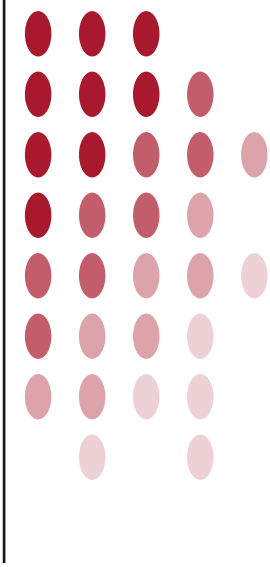
Capítulo 9
INTRODUCCIÓN A LOS MODELOS MIXTOS

Clasificación de efectos en fijos y aleatorios

Aproximaciones más versátiles, donde se modelan los términos fijos y aleatorios

Observaciones correlacionadas y medidas repetidas en el tiempo



Introducción a los modelos mixtos

Pedro M. Tognetti y Adriana Pérez

Los modelos lineales mixtos constituyen generalizaciones de los modelos lineales presentados en capítulos anteriores y ofrecen una gran flexibilidad para situaciones más complejas. El estudio detallado de los modelos mixtos supera el nivel de este libro y en consecuencia sólo presentaremos una breve introducción y ejemplos de aplicaciones habituales en ciencias ambientales y agropecuarias.

Existen numerosas situaciones que pueden ser abordadas desde los modelos mixtos. Una de las aplicaciones más frecuentes es la que permite cuantificar distintas fuentes de variabilidad aleatorias, además de la variación del error. Estos modelos, conocidos como de efectos aleatorios, proporcionan estimaciones de los componentes de varianza y son de amplio uso en mejoramiento genético animal y vegetal o en ensayos multiambientales (Ejemplo 9.1). Otra gran aplicación de los modelos mixtos es la que da cuenta de cierta estructura de correlación entre las observaciones (Ejemplo 9.2 y 9.3). Cuando se obtiene más de una observación por unidad experimental, o las unidades experimentales están agrupadas jerárquicamente, dichas observaciones no son independientes y no deberían ser consideradas verdaderas réplicas ya que aumentaría la probabilidad de error de tipo I. En otros casos, la falta de independencia entre las observaciones viene dada porque una misma unidad experimental es medida secuencialmente a lo largo del tiempo (Ejemplo 9.4). Ignorar la correlación existente entre las observaciones del mismo individuo llevaría a la pérdida de precisión en las estimaciones de los efectos fijos. Finalmente se mencionará otra aplicación que, si bien no responde formalmente a un modelo mixto, permite abordar mediante el modelado de matrices de covarianza una situación muy común que es la presencia de heterocedasticidad (Ejemplo 9.5).

El objetivo del capítulo es familiarizar al lector con los modelos lineales mixtos, facilitar el abordaje del tema en textos más avanzados y brindar herramientas parar econocer si un problema particular corresponde a este capítulo de la estadística. Esta introducción se enfoca en los modelos lineales mixtos para variables de distribución normal. Sin embargo, también puede servir para comenzar a

entender su generalización a situaciones en las que las variables siguen otros tipos de distribuciones de probabilid (Modelos Lineales Generalizados Mixtos, Bates *et al.* 2014, Bolker *et al.* 2009, Garibaldi *et al.* 2014).

9.1. ¿Por qué son mixtos los modelos? Efectos fijos y aleatorios

El nombre modelos *mixtos* alude a que los elementos del modelo combinan lo que se denomina *variables explicativas de efectos fijos* y *variables explicativas de efectos aleatorios* (Netter *et al.* 1990, Pinheiro y Bates 2000, Littell *et al.* 2007). En general, una variable explicativa es considerada como de efectos fijos si sus niveles fueron determinados explícitamente por el investigador y son de interés intrínseco para los objetivos del estudio. Por ejemplo, al comparar la tasa de ganancia de peso de bovinos Aberdeen Angus, Hereford y Shorthorn o la efectividad diferencial de cuatro estrategias de restauración, las conclusiones se aplicarán a esas tres razas bovinas o a las cuatro estrategias de restauración utilizadas en el estudio. En los capítulos anteriores se evidencia que todos los factores y tratamientos estudiados en los experimentos correspondían a ejemplos de efectos fijos, y el error experimental era el único componente aleatorio considerado.

En cambio, una variable explicativa se considera de efectos aleatorios si los niveles estudiados corresponden a una muestra aleatoria de una población de niveles para esa variable. Por ejemplo, sabemos que existen muchas especies de la familia de las leguminosas en el gran Chaco. Para evaluar el contenido de nitrógeno de los folíolos de las leguminosas de esa región, seleccionamos al azar 15 especies de esa familia. Las conclusiones se extenderán a la población de los niveles del factor (familia de las leguminosas) a partir de una muestra al azar de niveles (especies dentro de esa familia; p.ej. *Medreag lupina*, *Tipuana tipu*, *Acacia caven*, etc). Entonces, la decisión de tratar a una variable como fija o como aleatoria considera el procedimiento con que se generan los niveles del factor de interés, es decir una elección premeditada o una selección al azar, y el espacio de inferencia al que apuntan las conclusiones, que refiere a los niveles incluidos en el experimento solamente o a todos los niveles de ese factor que existen la población.

En síntesis, tres grandes preguntas ayudan a decidir si un factor debe considerarse fijo o aleatorio:

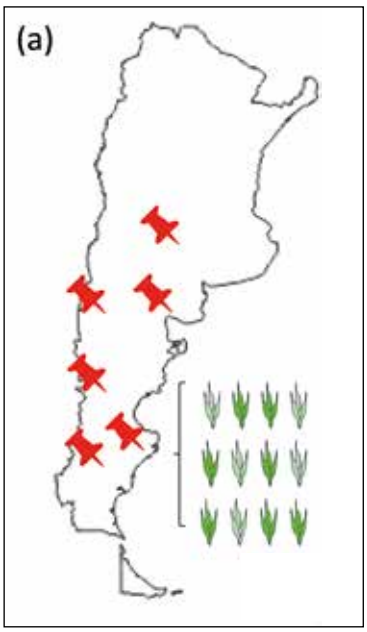
1. ¿Los niveles del factor se estipulan o surgen de elegirlos al azar?

2. ¿Las conclusiones se limitarán a los niveles del factor estudiado o se aplicarán a una población mayor de niveles? Por ejemplo, si concluiremos que hay diferencias en la adopción de tecnología entre las familias Pérez, Nakandakare y Osakia o que hay variabilidad en la adopción de tecnología entre familias de floricultores de Florencio Varela (de las cuales aquellas constituyen una muestra que participó en el estudio).
3. Si el experimento se repitiera, ¿se estudiarían nuevamente los mismos niveles del factor o se tomaría una muestra que, por azar, podría incluir otros niveles? (*i.e.* otra muestra de familias de floricultores de Florencio Varela)

Ejemplo 9.1. Poblaciones de pastos nativos

Poa ligularis es una gramínea perenne común en gran parte de la Argentina (Leva 2010, Leva *et al.* 2013). Es fuente de forraje en sistemas ovinos de Patagonia, por lo que se estudia su morfología y sus rasgos de respuesta al pastoreo en diferentes localidades. Esta especie es dioica, y se cree que los individuos machos y hembras tienen rasgos de crecimiento disímiles porque los machos deben destinar menos recursos a la producción de polen que las hembras a la producción de óvulos y desarrollo de los frutos (Graff *et al.* 2013).

En primer lugar, la hipótesis de investigación puede ser si los individuos de *P. ligularis* hembra y macho tienen macollos de diferente biomasa promedio. El sexo de las plantas es claramente un efecto fijo en términos estadísticos. Segundo, dado que *P. ligularis* se encuentra en amplias regiones de Argentina por debajo del paralelo 33° de latitud Sur, también interesa saber cómo son las plantas de *P. ligularis* en diferentes localidades. ¿Existe variabilidad en la biomasa de los macollos tomados de individuos de *Poa* que crecen en distintos puntos del país? En este caso, podemos considerar a las localidades como un efecto aleatorio. Entre todas las localidades en las que crece *P. ligularis*, podemos seleccionar al azar algunas (Figura 9.1). No nos interesa el efecto de una localidad en particular, sino nos interesa saber si la variación de la biomasa de los macollos entre localidades es mayor que cero. Finalmente, la variabilidad en la biomasa de macollos a través de las localidades podría ser diferente entre individuos macho e individuos hembra. Para responder a estas preguntas se plantea un muestreo donde en varias localidades se cosechan individuos macho y hembra de *P. ligularis* (Figura 9.1).



(b)

Figura 9.1: (a) Muestreo de individuos de *Poa ligularis* en diferentes localidades elegidas al azar dentro de la zona de distribución de la especie (área aproximada). Cada chincheta indica una localidad, y en cada una de esas localidades se recolectaron varios individuos. Solamente se detallan los individuos recolectados en una de las localidades, reunidos por un corchete. Adaptado de Leva et al. 2013. (b) Dibujo de una planta de *Poa ligularis* y sus partes (fuente: www.darwin.edu.ar).

9.2. El modelo mixto y sus parámetros

El modelo lineal mixto puede ser expresado como extensión del modelo lineal presentado anteriormente (Cap. 1) al cual se le agregan los efectos aleatorios, es decir:

Variable respuesta = efectos fijos + efectos aleatorios + error,

donde la porción **efectos fijos** representa a la suma de componentes fijos del modelo, lo que incluye efectos de factores (variables categóricas o variables explicativas de clasificación), covariables (variables explicativas cuantitativas) e interacciones.

Por otra parte, el término **efectos aleatorios** representa la suma de todos los posibles efectos aleatorios (digamos *'B'*), excluido el término de error experimental (ε). Se asume que, tanto los efectos aleatorios (*B*) como los errores (ε) tienen distribución normal y son independientes entre sí. Esto se puede expresar de la siguiente manera:

$$\begin{cases} Y_{ijk} = \mu + \tau_i + B_j + \varepsilon_{ijk} \\ \varepsilon_{ijk} \sim N\left(0, \sigma^2_{\varepsilon}\right) \\ B_j \sim N\left(0, \sigma^2_{B}\right) \\ Cov\left(\varepsilon_{ijk}; B_j\right) = 0 \end{cases} \qquad (9.1)$$

donde *i* representa los niveles para los efectos fijos, *j* los posibles niveles para los efectos aleatorios y *k* a las observaciones dentro de un grupo *ij-ésimo*. En los modelos mixtos puede existir correlación distinta de cero entre los efectos aleatorios ***B*** al igual que entre los errores ε, es decir:

$$Cov\left(B_j; B_{j'}\right) \neq 0 \tag{9.2}$$

$$Cov\left(\varepsilon_{ij}; \varepsilon_{i'j'}\right) \neq 0 \tag{9.3}$$

Si evaluamos una variable que se mide sobre el mismo individuo a través del tiempo, es esperable que exista una correlación temporal entre los valores de los ε_{ij}. Además, es razonable pensar que las unidades experimentales dentro de un bloque son más homogéneas entre sí que entre bloques. Entonces, podría existir una correlación diferente de cero entre las observaciones de unidades experimentales dentro de un bloque (Cap. 3 y Cap. 8). Para estas estructuras de agrupamiento en los datos, los modelos mixtos permiten estimar la correlación entre efectos aleatorios y entre errores.

Ejemplo 9.1. (continuación). Modelo mixto para poblaciones de Poa ligularis

Primero, en un modelo lineal con efectos fijos, la observación de biomasa por macollo (Y) de cualquier individuo de la población de *P. ligularis* en una localidad en particular podría modelarse de la siguiente manera:

$$\begin{cases} Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N\left(0, \sigma^2_{\varepsilon}\right) \\ Cov\left(\varepsilon_{ij}; \varepsilon_{i'j'}\right) = 0 \end{cases} \tag{9.4}$$

donde τ_i indica el efecto fijo del sexo *i-ésimo* de la planta y ε_{ij} denota la variación en el peso de los macollos no explicada por sexo de los individuos con distribución de probabilidades Normal. Aquí *i* representa el sexo (*i*: m, h) y *j* a las observaciones dentro de cada sexo (*j*: $1\ldots n_{hembras}$; $1\ldots n_{machos}$). Además, como las observaciones son independientes, la covarianza (y por lo tanto la correlación) entre errores vale cero. Es decir que la diferencia entre el valor observado y el esperado para un macollo cualquiera *ij* no está asociada linealmente con la diferencia en cualquier otro *i'j'*.

Segundo, al considerar que los individuos fueron recolectados en diferentes localidades, la biomasa de macollos en cada observación en un modelo mixto puede modelarse como:

$$\begin{cases} Y_{ijk} = \mu + \tau_i + B_j + \varepsilon_{ijk} \\ B_j \sim N\left(0, \sigma_B^2\right) \\ \varepsilon_{ijk} \sim N\left(0, \sigma_\varepsilon^2\right) \\ Cov\left(B_j; \varepsilon_{ijk'}\right) = 0 \end{cases} \quad (9.5)$$

donde τ_i indica el efecto fijo del sexo sobre la biomasa de los macollos y ε_{ijk} denota el error aleatorio como en el caso anterior. Entonces, B_j es una cantidad de biomasa que varía entre localidades (*j*) que se incorpora a cada observación. El subíndice '*j*' puede ser Carro Quemado, Santa Rosa, Ñacuñán, Chos Malal, Alí Curá, Jacobacci, Rawson, etc. (Figura 9.1). No hay interés en estudiar el efecto de cada una de esas localidades en particular, pero sí en conocer la variación entre y dentro de localidades en la biomasa de macollos. Entonces, B_j puede considerarse una variable aleatoria, con una distribución de probabilidades que suponemos Normal, con media 0 y varianza σ^2_B. Si la varianza σ^2_B es pequeña, significa que existe poca variabilidad en la biomasa promedio de macollos de plantas de igual sexo '*i*' entre las localidades. Notar que aquí se hace referencia a la variabilidad entre localidades y no a la diferencia de promedios de biomasa entre dos localidades en particular. En otras palabras, significa que la biomasa promedio de las plantas de, digamos, Jacobacci podría ser más o menos similar a las de otra localidad tomada al azar.

Con este modelo podemos poner a prueba algunas hipótesis. Por ejemplo, para evaluar si las plantas *P. liguaris* generan macollos con diferente biomasa entre machos y hembras, la hipótesis estaría referida a τ_i (hembra / macho). En cambio, para poner a prueba si existe variabilidad entre localidades en la biomasa promedio de macollos, la hipótesis asociada refiere a σ^2_B. Finalmente, los individuos de *P. ligularis* de una localidad podrían ser más parecidos entre sí que con los individuos de otras localidades. Esto involucra otros parámetros estadísticos que pueden estimarse en estos modelos: las covarianzas entre valores de biomasa de los individuos de *P. ligularis* a recolectar en un mismo ambiente.

Para estimar estos parámetros y conocer mas sobre la ecología de las poblaciones de *P. ligularis*, se realizó un muestreo en 6 localidades (Figura 9.1; Leva 2010, Leva *et al.* 2013). Primero, se halló que el peso promedio de los macollos macho fue significativamente mayor al peso promedio de los individuos hembra ($F_{1,160}$ = 10,16; P = 0,002). Mientras que el peso de los macollos de plantas hembra observado fue en promedio de 1,197 g, el peso promedio de los macollos de plantas macho observado fue en promedio 1,365 g (Figura 9.2). Estos resultados soportan la idea de los investigadores sobre la asignación diferencial de recursos entre plantas macho y hembra.

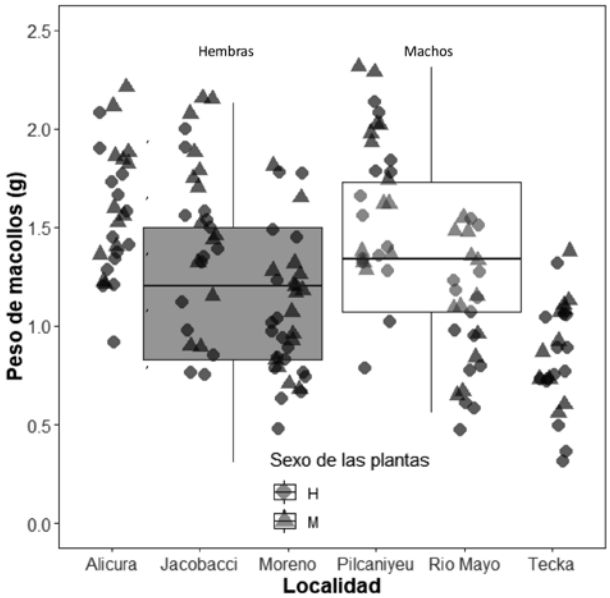


Figura 9.2: Peso de macollos de *Poa ligularis* recolectados en diferentes localidades de Patagonia. Las cajas presentan la comparación entre hembras (gris) y machos (blanco), e indican los dos cuartiles centrales del conjunto de datos. Los puntos son cada una de las observaciones de las plantas hembra (círculos) y macho (triángulos). Modificado de Leva 2010 y Graff *et al.* 2013

Por otra parte, el peso medio de los macollos obtenido a partir de la muestra de plantas de Jacobacci fue de 1,457 g, mientras que el peso medio de los macollos de todas las plantas muestreadas fue de 1,275 g. La cantidad 0,182 g (diferencia entre 1,457 y 1,275), está relacionada con la variable aleatoria *B*, "*efecto de la localidad*". Se supone que los B_j son independientes, con esperanza cero y varianza σ_B^2. Los estimadores de σ_B^2 y de σ_ε^2 son:

$$\hat{\sigma}_B^2 = 0{,}096 \text{ g}^2 \qquad \hat{\sigma}_\varepsilon^2 = 0{,}116 \text{ g}^2$$

Transformando la varianza a desvío estándar, los estimadores para las localidades y el error son 0,311g y 0,341g por macollo, respectivamente. Los intervalos de confianza fueron 0,163-0,595 g para el desvío estándar por localidad y 0,305-0,380 g para el desvío estándar del error aleatorio. El primer intervalo de confianza no incluye al cero, lo que sugiere la existencia de variabilidad en el peso de macollos entre las localidades. Cabe mencionar que estos resultados no se restringen a las seis localidades estudiadas sino a la población de localidades en el área de distribución de *Poa*, de la cual se extrajo una muestra aleatoria de seis localidades.

Los modelos mixtos también pueden incluir interacciones, pero su interpretación es diferente a la de los modelos con factores fijos exclusivamente (Schielzeth y Nakagawa 2013). Al considerar el ejemplo de *P. ligularis*, el sexo de las plantas es un factor fijo (τ_i) y la localidad es un factor aleatorio (B_j). Para el término de interacción (τB_{ij}) la hipótesis nula postula que la variabilidad entre localidades es idéntica entre machos y hembras. En términos generales, esto sugiere que la varianza asociada a B es la misma a través de los niveles del factor τ. Esta interpretación de la interacción difiere de la presentada anteriormente (Cap. 5), en donde para el caso de dos factores fijos, la hipótesis nula para la interacción postula que la respuesta media a los distintos niveles de un factor no difiere entre niveles del otro factor. Un aspecto común en el componente de interacción de los modelos de efectos fijos y los modelos mixtos es que siempre se debe poner a prueba la presencia de interacción entre factores en primera instancia y, si no se detecta interacción significativa, se puede proceder al estudio de los efectos principales de los factores (Cap. 5).

9.3. Modelos de regresión mixtos

Como se indicó, los modelos mixtos no solo se aplican a casos con variables explicatorias cualitativas. La versatilidad de esta aproximación permite incorporar efectos aleatorios a modelos de regresión y análisis de covariables, como pendientes y ordenadas al origen aleatorias. El siguiente ejemplo detalla la aplicación de un modelo mixto con una variable predictora cuantitativa continua.

Ejemplo 9.2. Producción de biomasa en pastizales

Si la productividad de los ecosistemas está limitada por la disponibilidad de recursos, como agua y nutrientes, se espera que su agregado altere las tasas de los procesos biológicos como la producción de biomasa o la descomposición. Supongamos que se diseñó un experimento para evaluar cómo responde la producción de biomasa de los pastizales pampeanos al agregado de nitrógeno. Para que las conclu-

siones abarquen a los "pastizales pampeanos", este experimento muestreó al azar sitios en toda la extensión de este territorio. Fueron contactados 10 investigadores en universidades e institutos de investigación en Corrientes, Entre Ríos, La Pampa, Buenos Aires y Santa Fe (algunas provincias tienen más de un sitio experimental), y se les propuso establecer un experimento coordinado. Sobre la base de un sorteo al azar de ubicaciones dentro de cada provincia, cada grupo se encargó de instalar el mismo experimento en un sitio. Como fueron sorteados, cada uno de los sitios puede considerarse como un bloque al azar dentro de los pastizales pampeanos. En cada uno de esos bloques, el experimento consistió en agregar dosis crecientes de nitrógeno (0, 2, 4, 6, 8 y 10 g N m^2 $año^{-1}$) en parcelas de 20 x 20 m. En estas parcelas se cosechó un área pequeña cada año, en el momento de máxima acumulación de biomasa (cambia según el pastizal). La biomasa verde acumulada en el pico de biomasa se relaciona estrechamente con la productividad primaria neta aérea. Recibimos los datos y nos disponemos a analizarlos (Figura 9.3). Fijamos el valor de alfa en 0,05, lo que implica que de antemano la probabilidad de rechazar equivocadamente una hipótesis nula verdadera es del 5%.

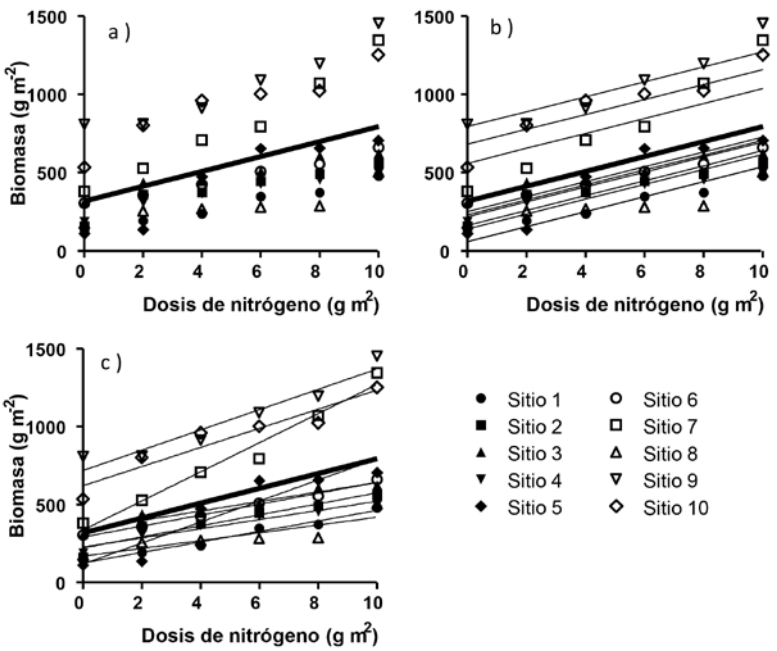


Figura 9.3: Respuesta en la producción de biomasa al agregado de diferentes dosis de nitrógeno en pastizales Pampeanos. a) Modelo lineal que considera el efecto fijo de la dosis de nitrógeno. b) Modelo mixto que considera el efecto fijo de la dosis de nitrógeno sobre la producción de biomasa e incorpora a la ordenada al origen como un efecto aleatorio. c) Modelo mixto que incluye efecto aleatorio para la ordenada al origen y para la pendiente entre sitios

Primero, según lo que aprendimos en capítulos anteriores (Cap. 6 y Cap. 7), analizamos estos datos con un modelo de regresión lineal simple (Figura 9.3.a). Según este modelo, la producción de biomasa a observar en la parcela *'i'* es igual a la suma la producción de biomasa sin agregado de nitrógeno (β_o), más el producto entre cambio en la biomasa por cada unidad de nitrógeno que se agrega (β_1) y la dosis correspondiente, más la variación dada por aspectos no considerados y que corresponden al error experimental (ε_i)

$$\begin{cases} Y_i = \beta_0 + \beta_1 Dosis_i + \varepsilon_i \\ \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right) \\ Cov\left(\varepsilon_i; \varepsilon_{i'}\right) = 0 \end{cases} \quad (9.6)$$

donde *i*:1... n. Con cualquier programa estadístico, la ordenada al origen se estima en 317,32 g · m^{-2}, y la pendiente en 47,69 g m^{-2} · gN^{-1} (Cuadro 9.1). El ANOVA para el modelo indica un $F_{1,58}$ = 20,89 y un valor p< 0,001, mucho menor que el alfa fijado de antemano de 0,05, por lo que rechazamos la hipótesis de que el agregado de nitrógeno no afecta la producción de biomasa. El coeficiente de determinación resultó en R^2 = 0,26 y el el cuadrado medio del error resultó en 76.198 g^2 · m^{-4}. Este último es el estimador insesgado de la varianza del error experimental (Cuadro 9.1).

Cuadro 9.1: Estimaciones puntuales para la media y la varianza de la ordenada al origen, la pendiente y el error en los tres modelos propuestos en el texto. a) Modelo con pendiente y ordenada al origen fijas. b) Modelo con pendiente fija y ordenada al origen aleatoria. c) Modelo con pendiente y ordenada al origen aleatorias. F/A indica si el término es fijo (F) o aleatorio (A)

	Pendiente y ordenada al origen fijas		Pendiente fija y ordenada al origen aleatoria		Pendiente y ordenada al origen aleatorias	
	F/A	Estimadores	F/A	Estimadores	F/A	Estimadores
Ordenada al origen	F	$\hat{\beta}_0$ 7,33 g · m^{-2}	A	$\hat{\beta}_0$ = 317,33 g · m^{-2} $\hat{\sigma}^2_{\beta_0}$ = 71.395 $g^2 \cdot m^{-4}$	A	$\hat{\beta}_0$ = 317,33 g · m^{-2} $\hat{\sigma}^2_{\beta_0}$ = 71.395 $g^2 \cdot m^{-4}$
Pendiente	F	$\hat{\beta}_1$ = 47,69 g · $m^{-2} \cdot g^{-1}$	F	$\hat{\beta}_1$ = 47,69 g · $m^{-2} \cdot g^{-1}$	A	$\hat{\beta}_1$ = 47,69 g · $m^{-2} \cdot g^{-1}$ $\hat{\sigma}^2_{\beta_1}$ = 461,12 $g^2 \cdot m^{-4} \cdot g^{-2}$
Error	A	$\hat{\sigma}^2_\varepsilon$ = 76.198 $g^2 \cdot m^{-4}$	A	$\hat{\sigma}^2_\varepsilon$ = 9.726 $g^2 \cdot m^{-4}$	A	$\hat{\sigma}^2_\varepsilon$ = 3.789 $g^2 \cdot m^{-4}$

De hecho, la Figura 9.3.a muestra que algunos símbolos (sitios) se encuentran sistemáticamente por encima (ver cuadrados abiertos, Sitio 7). Toda esa información sobre diferencias positivas en la producción se encuentra considerada en la varianza del error experimental. Si incorporamos el sitio como un factor fijo, tendríamos 10 parámetros más para estimar (uno por cada sitio). ¿Nos interesan particularmente los sitios? ¿Cómo podemos incorporar esa información a nuestro modelo?

Sin embargo, el modelo anterior es incorrecto, ya que no considera el hecho de que las observaciones provenientes de un mismo bloque (*i.e.* localidad) no son independientes. En cambio, los modelos mixtos admiten incorporar a las localidades como un efecto aleatorio y asumir una correlación entre las observaciones dentro de una localidad (ver abajo CCI). Un primer paso sería considerar el efecto del sitio, como una diferencia promedio en la producción que varía aleatoriamente de sitio en sitio, independientemente de la dosis (Figura 9.3b). En este modelo, tenemos:

$$\begin{cases} Y_i = \beta_0 + \beta_1 Dosis_i + B_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right) \\ B_j \sim N\left(0, \sigma_B^2\right) \\ Cov\left(\varepsilon_{ij}; B_j\right) = 0 \end{cases} \tag{9.7}$$

donde la producción de biomasa en la dosis *i*-ésima del sitio '*j*', es la suma de la biomasa promedio de todos los sitios sin fertilizante, más el producto entre el cambio en la biomasa promedio cuando la dosis de fertilizante aumenta una unidad por la dosis correspondiente aplicada, más la suma de dos términos aleatorios con distribución de probabilidades normal: el error aleatorio (ε_{ij}) y el cambio en la biomasa dado por el sitio (B_j). Es decir que cada sitio tomado al azar, suma o resta un valor de biomasa que es independiente de las dosis.

En forma similar a lo presentado en la ecuación 7.2, al reordenar la primera ecuación del modelo propuesto en 9.7 se obtiene:

$$Y_i = (\beta_0 + B_j) + \beta_1 Dosis_i + \varepsilon_{ij} \tag{9.8}$$

Ahora, la ordenada al origen es la suma de (β_o+B_j), o sea la suma de una constante más una variable aleatoria (ver Batista 2018, Pag 88: *Funciones lineales de variables aleatorias*). Indicado así, cada sitio tiene la misma

pendiente (β_1), pero su ordenada al origen es una variable aleatoria que tiene esperanza igual a β_o y varianza σ^2_B. Es importante recordar que la esperanza de B_j es cero, y su varianza, al ser diferente de cero, genera variabilidad en la biomasa promedio entre sitios. Este modelo mixto se conoce como "de intercepto y pendiente aleatoria".

Las estimaciones de la ordenada al origen y la pendiente no variaron respecto del primer modelo (Cuadro 9.1). Además, se obtuvieron las estimaciones para la varianza del error (= 9.726 $g^2 \cdot m^{-4}$) y la varianza de la ordenada al origen (σ^2_B = 71.395 $g^2 \cdot m^{-4}$). Acá sí hay diferencias con el primer modelo. Parte de la varianza que era del error, ahora es la varianza asociada a los sitios. Al observar la Figura 9.3.b, se nota que la mayor variabilidad está entre los sitios, ya que la dispersión de los puntos dentro de un sitio alrededor la estimación de la recta correspondiente es muy poca.

Un aspecto interesante surge al evaluar la varianza y covarianza de la variable respuesta. Intuitivamente, se puede concebir que los valores de producción de biomasa dentro de un sitio pueden estar correlacionados. De hecho, si la varianza explicada por los sitios es mucho más grande que la varianza del error, uno puede esperar que la correlación entre observaciones dentro de un sitio sea más alta. A este término se lo denomina coeficiente de correlación intraclase (CCI), y es el cociente entre la varianza asociada a los sitios respecto de la variación total. Es decir:

$$CCI = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_\varepsilon} \tag{9.9}$$

Para nuestro ejemplo, la correlación intraclase se estimó de la siguiente manera:

$$\widehat{CCI} = \frac{\hat{\sigma}^2_B}{\hat{\sigma}^2_B + \hat{\sigma}^2_\varepsilon} = \frac{71.395}{71.395 + 9.726} = 0,88$$

Un valor alto indica que hay una correlación muy alta entre las observaciones dentro un nivel del factor aleatorio, en este caso entre parcelas de una misma localidad. Visto de otra manera, indica que sería posible predecir los valores dentro de un sitio a partir de otras observaciones dentro del mismo sitio. Notar lo valioso de conocer esta y otras correlaciones en ciencias ambientales y agropecuarias (ver Capítulo 10).

Como tercer y ultima aproximación, podemos incorporar la posibilidad de que las pendientes varíen entre sitios (Figura 9.3.c). Esto es esperable, considerando que condiciones locales, como por ejemplo el tipo de vegetación, la fertilidad inicial, la capacidad de retención hídrica del suelo o el régimen de precipitaciones, puedan generar una respuesta diferencial al nitrógeno en cada sitio. No contamos con esa información biofísica, pero se puede indicar en el modelo que la pendiente cambie para cada sitio, de manera aleatoria. De ser así, el modelo debería incluir otro término aleatorio (aparte del error aleatorio y de la ordenada al origen aleatoria), pero que esté vinculado a la pendiente. Esto puede expresarse así:

$$\begin{cases} Y_i = \beta_0 + U_J \cdot Dosis_i + B_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right) \\ B_j \sim N\left(0, \sigma_B^2\right) \\ U_j \sim N\left(\beta_1, \sigma_U^2\right) \\ Cov\left(\varepsilon_i; B_j\right) = 0 \end{cases} \tag{9.10}$$

donde los dos términos aleatorios tienen distribución de probabilidad normal, descriptas por su media y varianza. Destacamos la distribución de probabilidad de la pendiente U_j, que a diferencia del error aleatorio y de la ordenada al origen, tiene una esperanza igual β_1.

Si pensamos que U_j puede escribirse como una suma $\left(\beta_1 + T_j\right)$, el mismo modelo en 9.10 podría reordenarse y escribirse como

$$\begin{cases} Y_i = (\beta_0 + B_j) + \left(\beta_1 + T_j\right) \cdot Dosis_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right) \\ B_j \sim N\left(0, \sigma_B^2\right) \\ T_j \sim N\left(0, \sigma_T^2\right) \\ Cov\left(\varepsilon_i; B_j\right) = 0 \end{cases} \tag{9.11}$$

Queda claro que la pendiente y la ordenada al origen varían en forma aleatoria entre sitios. Además, el valor esperado para la pendiente en el conjunto de

sitios debería ser numéricamente igual a la pendiente en los modelos anteriores. Al observar la Figura 9.3.c esta afirmación toma sentido, pero no debemos olvidar que la muestra que obtuvimos es una de todas las posibles. Lo que estimaremos es la variación en la pendiente, y no particularmente esas líneas (que son una realización de un valor de la variable). Este modelo mixto se conoce como "de intercepto y pendiente aleatoria".

Las estimaciones de los parámetros para este nuevo modelo se encuentran en el Cuadro 9.1. Las estimaciones de la esperanza de la pendiente ($\hat{\beta}_1$ = 47,69 $g \cdot m^{-2} \cdot g^{-1}$) y de la ordenada al origen ($\hat{\beta}_0$ = 317,33 $g \cdot m^{-2}$) resultaron numéricamente iguales a las de los modelos anteriores, soportando lo dicho anteriormente. La varianza de la pendiente fue estimada en $\hat{\sigma}^2_{\beta_1}$ = 461,12 $g^2 \cdot m^{-4} \cdot g^{-2}$. La estimación de la varianza del error fue de $\hat{\sigma}^2_{\varepsilon}$ = 3.789 $g^2 \cdot m^{-4}$. Es interesante notar que las estimaciones de la varianza del error fueron reduciéndose en valor en la medida que se incorporaron otros términos aleatorios al modelo. Dicho de otra manera, se pudo explicar parte del error experimental con términos que se incorporaron al modelo.

Caja 9.1: Más de modelos mixtos para curiosos

Los aspectos teóricos más profundos, los detalles metodológicos y las formas de comparación y evaluación de modelos exceden la profundidad de este texto. Sin embargo, tres puntos merecen especial atención al momento de abordar problemas con modelos mixtos (Pinheiro & Bates 2000, Zuur *et al.* 2009, Crawley 2013, Garibaldi *et al.* 2014).

En primer lugar, se utilizan frecuentemente matrices para plantear y analizar a los modelos estadísticos (Pinheiro y Bates 2000). Por ejemplo, en este libro se presentó una primera aproximación matricial a la regresión múltiple (Cap. 6). Este tipo de representación aplicada a modelos mixtos permite vincular en forma elegante cada observación con su componente fijo y aleatorio, pero además permite visualizar las posibles correlaciones entre estos componentes. Sin ser imprescindible, el álgebra permite una comprensión más profunda de las bondades de los modelos mixtos.

En segundo lugar, a diferencia de los modelos con efectos fijos de capítulos anteriores donde se utilizan cuadrados mínimos, en los modelos mixtos se utilizan métodos de estimación basados en máxima verosimilitud, particularmente máxima verosimilitud restringida (REML, por sus siglas en inglés: 'REstricted Maximum Likelihood'). Estas aproximaciones dificultan saber con precisión cuáles son los grados de libertad que se usan para poner a prue-

ba las hipótesis, y por eso deben ser tomados conociendo estas limitaciones (Pinheiro y Bates 2000, Zuur 2009, Garibaldi *et al.* 2014).

Finalmente, importa evaluar si la incorporación de más parámetros y la mayor complejidad mejoran sustancialmente la calidad predictiva o informativa del modelo. Para comparar dos o más modelos (Cap. 10), se utiliza generalmente el criterio de información de Akaike (AIC por sus siglas en ingles de '*Akaike Information Criterion*', Bunham y Anderson 2004; Anderson 2007). El AIC informa sobre el compromiso entre la complejidad del modelo y la bondad de ajuste. También se pueden calcular coeficientes (similares a los de determinación) que indican la proporción de la variación explicada por la parte fija y por la parte aleatoria (Nakagawa y Schielzeth, 2013).

9.4. Estructura de observaciones anidadas

Los diseños anidados son muy comunes en ciencias ambientales y agropecuarias. En un diseño anidado, cada nivel del factor anidado está asociado con un solo nivel del factor de jerarquía superior. Esto se diferencia de los experimentos con estructura factorial de tratamientos (Cap. 5), donde los niveles de un factor se 'cruzan' con los del otro (ver Schielzeth y Nakagawa 2013). Presentamos aquí un ejemplo de anidamiento.

Queremos comparar los resultados económicos de la adopción de estrategias de manejo adaptativo en sistemas de producción ovina en Santa Cruz y Chubut. Seleccionamos al azar cuatro departamentos en cada provincia, y obtenemos información económica (*i.e.* rentabilidad de las empresas) de 3 estancias al azar en cada departamento. Mientras que el departamento de Lago Buenos Aires está en Santa Cruz, el de Río Senguer se encuentra en Chubut. Es simple, no hay un "Río Senguer" en la provincia de Santa Cruz. Tampoco existe la misma estancia en cada Departamento. Al delinear un muestreo por estancias en departamentos y departamentos en las dos provincias, existe un anidamiento insoslayable dado por la naturaleza del sistema de estudio. Es interesante notar que a pesar de que la intención es comparar los resultados de rentabilidad entre Santa Cruz y Chubut, la información provista por las estancias dentro de los departamentos en cada una de las provincias resulta muy valiosa. Es importante considerar cuáles y cuántas son las fuentes de información independiente en cada nivel de la jerarquía (Hulbert 1984).

Un modelo con estructura de observaciones anidadas puede escribirse de la siguiente manera:

$$\begin{cases} Y_{ijk} = \mu + \tau_i + B_{(i)j} + \varepsilon_{(ij)k} \\ \varepsilon_{(ij)k} \sim N\left(0, \sigma^2_{\varepsilon}\right) \\ B_{(i)j} \sim N\left(0, \sigma^2_{B|\tau}\right) \\ Cov\left(\varepsilon_{(ij)k}; B_{(i)j}\right) = 0 \end{cases} \quad (9.12)$$

Donde Yijk es la rentabilidad de las empresas agropecuarias (%), τ_i es el efecto fijo del factor τ (e.g. Provincia) y $B_{(i)j}$ es el efecto aleatorio del nivel *j*-ésimo dentro de cada nivel de τ (Departamentos dentro de Provincia; indicado por el paréntesis: '*dentro de i*'). El error corresponde a la variabilidad entre las estancias dentro de un dado departamento y provincia. Con este diseño, las conclusiones pueden aplicarse a todas las estancias y todos los departamentos de ambas provincias.

Algunos autores denominan a estos modelos como jerárquicos o multinivel (Gelman y Hill 2007), debido a que existe una jerarquía estructural entre las unidades [biológicas, ambientales, sociales] que producen la información. Mas allá de este caso particular, existen diversas formas de anidamiento en términos de tipos de factores (fijos o aleatorios) y agrupamiento (factores cruzados, semi-cruzados o anidados); (Quinn y Keough 2002, Schielzeth y Nakagawa 2013).

Ejemplo 9.3. Rasgos funcionales de pastos nativos y exóticos

Dos grandes teorías intentan explicar la persistencia de especies exóticas en sistemas invadidos. Una sugiere que las especies exóticas son diferentes a las nativas en sus rasgos (e.g. altura de la planta, contenido de N en hoja, tasa de fotosíntesis, etc), mientras que la otra propone que son funcionalmente similares (Tecco *et al.* 2010). Resulta entonces interesante comparar grupos funcionales de plantas nativas y exóticas. Para evaluar estas ideas, se recolectaron hojas de ejemplares pertenecientes a especies de pastos nativos y exóticos en pastizales abandonados en la Pampa Interior (Cuadro 9.2). Tres especies nativas y tres especies exóticas fueron seleccionadas al azar de una lista de especies abundantes en estudios previos para la región. Para este ejemplo particular, podemos considerar que las especies son un factor "aleatorio", a pesar de su reducido número. El muestreo consistió en recorrer diferentes pastizales en la Pampa In-

terior, seleccionando al azar individuos de cada una de las seis especies. En cada individuo se midió la altura de inserción de la panoja (cm) y se recolectaron dos hojas por planta. En el laboratorio, midió el área foliar específica (superficie de tejido por unidad de masa seca de hoja, $mm^2\ mg^{-1}$) de cada una de las dos hojas que se recolectaron de diez individuos de las seis especies. El Cuadro 9.2 presenta los promedios por origen y especie para este conjunto de datos.

Cuadro 9.2: Promedio de del contenido de materia seca y del área foliar específica para pastos nativos y exóticos. El promedio de cada especie se generó a partir de valores provenientes de 2 hojas de individuos recolectados en 10 pastizales abandonados

Grupo	Especie	Área foliar específica ($mm^2\ mg^{-1}$)	Altura de la panoja (cm)
Nativos	*Brisa subaristata*	12,889	52,8
	Bromus unioloides	15,410	106,0
	Melica brasiliana	22,565	61,7
	Promedio	16,955	73,5
Exóticos	*Dactylis glomerata*	13,370	119,0
	Festuca arundinacea	10,545	136,2
	Phalaris aquatica	17,288	140,0
	Promedio	13,734	131,7

¿Cuáles son las fuentes de información independiente para comparar pastos nativos vs. pastos exóticos? ¿Cada hoja? ¿Cada individuo? ¿Cada especie? Una opción de comparación es promediar todas las hojas para pastos nativos y para pastos exóticos y hacer una prueba t-Student. Sin embargo, la información aportada por los individuos 'dentro' de una especie de pasto no es completamente independiente a la hora de comparar pastos nativos y exóticos.

En el caso del modelo para el área foliar específica de las hojas, el modelo mixto que considera la estructura de anidamiento dada por el muestreo podría escribirse así:

$$\begin{cases} Y_{ijkl} = \mu + \tau_i + B_{(i)j} + C_{(ij)k} + \varepsilon_{(ijk)l} \\ \varepsilon_{(ijl)k} \sim N(0, \sigma^2_{\varepsilon}) \\ C_{(ij)k} \sim N\left(0, \sigma^2_{C|B|\tau}\right) \\ B_{(i)j} \sim N\left(0, \sigma^2_{B|\tau}\right) \\ Cov\left(C_{(ij)k}; \varepsilon_{(ijk)l}\right) = 0 \\ Cov\left(B_{(i)j}; C_{(ij)k}\right) = 0 \\ Cov\left(B_{(i)j}; \varepsilon_{(ijk)l}\right) = 0 \end{cases} \quad (9.13)$$

donde Y_{ijkl} es el área foliar específica a observar en la hoja '*l*', de la planta '*k*', de la especie '*j*', que corresponde al origen '*i*' (nativo / exótico). También se detallan las distribuciones de probabilidades de los efectos aleatorios correspondientes a las *j*-especies dentro del origen (nativo exótico) y a los *k*-individuos dentro de las especies ($B_{(i)j}$ y $C_{(ij)k}$ respectivamente) y la distribución de probabilidades del error experimental ($\varepsilon_{(ijk)l}$). En este ejemplo, el error experimental está asociado a las variaciones entre las dos hojas dentro de cada individuo. Invitamos al lector a escribir el mismo modelo para el análisis de la altura de las plantas, en dónde se realizó una sola medición por individuo, y no hay dos lecturas por individuo de cada especie.

¿Resultados? Sobre la base de este conjunto de datos con dos hojas, de diez individuos de seis especies de pastos (¡de estación fría!), no tenemos evidencias suficientes para rechazar la hipótesis que dice que el área foliar específica promedio de las hojas no difiere entre estos dos grupos (alfa = 0,05; valor p = 0,4091). Al igual que en cualquier prueba de hipótesis, tampoco podríamos aseverar sin error que el área foliar específica promedio es igual entre pastos de invierno nativos y pastos de invierno exóticos (revisar Error tipo II en Capítulo 1).

¿La comparación de promedios es el único resultado? No. El modelo también brinda información sobre la varianza dada por las especies nativas y exóticas, y para la varianza originada por la variabilidad de los individuos '*dentro*' de una especie. En nuestro ejemplo, el estimador insesgado de la varianza entre las especies fue de $\hat{\sigma}^2_{(i)j} = 17{,}31\ mm^4 \cdot mg^{-4}$, la varianza estimada para los individuos dentro de las especies fue de $\hat{\sigma}^2_{(ij)k} = 5{,}92\ mm^4 \cdot mg^{-4}$, mientras que el del error experimental que mide la variabilidad entre hojas de un mismo individuo fue de $\hat{\sigma}^2_{(ijk)l} = 8{,}76\ mm^4\ mg^{-2}$. Con estos resultados, se evidencia por un lado que la variabilidad dada por las especies fue más grande que la de individuos dentro de las especies. Esto tiene cierta lógica, al pensar la comparación entre especies y dentro de las especies. Por el otro lado, la variabilidad entre las dos hojas tomadas del mismo individuo es muy similar a la dada por los individuos dentro de una misma especie.

Invitamos al lector a conducir un ANOVA con los promedios por especie brindados en el Cuadro 9.2. Es decir, usando cada especie como una repetición para el factor "origen" con dos niveles (nativo/exótico). Los resultados del valor p son idénticos para el cuadrado medio del error y muy similares para el estimador de la varianza de especies (salvando el redondeo y el diferente método de estimación; recordar elevar al cuadrado o usar raíz, según corresponda). Entonces, ¿qué ganamos por sobre el ANOVA? Primero, comprender cuáles son las fuentes de información que debemos usar para cada pregunta que queramos responder. Hay un nivel de aleatorización y repetición que podría responder a la pregunta sobre el origen (*i.e.* las especies) y otro nivel de aleatorización que podría responder a las preguntas

para comparar especies (*i.e.* individuos por especie). Segundo, si el lector realiza el ANOVA para la variable altura considerando cada especie como una repetición, encontrará que existen diferencias significativas en la altura, pero no tiene información acerca de la variabilidad entre individuos incluidos dentro del nivel de especie. No se trata solamente de un valor p para comparar promedios: la varianza y la covarianza son fuentes de información muy valiosas. En este ejemplo vimos que la varianza dada por las "especies" es más importante que la dada por los "individuos" para cada especie.

9.5. Medidas repetidas en el tiempo

Esta sección pone el foco en los experimentos dirigidos a evaluar y comparar tendencias de la variable respuesta en el tiempo a partir del registro de observaciones repetidas en la misma unidad experimental a través del tiempo. Estos experimentos son denominados experimentos con datos longitudinales o curvas de crecimiento, y los registros a obtener en una misma unidad experimental no son independientes. En consecuencia, el análisis debe considerar la estructura de covariación entre observaciones dentro de las unidades experimentales. En estos experimentos, cada unidad experimental aporta un vector de observaciones a través del tiempo. En el Capítulo 10 se aborda el problema homólogo para observaciones repetidas en el espacio.

Ejemplo 9.4. Restauración de suelos degradados en las Sierras Grandes de Córdoba mediante reforestación con tabaquillo

Gran parte de los suelos de las Sierras Grandes de Córdoba, Argentina, se encuentra erosionada como consecuencia de los fuegos intencionales y la ganadería. Una opción para la restauración de estos ambientes degradados es la reforestación con especies arbóreas nativas como el tabaquillo (*Polylepis australis*) (Renison; *et al.* 2005). Se llevó a cabo un experimento manipulativo a campo a fin de identificar técnicas que permitiesen mejorar el crecimiento de plantines de *P. australis* en un área degradada y con pendiente de las Sierras Grandes. Particularmente, los investigadores estaban interesados en determinar la eficacia de la construcción de terrazas, que podrían actuar como barrera estabilizando el suelo y deteniendo los procesos de erosión. Se delimitaron 30 sitios en el área de estudio. En la mitad de ellos, elegidos al azar, se construyó una terraza de 60x60 cm, delimitada por piedras. En cada sitio se implantó un plantín producido mediante propagación vegetativa, de aproximadamente 8 cm

de altura. Al cabo de 1, 2 y 3 meses se midió la altura de cada plantín (desde su base hasta la yema apical más distante). En la Figura 9.4 se presentan los promedios de altura por tratamiento y tiempo para este conjunto de datos.

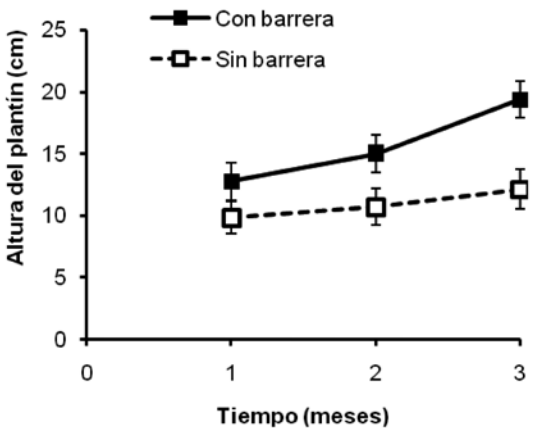


Figura 9.4: Variación de la altura de los plantines de *P. australis* a lo largo del tiempo en sitios con y sin barrera. Se muestra la media y el desvío estándar de 15 plantines

Se cuenta con 90 observaciones de altura, dadas por 3 mediciones secuenciales en cada uno de los 15 plantines asignados a cada uno de los dos tratamientos. Si bien existen dos factores (tratamiento y tiempo) que están cruzados y pueden afectar la altura de los plantines, no se trata de un diseño factorial clásico. Las 3 observaciones obtenidas en cada plantín no son independientes y probablemente estén afectadas por condiciones propias de cada individuo. Ignorar esa información llevaría a una sobreestimación del error residual, ya que incluiría la variación debida a los plantines. Como ya vimos en este capítulo, una alternativa es incluir en el modelo un factor aleatorio que induce una estructura de correlación entre las observaciones que comparten el mismo valor realizado del efecto aleatorio, en este caso el plantín. El modelo propuesto para este caso es:

$$\begin{cases} Y_{ijkl} = \mu + \tau_i + \beta_i + \tau\beta_{ij} + B_{(i)k} + \varepsilon_{ijk} \\ \varepsilon_{ijk} \sim N\left(0, \sigma_\varepsilon^2\right) \\ B_{(i)k} \sim N\left(0, \sigma_B^2\right) \\ Cov\left(\varepsilon_{ijk}; B_{(i)k}\right) = 0 \end{cases} \tag{9.14}$$

donde τ_i indica el efecto fijo del tratamiento (terraza) sobre la altura de los plantines, β_j el efecto fijo del tiempo, $\tau\beta_{ij}$ la interacción fija tratamiento-tiempo, $B_{(i)k}$ el efecto aleatorio del plantín y ε_{ijk} denota el error aleatorio.

Al ajustar el modelo con los datos se encontró una interacción significativa tratamiento-tiempo (valor p = 0,023). Por lo tanto, no se analizaron los efectos principales, *i.e.* el efecto del tratamiento independientemente del tiempo y viceversa. Se procedió a efectuar comparaciones *post-hoc* de interacción, detectándose una altura significativamente mayor de los plantines ubicados en sitios con terraza con respecto a los ubicados en sitios sin terraza a los dos y tres meses de iniciado el ensayo (valor p<0,05). Estos hallazgos proporcionarían evidencia de la eficacia de la construcción de terrazas en laderas de las Sierras de Córdoba para favorecer el crecimiento de plantines de *P. australis.* Puede, al igual que en los ejemplos vistos anteriormente, estimarse la varianza aportada por los plantines y la varianza residual, aunque estos parámetros son raramente de interés en este tipo de diseño.

Existen otras alternativas para modelar este tipo de datos como los llamados modelos marginales, que no incluyen efectos aleatorios, pero en los cuales se explicita la estructura de correlación de los errores dentro de cada individuo mediante una matriz de covarianza. Existen distintas estructuras para esta matriz, como simetría compuesta, auto-regresiva de orden 1, desestructurada, etc. Para una descripción más detallada de estos modelos ver Zuur *et al.* (2009).

9.6. Varianzas heterogéneas

Un supuesto central del ANOVA es que la varianza del error debe ser homogénea entre tratamientos (Cap. 4). En un modelo sencillo,

$$\begin{cases} Y_{ijk} = \mu + \tau_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N\left(0, \sigma^2_{\varepsilon}\right) \\ Cov\left(\varepsilon_{(ij)k}; \varepsilon_{(ij)k'}\right) = 0 \end{cases} \quad (9.15)$$

Una forma relajar este supuesto es considerar que los errores pueden modelarse como:

$$\varepsilon_{ij} \sim N\left(0, \sigma\ W_i\right) \quad (9.16)$$

donde W_i es un término que afecta la varianza, y que se puede estimar a partir de los datos. En otras palabras, las estructuras de varianzas permiten estimar varianzas diferentes para los niveles de un factor fijo (e.g. tratamientos) o covariables, y así resolver el problema de la heterocedasticidad. Solo detallaremos dos ejemplos de uso muy frecuente.

En el primero, la varianza es una función lineal de una covariable, y se lo llama de 'varianzas fijas'. En términos del modelo propuesto en la ecuación 9.16,

$$\varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon}^2 x_{ij}\right) \tag{9.17}$$

Por ejemplo, en la producción de animales bajo creep-fedding (complemento entre alimento balanceado y leche materna en ganado; ver ejemplo 7.1), podemos imaginar que la varianza en la ganancia de peso de terneros alimentados con dos tipos diferentes de suplementos podría ser una función lineal del peso inicial o de los litros de leche que produce la madre.

Notar que la variable respuesta es cuánto peso ganan los terneros como resultado de dos dietas diferentes, pero que la producción de leche o el peso inicial afectan la variabilidad en ese peso, y no necesariamente al valor promedio.

El segundo caso es el de varianzas por grupo o 'identidad' del tratamiento. En este caso, la varianza esta multiplicada por una constante que puede cambiar por estrato, grupo o tratamiento:

$$\varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon}^2 C_{ij}\right) \tag{9.18}$$

Un ejemplo podría ser la aplicación de fertilizante sobre el tamaño o biomasa de individuos de una población vegetal. El fertilizante podría aumentar la biomasa de los individuos dominantes, mientras que los subordinados son relegados. Esto hace que aumente la heterogeneidad con el fertilizante en comparación a un tratamiento sin fertilizante.

En cualquiera de los casos descriptos, una vez ajustado el modelo con varianzas heterogéneas, debe evaluarse el supuesto de que la varianza del error experimental ha sido apropiadamente ajustada (Zuur *et al.* 2009). Esto se refleja en un gráfico de residuales vs. predichos sin patrones ni efecto "embudo", como se presentó en el Cap. 4. Otros casos pueden emplearse cuando la varianza es una potencia de una covariable o función del exponente de una covariable.

Ejemplo 9.5. Leñosas en cultivos de verano

El establecimiento de malezas leñosas en sistemas cultivados puede estar controlado por la competencia que genera el cultivo y por el efecto del manejo agrícola del lote (fertilización, herbicidas, etc.). Si la sembradora o la fumigadora dejan áreas sin tratar, las leñosas podrían crecer más, lo que a su vez podría determinar que una cosechadora las esquivara. Este proceso reiterado genera un problema de lignificación a largo plazo.

Melina estaba interesada en este tema y estudió la importancia de la competencia generada por el cultivo de maíz sobre el tamaño de plantas de acacia negra (*Gleditsia triacanthos*). En un experimento a campo, manipuló el nivel de competencia sobre 36 arbolitos trasplantados al inicio del cultivo regulando la densidad de plantas de maíz en tres valores (1, 3.5 y 7 plantas m^{-2}) en 12 lotes diferentes por cultivo. Estos lotes no tenían un interés en particular y, al seleccionarlos al azar entre varios campos cercanos a Carlos Casares, Buenos Aires, los consideró como bloques aleatorios.

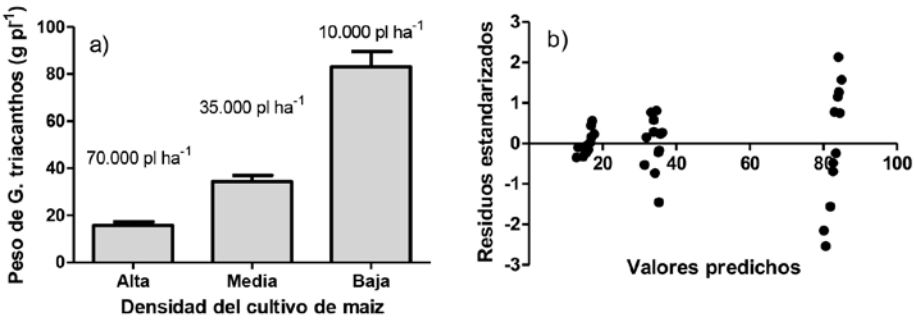


Figura 9.5: Resultados del experimento de remoción de plantas de maíz sobre la biomasa de individuos de *G. triacanthos*. El gráfico de la izquierda muestra los promedios y un error estándar de la media. El gráfico de la derecha muestra la dispersión de los residuales por tratamiento. Los puntos fueron levemente desplazados alrededor de los valores predichos (122,1g, 53,6g y 19,8g, respectivamente) por claridad

Según lo predicho, la biomasa promedio de acacia negra aumentó con la reducción en la densidad de plantas de maíz (Figura 9.5a). Las plantas de la leñosa pesaron en promedio menos de 20 gramos cuando la densidad de siembra fue de 7 plantas de maíz por m^2 (70.000 plantas por hectárea), y pesaron más de 80 gramos cuando la densidad fue de 1 planta de maíz por m^2 (ver resultados más abajo). Sin embargo, también se observó una mayor variabilidad en las plantas que crecieron a baja densidad de maíz (barras de error de la Figura 9.5a y 9.5b). Muchas de las diferencias observadas entre la biomasa de plantas

fueron mayores entre aquellas que crecieron en baja que en alta densidad del maíz. Esto es evidente al explorar los residuales (estandarizados en este caso). La Figura 9.5b muestra los residuales en función de los valores predichos (Ver Cap. 4). A continuación, exploramos los resultados de este análisis y la forma de resolver el problema de heterogeneidad de varianzas.

Aquí importa notar el estimador del desvío estándar ($\hat{\sigma}_{\varepsilon}$), que es la raíz cuadrada del estimador de la varianza del error en gramos por planta ($14,03\, g \cdot pl^{-1} = \hat{\sigma}_{\varepsilon} = \sqrt{\hat{\sigma}^2_{\varepsilon}}$). Es decir que la varianza del error se estimó en 196,84 g^2 pl^{-2}. Cuando se quiso poner a prueba la hipótesis nula que afirma que la densidad del cultivo de maíz no tiene un efecto sobre la biomasa promedio de las plantas de gleditsia se encontró con el problema de que aparentemente las varianzas son diferentes entre tratamientos. Como vimos anteriormente, la falta de cumplimiento de los supuestos genera cambios en las probabilidades de error (Tipo I y II), que varían según la robustez del diseño. En otros términos, los valores p obtenidos pueden estar distorsionados y generar conclusiones erróneas con más frecuencia. Por el otro lado, la existencia de varianzas diferentes entre tratamientos es un patrón interesante en sí mismo, y vale la pena modelarlo y describirlo. El siguiente paso es intentar resolver la heterogeneidad de varianzas. En el Capítulo 4 se indicó que podríamos cambiar la escala de observación transformando los datos. En este capítulo presentamos una solución alternativa que implica indicar en el modelo que los valores de varianza correspondientes a los tres tratamientos pueden diferir. Es decir que se estima una varianza común ($\hat{\sigma}_{\varepsilon}{}^2$) y se la multiplica por un coeficiente (W_i; ecuación 9.16), según a que grupo pertenezca. Entonces se estima la varianza para el tratamiento de referencia, y los factores indican cuantas veces más grandes o más chicas son las varianzas de los grupos restantes. El coeficiente para el primer grupo es siempre 1.

Para el nuevo modelo del ejemplo de acacia negra, el estimador de la varianza fue de $\hat{\sigma}^2_{\varepsilon}$ = 21,43 $g^2 \cdot planta^{-2}$. Esta estimación corresponde a varianza del grupo de alta densidad de plantas de maíz, es el numerador común de las varianzas para los otros grupos. Podemos llamarla $\hat{\sigma}^2_{Alta\ densidad}$. Para el grupo medio, el programa estadístico arrojo un factor de 2,03, por lo tanto, la varianza estimada para densidad media sería:

$$\hat{\sigma}^2_{Media\ densidad} = 2,03 \text{ x } \hat{\sigma}^2_{Alta\ densidad} \qquad (9.19)$$

numéricamente la estimación fue la siguiente:

$$\hat{\sigma}^2_{Media\ densidad} = 2,03 \text{ x } 21,43\ g^2 \cdot pl^{-2} = 43,58\ g^2 \cdot pl^{-2} \qquad (9.20)$$

Por lo tanto, el grupo de plántulas de acacia negra plantadas en cultivos de maíz con densidad media fue dos veces más variable que las plantas que se establecieron en alta densidad. Finalmente, la varianza estimada para el tratamiento de baja densidad de plantas de maíz fue

$$= 4{,}70 \text{ x } \hat{\sigma}^2_{Alta\ densidad} = 4{,}70 \text{ x } 21{,}43 \text{ g}^2 \cdot \text{pl}^{-2} = 100{,}75 \text{ g}^2 \cdot \text{pl}^{-2} \qquad (9.21)$$

Este modelo permitió ajustar la variabilidad de los residuos (Figura 9.6), poner a prueba la hipótesis de las diferencias de peso promedio de plantas de acacia negra bajo diferentes densidades de maíz y obtener información adicional con respecto al fenómeno de interés. Por un lado, para este nuevo modelo el valor P<0,001 permitió rechazar la hipótesis de que no hay diferencias en la biomasa de plantas de acacia negra bajo diferentes densidades de maíz. Por el otro, la reducción en la densidad de siembra no solo aumentaría la biomasa media de las plántulas de acacia negra, sino que además aumentaría también su varianza.

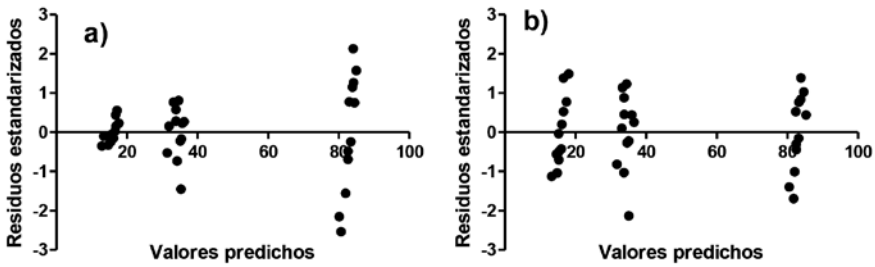


Figura 9.6: Residuos estandarizados para el modelo que considera una única varianza (a) y para el modelo que considera varianzas heterogéneas (b). La estandarización pesa a los residuales por la varianza estimada, según sea única o una por cada grupo

9.7. Consideraciones finales

De estos ejemplos de aplicación se desprende que los modelos mixtos son muy versátiles. Por un lado, en esta sección incorporamos nuevas herramientas que nos permiten explicar una mayor proporción de la variabilidad de un proceso de interés. Con algo de paciencia, podemos hacer que esta herramienta incluya diferentes parámetros que reflejen algún componente de la variación en los sistemas biológicos que sea de nuestro interés. Por el otro, tambien se indicó que un modelo mixto no es un modelo sin supuestos (Zuur *et al.* 2016), y esta herramienta no exime de la necesidad de validarlos antes de poner a prueba las hipótesis de interés. Numerosos aspectos de las ciencias agronómicas y ambientales utilizan modelos mixtos. En mejoramiento genético animal se modela la performance de un reproductor para un atributo de interés en función de efectos sistemáticos (como, por ejemplo, la edad a la que se toma la medición) y de su mérito genético. Los efectos sistemáticos son los efectos fijos del

modelo mixto y el mérito genético el efecto aleatorio. La matriz de covarianzas del modelo mide las correlaciones entre el mérito genético de diferentes individuos y sus elementos son una función del grado de parentesco entre ambos (Cantet *et al.* 2000). Bajo este modelo, las predicciones BLUP (por 'best linear unbiased prediction') constituyen una estimación del mérito genético de un reproductor y se utilizan como una herramienta de selección. En los catálogos de reproductores se las denomina valores de cría predichos. En la jerga de los mejoradores animales, se dice que el valor de cría no se 'estima', ya que no es un efecto fijo, sino que al ser un efecto aleatorio se lo 'predice'. En realidad, se estima el valor realizado de la variable aleatoria.En la producción de cultivos extensivos, se busca una combinación de genotipo y ambiente particular, y para ello se prueban tratamientos que combinan genotipos y ambientes como factores fijos. En forma alternativa, los modelos mixtos permiten estimar cuánta varianza está dada por los genotipos y cuánta se puede explicar por el ambiente (de la Vega y Chapman 2010). En estudios longitudinales de contaminación, el muestreo puede repetirse en el tiempo y considerar una jerarquía de muestreo. En relevamiento de recursos naturales, se debe evaluar cuáles son los criterios de agrupamiento al obtener información proveniente de la muestra obtenida durante el trabajo de campo (Manly 2009). Los modelos mixtos son una herramienta muy potente para abordar diseños con cierto tipo de organización jerárquica, como muestreos anidados, parcelas divididas o desbalance entre bloques.

Finalmente, la obtención de información y la profundización del conocimiento sobre un proceso no tienen costo cero. En modelos mixtos, el costo de modelar varianzas, correlaciones y otros términos no es 'gratis', y ese costo se 'paga' con grados de libertad. Modelos más complejos requieren mayor número de observaciones para generar estimaciones confiables. Por un lado, el costo de incluir más parámetros se refleja en la dificultad de cálculo, estimación e interpretación de los resultados. ¿Cuántos parámetros deben estimarse para modelar la correlación temporal y la heterogeneidad de varianzas de un experimento con factorial de 3 x 2 tratamientos en 10 fechas de medición? Muchísimos. ¿Cuál es el significado biológico, agronómico o ambiental de cada parámetro? No debemos dejar de hacer este tipo de preguntas cuando analizamos nuestros resultados. Por otro lado, el costo de tener más parámetros implica incluir una mayor cantidad de piezas de información independiente que den potencia a nuestras afirmaciones. Identificar las verdaderas piezas de información independiente sigue siendo crucial en estadística. Las repeticiones son tan necesarias y valiosas en estos modelos mixtos como en los modelos más simples presentados en el inicio de este texto y en capítulos previos. Un modelo complicado no salva un mal diseño o la falta de información.

EJERCICIOS DE APLICACIÓN

9.1. En los siguientes dos ejemplos, interprete cuáles son efectos fijos y aleatorios. Discuta posibles estructuras jerárquicas, medidas repetidas, fuentes de correlación u otros factores que deban ser considerados especialmente. Puede generar un modelo con subíndices para indicar los niveles. Plantee un posible croquis del diseño experimental y un esquema de posibles resultados.

a. Como miembro de una agencia estatal, Ud. está a cargo de entender cómo se modifica la carga de plomo en los arroyos tributarios del Río Matanza-Riachuelo según la distancia a la Ciudad de Buenos Aires. Para ello, realizó un muestreo de cinco localidades cualquiera: Cañuelas (■), Las Heras (▲), Marcos Paz (●), La Matanza (★), Esteban Echeverría (◆) (Figura. 9.7). En cada punto marcado en el mapa se tomó una muestra compuesta por tres probetas de 200 ml. El contenido de las tres probetas fue analizado químicamente en forma independiente. ¿Cómo puede diagramar el análisis de datos para evaluar los cambios en la concentración promedio de plomo de los arroyos subsidiarios del Rio Matanza como función de la distancia al puerto de Buenos Aires?

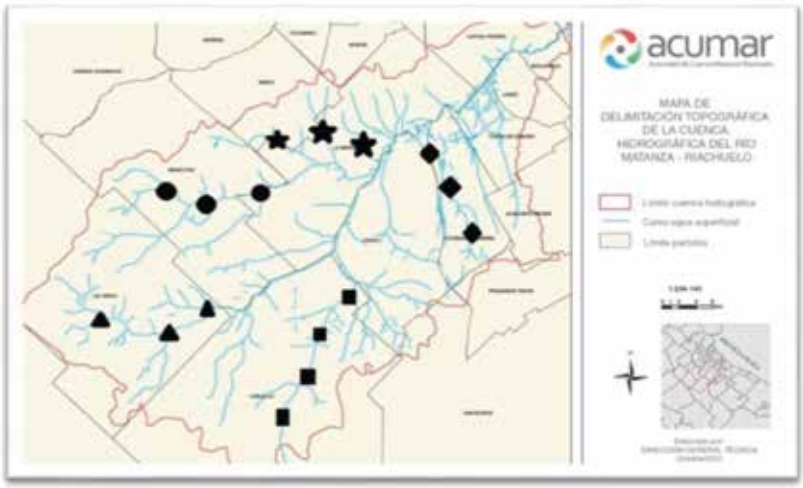


Figura E.9.1: Mapa de delimitación topográfica de la cuenca del Río Matanza-Riachuelo

b. Usted acaba de rendir su última materia y comienza a trabajar en el trabajo final de la carrera centrado en el proceso de sucesión secundaria en campos de cultivo. Recibe una base datos registrados durante 25 años en ocho lotes de 1 ha. Desea estudiar cómo cambia la diversidad florística a partir del cese de actividades agrícolas. Los lotes pertenecían a compañías diferentes y todos fueron usados bajo rotaciones Trigo/Soja-Maíz por más de 12 años antes del cese de actividades. Una vez abandonados, se realizaron muestreos anuales en cada uno de esos lotes. El

muestreo consistía en aleatorizar 20 marcos de $1m^2$ en cada lote, en los que se registraban todas las especies vegetales presentes. Con esta información se generaba un valor de diversidad de especies por cada cuadrante/marco de 1m2. ¿Cómo puede abordar el análisis de datos para conocer los cambios en la diversidad de especies en función del tiempo de abandono agrícola?

9.2. La selección de genotipos para la producción de girasol requiere un abordaje regional. Una interacción genotipo por ambiente muy fuerte podría ser un impedimento para el mejoramiento genético. Esta inquietud motivó un trabajo que Ud. recibe para revisar. El trabajo condensa una red de ensayos de un grupo de empresas. Con la descripción del modelo, el mapa y la tabla de estimaciones, responda las preguntas que siguen.

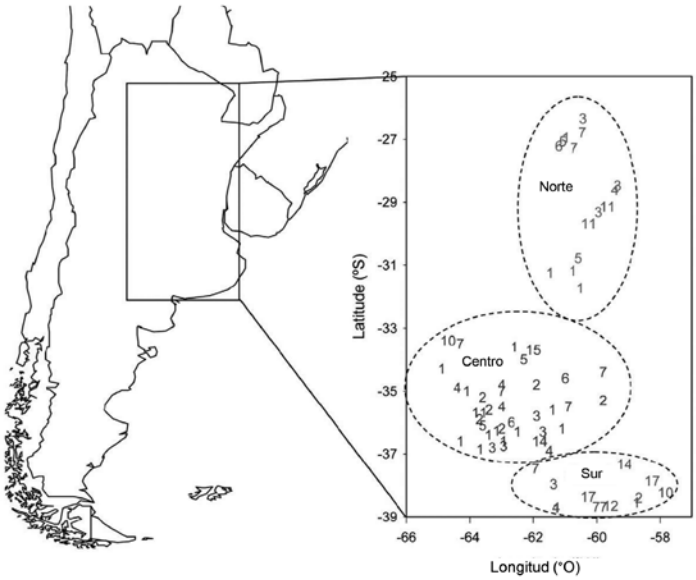


Figura E.9.2: Ubicación geográfica de la red de ensayos de evaluación de rendimiento aceitero del cultivo de girasol en tres zonas de producción de Argentina. Cada zona o mega-ambiente (Norte, Centro, Sur) reúne varias estaciones experimentales o localidad (números en las elipses) en las que se evalúan diferentes genotipos de girasol durante varios años. Los números indican la cantidad de años en que se llevaron a cabo ensayos en cada localidad. En general los genotipos no están replicados en los diferentes campos todos los años. Por otro lado, en cada estación experimental puede haber bloques. Adaptado de de la Vega y Chapman 2010. Crop Science, 50, 574-583

En la sección de métodos se explica el modelo estadístico que utilizaron:

"… tal que la observación del fenotipo y_{ijmn} en el hibrido i-ésimo en el bloque incompleto n-ésimo, del sitio m-ésimo en el ambiente j-ésimo fue modelado como: $y_{ijmn} = \mu + e_j + (r/e)_{jm} + (b/r/e)_{jmn} + gi + (ge)_{ij} + \varepsilon_{ijmn}$, donde μ es la media general; e_j

es el efecto fijo del ambiente j-ésimo; $(r/e)_{jm}$ es el efecto aleatorio del sitio m-ésimo en el ambiente j-ésimo [~N (0, σ^2_r)] con m = 1, …, r; $(b/r/e)_{jmn}$ es el efecto aleatorio del bloque n-ésimo, anidado en la réplica m-ésimo en el ambiente j-ésimo [~N (0, σ^2_b)] con n = 1,…,b; g_i es el efecto aleatorio del híbrido i [~N (0, σ^2_g)] , con i = 1,…,g; $(ge)_{ij}$ es el efecto aleatorio de la interacción entre el híbrido i y el ambiente j [~N (0, σ^2_{ge})]; y ε_{ijmn} es el efecto de error aleatorio para el híbrido i en el bloque n, de la réplica m en el ambiente j (error experimental [~N (0, $\sigma^2_{\varepsilon(j)}$)]."

Cuadro E.9.1: Estimación de los componentes de varianza para el rendimiento en aceite (kg ha-1) derivados de las tres zonas (e), híbridos (g), réplicas (r), bloques incompletos (b) y residuales (ε) en los ensayos coordinados. La última línea indica los rendimientos promedio de las regiones (kg ha-1)

Fuente de variación	Región Norte	Región Centro	Región Sur
r/e	4451	2671	1942
b/r/e	2543	0	4282
g	5542	10601	4877
ge	10632	14408	7527
ε	39519	45643	34568
Rendimiento	1136	1544	1375

a. Interprete los componentes del modelo. ¿Cuáles factores son considerados fijos y cuales son aleatorios? Indique por qué cree que es así.
b. Genere un croquis para una de las zonas, donde se indiquen los campos experimentales, las parcelas, los genotipos, etc.
c. Interprete los componentes de varianza del Cuadro 9.1
d. "Una interacción genotipo por ambiente muy grande podría ser un impedimento para el mejoramiento genético del cultivo en la región". Explique.
e. ¿Dónde está resumida la información de rendimiento los diferentes años? ¿Dónde está resumida parte de esa variabilidad?

9.3. Los grandes herbívoros, además de provocar efectos directos sobre los pastos forrajeros, también impactan de manera indirecta sobre la lignificación de los pastizales; efecto mucho menos estudiado que el anterior. Por un lado, el ganado puede atacar a las leñosas y limitar su expansión; por el otro, puede modular la competencia que les generan las plantas forrajeras vecinas. Melina J. diseñó un experimento factorial en bloques (Cuadro 9.2) para evaluar estos efectos en el contexto de su trabajo de tesis de final de carrera. Los factores estudiados fueron el nivel de competencia (tres niveles) y la defoliación de las plantas de gleditsia (dos niveles). Para este ensayo:

a. Genere un esquema del experimento.
b. Produzca un gráfico con los resultados del experimento. Interprete.
c. Utilizando un programa estadístico (e.g. InfoStat) analice estos datos a través de un modelo mixto con una estructura factorial de tratamientos y bloques como factor aleatorio.

d. Explore los residuales estandarizados (relativizados a la varianza, es una de las posibles opciones de residuales). Interprete.
e. Evalúe la necesidad de incorporar otros términos al modelo y presente los nuevos resultados. Compare e interprete los valores de AIC de los dos modelos. Los datos se presentan en el Cuadro E.9.2.

Cuadro. E.9.2: Biomasa de G. triacanthos después de tres meses creciendo en macetas de 20l. El diseño consistió en un experimento factorial con dos niveles de defoliación (sin corte –D; con corte quincenal +D) y tres niveles de competencia (sin pastura, pastura cortada con frecuencia quincenal, pastura intacta). '--' indica que la planta de gleditsia murió

	Sin pastura		Pastura cortada		Pastura intacta	
Bloque	**-D**	**+D**	**-D**	**+D**	**-D**	**+D**
1	15.017	1.946	1.534	0.794	--	--
2	22.571	1.939	0.778	--	0.274	0.137
3	17.409	1.939	1.555	0.151	0.352	0.152
4	21.599	1.947	0.508	--	--	0.232
5	23.083	1.942	1.495	0.173	0.243	0.201
6	21.877	1.947	0.692	0.060	0.212	--
7	45.847	1.942	1.543	--	0.359	--
8	26.924	1.944	1.943	0.197	--	0.166
9	10.486	1.940	1.517	0.070	0.578	--
10	25.970	1.936	1.820	0.090	0.556	0.112

9.4. En una actividad innovadora de cría de ñandúes se desea estimar el peso promedio individual (kg/animal) en distintos momentos. Para ello se toman 10 ñandúes machos y 10 hembras recién nacidos en nidos ubicados al azar en diferentes lotes de campos de la depresión del Salado. Cada individuo es pesado a los 10, 50 y 100 días de edad.
   a. ¿Considera que la información obtenida en las distintas fechas es independiente? ¿Por qué?
   b. Si su respuesta fue no, plantee los supuestos necesarios para satisfacer la independencia.
   c. Resuma esta información en un modelo mixto que considere el efecto de los diferentes factores. Recuerde incorporar la distribución de probabilidades de los factores aleatorios.
   d. Si los datos se hubiesen analizado ignorando la falta de independencia entre las observaciones ¿cómo cree que hubiese sido la varianza residual comparada con la del modelo mixto?