7-2.

Ludbrook, J. 2012. A primer for biomedical scientists on how to execute model II linear regression analysis. Clinical and Experimental Pharmacology and Physiology **39**(4):329-335. https://doi.org/10.1111/j.1440-1681.2011.05643.x.

Marti, J., and G. A. Slafer. 2014. Bread and durum wheat yields under a wide range of environmental conditions. Field Crops Research **156**:258-271. https://doi.org/10.1016/j.fcr.2013.10.008.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Smith, R. J. 2009. Use and misuse of the reduced major axis for line-fitting. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists **140**(3):476-486. https://doi.org/10.1002/ajpa.21090.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. Freeman. New York. US.

Taskinena, S., and D. I. Wartona. 2013. Robust tests for one or more allometric lines. Journal of Theoretical Biology **333**(21):38-46. https://doi.org/10.1016/j.jtbi.2013.05.010.

Warton, D. I., I. J. Wright, D. S. Falster, and M. Westoby. (2006). Bivariate line-fitting methods for allometry. Biological Reviews **81**(2):259-291. https://doi.org/10.1017/S1464793106007007.

Warton, D. I., R. A. Duursma, D. S. Falster, and S. Taskinen. 2012. smatr 3-an R package for estimation and inference about allometric lines. Methods in Ecology and Evolution **3**(2):257-259. https://doi.org/10.1111/j.2041-210X.2011.00153.x.

Yoda, K. 1963. Self-thinning in overcrowded pure stands under cultivated and natural conditions (Intraspecific competition among higher plants. XI). J. Inst. Polytech. Osaka City Univ Ser D **14**:107-129.