Ecología Austral 30: 438-453 Diciembre 2020
Asociación Argentina de Ecología
https://doi.org/10.25260/EA.20.30.3.0.1131

Cuando la variabilidad varía: Heterocedasticidad y funciones de varianza

Facundo J. Oddi[1,2,✉]; Fernando E. Miguez[3]; Guido G. Benedetti[1] & Lucas A. Garibaldi[1,2]

[1] *Universidad Nacional de Río Negro. Instituto de Investigaciones en Recursos Naturales, Agroecología y Desarrollo Rural. San Carlos de Bariloche, Río Negro, Argentina.* [2] *Consejo Nacional de investigaciones Científicas y Técnicas (CONICET). Instituto de Investigaciones en Recursos Naturales, Agroecología y Desarrollo Rural. San Carlos de Bariloche, Río Negro, Argentina.* [3] *Department of Agronomy, Iowa State University, Ames, IA, USA.*

Resumen. La variabilidad es una característica inherente al mundo que nos rodea. Cuantificarla es clave para comprender muchos de los procesos de interés para las ciencias ambientales y sociales como, por ejemplo, la adaptación de las especies al cambio climático o la desigualdad social. Para cuantificar la variabilidad se suele usar la varianza, uno de los parámetros de la distribución normal. Sin embargo, los modelos lineales clásicos asumen que la varianza es constante (supuesto de homocedasticidad) y se preocupan sólo por los cambios en las tendencias promedio. Es posible extender los modelos clásicos y relajar el supuesto de homocedasticidad mediante funciones de varianza, muy poco difundidas y abordadas por los textos en español. En esta ayuda didáctica nos proponemos introducir las funciones de varianza en modelos lineales desde un enfoque teórico-aplicado. Comenzamos introduciendo un problema real en el que se espera que la varianza no sea constante, y lo acompañamos con un ejemplo simulado. Posteriormente, planteamos el modelo lineal clásico y discutimos cómo se lo puede extender para modelar la heterocedasticidad. A continuación, explicamos algunas de las funciones de varianza y las aplicamos al caso real y a los datos simulados. Para ello hacemos uso de la función gls() del paquete *nlme* de R y proveemos el código para la reproducción del análisis. También exponemos otras opciones disponibles en R para tratar con datos heterocedásticos. Esperamos que este artículo brinde las bases para que profesionales y científicos con conocimientos estadísticos básicos comiencen a utilizar funciones de varianza y amplíen el conjunto de herramientas para analizar sus datos.

[Palabras clave: modelos lineales generales, mínimos cuadrados generalizados, selección de modelos, modelos anidados, criterios de información, AIC, función gls, R]

Abstract. When variability varies: Heteroscedasticity and variance functions. Variability is inherent to the world around us. Its quantification is essential to understand processes of interest in environmental and social sciences, such as adaptation of species to climate change or social inequality. Variance, one of the parameters of the normal distribution, is commonly used to quantify variability. Classical linear models assume that variance is constant (homoscedasticity assumption), while focusing only on changes in average trends. It is possible to extend classical models and relax the assumption of homoscedasticity through variance functions. However, these functions are scarcely used and we often lack examples in the Spanish-written scientific literature. In this paper, we introduce variance functions in linear models from a theoretical-applied approach. We begin by introducing a real problem where heteroscedasticity is expected, which is accompanied by one simulated example. Subsequently, we formulate the classical linear model and discuss how it can be extended to model heteroscedasticity. Then, we explain some of the variance functions and apply them to the real case and the simulated data. We use the gls() function of the *nlme* package in R, and provide scripts that make data analyses reproducible. Additionally, we describe other options available in R for dealing with heteroscedastic data. We expect this paper will provide a guide for using variance functions and will expand the toolbox of scientists with basic statistical knowledge.

[Keywords: general linear models, generalized least squares, model selection, nested models, information criteria, AIC, function gls, R]

Editor asociado: Diego Vázquez
✉ foddi@unrn.edu.ar

Recibido: 8 de Abril de 2020
Aceptado: 12 de Agosto de 2020

Introducción

La variabilidad es una característica inherente al mundo que nos rodea. Todos los patrones que observamos en la naturaleza y en las sociedades presentan algún grado de variabilidad debido a la complejidad de los procesos que los forman. A modo de ejemplo, la temperatura del aire varía a lo largo de un día, y esta variación es diferente entre días (muy diferente si comparamos la de un día de verano con la de uno de invierno), entre sitios e, incluso, en un mismo sitio (e.g., a uno u otro lado de un árbol). La biología tiene gran interés en estudiarla (Hallgrímsson and Hall 2005) dadas sus implicancias fisiológicas, ecológicas y evolutivas (Møller and Jennions 2002). Los ecólogos, en particular, estudian la variabilidad en diferentes niveles de organización biológica (i.e., organismos, poblaciones, comunidades, ecosistemas) (Whitham et al. 2006) e intentan explicar sus causas (Poorter et al. 2009). Lo mismo sucede en el ámbito social, por ejemplo, cuando los economistas analizan por qué la desigualdad del ingreso varía entre países (Li et al. 1998) o a qué se debe la volatilidad (variabilidad) de los precios en el mercado financiero (Shiller 1989), o cuando los sociolingüistas estudian la variación del lenguaje (Tagliamonte 2006).

Un estadístico que se usa a menudo para cuantificar la variabilidad es la varianza (Oddi et al. 2018). La varianza mide el alejamiento de un conjunto de datos con respecto al promedio (mientras más alejados, los datos presentan más diferencia entre sí y la variabilidad es mayor). Esta cantidad es central a la estadística ya que los modelos clásicos (regresión lineal, ANOVA) se basan en la distribución normal (Quinn and Keough 2002), cuyos parámetros son la media y la varianza. Mientras que los cambios en la media son vistos como determinísticos y explicados por uno o varios de los predictores incluidos en el modelo estadístico, la varianza cuantifica la variabilidad aleatoria; en otras palabras, aquella debida a otras causas (las no consideradas en la componente determinística). Los profesionales de las ciencias sociales y ambientales comprenden bien esto; en general, se entrenan en este tipo de modelos para contestar sus preguntas. No obstante, estos modelos tienen supuestos restrictivos que muchas veces no se cumplen; entre ellos, que la varianza es la misma para cualquier valor de la media (homocedasticidad). Cuando la varianza no es constante (Figura 1), se dice que existe heterocedasticidad e implica un problema para el análisis, ya que las probabilidades obtenidas en las pruebas de significancia (valores *P*) no son las que corresponden (i.e., los errores estándar son incorrectos y los estadísticos ya no siguen las distribuciones teóricas que sustentan estas pruebas) (Zuur et al. 2009). Luego, las inferencias que hagamos a partir de un modelo estadístico que viola los supuestos en los que se basa no serán válidas. Para corregir la heterocedasticidad se suelen utilizar transformaciones tales

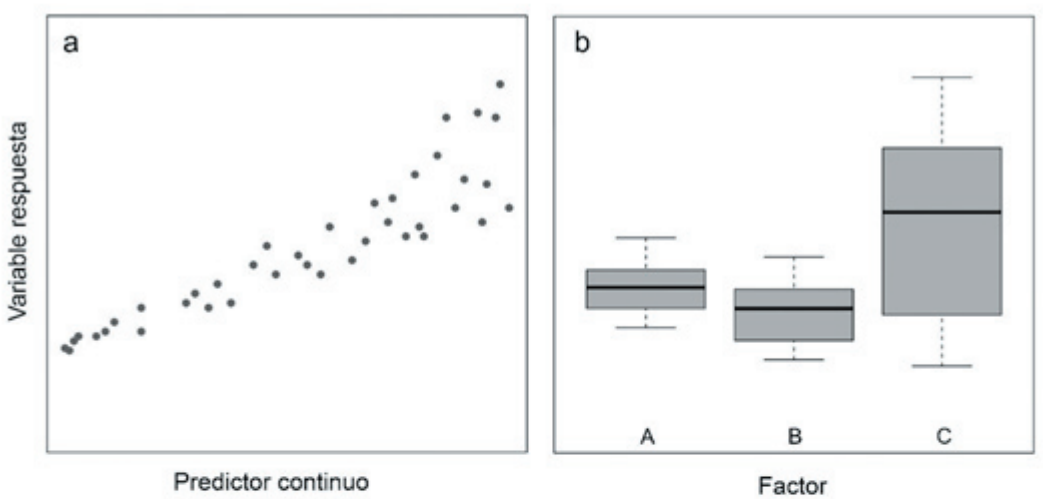


Figura 1. Ejemplos de patrones donde se observa heterocedasticidad en una variable respuesta. a) La variabilidad de la variable respuesta aumenta con el valor de un predictor continuo. b) La variabilidad difiere entre los niveles (A, B, C) del factor.

Figure 1. Patterns where heteroscedasticity is observed on a response variable. a) Variability of the response variable increases with continuous predictor. b) Variability differs among factor levels.

como logaritmos, senos o raíces (O'Hara and Kotze 2010; Warton and Hui 2011), pero esto puede desvirtuar la relación originalmente estudiada (Bolker 2008). Este aspecto es clave porque al trasformar los datos los parámetros del modelo pierden interpretabilidad, lo cual podría generar confusión en el análisis. Incluso, algunas transformaciones producen sesgos cuando las estimaciones son re-transformadas, es decir, cuando realizamos la operación inversa para volver a la escala original (Neyman and Scott 1960). Una forma de flexibilizar el análisis clásico es incluir funciones de varianza y modelar la heterocedasticidad. Además de posibilitar trabajar con datos heterocedásticos, estos modelos permiten comprender qué factores afectan la varianza, lo cual en muchos problemas es tan o más importante que la variación en las tendencias promedio (Schielzeth and Nakagawa 2013). Incluso, en muchas ocasiones, la aparente falta de normalidad se debe a la heterocedasticidad, un problema que se corrige al modelar la varianza. Sin embargo, no es común que las funciones de varianza se enseñen en los cursos de grado, lo que limita su comprensión e implementación. Además, existen muy pocos textos en español que introduzcan estos modelos (Garibaldi et al. 2019), lo que dificulta que los profesionales de habla hispana los usen.

El objetivo de esta ayuda didáctica es introducir las funciones de varianza en modelos lineales. Para ello, comenzamos introduciendo un caso concreto en el que se espera que la varianza no sea constante, y lo acompañamos con un ejemplo simulado. Luego planteamos el modelo lineal clásico y discutimos cómo se lo puede extender para modelar la heterocedasticidad. A continuación mostramos algunas de las funciones de varianza que se pueden usar, y finalmente las aplicamos a los ejemplos de trabajo. Para ello hacemos uso de la función gls() del paquete *nlme* y proveemos el código para la reproducción del análisis. También exponemos otras herramientas actualmente disponibles en R para tratar con datos heterocedásticos (Caja 1). Esperamos que este artículo brinde las bases para que los profesionales de las ciencias sociales y ambientales con conocimientos estadísticos básicos (i.e., distribuciones por muestreo, prueba de hipótesis, ANOVA, regresión lineal) comiencen a utilizar funciones de varianza y amplíen el conjunto de modelos para analizar sus datos.

Ejemplos de trabajo

Utilizaremos dos enfoques para ilustrar el ajuste en R de modelos con funciones de varianzas: a) abordaremos un caso real en el que se observa heterocedasticidad y b) simularemos datos heterocedásticos (Pinheiro and Bates 2000). Lo interesante de la simulación es que conocemos los verdaderos valores de los parámetros del modelo del cual surgen los datos de la muestra (es decir, conocemos la población) y, de esta forma, cuán buenas son las inferencias de los modelos que ajustamos. En general, los científicos utilizan la estadística para tomar decisiones sobre poblaciones que conocen en su totalidad (Garibaldi et al. 2017). El ejemplo con datos reales nos servirá para mostrar cómo las funciones de varianza ayudan a resolver una situación en la cual el modelo lineal clásico no cumple con el supuesto de homocedasticidad. A continuación introducimos brevemente su marco conceptual (el análisis estadístico se desarrolla en el archivo *volatilidad.R* del material suplementario) y luego describimos un ejemplo de trabajo con datos simulados (la simulación y el posterior análisis estadístico es desarrollada en el archivo *simulacion.R* (ver Material Suplementario).

Volatilidad económica (datos reales)

El bienestar de las personas y de las sociedades en su conjunto está determinado, en gran medida, por su estabilidad económica (Howell and Howell 2008). Un concepto muy relacionado, y que volvió a cobrar relevancia por las frecuentes crisis económicas, es el de la volatilidad (Wolf 2005; Fanelli 2008). Se refiere a la inestabilidad o fluctuación de una variable económica respecto a su tendencia. Cuantitativamente se la define como el desvío estándar de la variable en un período de 5 a 10 años. Por ejemplo, podemos calcular la volatilidad (v) del PBI de la Argentina entre 2014 y 2019 como:

$$v = \sum_{i=2014}^{2019} \sqrt{\frac{(PBI_i - \overline{PBI})^2}{5}}$$

De alguna manera, la volatilidad se asocia con la incertidumbre sobre el futuro; altos niveles de volatilidad implican permanentes cambios estructurales que disminuyen la capacidad de anticiparse a los eventos y dificultan la toma de decisiones (Dehn 2000). La volatilidad afecta el crecimiento económico (Ramey and Ramey 1995; Loayza and Hnatkovska 2005) y

Caja 1. R y otras formas de abordar la heterocedasticidad

La heterocedasticidad se presenta en diferentes contextos, desde comparaciones experimentales de diferentes tratamientos, estudios observacionales, hasta meta-análisis. De la misma manera, existen diferentes formas de abordar el incumplimiento del supuesto de homogeneidad de varianzas. Es importante mencionar que los abordajes conceptuales son independientes del programa informático utilizado (West et al. 2014) y que estos últimos facilitan, o no, su implementación. Así, además de R, existen otros programas que cuentan con paquetes que permiten modelar las varianzas, tales como SAS (http://www.sas.com/), SPSS (https://www.ibm.com/analytics/spss-statistics-software) o Stata (https://www.stata.com/). También existen programas como InfoStat (https://www.infostat.com.ar/), desarrollado por docentes-investigadores de la Universidad Nacional de Córdoba (UNC), que se vinculan a R y permiten ajustar funciones de varianza desde un menú que evita el uso de códigos de programación. La gran ventaja de R se encuentra en que es un entorno informático gratuito y de código abierto, es decir permite su libre instalación, uso, actualización, modificación y distribución, además de su amplia difusión entre los usuarios de la estadística. Si bien la flexibilidad de las funciones de varianzas estándar incluidas en el paquete *nlme* permite resolver gran cantidad de situaciones en donde la varianza no es constante, R provee más herramientas para abordarlas. Incluso la función gls() permite trabajar con funciones definidas por el usuario (Pinheiro and Bates 2000). Mientras que esto otorga gran flexibilidad, requiere un mayor esfuerzo en la escritura del código y plantea una dificultad extra para lograr la convergencia del ajuste. Con las funciones lm() y nls() incluidas en el paquete *base* también es posible ajustar modelos heterocedásticos haciendo uso del argumento weigths. En este caso, la contribución de cada observación es ponderada con un 'peso' (w_i) que se asume conocido (i.e., no hay que estimarlo) y podría usarse el valor de un predictor si se observa que la varianza aumenta con el mismo. Básicamente, la idea es que los residuos que presenten mayor alejamiento respecto a la tendencia lineal pesen menos en la estimación de la varianza. A estos modelos se les suele llamar 'regresiones ponderadas' porque los parámetros se estiman minimizando la sumatoria de los residuos cuadrados ponderados $([y_i - \hat{y}_i]^2 * w_i)$ y son conceptualmente similares a la función de varianza *varFixed* (Gałecki and Burzykowski 2013). Otra posibilidad es usar la función lmvar() en el paquete del mismo nombre (Posthuma Partners 2019) o la función remlscore() del paquete *statmod* (Smyth 2002), ambas con ajustes por máxima verosimilitud. En estudios de meta-análisis, se puede usar el paquete *metafor* (Viechtbauer 2010) y ponderar por la varianza observada en cada estudio. La función nlreg() del paquete del mismo nombre (Brazzale 2005) también permite escribir la fórmula de la función de varianza desde el argumento weigths, lo cual brinda flexibilidad pero, de nuevo, la convergencia del modelo es más compleja (el ajuste usa el criterio de máxima verosimilitud). Estos son sólo algunos ejemplos. R es un programa muy versátil, dinámico y con continuas contribuciones y actualizaciones, por lo que existen muchas formas de incluir funciones de varianza, incluso la posibilidad de que sea el usuario quien desarrolle el código de todo el modelo (i.e., sin usar las funciones de los paquetes) y estime los parámetros mediante optimización numérica con herramientas como optim(). Es importante mencionar también que en algunas situaciones las trasformaciones podrían resultar útiles, e incluso una mejor alternativa que el modelado de la varianza (ver Xiao et al. 2011). Finalmente, la heterocedasticidad podría ser modelada por distribuciones de probabilidad en las que el cambio de la varianza con la media es una propiedad intrínseca de la distribución como en la Gamma o, en el caso de que la variable respuesta sea un conteo, la distribución de Poisson o la binomial negativa. Esta alternativa implicaría que nos movamos al mundo de los modelos lineales generalizados (Nelder and Wedderburn 1972), siendo la función glm() del paquete *base* la primera opción que nos brinda R para trabajar dentro de este marco de modelado.

muestra una alta correlación con el número de episodios de crisis que experimenta un país (Wolf 2004). Influye sobre el bienestar, ya que aumenta también la volatilidad del consumo (Kose et al. 2003). Por lo tanto, su estudio es de interés tanto teórico como empírico.

La base de datos de este ejemplo fue generada a partir de la información disponible en los Indicadores de Desarrollo Mundial (WDI) del Banco Mundial (https://tinyurl.com/ybrr5uyr). De este catálogo extrajimos el crecimiento anual del PBI y del consumo de hogares de 143 países en el período 2010-2018. Con esta información calculamos la volatilidad del consumo (vc), la variable respuesta de nuestro problema, y la volatilidad del ingreso (vi), el predictor. Estas variables representan el desvío estándar del ingreso y del consumo para el período de análisis. La unidad de análisis es el país, al cual le asignamos el correspondiente valor de volatilidad en el ingreso y de volatilidad en el consumo; es decir, estamos tratando con datos observacionales de corte transversal. Los datos se encuentran en el archivo *volatilidad.txt* (ver Material Suplementario).

En este ejemplo evaluaremos si la volatilidad del ingreso a nivel país tiene un efecto sobre la volatilidad del consumo. De acuerdo con el marco conceptual, esperamos que a mayores niveles de volatilidad en el ingreso haya menos estabilidad y, por lo tanto, mayor volatilidad en el consumo. Ahora bien, los países que son productivamente fluctuantes, en años 'malos' pueden sostener su consumo accediendo al crédito. En otras palabras, un país que accede al crédito podría tener un consumo relativamente estable (baja volatilidad) a pesar de su alta volatilidad en el ingreso. De esta forma, a altos niveles de volatilidad en el ingreso habría mayor variabilidad entre países debido a las diferentes estrategias financieras. Esperamos que tanto el promedio como la varianza de la volatilidad del consumo aumenten con la volatilidad en el ingreso. Para analizar si los datos apoyan nuestras hipótesis económicas plantearemos un modelo en el que tanto la media (μ) como la varianza (σ^2) de la volatilidad en el consumo sean una función (f) de la volatilidad en el ingreso: μ=f(vi); σ^2=f(vi). La media será modelada a partir de una relación lineal entre vc y vi, y la varianza por alguna de las funciones de varianza que se detallan más adelante.

Crecimiento vegetal (datos simulados)

En este ejemplo ilustramos un caso donde la varianza difiere entre grupos o niveles de un factor. Para esto simulamos el crecimiento de dos especies vegetales (Sp_1, Sp_2) en dos sitios (norte y sur). Los datos son simulados suponiendo que la Sp_2 tiene un mayor crecimiento que la Sp1, y que en el norte éste es más alto y más variable que en el sur (Caja 2).

Modelo lineal clásico

Una forma de escribir este modelo, existen muchas otras (ver Caja 3), es la siguiente:

$$y_i \sim \mathcal{N}(\mu_i; \sigma^2)_{independientes}$$

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{p-1i}$$

donde y representa la variable dependiente o de interés (i.e., aquella sobre la que se quiere analizar la variabilidad) y x_1, x_2, … x_{p-1} son las variables independientes, regresoras o predictoras (i.e., aquellas con las cuales se pretende explicar la variabilidad de y) registradas en la unidad experimental i(i=1,…, n), y p es el número de coeficientes de regresión incluido el intercepto (β_0). Por ejemplo, en el problema de la volatilidad económica el modelo lineal clásico queda expresado como:

$$vc_i \sim \mathcal{N}(\mu_i; \sigma^2)_{independientes}$$

$$\mu_i = \beta_0 + \beta_1 vi_i$$

donde i es el país. Nótese que el modelo tiene p+1 parámetros; esto es, p coeficientes de regresión más la varianza (σ^2). Así, en nuestro ejemplo el modelo consta de tres parámetros: dos coeficientes de regresión (la ordenada al origen y la pendiente) y la varianza. El supuesto de homocedasticidad queda implícito en la formulación del modelo al no incluir el subíndice i en la varianza. El modelo también supone normalidad (*N*), linealidad (μ_i es modelada por una suma de parámetros) e independencia de las observaciones.

Extendiendo el modelo lineal: Modelos con heterocedasticidad

Lo que plantea el modelo anterior es que la variabilidad total de y puede ser expresada mediante distribuciones normales cuyas medias son una función de uno o varios predictores (esta función es conocida como la componente fija del modelo), es decir, es

Caja 2. SIMULACIÓN

Aquí ilustramos un caso en el que la varianza difiere entre grupos o niveles de un factor. Para esto, simulamos el crecimiento (c) de dos especies vegetales (Sp_1, Sp_2) en dos sitios (norte y sur). La simulación asume que utilizamos un diseño multifactorial cruzado de 2x2 (2 factores con 2 niveles cada uno=4 tratamientos) con 20 replicaciones por tratamiento (n=80). La media (μ) de c, la variable respuesta, es una función lineal que responde al diseño de muestreo: $\mu=\beta_0+\beta_1 Sp_2+\beta_2$ sur$+\beta_3$ (Sp_2:sur). Aquí Sp_2 es un predictor dicotómico que vale 1 si el crecimiento se registra en la segunda especie y 0 en caso contrario, y de igual manera en el caso del predictor sur (1 si c se mide en el sur y 0 en el norte). β_0 representa la media del crecimiento de Sp_1 en el norte (tratamiento base) y β_1, β_2, y β_3, son diferencias promedios con respecto a β_0 y expresan los efectos principales de la especie y el sitio, y el efecto de la interacción entre especie y sitio, respectivamente. En la simulación se supone que:

- el crecimiento sigue una distribución normal
- en el norte, el crecimiento promedio de la Sp_1 es 15 cm/año (β_0=15),
- en el sur, el crecimiento es, en promedio, 7 cm/año menor que en el norte (β_1=-7),
- la Sp_2 crece, en promedio, 1 cm/año más que la Sp_1 (β_2=1),
- no hay interacción entre sitio y especie (β_3=0),
- en el norte, la varianza del crecimiento es de 25 cm^2/año^2 (desvío estándar=5 cm/año)
- en el sur, la varianza es 25 veces menor que en el norte (varianza=1 cm^2/año^2; desvío estándar=1 cm/año).

Análisis

Figura C2-1. Comparación de los residuales del modelo lineal clásico (varianza constante) con los del modelo que asume una varianza por sitio (modelo *varIdent*).

Figure C2-1. Comparison of the residuals of the classical linear model (constant variance) with those of the model that assumes a variance per site (*varIdent* model).

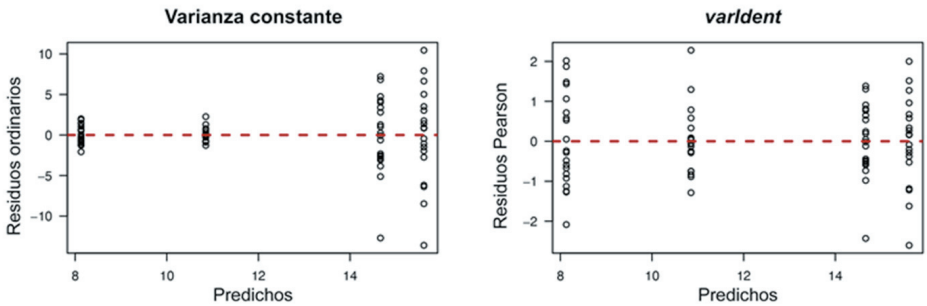


Tabla C2-1. Comparación de la bondad ajuste (mediante un test de cociente de verosimilitudes) de ambos modelos.

Table C2-1. Comparison of goodness of fit (using a likelihood ratio test) of both models.

Función de varianza	Número de parámetros	-log verosimilitud	LRT	Valor *P*
Varianza constante	5	-214.5	76.1	<0.001
varIdent	6	-176.5		

Tabla C2-2. Comparación de los coeficientes que estiman ambos modelos.

Table C2-2. Comparison of the coefficients estimated by both models.

	Varianza constante		*varIdent*	
Parámetro	Estimación puntual	Error estándar	Estimación puntual	Error estándar
β_0=15	14.7	0.8	14.7	1.2
β_1=-7	-6.5	1.2	-6.5	1.2
β_2=1	0.9	1.2	0.9	1.7
β_3=0	-0.2	1.7	-0.2	1.7

Tabla C2-3. Comparación de las inferencias sobre los efectos fijos mediante la prueba *F* secuencial (gln=grados de libertad del numerador del estadístico).

Table C2-3. Comparison of inferences about fixed effects using the sequential F test (gln = degrees of freedom of the numerator of the statistic).

	Varianza constante			*varIdent*		
Efecto	gln	*F*	Valor *P*	gln	*F*	Valor *P*
sitio	1	62.38	<0.001	1	62.38	<0.001
especie	1	0.98	0.325	1	5.54	0.021
sitio : especie	1	0.02	0.893	1	0.02	0.893

explicada por los predictores en términos de cambios promedios. Este modelo puede ser extendido para que la varianza de estas distribuciones (la componente aleatoria) también sea una función de predictores. A estas funciones se las llama 'funciones de varianza' y el modelo estadístico, a veces llamado "modelo lineal extendido" (Pinheiro and Bates 2000), puede expresarse como:

$$y_i \sim \mathcal{N}\left(\mu_i;\ \sigma_i^2\right)_{independientes}$$

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{p-1i}$$

$$\sigma_i^2 = \sigma^2 f^2\left(z_{1i}, z_{2i}, \ldots z_{qi};\ \delta\right)$$

donde z_1, z_2,…z_q son los predictores de la varianza (estos pueden ser los mismos que modelan la media u otros, incluso el predictor de varianzas puede ser el valor esperado de y, es decir μ_i), δ representa el vector de parámetros que afectan a la varianza (aquellos asociados a los predictores z), y f es la función de varianza propiamente dicha (Pinheiro and Bates 2000). Siguiendo con nuestro ejemplo, expresamos el modelo extendido como:

$$vc_i \sim \mathcal{N}\left(\mu_i;\ \sigma_i^2\right)_{independientes}$$

$$\mu_i = \beta_0 + \beta_1 vi_i$$

$$\sigma_i^2 = \sigma^2 f^2\left(vi_i;\ \delta\right)$$

Como se observa en la tercera ecuación del modelo, la varianza correspondiente a una observación i es obtenida multiplicando una varianza base (σ^2) por el cuadrado de f (esto asegura que la varianza sea positiva). Aunque por convención se la denomine función de varianza, como f está elevada al cuadrado lo que se modela es la desviación estándar. La varianza base es la varianza de y para cierto valor de los predictores y su interpretación depende de la función usada. Es decir, σ^2 es el parámetro que escala una varianza cuyos cambios relativos son determinados tanto por la forma de f, como por los valores de δ (Gałecki and Burzykowski 2013).

La diferencia sustancial con el modelo lineal clásico es el subíndice i en el parámetro de varianza (primera ecuación), y esto implica que la observación y_i proviene de una distribución normal cuya varianza, al igual que la media, no es general sino específica. En otras palabras, hemos relajado el supuesto de homocedasticidad (Caja 3). De acuerdo a este modelo, los predictores ahora no sólo explican los cambios promedios sino además la variabilidad en torno al promedio. En consecuencia, el número de parámetros del modelo es p+1 (σ^2) más los que incluye la función de varianza (los incluidos en el vector δ; ver siguiente sección). Si bien la función de varianza puede ser no lineal, se considera que el modelo es lineal debido a que la media varía según una función que es una suma de parámetros, es decir, la linealidad se encuentra en la segunda ecuación del modelo ($\mu_i=\beta_0+\beta_1 x_{1i}+\ldots$).

FUNCIONES DE VARIANZA ESTÁNDAR

En esta sección mostraremos las funciones de varianzas implementadas en el paquete *nlme* (Pinheiro et al. 2016) de R (R Core Team 2018) (Tabla 1). Mientras que, en principio, cualquier función f podría plantearse para modelar la varianza, las presentadas aquí son funciones estándar (Pinheiro and Bates 2000) que facilitan el ajuste del modelo, es decir la estimación de sus parámetros. Esto es un aspecto importante porque estos modelos en general son estimados maximizando la verosimilitud de los datos dado el modelo (i.e., métodos de 'máxima verosimilitud' o 'máxima verosimilitud restringida') mediante optimización numérica (Gałecki and Burzykowski 2013).

varIdent

Se aplica cuando existe heterocedasticidad entre grupos o niveles de un factor. La varianza es modelada por un predictor categórico. En un factor de J niveles, la función es dada por la siguiente expresión:

$$\sigma_{ij}^2 = \sigma^2\ \delta_j^2$$

donde σ^2 es la varianza del nivel de referencia y δ_j representan cocientes entre las desviaciones estándar del nivel j(j=1,… , J) y el de referencia. es decir, δ_j indica cuanto más grande o chica es la variabilidad del grupo j en relación al que se usa como referencia (σ_{ij}^2 se lee como la varianza de la unidad experimental i en el nivel j; cuando i pertenece al nivel de referencia entonces $\delta_j=1$). Por lo tanto, los parámetros δ asumen valores mayores a cero (Gałecki and Burzykowski 2013). Si suponemos que el predictor z es un factor de 2 niveles, como el

Tabla 1. Funciones de varianza estándar implementadas en R. En la función gls() del paquete *nlme* se indican dentro del argumento weights (el código es escrito suponiendo que 'z1', 'z2' y 'z3' son vectores definidos en R con los valores de predictores utilizados para modelar la varianza). Para una implementación completa de la función gls(), ver código de R en los ejemplos.

Table 1. Standard variance functions in R. In the gls() function of the *nlme* package they are indicated in the weights argument (in this Table the code is written assuming that 'z1', 'z2' and 'z3' are vectors defined in R with the values of the variance predictors). For an implementation of the gls() function, see R scripts.

	Código en R	Descripción
Varianza constante	weigths = NULL	Una única varianza para todas las observaciones (modelo lineal clásico). NULL es la opción por defecto en la función gls()
varIdent	weigths = varIdent(form = ~1|z1)	Una varianza por grupo o tratamiento. El nivel de referencia de la función de varianza puede no coincidir con el de la función de la media
varFixed	weigths = varFixed(~z2)	La varianza aumenta en forma proporcional a un predictor cuantitativo
varPower	weigths = varPower(form = ~z2)	Relación de potencia entre la varianza y un predictor cuantitativo. Esta función también permite modelar la varianza en función de la media de la variable respuesta. El código para indicar esto es weigths = varPower(.)
varConsPower	weigths = varConsPower(form = ~z2)	La varianza cambia con un predictor cuantitativo de acuerdo a una relación que incluye una constante y una potencia. Permite modelar la varianza en función de la media
varExp	weigths = varExp(form = ~z2)	La relación entre la varianza y el predictor cuantitativo es exponencial. Permite modelar la varianza en función de la media
varComb	weigths = varComb(varIdent(~z1), varExp(form = ~z2))	Combina dos o más funciones de varianza
	weigths = varComb(varIdent(form = ~1|z1), varFixed(~z2))	
	weigths = varComb(varPower(form = ~z2), varExp(form = ~z3))	

sitio con los niveles norte y sur en el ejemplo simulado del crecimiento vegetal, una forma alternativa de expresar esta función sería (Oddi et al. 2019): $\sigma_i^2 = \sigma^2 (\text{norte}_i + \delta_1 \text{sur}_i)^2$. Aquí σ^2 es la varianza en el norte (el nivel de referencia) y δ_1 es el cociente entre las desviaciones estándar en el sur y la deviación estándar en el norte. Los niveles norte y sur son variables binarias (*dummy*) con valor 1 en el caso de que la unidad experimental i pertenezca al grupo, y 0 en caso contrario, lo cual responde a una de las parametrizaciones que existen para tratar con variables categóricas, y la que R usa por defecto. Es posible cruzar dos factores e incluirlos en la función (en el ejemplo mencionado podríamos cruzar el factor sitio, de 2 niveles, con el factor especie, también de 2 niveles, obteniendo 2x2=4 grupos). El número de parámetros de la función *varIdent* es igual a la cantidad de grupos, 2 en el caso del ejemplo donde la varianza cambia entre sitios: σ, δ_1. Cuando todos los δ_j=1, las desviaciones estándar de los niveles son iguales y la varianza es constante. Por lo tanto, el modelo lineal clásico está anidado en (es un caso especial de) un modelo que incluya *varIdent* como función de varianza. Como se explica en la sección sobre comparación de modelos, el anidamiento determina el tipo de análisis que podemos hacer para evaluar las funciones de varianza (Pinheiro and Bates 2000).

varFixed

Se utiliza cuando la varianza aumenta proporcionalmente con un predictor cuantitativo:

$$\sigma_i^2 = \sigma^2 |z_i|$$

donde z es un predictor cuantitativo y σ^2 es la varianza de y cuando $|z|$=1. Esta sería la función de varianza indicada si la variabilidad de la volatilidad en el consumo aumentara en forma proporcional con la volatilidad en el ingreso. Como puede observarse, esta función de varianza es lineal (intercepto=0 y pendiente=σ^2) y el número de parámetros es el mismo que en el caso del modelo lineal clásico (p+1) con lo cual (en términos del número

de parámetros), no agrega complejidad al modelo. La función no puede reducirse a una varianza constante y de esta forma el modelo lineal clásico no está anidado en el modelo con *varFixed*.

varPower

Esta función modela varianzas que son una potencia del valor absoluto de un predictor cuantitativo.

$$\sigma_i^2 = \sigma^2 |z_i|^{2\delta}$$

donde σ^2 es la varianza de y cuando z=1, y δ modela la relación de potencia entre la varianza y el predictor. La función es aplicable a predictores que asumen valores diferentes de cero (cuando z=0 y $\delta>0$, $\sigma_i^2=0$; cuando z=0 y $\delta<0$, $\sigma_i^2=\infty$). El parámetro δ puede tomar cualquier valor real y esto permite modelar varianzas que aumentan o decrezcan al aumentar el predictor. Cuando $\delta=0$, la varianza es constante, y cuando $\delta=0.5$ llegamos a una varianza fijada. De esta forma, tanto un modelo lineal clásico como uno con *varFixed* se anidan dentro de un modelo que incluye la función

Caja 3. NOTACIÓN MATRICIAL

Muchos de los conceptos tras los modelos estadísticos quedan más claros cuando se utiliza la notación matricial. En el caso de la comparación entre el modelo lineal clásico con el extendido, el aspecto central se encuentra en la matriz de varianzas y covarianzas. El lector no interesado puede omitir su lectura.

Modelo lineal clásico (homocedasticidad)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$
$$\mathbf{e} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{\sigma}^2\mathbf{I})$$

$$\underset{(n\,x\,1)}{\mathbf{Y}} = \underset{(n\,x\,p)}{\mathbf{X}} * \underset{(p\,x\,1)}{\boldsymbol{\beta}} + \underset{(n\,x\,1)}{\mathbf{e}}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ 1 & x_{13} & x_{23} & \dots & x_{p3} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ \dots \\ e_n \end{bmatrix}$$

$$\boldsymbol{\sigma}^2\mathbf{I} = \boldsymbol{\sigma}^2 * \underset{(n\,x\,n)}{\text{matriz de correlacion}} = \underset{(n\,x\,n)}{\text{matriz de varianzas y covarianzas}}$$

$$\boldsymbol{\sigma}^2\mathbf{I} = \boldsymbol{\sigma}^2 * \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & \mathbf{1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma}^2 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \boldsymbol{\sigma}^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \boldsymbol{\sigma}^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\sigma}^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \boldsymbol{\sigma}^2 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & \boldsymbol{\sigma}^2 \end{bmatrix}$$

El modelo supone independencia, esto es, la correlación entre dos residuales no se estima sino que se asume igual a cero. Esto queda materializado al especificar una matriz identidad (I) como matriz de correlación, es decir, una matriz cuyos elementos en la diagonal valen 1 y por fuera de ella valen 0. El supuesto de homocedasticidad del modelo clásico se encuentra en la diagonal de la matriz de varianzas y covarianzas, cuyos elementos valen todos σ^2, lo cual resulta de la multiplicación entre la varianza (un escalar) y la matriz de correlación.

(continuación Caja 3)

Modelo lineal extendido (heterocedasticidad)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$
$$\mathbf{e} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{\sigma}^2\mathbf{W})$$

matriz de correlacion $(n \times n)$ **matriz de varianzas y covarianzas** $(n \times n)$

$$\boldsymbol{\sigma}^2\mathbf{W} = \boldsymbol{\sigma}^2 * \begin{bmatrix} w_1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & w_2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & w_3 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & w_4 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & w_{n-1} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & w_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \sigma_{n-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & \sigma_n^2 \end{bmatrix}$$

Este modelo también asume independencia pero relaja el supuesto de homocedasticidad. La matriz de correlación es una matriz diagonal cuyos elementos resultan de la función de varianza. Los elementos de la diagonal de la matriz de varianzas y covarianzas ahora varían, indicando que cada error proviene de una distribución normal con una varianza especifica (σ_i^2).
Cuando existe correlación en el tiempo o en el espacio, la función gls() también permite, desde el argumento correlation, modelar la estructura de covarianzas, es decir, los elementos por fuera de la diagonal. En el caso que la falta de independencia sea inducida por un factor de agrupamiento (e.g., los frutos se agrupan en árboles, los cuerpos de agua en cuencas, los estudiantes en cursos, las ciudades en regiones, las medidas que se repiten en un mismo individuo, etc.), podemos recurrir a los modelos de efectos mixtos e incluir este factor como un efecto aleatorio (entre las alternativas más utilizadas encontramos las funciones lme() del paquete *nlme* y lmer() del paquete *lme4*) (Bates et al. 2015). En ambos casos estamos operando sobre los elementos por fuera de la diagonal, pero mientras los modelos de efectos mixtos se aplican a datos jerárquicos (la dependencia responde a una variable categórica y se asigna una misma covarianza a todas las observaciones dentro del grupo), en los de correlación lo que interesa es 'modelar' la estructura (la covarianza es determinada por una función) y esto puede hacerse mediante variables cuantitativas, como la distancia o el tiempo, o categóricas (Pinheiro and Bates 2000). En efecto, podríamos ajustar un modelo de efectos mixtos con correlación temporal/espacial y heterocedasticidad (para esto podríamos usar la función lme(), y dentro de ésta definir una estructura de correlación desde el argumento correlation y una función de varianza desde el argumento weights). Si bien la discusión de estos conceptos excede lo abordado en este artículo, creemos importante diferenciar cada caso.

varPower. La función agrega un parámetro así que el modelo tiene p+2 parámetros. Si se hubiera registrado un factor con J niveles (j=1, … , J), es posible extender la función para que δ varíe por nivel [$\sigma_{ij}^2 = \sigma^2 |z_i|^{2\delta_j}$], con lo cual, en este caso, estaríamos modelando la varianza mediante predictores cuantitativos y categóricos. Siguiendo con el ejemplo económico, la varianza podría ser modelada de acuerdo a una relación de potencia con la volatilidad en el ingreso como predictor continuo (z) y esta función podría variar entre dos grupos (j) de países (i), por ejemplo entre los que acceden al crédito y los que no. Al generalizar la función, el número de parámetros del modelo es p+1+J.

varExp

En este caso la varianza es modelada de acuerdo a una relación exponencial con un predictor cuantitativo:

$$\sigma_i^2 = \sigma^2 e^{2\delta z_i}$$

donde δ controla la relación exponencial. Cuando δ=0 se obtiene una varianza contante así que este modelo anida al modelo clásico. El número de parámetros del modelo es p+2 o, en forma general, p+1+J si se permite que δ varíe por grupo [$\sigma_{ij}^2 = \sigma^2 e^{2\delta_j z_i}$].

varConstPower

La varianza es proporcional a una constante más una potencia del valor absoluto de un predictor cuantitativo:

$$\sigma_i^2 = \sigma^2 (\delta_1 + |z_i|^{\delta_2})^2$$

donde δ_1 representa la constante y δ_2 determina la forma potencial de la relación. Cuando δ_1=0, se llega a *varPower*, si además δ_2=0, o δ_2=0.5, se obtienen la varianza contante y el modelo *varFixed*, respectivamente. Esta función también permite estimar parámetros de varianza para diferentes grupos [$\sigma_{ij}^2 = \sigma^2 (\delta_{1j} + |z_i|^{\delta_{2j}})^2$] y el número de parámetros del modelo es p+1+2J.

varComb

Combina, mediante multiplicación, dos o más funciones de varianza. En el ejemplo económico, si la varianza aumentase linealmente con la volatilidad en el ingreso, pero este aumento difiriese entre grupos de países, los que acceden y los que no acceden al crédito, se podrían combinar las funciones *varFixed* y *varIdent*. De la misma forma, si el aumento lineal de la varianza dependiese de otro predictor continuo, como por ejemplo el tamaño poblacional del país, podrían combinarse *varFixed* con *varPower*. El número parámetros del modelo dependerá de las funciones que se combinan.

Comparación de modelos y selección de la función de varianzas

Al modelar la varianza estamos flexibilizando el modelo lineal clásico a costa de complejizar el análisis. Entonces, siguiendo el principio de parsimonia (Garibaldi et al. 2017), sólo deberíamos incluir funciones de varianza en caso de ser necesario, esto es, si su inclusión mejora sustancialmente el ajuste de modelo. Ahora bien, ¿Cómo determinarlo? Para esto existen herramientas gráficas y analíticas que se utilizan complementariamente.

Los supuestos de los modelos estadísticos se exploran gráficamente sobre los residuales (para una implementación práctica ver los códigos de R del material suplementario). Un residual ($\mathbf{e}_i$) es la diferencia entre la observación ($\mathbf{y}_i$) y el valor predicho por el modelo ($\widehat{\mathbf{y}_i}$) para esa observación ($\mathbf{e}_i = \mathbf{y}_i - \widehat{\mathbf{y}_i}$). Si al graficar los residuales *versus* los predichos, o *versus* alguno de los predictores, observamos que éstos no se distribuyen aleatoriamente, es decir, que muestran un patrón, esto podría deberse a la existencia de diferentes varianzas (un ejemplo típico de un patrón que resulta de la heterocedasticidad es el de 'embudo'). Para evaluar si el modelado de la varianza remueve el patrón se utilizan los mismos gráficos pero con los residuos de Pearson ($\mathbf{ep}_i$). Los residuos de Pearson son residuos relativos al desvío estándar ($\mathbf{s}_i$) que, en estos modelos, le corresponde según la función de varianza ($ep_i = \left[y_i - \widehat{y_i}\right] / s_i$).

En cuanto a las herramientas analíticas, las pruebas de Levene y la de Bartlett son algunas de las que podrían usarse para evaluar el cumplimiento del supuesto de homocedasticidad (Quinn and Keough 2002). Sin embargo, un enfoque superador es ajustar modelos con diferentes estructuras de varianzas y compararlos mediante alguna medida que indique cuál de los modelos presenta mejor bondad de ajuste (Gałecki and Burzykowski 2013). Para modelos anidados, una posibilidad es utilizar una prueba de cociente de verosimilitudes (LRT por sus siglas en inglés). Asintóticamente, el LRT muestral sigue aproximadamente una distribución de χ^2 (Chi cuadrado) con grados de libertad igual a la diferencia en el número de parámetros de los modelos que se comparan (la comparación es siempre entre dos modelos). La hipótesis nula detrás del LRT establece que ambos modelos presentan la misma verosimilitud, es decir, el mismo grado de ajuste a los datos. De esta forma, se selecciona el modelo más complejo sólo en caso de que el valor *P* obtenido esté por debajo del nivel de significancia establecido para el análisis. Por ejemplo, dado un conjunto de datos, ajustamos un modelo que asume varianza constante y otro que supone que existen diferentes varianzas para diferentes niveles de un factor (*varIdent*). Luego, si el valor *P* asociado al LRT es suficientemente pequeño, entonces tendríamos evidencia estadística para concluir

Caja 4. Modelado de la volatilidad económica

Resultados del análisis estadístico desarrollado en el archivo *volatilidad.R* (ver Material Suplementario).

Selección de la función de varianza

Tabla C4-1. Bondad de ajuste de las diferentes estructuras de varianza evaluadas.

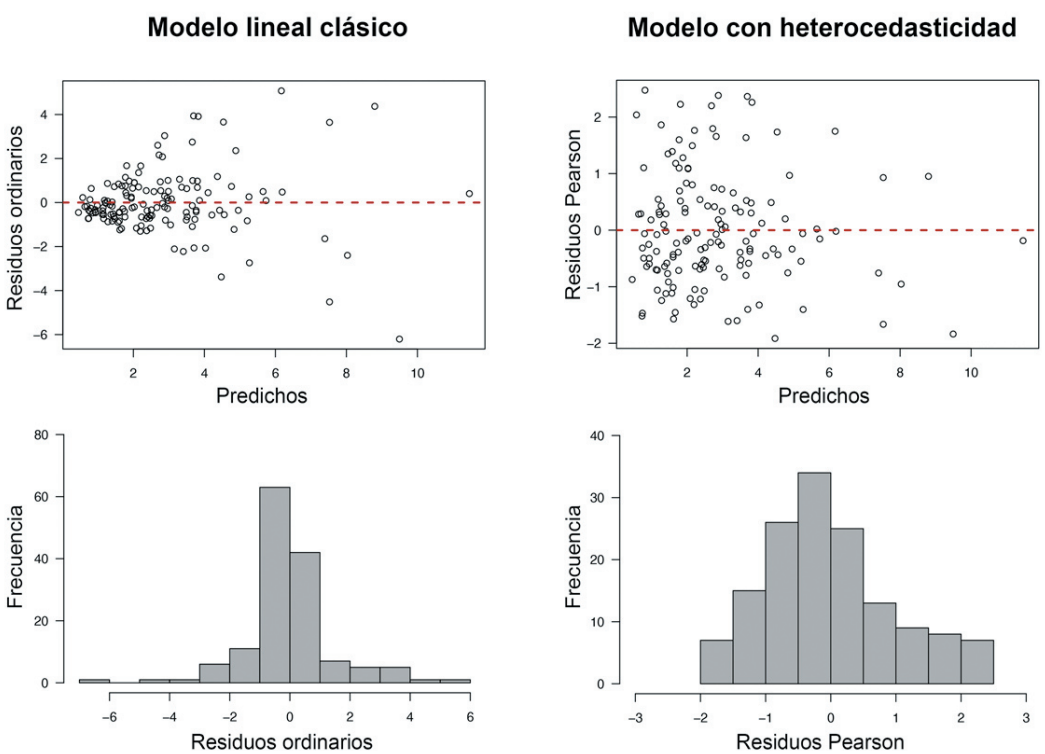
Tabla C4-1. Bondad de ajuste de las diferentes estructuras de varianza evaluadas.

Función de varianza	Número de parámetros	log verosimilitud	AIC	Delta AIC
varPow	4	-208.6	425.3	-
varConstPower	5	-208.2	426.3	1.0
varExp	4	-213.5	435.0	9.7
varFixed	3	-220.5	447.0	21.7
varianza constante	3	-259.2	524.3	99.0

Análisis de residuales

Figura C4-1. Comparación entre los residuales del modelo lineal clásico (asume varianza constante) y el modelo con heterocedasticidad modelada por la función de varianza *varPower*.

Figure C4-1. Comparison between the residuals of the classical linear model (assumes constant variance) and the model with heteroscedasticity modeled by the variance function *varPower*.



Ajuste del modelo seleccionado

Tabla C4-2. Parámetros estimados en el modelo de mejor bondad de ajuste (*varPower*) (Tabla C3-1).

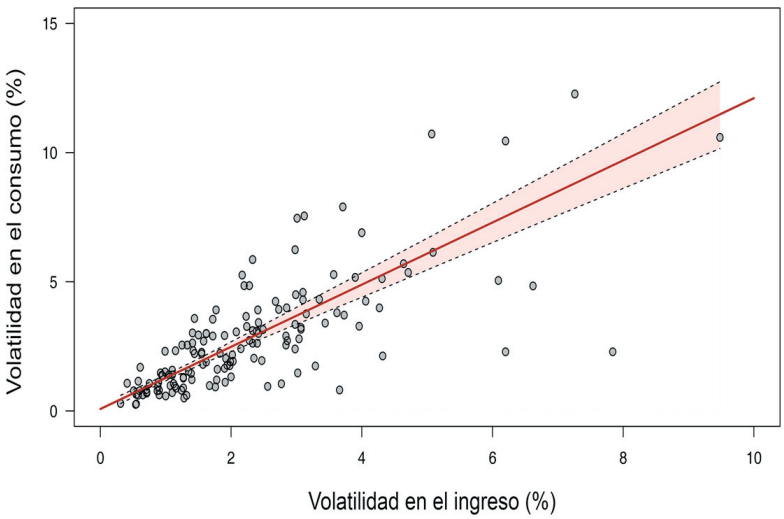
Table C4-2. Parameters estimated in the best goodness of fit (varPower) model (Table C3-1).

Parámetro	Estimación puntual	Intervalo de confianza (95%)
β_0	0.08	-0.12 ; 0.27
β_1	1.20	1.05 ; 1.36
σ^2	0.57	0.49 ; 0.67
δ_1	0.93	0.76 ; 1.10

(continuación Caja 4)

Figura C4-2. Valores observados (puntos grises), recta de ajuste (línea roja) del modelo *varPower*, e intervalos de confianza de la media de la volatilidad en el consumo (área rosada).

Figure C4-2. Observed values (gray points), fit line (red line) of the *varPower* model, and confidence intervals of the mean volatility in consumption (pink area).



que existe heterocedasticidad. En el caso de modelos no anidados, una alternativa es compararlos mediante criterios de información, por ejemplo el Criterio de Información de Akaike (AIC). El AIC es una medida de parsimonia que integra tanto el ajuste a los datos (verosimilitud) como la complejidad del modelo (número de parámetros) (ver Garibaldi et al. 2017 para una introducción didáctica a este tipo de inferencia). Este sería el caso cuando el modelo clásico es comparado con uno que supone que la varianza aumenta linealmente con el predictor (*varFixed*). A diferencia del LRT y de otras formas de comparar modelos bajo el enfoque frecuentista, el AIC permite comparar más de dos modelos a la vez, estén éstos anidados o no.

Funciones de varianza en los ejemplos trabajados

Volatilidad económica (datos reales)

Al explorar los datos observamos que la volatilidad del consumo tiende a aumentar con la del ingreso y, de la misma forma, la variación entre países. Esto es confirmado al ajustar el modelo estadístico (Caja 4). En primer lugar, todos los modelos con funciones de varianza presentan una sensible reducción en el AIC respecto al modelo lineal clásico (se utilizó un marco inferencial de criterios de información porque no todos los modelos ajustados estaban anidados) (Tabla C4-1), lo cual representa una evidencia sólida respecto a la existencia de heterocedasticidad. En particular, el mejor ajuste se logra con una estructura de varianzas *varPower*. El análisis visual de residuales sugiere que esta función modela adecuadamente la heterocedasticidad y corrige la falta de normalidad (Figura C4-1), esto último apoyado además por pruebas estadísticas (ver script en R). La función *varConstPower* presenta un soporte similar (los modelos se consideran similares si su diferencia en AIC es menor a 2 y en este caso es de 1 unidad) y también podría utilizarse para modelar los datos. Si decidiéramos usar esta función, aumentaríamos la verosimilitud de los datos (log verosimilitud) (Tabla C4-1) a costa de un modelo de mayor complejidad (número de parámetros, Tabla C4-1). En segundo lugar, el intervalo de confianza de la pendiente no incluye el cero (Tabla C4-2), es decir, existe evidencia para concluir que la volatilidad en el ingreso influye sobre la del consumo (Figura C4-2). En efecto, al remover la volatilidad en el ingreso el AIC aumenta considerablemente (ver script), soportando la conclusión anterior. La pendiente

estimada es 1.2 e implica que la volatilidad del consumo aumenta más de un punto por cada punto de volatilidad en el ingreso.

Crecimiento vegetal (datos simulados)

El modelo lineal clásico viola los supuestos de homocedasticidad y de normalidad, problema que se corrige al ajustar un modelo con una varianza por sitio (Figura C2-1; ver script), lo cual se traduce en una clara mejora en la bondad de ajuste de acuerdo con la prueba de cociente de verosimilitudes (dado que los dos modelos puestos a prueba están anidados, las inferencias fueron hechas desde un enfoque frecuentista) (Tabla C2-1). Si bien la estimación puntual de los parámetros de la componte fija es muy similar, el modelo lineal clásico estima una única varianza residual (s=3.8) (ver script) y, de esta forma, existen diferencias entre modelos en los errores estándar de b_0 y b_1 (Tabla C2-2). Uno de los mayores problemas que se observan es que al asumir una varianza común, se reduce notablemente el valor del estadístico F asociado al efecto de la especie y, por ende, aumenta su valor *P* (Tabla C2-3). Consecuentemente, con el modelo lineal clásico cometeríamos un error de tipo II al concluir que no hay efecto de la especie. En cambio, al modelar la varianza afirmaríamos que no existe interacción y detectaríamos los efectos principales del sitio y de la especie (Tabla C2-3). En otras palabras, la inclusión de la función de varianza nos permitiría realizar inferencias correctas sobre el proceso que gobierna el crecimiento vegetal. Es importante tener en claro que esta simulación (i.e., los 80 datos) representa una realización de infinitas posibles y que los resultados del análisis variarán de acuerdo a la muestra obtenida.

Consideraciones finales

Las funciones de varianza son generalmente introducidas dentro del marco de los modelos de efectos mixtos (Pinheiro and Bates 2000; Zuur et al. 2009; Gałecki and Burzykowski 2013). De hecho, en R las funciones de varianza estándar forman parte de funciones programadas para ajustar modelos de efectos mixtos (por ejemplo lme(), nlme()). Como consecuencia, en general sólo se incluyen en los contenidos de cursos de estadística de nivel avanzado. Esto contribuye a que el marco conceptual de las funciones de varianza pueda presentarse confuso (Zuur et al. 2009) o incluso desconocido para quienes no profundizan sus conocimientos estadísticos en instancias de posgrado. En esta ayuda didáctica introdujimos las funciones de varianza, considerándolas un primer paso en la extensión del modelo lineal clásico. Técnicamente, sólo modificamos la diagonal de la matriz de varianzas y covarianzas (Caja 3). Los modelos de efectos mixtos, así como los de correlación temporal o espacial, permiten dar un paso más y trabajar por fuera de la diagonal de la matriz, es decir, relajar también el supuesto de independencia (Caja 3).

Como ya discutimos, los modelos con heterocedasticidad brindan gran flexibilidad para abordar diferentes tipos de problemáticas del ámbito científico. Creemos importante que la decisión de incluir una función de varianza sea acompañada por el conocimiento del problema que se aborda. En este sentido, deberíamos preguntarnos si es razonable suponer heterocedasticidad; de no ser así la función de varianza podría modelar los datos de la muestra pero carecer de validez para inferencias poblacionales. Por otro lado, cuando tenemos hipótesis conceptuales respecto a la heterocedasticidad, o en situaciones en las que es claro que se necesita modelar la varianza, el tamaño muestral requerido para para trabajar con un determinado nivel de precisión podría ser elevado (mucho mayor en comparación a los modelos clásicos). En la práctica esto puede limitar la aplicación de algunas funciones de varianza.

Este material constituye el primer texto en español en la temática y hemos optado por un enfoque teórico-práctico. Es por ello que lo consideramos una contribución de relevancia para que muchos profesionales de las ciencias ambientales y sociales amplíen el conjunto de herramientas estadísticas del que disponen para contestar a sus preguntas.

Agradecimientos. Esta contribución ha sido financiada con fondos de la Universidad Nacional de Río Negro (PI 40-B-728, PI 40-B-567), la Agencia Nacional de Promoción Científica y Tecnológica (PICT 2016 – 0305, PICT-2018-00941) y del Belmont Forum y BiodivERsA mediante el llamado 2017-2018 para propuestas de investigación (bajo el Programa BiodivScen ERA-Net COFUND con financiamiento de AEI, NWO, ECCyT y NSF).

Referencias

Bolker, B. 2008. Ecological models and data. Princeton University Press, Princeton.

Brazzale, A. R. 2005. hoa: An R package bundle for higher order likelihood inference. Rnews, 5/1 May 2005, 20-27. ISSN 609-3631. URL: www.r-project.org/doc/Rnews/Rnews_2005-1.pdf.

Dehn, J. 2000. The effects on growth of commodity price uncertainty and shocks. Policy Research Working Paper No. 2455. World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-2455.

Fanelli, J. 2008. Macroeconomic volatility, institutions and financial architectures: the developing world experience. Palgrave Macmillan, London. https://doi.org/10.1057/9780230590182.

Garibaldi, L. A., F. Aristimuño, F. Oddi, and F. Tiribelli. 2017. Inferencia multimodelo en ciencias sociales y ambientales. Ecología Austral **27**:348-363. https://doi.org/10.25260/EA.17.27.3.0.513.

Garibaldi, L. A., F. J. Oddi, F. Aristimuño, and A. N. Behnisch. 2019. Modelos estadísticos en lenguaje R. Editorial UNRN, Viedma.

Gałecki, A., and T. Burzykowski. 2013. Linear Mixed-Effects Models Using R. A Step-by-Step Approach. Springer Texts in Statistics, Springer Science+Business Media, New York. https://doi.org/10.1007/978-1-4614-3900-4.

Hallgrímsson, B., and B. Hall. 2005. Variation: a Central Concept in Biology. Elsevier Academic Press, Boston.

Howell, R. T., and C. J. Howell. 2008. The relation of economic status to subjective well-being in developing countries: A meta-analysis. Psychological Bulletin **134**:536-560. https://doi.org/10.1037/0033-2909.134.4.536.

Kose, M. A., E. S. Prasad, and M. E. Terrones. 2003. Financial Integration and Macroeconomic Volatility. IMF Staff Papers **50**(1). https://doi.org/10.5089/9781451846997.001.

Li, H., L. Squire, and H. F. Zou. 1998. Explaining international and intertemporal variations in income inequality. The Economic Journal **108**:26-43. https://doi.org/10.1111/1468-0297.00271.

Loayza, N., and V. V. Hnatkovska. 2004. Volatility and Growth. Policy Research Working Paper No. 3184. World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-3184.

Møller, A. P., and M. D. Jennions. 2002. How much variance can be explained by ecologists and evolutionary biologists? Oecologia **132**:492-500. https://doi.org/10.1007/s00442-002-0952-2.

Nelder, J. A., and Wedderburn R. W. M. 1972. Generalized Linear Models. Journal of the Royal Statistical Society, Series A **135**:370-384. https://doi.org/10.2307/2344614.

Neyman, J., and E. L. Scott. 1960. Correction for bias introduced by a transformation of variables. The Annals of Mathematical Statistics **31**:643-655. https://doi.org/10.1214/aoms/1177705791.

Oddi, F. J., F. J. Aristimuño, C. Coulin, and L. A. Garibaldi. 2018. Ambigüedades en términos científicos: sobre el uso del "error" y el "sesgo" en estadística. Ecología Austral **28**:525-536. https://doi.org/10.25260/EA.18.28.3.0.680.

Oddi, F. J., F. Miguez, L. Ghermandi, L. O. Bianchi, and L. A. Garibaldi. 2019. A nonlinear mixed-effects modelling approach for ecological data: Using temporal dynamics of vegetation moisture as an example. Ecology and Evolution **9**:10225-10240. https://doi.org/10.1002/ece3.5543.

O'Hara, R. B., and D. J. Kotze. 2010. Do not log-transform count data. Methods in Ecology and Evolution **1**:118-122. https://doi.org/10.1111/j.2041-210X.2010.00021.x.

Pinheiro, J. C., and D. M. Bates. 2000. Mixed-effects models in S and SPLUS. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4419-0318-1.

Pinheiro, J. C., D. M. Bates, S. DebRoy, D. Sarkar, and R Core Team. 2016. _nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-126. URL: CRAN.R-project.org/package=nlme.

Poorter, H., Ü. Niinemets, L. Poorter, I. J. Wright, and R. Villar. 2009. Causes and consequences of variation in leaf mass per area (LMA): a meta-analysis. New Phytologist **182**:565-588. https://doi.org/10.1111/j.1469-8137.2009.02830.x.

Posthuma Partners. 2019. lmvar: Linear Regression with Non-Constant Variances. R package version 1.5.2. URL: CRAN.R-project.org/package=lmvar.

Quinn, G. P., and M. J. Keough. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, New York. https://doi.org/10.1017/CBO9780511806384.

R Core Team. 2018. R: A language and environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. URL: www.R-proje ct.org.

Ramey, G., and V. A. Ramey. 1995. Cross-country evidence on the link between volatility and growth (No. w4959). National bureau of economic research. The American Economic Review **85**:1138-1151. https://doi.org/10.3386/w4959.

Schielzeth, H., and S. Nakagawa. 2013. Nested by design: Model fitting and interpretation in a mixed model era. Methods in Ecology and Evolution **4**:14-24. https://doi.org/10.1111/j.2041-210x.2012.00251.x.

Shiller, R. J. 1989. Market volatility. MIT press, Cambridge.

Smyth, G. K. 2002. An efficient algorithm for REML in heteroscedastic regression (remlscore, randomizedBlock, and mixedModel2 functions). Journal of Computational and Graphical Statistics **11**:836-847. https://doi.org/10.1198/106186002871.

Tagliamonte, S. A. 2006. Analysing Sociolinguistic Variation (Key Topics in Sociolinguistics). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511801624.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software **36**:1-48. URL: www.jstatsoft.org/v36/i03. https://doi.org/10.18637/jss.v036.i03.

Warton, D., and F. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology **92**:3-10. https://doi.org/10.1890/10-0340.1.

West, B. T., K. B. Welch, and A. Gałecki. 2014. Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman and Hall, Boca Raton. https://doi.org/10.1201/b17198.

Whitham T. G., J. K. Bailey, J. A. Schweitzer, S. M. Shuster, R. K. Bangert, C. J. LeRoy, E. V. Lonsdorf, G. J. Allan, S. P. DiFazio, B. M. Potts, D. G. Fischer, C. A. Gehring, R. L. Lindroth, J. C. Marks, S. C. Hart, G. M. Wimp, and S. C. Wooley. 2006. A framework for community and ecosystem genetics: from genes to ecosystems. Nature Reviews Genetics **7**:510-523. https://doi.org/10.1038/nrg1877.

Wolf, H. 2004. Accounting for consumption volatility differences. IMF Staff Papers **51**:109-125.

Wolf, H. 2005. Volatility: Definitions and Consequences. Pp. 45-64 *en* J. Aizenman and B. Pinto (eds.). Managing Economic Volatility and Crises: A Practitioner's Guide. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511510755.004.

Xiao, X., E. P. White, M. B. Hooten, and S. L. Durham. 2011. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. Ecology **92**:1887-1894. https://doi.org/10.1890/11-0538.1.

Zuur, A. F., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed effects models and extensions in ecology with R. Springer, New York. https://doi.org/10.1007/978-0-387-87458-6.