

BIOMETRÍA II

CLASE 1

EL MODELO LINEAL

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA



Efecto de la exposición postnatal a etanol sobre el volumen del cerebro en ratones

2

- La exposición intrauterina al etanol causa alteraciones cognitivas y conductuales persistentes
- Se desea estudiar los efectos neuroestructurales asociados a esta exposición en ratones
- 18 ratones de 7 días (equivalente al 3er trimestre de gestación en humanos) fueron divididos al azar en 3 grupos de igual tamaño. A cada grupo se le aplicó uno de los siguientes tratamientos: a) Solución salina, b) Etanol 1 g/kg, c) Etanol 2 g/kg. A los 82 días se determinó el volumen cerebral por resonancia magnética (en cm³)
 - Unidad experimental
 - Variable respuesta VR (Y)
 - Variable explicativa VE (X)
 - Réplicas

Modelo?

Modelos

3

- Simplificaciones de la realidad
- Todos los modelos son incorrectos...
- ...pero algunos modelos son más útiles que otros
- El modelo correcto no puede ser conocido con exactitud
- Cuanto más simple sea un modelo (menos parámetros), mejor (**Principio de parsimonia**)

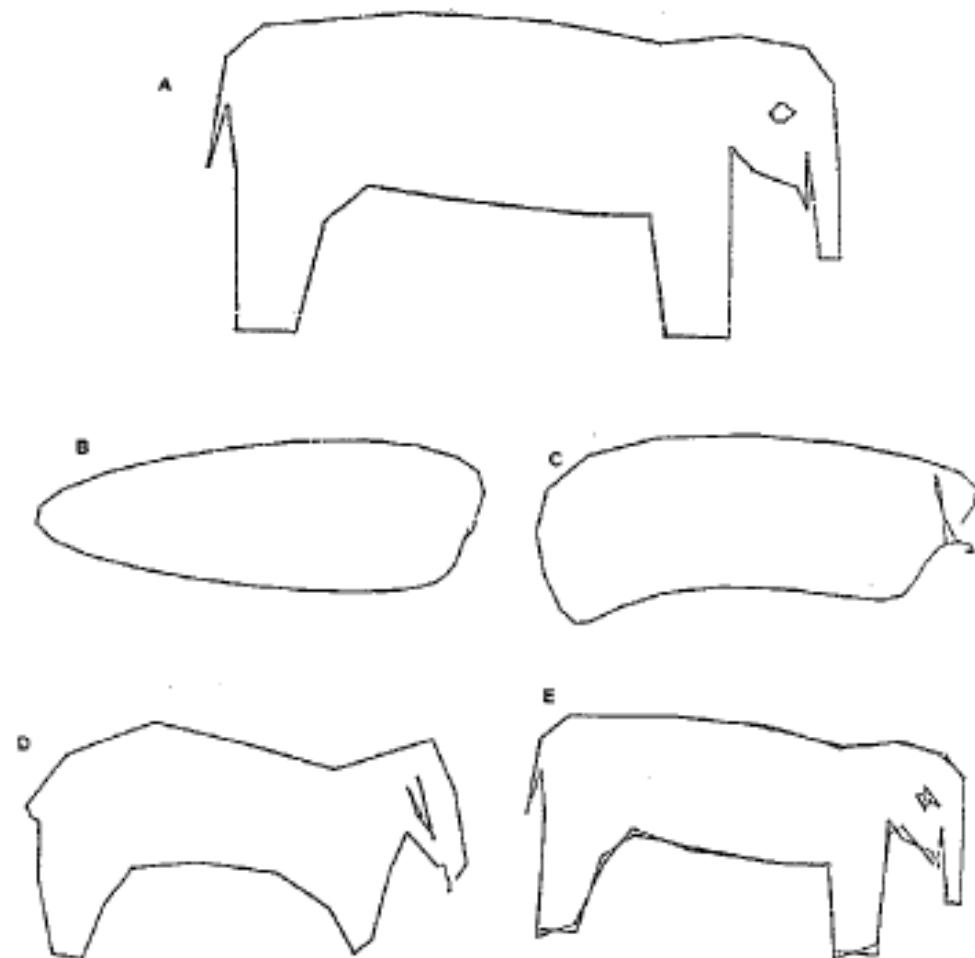


FIGURE 1.2. "How many parameters does it take to fit an elephant?" was answered by Wel (1975). He started with an idealized drawing (A) defined by 36 points and used least squares Fourier sine series fits of the form $x(t) = \alpha_0 + \sum \alpha_i \sin(it\pi/36)$ and $y(t) = \beta_0 + \sum \beta_i \sin(it\pi/36)$ for $i = 1, \dots, N$. He examined fits for $K = 5, 10, 20$, and 30 (shown in B-E) and stopped with the fit of a 30 term model. He concluded that the 30-term model "may not satisfy the third-grade art teacher, but would carry most chemical engineers into preliminary design."

Burnham, K. P., & Anderson, D. (2003). Model selection and multi-model inference. *A Practical information-theoretic approach*. Springer.

Modelo estadístico

4

Es una expresión matemática que indica cómo una variable aleatoria (VR, Y), con una distribución de probabilidades dada, se relaciona con una o más variables predictoras o explicativas (VE, X) consideradas en el diseño experimental

$$Y = \text{función } (X \text{ o } X_s)$$

$$Y \sim X$$

Modelos lineales

5

- Modelos lineales **en los parámetros**! La linealidad se refiere a los parámetros, no a la X. Los parámetros aparecen sumando; ningún parámetro aparece como exponente o multiplicado o dividido por otro parámetro.
- La VR es una combinación lineal de las VE

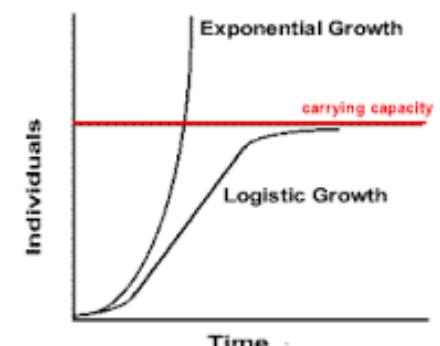
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

- En un modelo **no lineal**: los parámetros aparecen en la ecuación en forma no-lineal

$$Y_i = \beta_0 e^{\beta_1 X_i} + \varepsilon_i$$



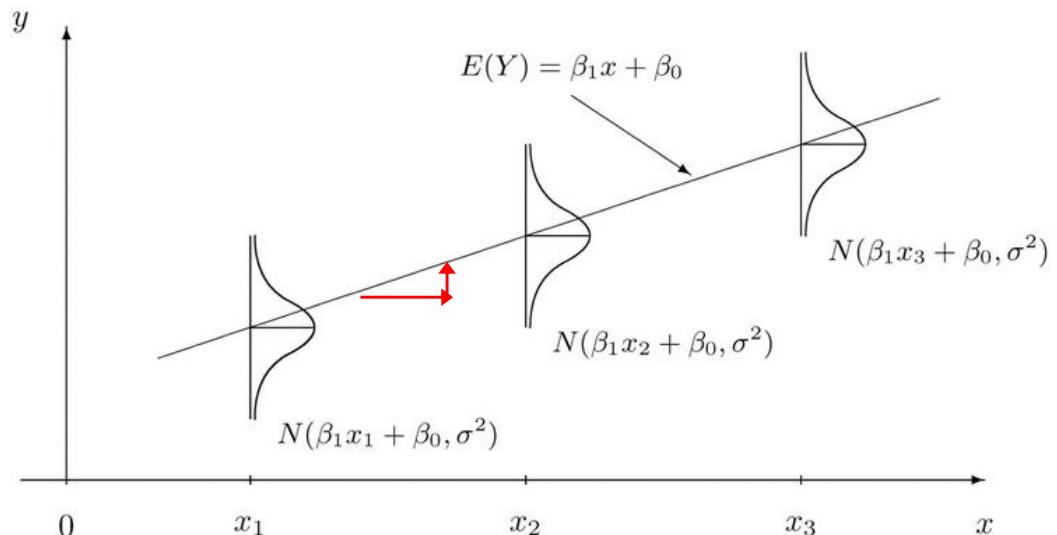
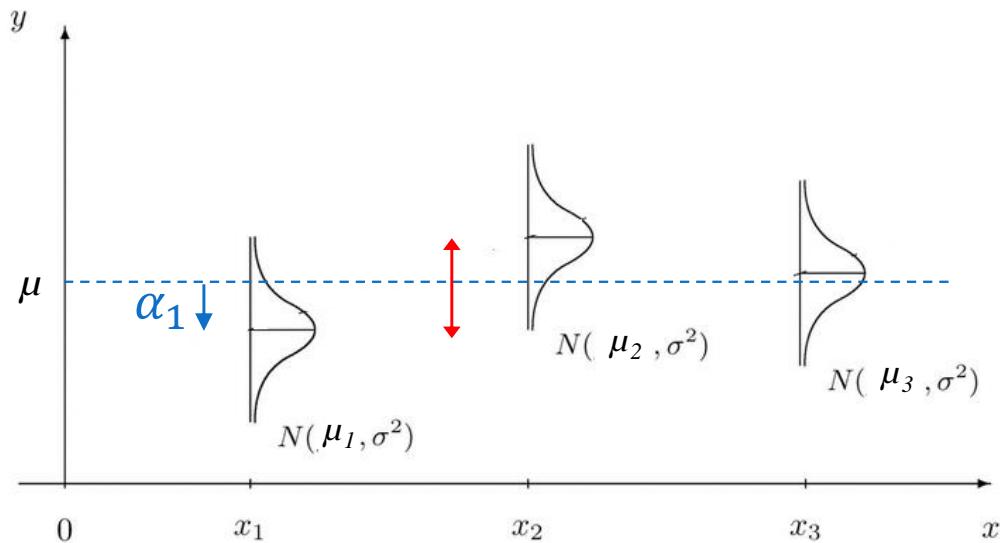
Parametrización del modelo:

X cuali (modelo de comparación de medias) o

X cuanti (modelo de regresión)?

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

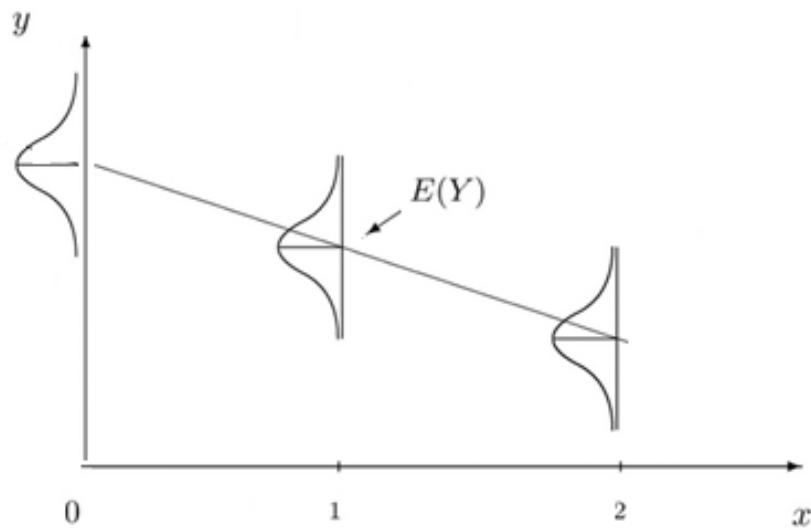


- ✓ En la comparación de medias ("Anova") las VE se denominan **factores** y se las trata como cualitativas. La magnitud del efecto se mide como **diferencia de medias**
- ✓ En Regresión las VE son cuantitativas. La magnitud del efecto se mide mediante **pendientes** o **coeficientes de regresión**. Las VE cualitativas pueden ser incluidas previo transformación en variables **indicadoras** o dummy

Regresión lineal



Parametrización del modelo



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

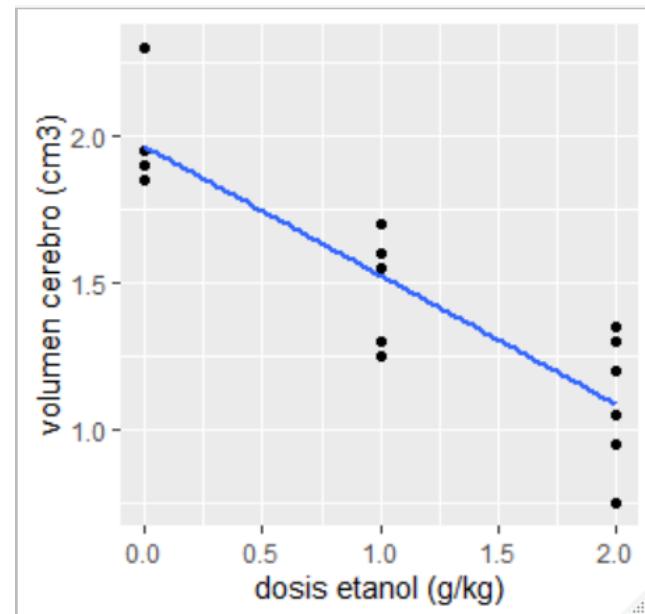
$$E(Y_i) = \beta_0 + \beta_1 X_i \quad Y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

Dado un valor de X , la esperanza de Y queda determinada únicamente (componente **sistemático**).

Existe variación aleatoria (error) que responde a una distribución de probabilidades (componente **aleatorio**)

Modelo con **3** parámetros

- β_0 es el valor esperado de Y cuando X vale 0
- β_1 es el cambio esperado en Y por cada aumento unitario en X
- σ^2 es la varianza de Y para cada valor de X , común a todos



	etanol	vol.mean	vol.sd
1	0	1.975	0.1635543
2	1	1.500	0.1816590
3	2	1.100	0.2280351

Ecuación estimada?
Interpretación de intercepto y pendiente?

$$E(Y_i) = \beta_0 + \beta_1 \text{etanol}_i$$

```
m1<-lm(vol~etanol, bd)
summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96250	0.06999	28.038	4.96e-15 ***
etanol	-0.43750	0.05422	-8.069	4.96e-07 ***
<hr/>				

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1878 on 16 degrees of freedom
 Multiple R-squared: 0.8028, Adjusted R-squared: 0.7904
 F-statistic: 65.12 on 1 and 16 DF, p-value: 4.956e-07

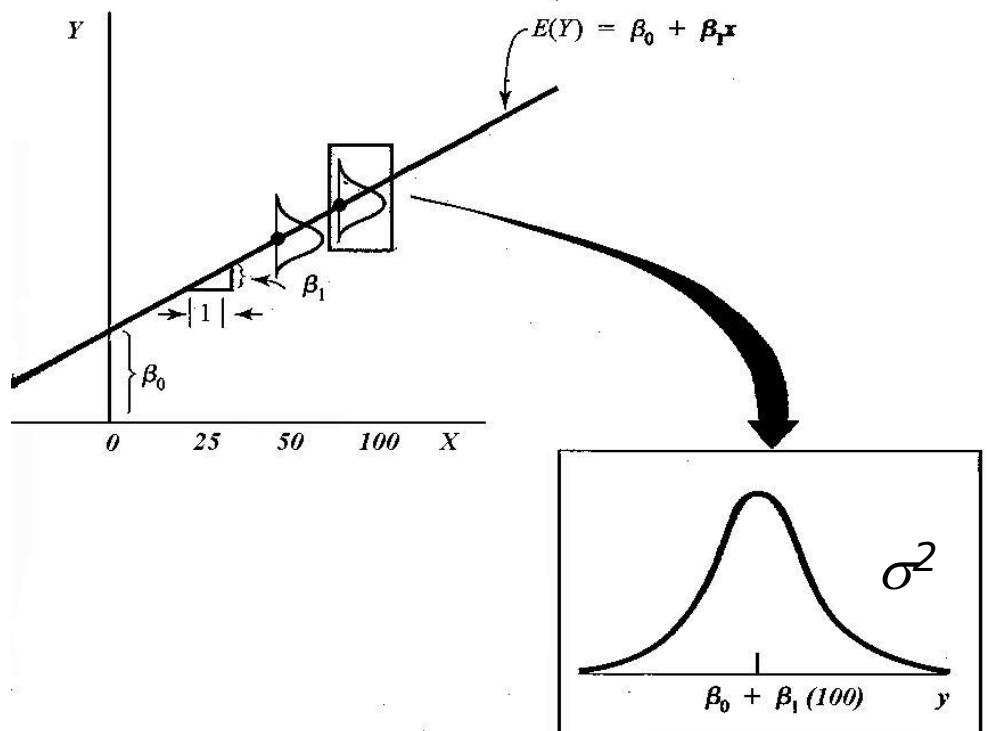
ratones_etanol.csv

S_e estimador de σ

Supuestos del modelo

9

- Para cada valor de X existe una subpoblación de Y
 - La media de cada una de estas subpoblaciones es $E_{Y/X} = \beta_0 + \beta_1 X_i$ (linealidad)
 - La distribución de cada subpoblación es normal $Y_{i/X} \approx NID(\mu_{Y/X}, \sigma^2)$
 - las varianzas de las subpoblaciones son iguales, es decir que el modelo asume una varianza constante σ^2 , sin importar el nivel de X $\text{Var}[Y/X] = \sigma^2$



No es necesarios para estimar los parámetros pero sí para que la inferencia sea válida

Inferencia sobre los coeficientes de regresión

10

$H_0: \beta_1 = 0$ la variación de Y **no se explica** linealmente por la variación de X

$H_1: \beta_1 \neq 0$ la variación de Y **sí se explica** linealmente por la variación de X

Dos opciones (equivalentes)

- A) Test t para β_i (en `summary`)
- B) Anova (en `anova(modelo)`)

Test t para β_i

11

Se basa en la distribución del estimador $\hat{\beta}_1$

Si la distribución de $Y_{/X}$ es normal, $\hat{\beta}_1$ sigue una distribución normal, con esperanza

$$\beta_1 \text{ y EE} = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}$$

EE: **error estándar** (de un estimador). Es una medida de la precisión en la estimación del parámetro

Se demuestra que $\hat{\beta}_1$ sigue una distribución aproximadamente normal cuando n es grande (extensión del Teorema Central del Límite)

$$t_{n-k-1} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S_e^2}{\sum(x_i - \bar{x})^2}}}$$

$$t_{n-k-1} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S_e^2}{\sum(x_i - \bar{x})^2}}} \quad k = \text{cantidad de VE}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.96250	0.06999	28.038	4.96e-15	***
etanol	-0.43750	0.05422	-8.069	4.96e-07	***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

Anova

12

Se basa en descomponer la variabilidad de la VR en sus distintas fuentes:

- Variabilidad explicada por la/s VE
- Variabilidad no explicada o aleatoria (error /residual)

El estadístico es F (cociente de varianzas) y su distribución es F de Fisher (GL numerador, GL denominador)

Prueba t y anova
son equivalentes
($t^2 = F$)

Anova

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Varianzas

variación controlada, impuesta por el investigador

13

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	F
Explicada por las VE	$\sum (\hat{y}_i - \bar{y}_i)^2$	k	\underline{SCexpl} \underline{GLexpl}	\underline{CMexpl} $CMerror$
No explicada por las VE, aleatoria o error	$\sum (y_i - \hat{y}_i)^2$	$n-k-1$	$\underline{SCerror}$ $\underline{GLerror}$	
Total	$\sum (y_i - \bar{y})^2$	$n-1$		Variación aleatoria o no controlada Estima σ^2

anova(m1)

Analysis of Variance Table

Response: vol

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
etanol	1	2.29688	2.29688	65.116	4.956e-07 ***						
Residuals	16	0.56437	0.03527								

signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

k = cantidad de VE

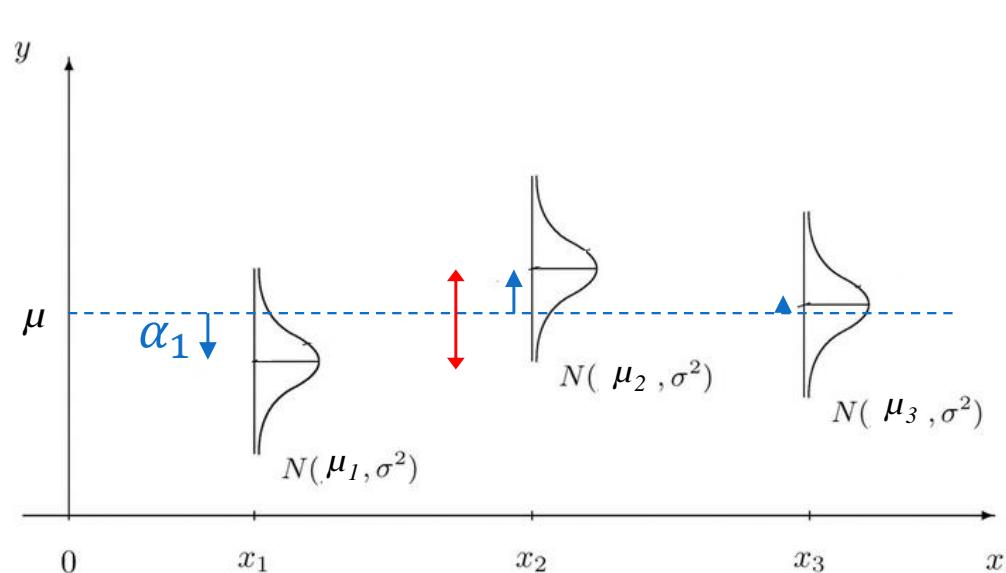


Efecto de la exposición postnatal a etanol sobre el volumen del cerebro en ratones

14

Modelo de comparación de medias

Parametrización



$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$

$$E(Y_i) = \mu + \alpha_i \quad Y_{ij} \sim \text{Normal}(\mu_i, \sigma^2)$$

Dado un nivel de X, la esperanza de Y queda determinada únicamente (componente **sistemático**).

Existe variación aleatoria (error) que responde a una distribución de probabilidades (componente **aleatorio**)

Modelo con **4** parámetros:

- α_i es el efecto del nivel *i*-ésimo del factor sobre la esperanza de Y. Alternativamente puede pensarse en μ_i , la esperanza de Y para cada nivel del factor
- σ^2 es la varianza de Y para cada nivel del factor, común a todos

Los **supuestos** de este modelo son exactamente los mismos que para el modelo de regresión, salvo que aquí no aplica el supuesto de linealidad

Inferencia sobre los efectos α_i

16

$H_0: \alpha_i = 0$ no existe efecto del factor // las medias poblacionales de los grupos son iguales

$H_1: \text{Al menos un } \alpha_i \neq 0$ existe efecto del factor // al menos un grupo difiere en su media poblacional

Una opción: [Anova](#) (en `anova(modelo)`)

La variación en la VR se partitiona en:

- variación explicada por la VE (factor/tratamientos)
- variación no explicada o error

No hay una prueba t equivalente

Y luego aplicar un método de comparaciones

ANOVA

H_0 : Todos los $\alpha_i = 0$
 H_1 : Algun $\alpha_i \neq 0$

Varianzas

variación controlada, impuesta por el investigador

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	F
Entre Tratamientos /grupos	$\sum n_i (\bar{y}_i - \bar{y})^2$	$t-1$	$\frac{S_{C\text{trat}}}{G_{L\text{trat}}}$	$\frac{C_{M\text{trat}}}{C_{M\text{error}}}$
Dentro de tratamientos o error	$\sum (y_{ij} - \bar{y}_i)^2$	$(n_i-1)t = n-t$	$\frac{S_{C\text{error}}}{G_{L\text{error}}}$	
Total	$\sum (y_{ij} - \bar{y})^2$	$n-1$		

Variación aleatoria o no controlada
Estima σ^2

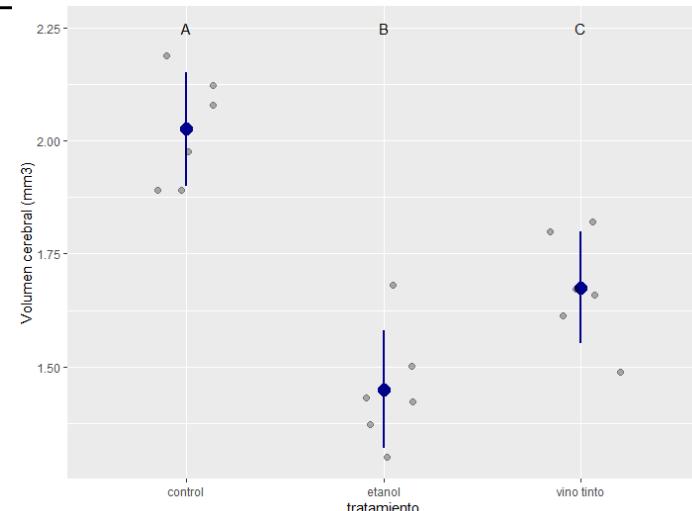
```
m2<-lm(vol~tratamiento, bd)
anova(m2)
```

Analysis of Variance Table

Response: vol

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
tratamiento	2	1.0075	0.50375	31.656	4.14e-06	***
Residuals	15	0.2387	0.01591			

S_e^2 estimador de σ^2



Pero, podemos analizar estos datos con un modelo de regresión?

18

- Las regresiones solo admiten VE cuantitativas
- Las v. cualitativas deben ser codificadas numéricamente para poder ser incluidas en la regresión ([v. auxiliares, indicadoras o dummy](#))

Tratamiento	volumen
etanol	2.4
etanol	3.3
etanol	2.4
etanol	1.4
etanol	2.6
vino tinto	3.6
vino tinto	1.1
vino tinto	3.5
vino tinto	3.6
vino tinto	3.4
control	2
control	1.3
control	4.6
control	1.7
control	2.2

Variables auxiliares,
indicadoras o "Dummy"

VE modelada
cualitativa

Una de las variables
auxiliares no aporta
información novedosa
ya que puede deducirse
a partir de las otras dos
(nivel de referencia)

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

$$E(Y_i) = \beta_0 + \beta_1 etanol_i + \beta_2 vino\ tinto_i$$

Cuando el tratamiento es el control

$$E(Y_i) = \beta_0 = \mu_{control}$$

Cuando el tratamiento es etanol

$$E(Y_i) = \beta_0 + \beta_1 = \mu_{etanol}$$

Cuando el tratamiento es vino tinto

$$E(Y_i) = \beta_0 + \beta_2 = \mu_{vino\ tinto}$$

Modelo de 4 parámetros:

- β_0 es el valor esperado del nivel de referencia (control)
- β_1 es la diferencia de medias entre el tratamiento con etanol y el control
- β_2 es la diferencia de medias entre el tratamiento con vino tinto y el control (control)
- σ^2 es la varianza de Y para tratamiento, constante

tratamiento	vol.mean	vol.sd	vol.
control	2.025	0.1246996	
etanol	1.450	0.1308434	
vino tinto	1.675	0.1227599	

Es el resumen de un modelo de regresión, donde las VE cuali son convertidas en v.indicadoras

$$E(Y_i) = \beta_0 + \beta_1 \text{dosis1}_i + \beta_2 \text{dosis2}_i$$

```
m2<-lm(vol~tratamiento, bd)
summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.02500	0.05150	39.321	< 2e-16 ***
tratamiento etanol	-0.57500	0.07283	-7.895	1.01e-06 ***
tratamiento vino tinto	-0.35000	0.07283	-4.806	0.000231 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1261 on 15 degrees of freedom
 Multiple R-squared: 0.8085
 F-statistic: 31.66 on 2 and 15 DF, p-value: 4.14e-06

Magnitud del efecto
 (diferencia de medias con el grupo de referencia)

EE para la
 diferencia de
 medias

S_e estimador de σ

- Salvo cuando hay solo dos niveles, no hay una prueba “global” sobre el efecto de la VE
- Los coeficientes son diferencias de medias con respecto al nivel de referencia; no se informan otras comparaciones
- No son comparaciones ortogonales
- No controlan el error global



Inferencia sobre la diferencia de medias

```
m2<-lm(vol~tratamiento, bd)
```

```
anova(m2)
```

Analysis of Variance Table

Response: vol

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
tratamiento	2	1.0075	0.50375	31.656	4.14e-06 ***
Residuals	15	0.2387	0.01591		

Prueba global: Al menos una media poblacional difiere significativamente del resto

```
emmeans(m2, pairwise ~ tratamiento)
```

```
$emmeans  
tratamiento emmean      SE df lower.CL upper.CL  
control       2.02 0.0515 15     1.92    2.13  
etanol        1.45 0.0515 15     1.34    1.56  
vino tinto    1.68 0.0515 15     1.57    1.78
```

Confidence level used: 0.95

Magnitud del efecto
(diferencia entre medias)

```
$contrasts  
contrast      estimate      SE df t.ratio p.value  
control - etanol  0.575 0.0728 15  7.895 <.0001  
control - vino tinto 0.350 0.0728 15  4.806 0.0006  
etanol - vino tinto -0.225 0.0728 15 -3.089 0.0193
```

Comparaciones de Tukey

P value adjustment: tukey method for comparing a family of 3 estimates

Volviendo al ejemplo de regresión (VE cuantitativa)

22

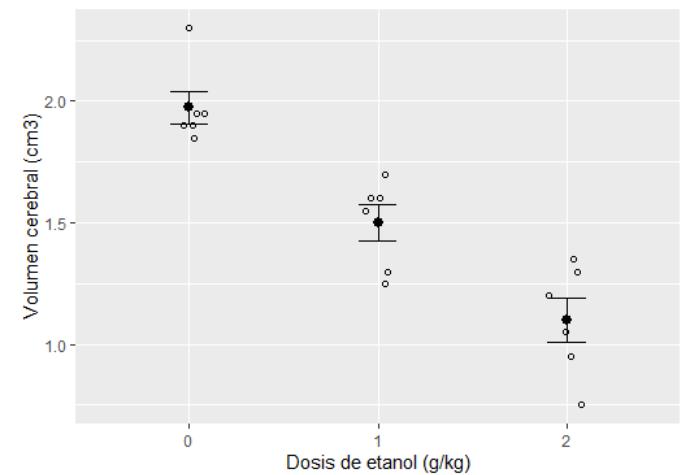
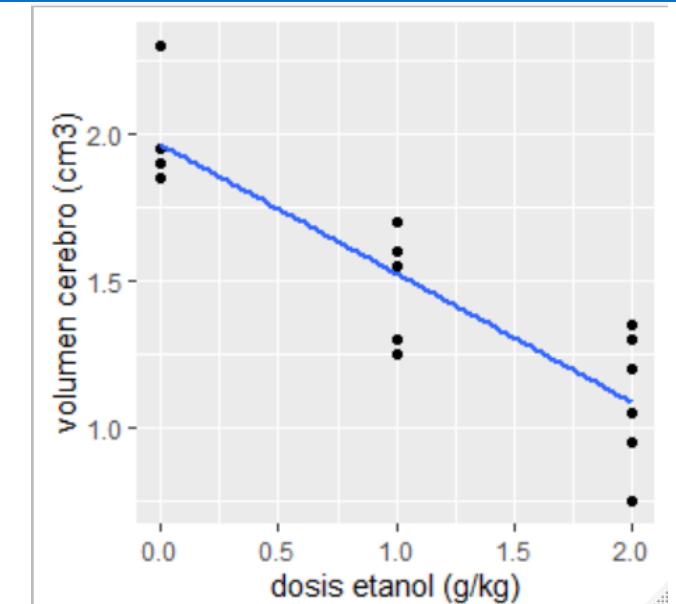
- Y si la relación no fuese lineal?
- Y si considerase poco prudente ajustar una regresión lineal con solo 3 niveles de X?
- Y si me interesase la diferencia de medias entre dosis (tratamientos)?

Es decir, si se desea incluir a una VE cuantitativa como **cualitativa**:

Modelo de comparación de medias

`m3<-lm(vol~factor(etanol), bd)`

Convierte a la
variable en
cualitativa



Inferencia sobre los efectos α_i

$H_0: \alpha_i = 0$ no existe efecto del factor // las medias poblacionales de los grupos son iguales

$H_1: \text{Al menos un } \alpha_i \neq 0$ existe efecto del factor // al menos un grupo difiere en su media poblacional

[anova\(m3\)](#)

Analysis of Variance Table

Response: vol

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
factor(etanol)	2	2.30250	1.15125	30.906	4.786e-06 ***
Residuals	15	0.55875	0.03725		

[emmeans\(m3, pairwise ~ factor\(etanol\)\)](#)

\$emmeans

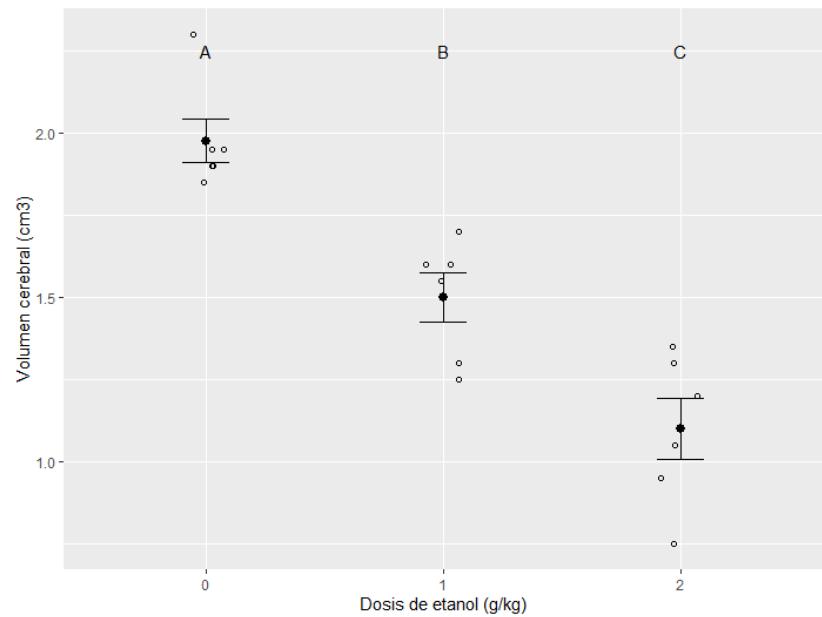
etanol	emmmean	SE	df	lower.CL	upper.CL
0	1.98	0.0788	15	1.807	2.14
1	1.50	0.0788	15	1.332	1.67
2	1.10	0.0788	15	0.932	1.27

Confidence level used: 0.95

\$contrasts

	contrast	estimate	SE	df	t.ratio	p.value
0 - 1		0.475	0.111	15	4.263	0.0019
0 - 2		0.875	0.111	15	7.852	<.0001
1 - 2		0.400	0.111	15	3.590	0.0071

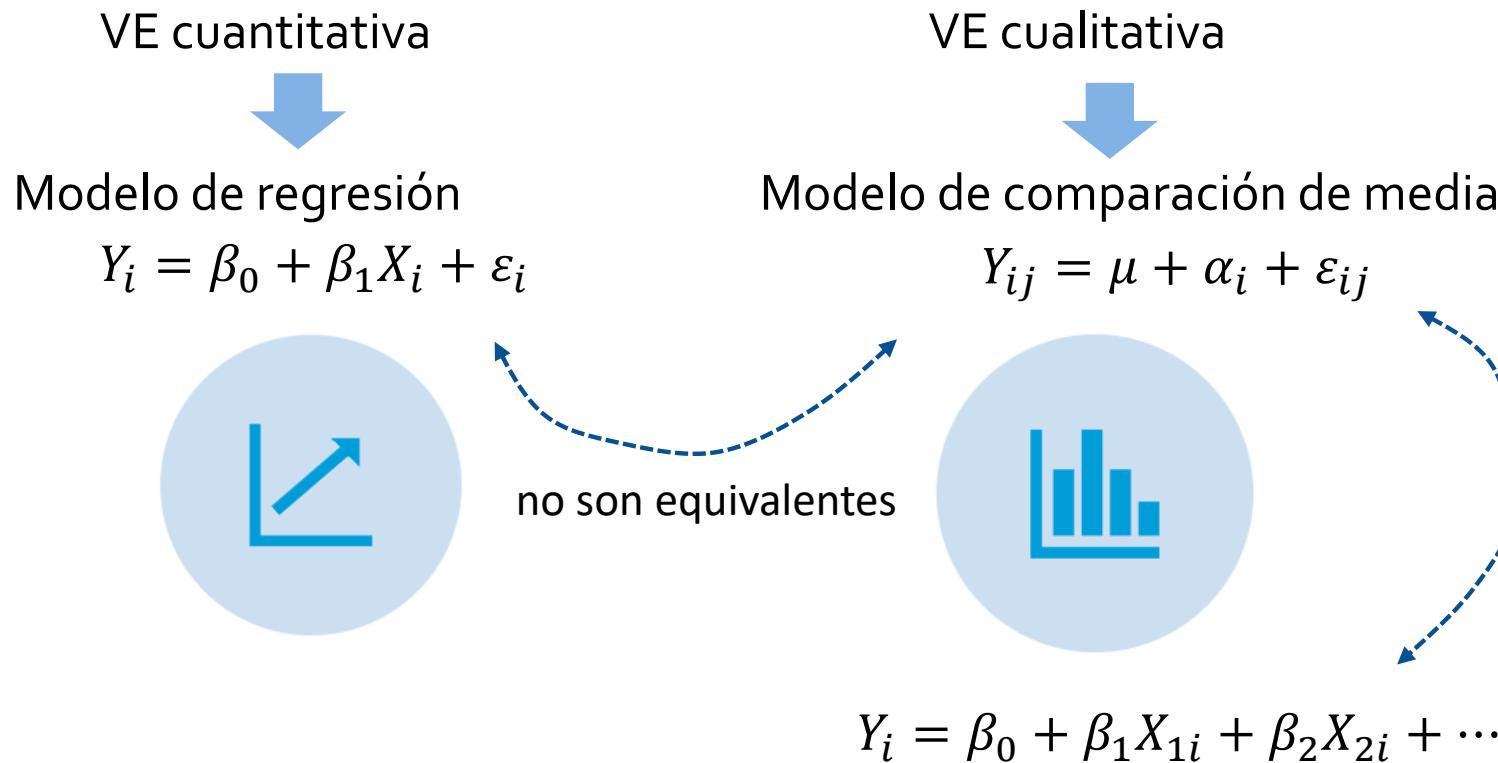
P value adjustment: tukey method for comparing a family of 3 estimates



En resumen

1m(VR~VE)

24



- Más parsimonioso (menos parámetros)
- Permite interpolar
- Modelos lineales en los parámetros
- Primera opción si la VE es cuanti
- Más parámetros
- Inferencia solo para los niveles estudiados
- No implica una función entre Y y X
- Primera opción si la VE es cuali

Para la próxima

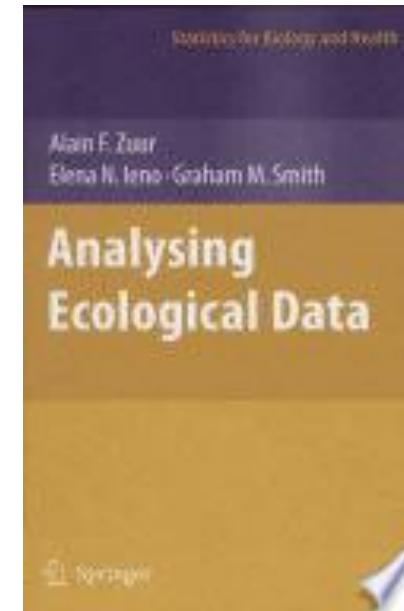
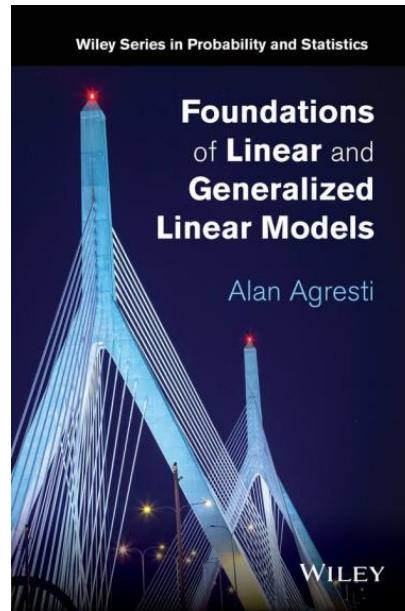
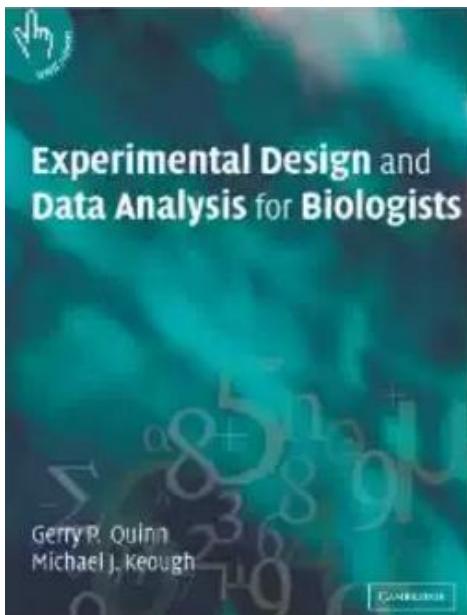
25

- Leer Perelman S y Garibaldi L. 2019. Capítulo 1. Introducción a la estadística experimental.
- Responder ejercicios 1.1, 1.3 y 1.4

Bibliografía general

26

- Quinn, G. P., & Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge University Press
- Agresti A. (2015). Foundations of Linear and Generalized Linear Models . Wiley
- Zuur, A., Ieno, E. N., & Smith, G. M. (2007). Analyzing ecological data. Springer Science & Business Media.
- Faraway, JJ (2002) Practical regression and Anova using R



BIOMETRÍA II

CLASE 2

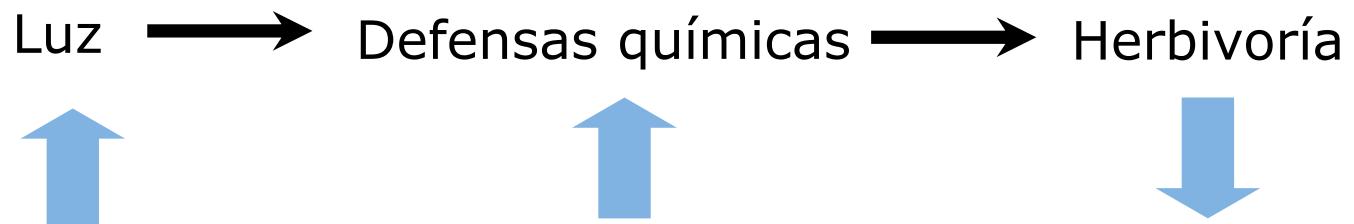
DISEÑOS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Hipótesis del balance carbono-nutrientes (Bryant et al, 1983)

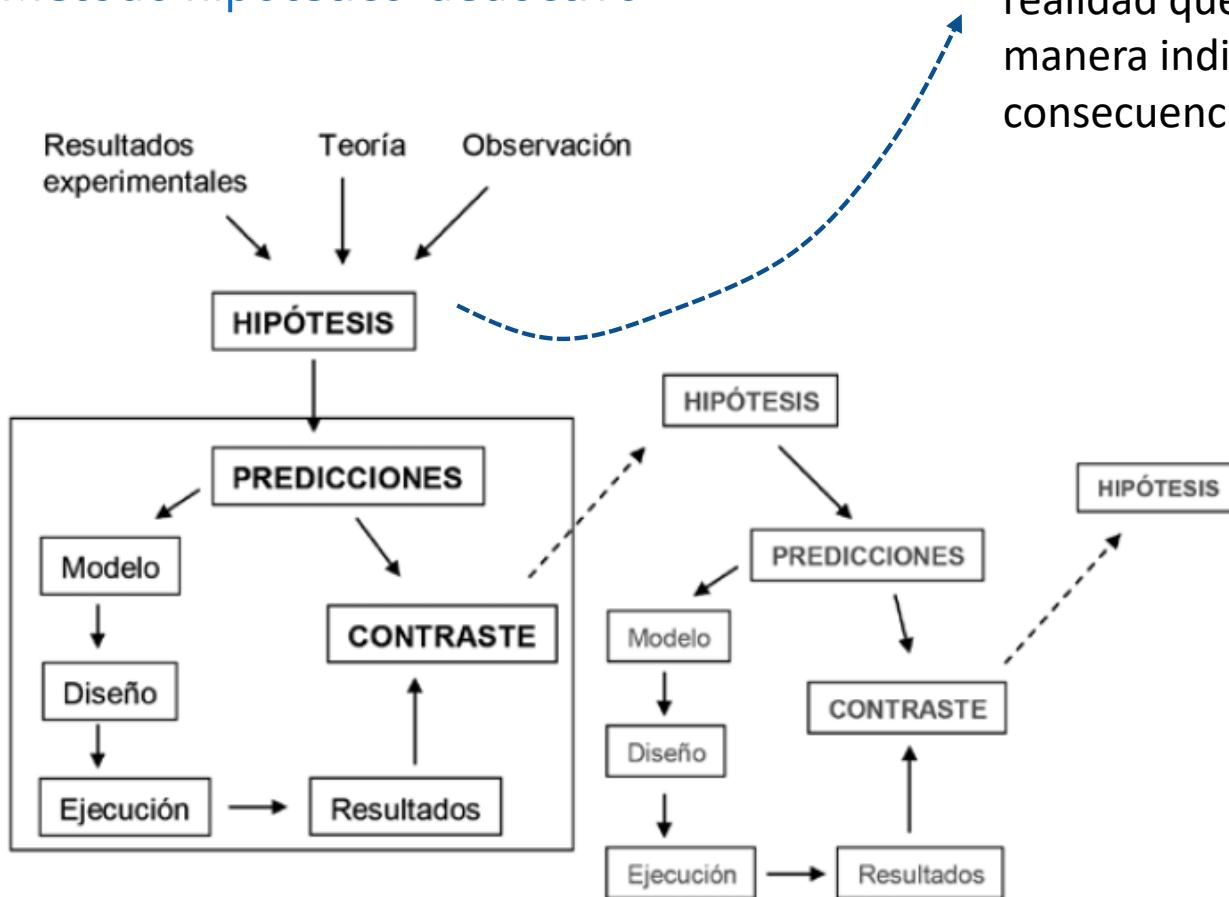
2

- Explica las variaciones intraespecíficas en los niveles de herbivoría de las plantas como una consecuencia de la variación en la disponibilidad de recursos
- La **hipótesis** propone que el recurso que supere necesidades de crecimiento (i.e., el que se encuentre en exceso para la planta) es derivado a la producción de defensas químicas
- La **predicción** es que plantas con alta disponibilidad de luz usarán el exceso de C para producir defensas químicas carbonadas (fenoles, taninos). Esto disminuiría la palatabilidad de las hojas para los herbívoros y en consecuencia, los niveles de defoliación de la planta



Método hipotético-deductivo

Las **hipótesis** son afirmaciones sobre la realidad que solo pueden verificarse de manera indirecta, es decir por alguna de sus consecuencias (**predicciones**)



Esquema general de los componentes de un programa de investigación. El recuadro marca los límites del experimento. Del contraste de las predicciones de las hipótesis surgen conclusiones que generan nuevas hipótesis

Un posible ensayo



4

- No existen estudios sobre la hipótesis del balance carbono-nutrientes en bosques australes
- Se decide efectuar un estudio en un bosque mixto ubicado en el Parque provincial Llao-Llao, particularmente sobre el arbusto endémico *Berberis buxifolia* (calafate)
- De todos los ejemplares de *Berberis buxifolia* (calafate) con alta disponibilidad lumínica se seleccionaron al azar 6
- De todos los ejemplares de *Berberis buxifolia* (calafate) con baja disponibilidad lumínica se seleccionaron al azar 6
- Al cabo de un tiempo se midió el nivel de herbivoría (como % de follaje consumido)

Otro posible ensayo

5



- El estudio se desarrolló en un bosque mixto ubicado en el Parque provincial Llao-Llao
- Se seleccionaron al azar 12 ejemplares de *Berberis buxifolia* (calafate), que se dividieron al azar en dos grupos de igual tamaño
- A un grupo se le asignó alta disponibilidad lumínica y al otro baja
- Al cabo de un tiempo se midió el nivel de herbivoría (como % de follaje consumido)

Causalidad vs asociación

6

- Los estudios experimentales bien diseñados y analizados proveen fuerte evidencia sobre relaciones **causales**. Permiten hablar del **efecto** de una variable sobre otra
- Los estudios observacionales son los únicos abordajes posibles cuando los tratamientos no pueden ser asignados aleatoriamente por la naturaleza del tratamiento, por razones éticas, etc. Permiten hablar de **asociación** entre variables
- ¿Cómo podemos establecer una relación de causalidad en esos casos?
 - Si la asociación es fuerte
 - Si los resultados son consistentes (distintos estudios, en distintas poblaciones llegan a los mismos resultados)
 - Si dosis mayores están asociadas a respuestas mayores
 - Si la supuesta causa precede al efecto en el tiempo

Criterios de causalidad de
Bradford Hill

Elementos de un ensayo

7

- **Unidad experimental o individuo:** es el material experimental que recibe un tratamiento y de la cual se obtiene una observación independiente
- **Variable respuesta o dependiente (Y):** es la respuesta del sistema que se va a evaluar. Su comportamiento es aleatorio e interesa estudiar si depende de otra/s variable/s llamadas explicativas
- **Variable explicativa, predictora o independiente (X):** es la que interesa estudiar si afecta a la variable respuesta; sus valores (niveles o tratamientos) son controlados (fijados por el investigador). Si es cualitativa se la conoce también como **factor**

¿En nuestro ejemplo?

Tratamiento control

8

- Es indispensable para evaluar el efecto de los tratamientos experimentales
- Puede consistir en la ausencia de tratamiento o en la aplicación del tratamiento habitual
- Debe considerarse que las UE asignadas al control deben diferir del resto sólo en el factor que interesa comparar

¿Cuál sería el control?

- Se desea determinar si la vitamina E previene los accidentes cerebro-vasculares
- Se desea analizar en ratas el efecto sobre los niveles de agresividad de lesiones producidas en una zona de la corteza cerebral

Diseño experimental

9

- Involucra determinar la forma en la que los tratamientos son asignados a las UE y la elección del tamaño muestral. El diseño experimental determina el modelo estadístico que permitirá poner a prueba la hipótesis de investigación

Al diseñar un experimento se debe tener en cuenta:

- que el experimento necesita repeticiones independientes de las unidades experimentales
- que la decisión de la cantidad de réplicas no es trivial
- que los diseños balanceados deberían ser preferidos a los no balanceados
- que el control del error experimental mejora la precisión
- el nivel de **independencia** entre las observaciones

Efecto de la actividad ganadera sobre las aves de humedal

- **Metodología:** Se seleccionaron 2 áreas representativas de un humedal similares en cuanto al clima y al régimen hidrológico pero que diferían sustancialmente en cuanto a la carga animal (ganado) presente: una con alta carga (A) y otra con baja carga (B). En cada una de las áreas (áreas A y B) se dispusieron 20 puntos de muestreo ubicados regularmente cada 150 metros en forma de grilla. En cada punto se estimó la diversidad de aves.
- **Resultado:** la diversidad media en el sitio B fue significativamente mayor que en el sitio A ($p=0,001$)
- **Conclusión:** La actividad ganadera afecta negativamente a la comunidad de aves de los humedales.



Seudorreproducción (Hurlbert, 1984)

11

- Consiste en considerar como réplicas independientes a observaciones que no lo son
- Es un problema de escala
- No se debe confundir réplicas verdaderas con **submuestras** (varias observaciones en la misma unidad experimental, consideradas así en el modelo estadístico)
- Se puede originar por un mal diseño o por un mal análisis



Ver Heffner et al (1996). *Pseudoreplication revisited. Ecology, 77:2588*

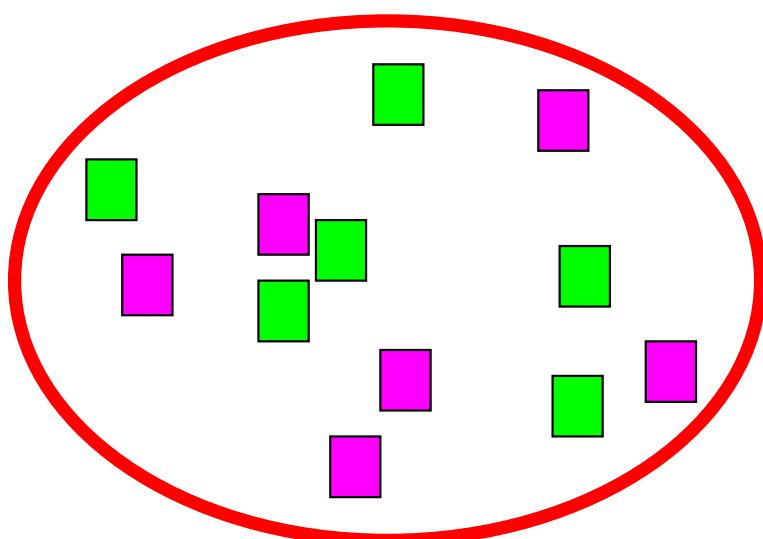
Volvamos a nuestro estudio



12

Efecto de la disponibilidad lumínica sobre la herbivoría

Diseño 1: Se eligieron 12 ejemplares que se dividieron al azar en dos grupos: uno fue expuesto a disponibilidad lumínica alta mientras que el otro a baja.



Id	Tratamiento	Nivel de herbivoría
1	Alta	9
2	Alta	14
.		.
11	Baja	10
12	Baja	42

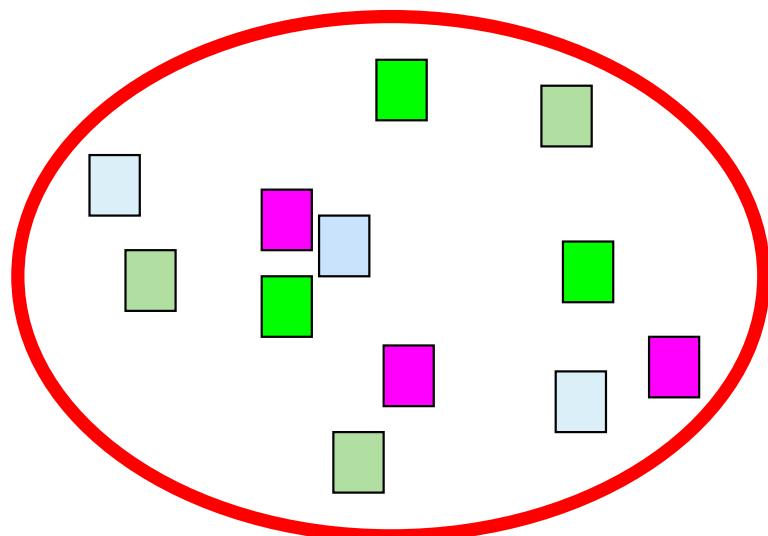
Volvamos a nuestro estudio



13

Efecto de la disponibilidad lumínica sobre la herbivoría

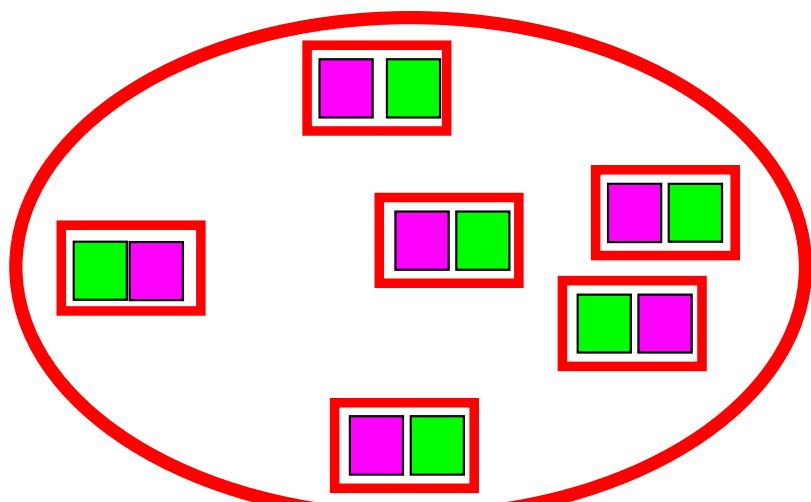
Diseño 2: Se eligieron 12 ejemplares que se dividieron al azar en cuatro grupos. A cada grupo se le asignó una combinación de disponibilidad lumínica (alta/baja) y de exclusión de herbívoros grandes (sí/no)



Id	Trata-miento luz	Tratamiento exclusión	Nivel de herbivoría
1	Alta	sí	9
2	Alta	sí	14
3	Alta	sí	.
4	Alta	no	
5	Alta	no	
11	Baja	no	10
12	Baja	no	42



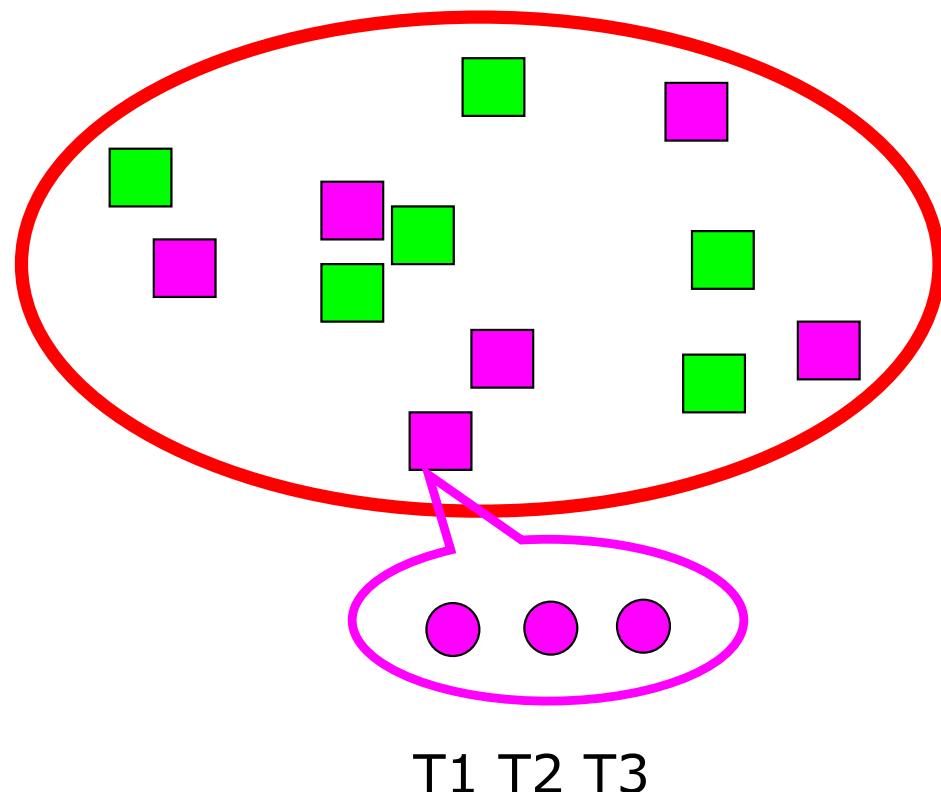
- Diseño 3: Se seleccionaron 6 sectores. En cada uno se eligieron dos ejemplares, uno cualquiera fue sometido a disponibilidad lumínica alta y el otro a baja.



Id	Trata-miento	Bloque	Nivel de herbivoría
1	Alta	1	9
2	Baja	1	14
.		.	.
11	Alta	6	10
12	Baja	6	32



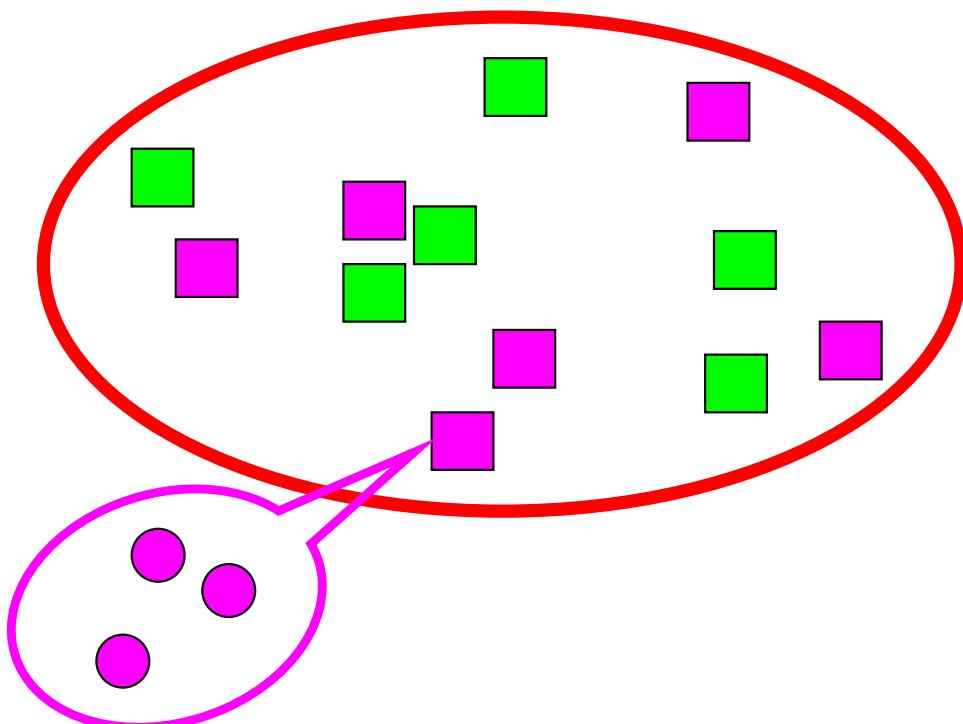
Diseño 4: Se eligieron 12 ejemplares que se dividieron al azar en dos grupos: uno fue expuesto a disponibilidad lumínica alta mientras que el otro a baja. Se midió la herbivoría al inicio, al mes y a los 2 meses



Id	Trata-miento	Ejemplar	Tiempo	Nivel de herbivoría
1	Alta	1	0	9
2	Alta	1	1	14
3	Alta	1	2	25
35	Baja	12	1	38
36	Baja	12	2	52



Diseño 5: Se eligieron 12 ejemplares que se dividieron al azar en dos grupos: uno fue expuesto a disponibilidad lumínica alta mientras que el otro a baja. Se midió la concentración de alcaloides en 3 hojas elegidas al azar de cada arbusto

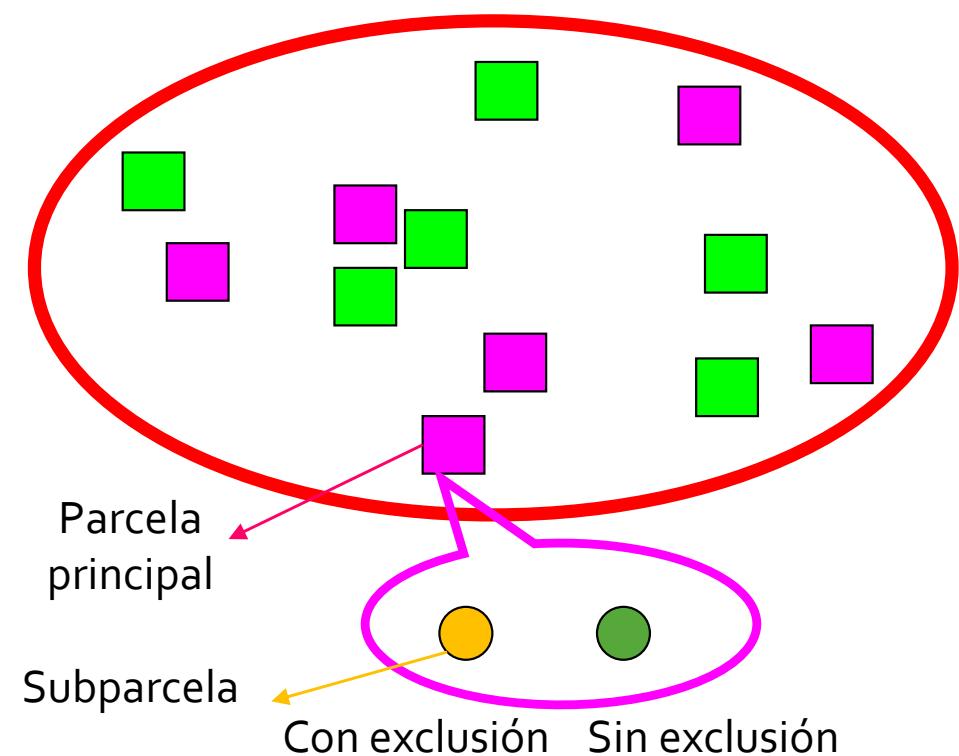


Id	Trata-miento	Ejem-plar	Alcaloides
1	Alta	1	29
2	Alta	1	14
3	Alta	1	25
35	Baja	12	18
36	Baja	12	32

Diseño 6: Se eligieron 12 ejemplares que se dividieron al azar en dos grupos: uno fue expuesto a disponibilidad lumínica alta mientras que el otro a baja.



En cada ejemplar se protegió parte de la copa de los herbívoros grandes y otra parte, no (exclusión sí/no)



Id	Trat. luz	Trat. exclusión	Ejemplar	Nivel de herbivoría
1	Alta	sí	1	9
2	Alta	no	1	14
3	Alta	sí	2	11
4	Alta	no	2	15
5	Alta	sí	3	
23	Baja	sí	12	10
24	Baja	no	12	42

Algunos diseños

¿independencia entre las observaciones?

18

Diseño Completamente aleatorizado (DCA) – Diseño factorial

- ✓ los tratamientos son asignados al azar a las UE. Todas las observaciones son independientes. No es recomendable cuando las UE son heterogéneas (mucho error experimental)

Diseño de Bloques al azar (DBA)

- ✓ el experimentador agrupa las UE en bloques homogéneos y luego asigna al azar los tratamientos a las UE dentro de cada bloque. Más eficiente que DCA cuando las UE son heterogéneas

Diseño de medidas repetidas (DMR)

- ✓ los tratamientos son asignados al azar a las UE. Cada UE es medida a lo largo del tiempo

Diseño anidado

- ✓ Cualquier diseño en el que haya submuestreo (más de una observación por UE). Mejora la precisión en la estimación de la respuesta de cada UE, pero no aumenta la cantidad de verdaderas réplicas

Diseño de parcela dividida

- ✓ Los tratamientos se asignan secuencialmente

Bibliografía específica

19

Perelman S y Garibaldi L. 2019. Capítulo 1. Introducción a la estadística experimental. En Experimentación y modelos estadísticos. Editorial Facultad de Agronomía, Universidad de Buenos Aires

Festing, M. F., & Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR journal*, 43(4), 244-258.

Guía de diseños: A Field Guide to Experimental Designs
<http://www.tfrec.wsu.edu/ANOVA/index.html>

MODELOS SEGÚN LA DISTRIBUCIÓN DE PROBABILIDADES DE LA VARIABLE RESPUESTA



Efecto de la disponibilidad lumínica sobre el nivel de herbivoría en *Berberis buxifolia*



21

¿Cómo medimos el nivel de herbivoría (VR)?

- Área foliar dañada (cm^2)
- Cantidad de hojas con signos de herbivoría en una muestra aleatoria de 10 hojas
- Planta con signos de herbivoría (sí/no)
- etc

¿Por qué es importante conocer la distribución de probabilidades de una VA?

Para decidir qué modelos son los adecuados para modelarla

Modelos lineales generales vs modelos lineales generalizados

22

Modelo lineal general

$$\varepsilon_i \approx NID(0, \sigma^2)$$

- ✓ VR cuantitativa continua con distribución normal, varianza constante y observaciones independientes
- ✓ Estimación de los parámetros por **cuadrados mínimos / máxima verosimilitud**
 - Anova, regresión lineal y polinómial, ancova

Modelos lineales generalizados (GLM)

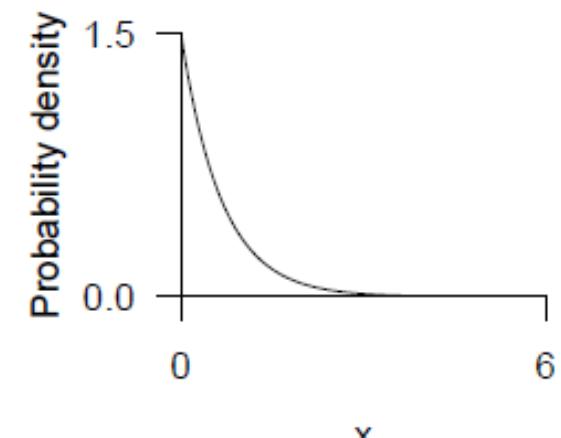
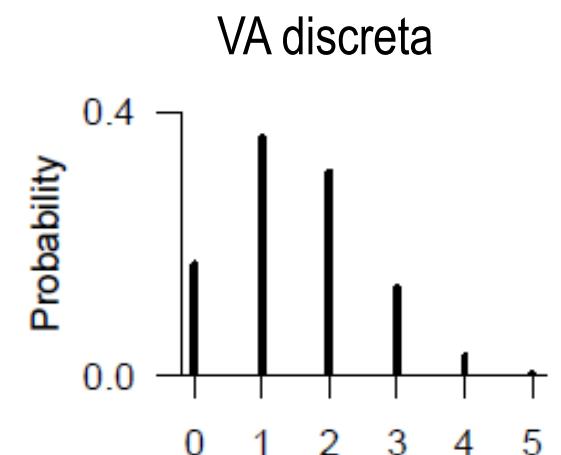
- ✓ VR cualitativa o cuantitativa discreta o cuantitativa continua con distribución de probabilidades de la familia exponencial
 - Regresión logística: VR dicotómica (Sí/No), distribución Bernoulli
 - Regresión binomial: VR discreta (cantidad de éxitos en una muestra n)
 - Regresión de Poisson : VR discreta (conteos)
 - Otras: distribución normal (el modelo lineal general es un caso particular de los GLM); distribución gamma, etc
- ✓ Estimación de los parámetros por **máxima verosimilitud**

VARIABLE RESPUESTA	TIPO DE VARIABLE y COTAS	POTENCIAL DISTRIBUCION DE PROBABILIDADES
Área foliar dañada (cm^2)		
Cantidad de hojas con signos de herbivoría en una muestra aleatoria de 10 hojas		
Planta con signos de herbivoría (sí/no)		
Número de eventos de actividad de herbívoros		
Área foliar dañada / Área foliar total		
Concentración de fenoles en la planta (asimétrica positiva)		
Tiempo de sobrevida		

Distribuciones de probabilidad

24

- El rango de valores de una variable aleatoria (VA) se denomina **dominio**
- Existen distintos **modelos teóricos** que describen cómo se distribuyen las probabilidades para los distintos valores de la VA.
- Una distribución de probabilidades tiene **parámetros** (valores que la caracterizan)
- La distribución de probabilidades en una VA discreta asocia cada valor que esta puede tomar a un valor de probabilidad. La **función de probabilidad** es la fórmula que permite calcular la probabilidad para cada valor de la VA
- En VA continuas la **función de densidad de probabilidad** describe el comportamiento de la VA y la probabilidad de que la VA tome valores entre a y b equivale al área bajo la curva de la función de densidad
- Las distribuciones tienen una **esperanza** y una **varianza**

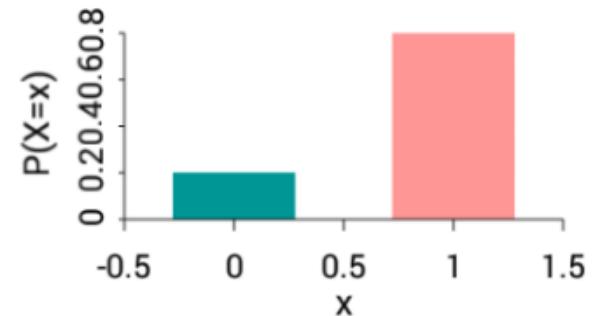
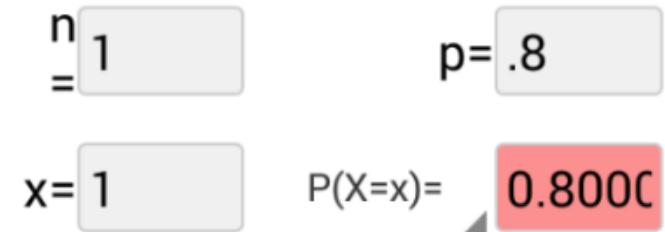


Distribuciones de probabilidad para variables discretas

25

Bernoulli

- VA dicotómica (presencia/ausencia; éxito /fracaso). Por convención se asigna 0 al fracaso y 1 al éxito
- Parámetros: π = probabilidad de éxito,
 $1 - \pi$ = probabilidad de fracaso
- Función de probabilidad:
 $P(Y = y) = \pi^y(1 - \pi)^{1-y}$ donde $y = 0, 1$
- Dominio: entre 0 y 1
- Esperanza = π Varianza = $\pi(1 - \pi)$
- Ej: la probabilidad de que una semilla germine es 0,8. Interesa modelar si una semilla germina o no



Distribuciones de probabilidad para variables discretas

26

Binomial

- VA: cantidad de éxitos en n repeticiones.
Conteos en n ensayos (n debe ser un número natural)
- Las repeticiones deben ser independientes; la probabilidad de éxito en cada repetición (p) debe permanecer constante
- Función de probabilidad: $P(Y = y) = {}_nC_y \pi^y (1 - \pi)^{1-y}$
- Dominio: entre 0 y n
- Parámetros: π y n
- Esperanza = πn Varianza = $\pi (1 - \pi)n$
- Ej: Se siembran 10 semillas. La probabilidad de que una semilla germine es 0,8. Interesa modelar la cantidad de semillas que germinan en 10

$$X \sim \text{Bin}(n, p)$$

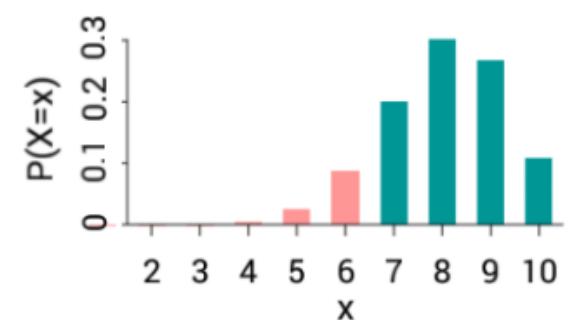
$$n = 10$$

$$p = .8$$

$$x = 6$$

$$P(X \leq x) =$$

$$0.1208$$



`rbinom(n=x, size=x, prob=x) # simula n observaciones de una VA binomial`

Distribuciones de probabilidad para variables discretas

27

Poisson

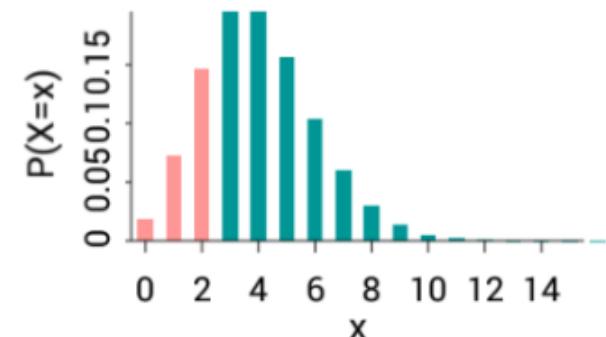
- VA: cantidad de eventos en un continuo de espacio o tiempo. Conteos por unidad de tiempo / volumen, etc
- Los eventos se producen al azar y de manera independiente en el continuo, con una esperanza o media determinada (λ)
- Función de probabilidad: $P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$
- Dominio: entre 0 e infinito
- Esperanza = λ Varianza = λ
- Ej: Cierta especie de gramínea crece en un pastizal a una densidad de 4 plantas por m^2 . Interesa modelar la cantidad de plantas por m^2

$X \sim \text{Pois}(\lambda)$

$\lambda = 4$

$x = 2$

$P(X \leq x) =$ 0.2381



`rpois(n=x, lambda=x) # simula n observaciones de una VA Poisson`

Distribuciones de probabilidad para variables continuas

28

Normal

- VA continua
- Función de densidad de probabilidad:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Dominio: entre $-\infty$ y $+\infty$
- Parámetros: esperanza (μ) y desvío estándar(σ)

$$X \sim N(\mu, \sigma)$$

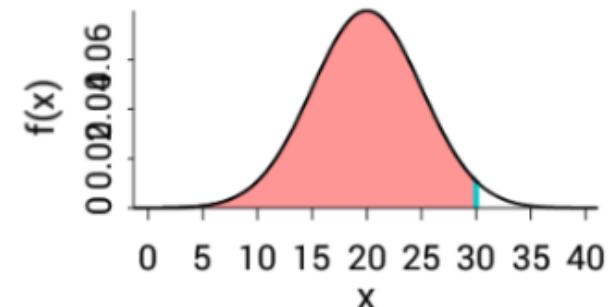
$$\mu = 20$$

$$\sigma = 5$$

$$x = 30$$

$$P(X < x) =$$

$$0.9772$$



`rnorm(n = x, mean = x, sd = x)` # simula n observaciones de una variable con distribución normal

Distribuciones de probabilidad para variables continuas

29

Lognormal

$$X \sim \text{LogN}(\mu, \sigma)$$

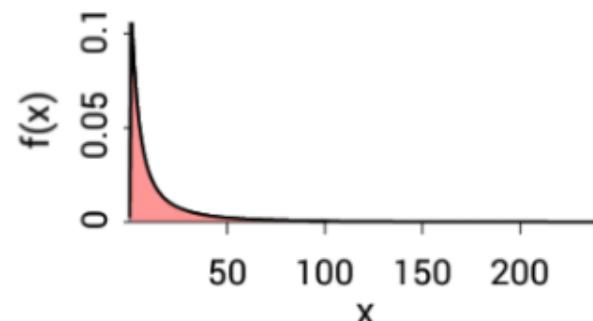
$$\mu = 2$$

$$\sigma = 1.5$$

$$x = 100$$

$$P(X < x) =$$

$$0.9587$$



Gamma

$$X \sim \text{Gamma}(\alpha, \beta)$$

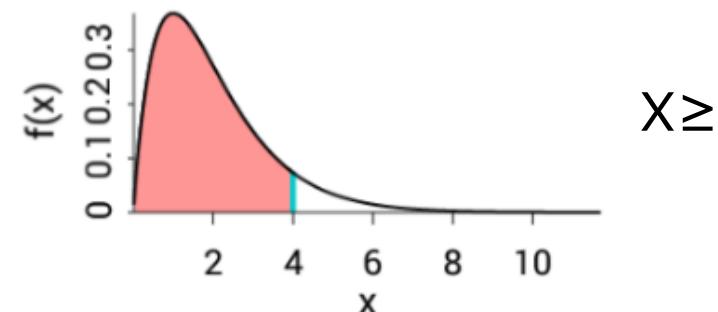
$$\alpha = 2$$

$$\beta = 1$$

$$x = 4$$

$$P(X < x) =$$

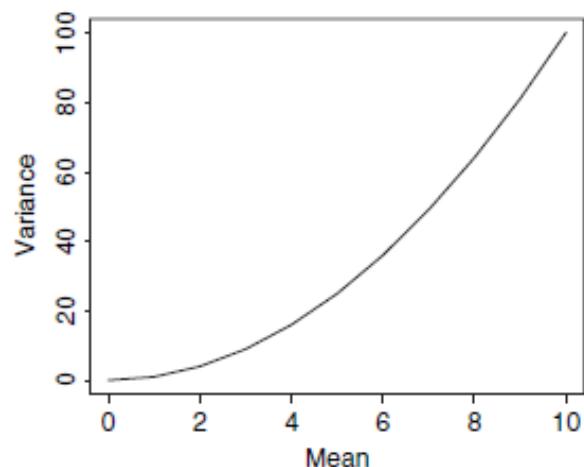
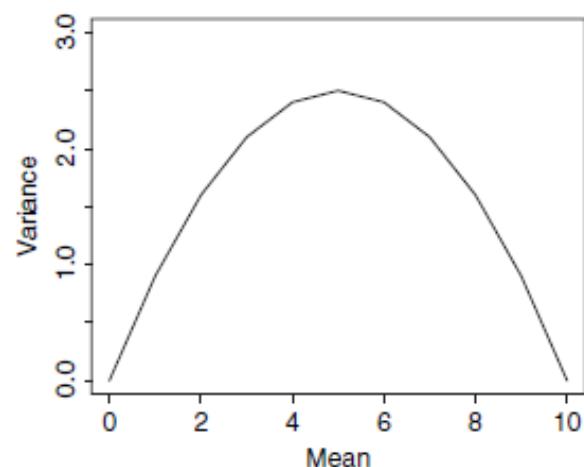
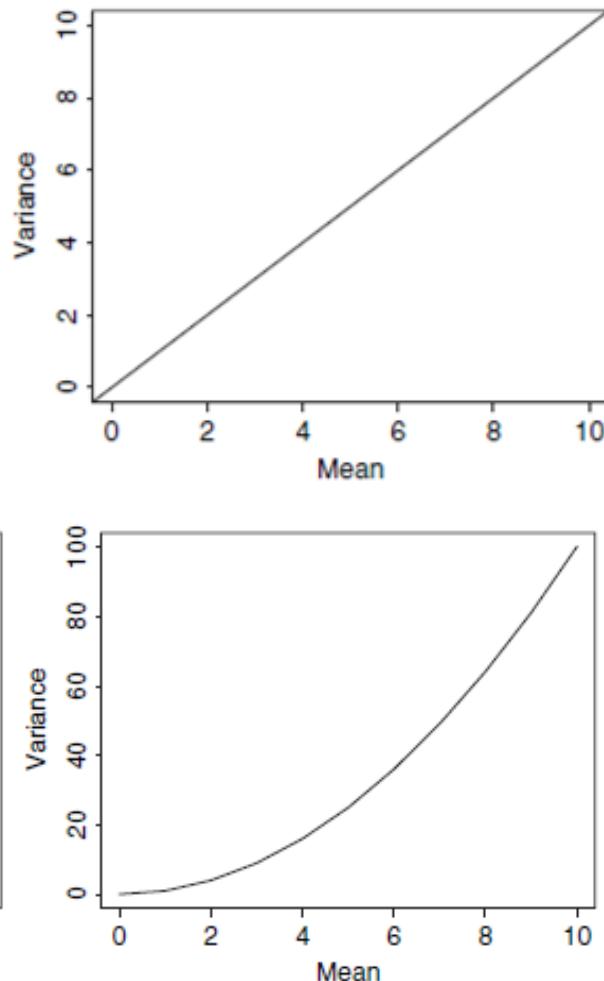
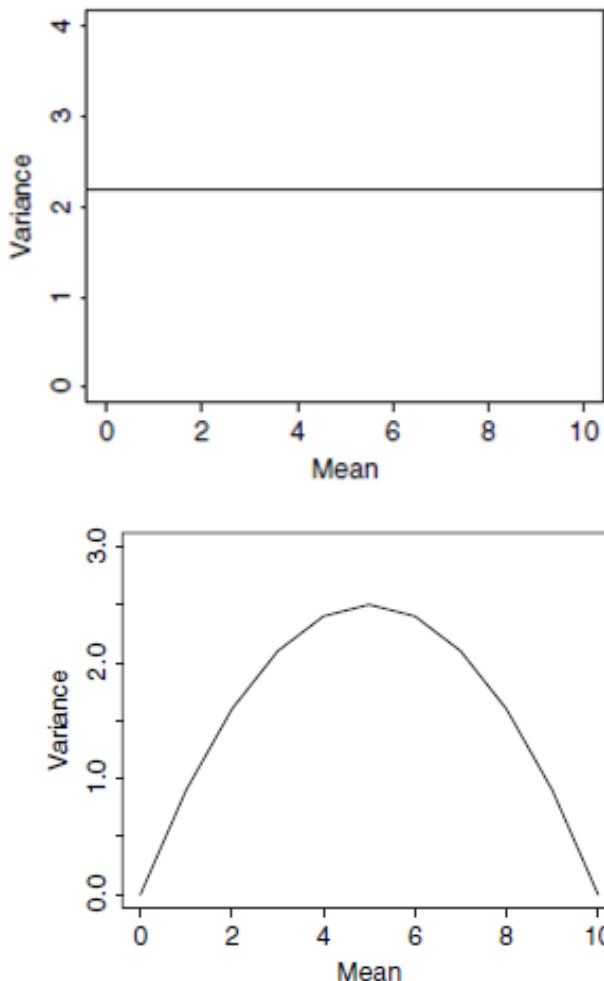
$$0.9084$$



- Y además están las distribuciones de probabilidad para estadísticos, como la normal estandarizada, la t de Student, chi cuadrado, F de Fisher...

Relación entre esperanza y varianza

30



Normal
Poisson
Binomial
Gamma



ESTIMADORES



Estimación de parámetros

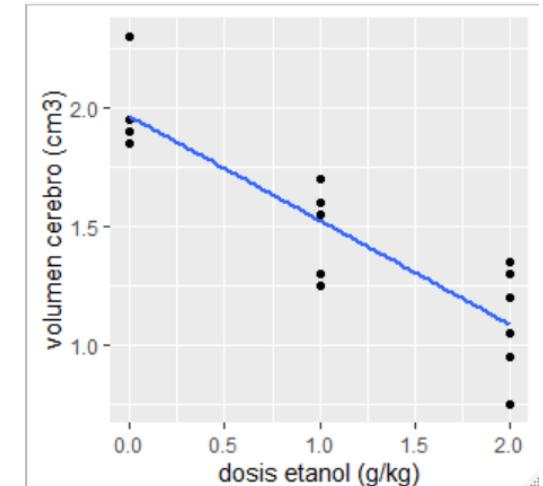


32

Modelo de regresión

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96250	0.06999	28.038	4.96e-15 ***
etanol	-0.43750	0.05422	-8.069	4.96e-07 ***



Modelo de comparación de medias

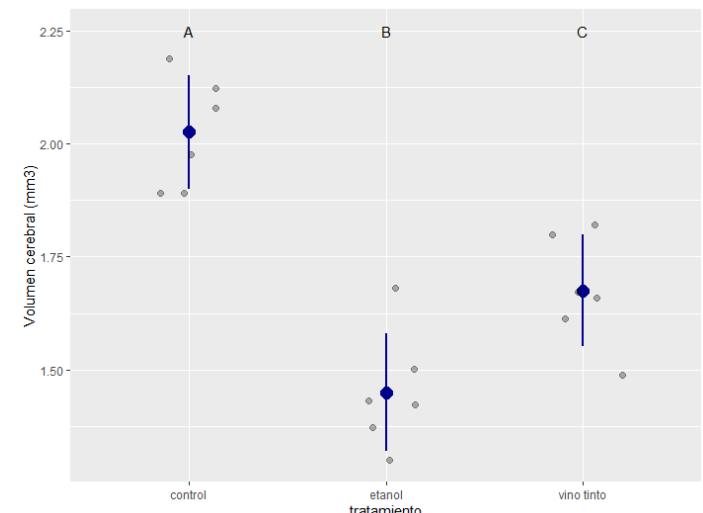
\$emmeans

tratamiento	emmmean	SE	df	lower.CL	upper.CL
control	2.02	0.0515	15	1.92	2.13
etanol	1.45	0.0515	15	1.34	1.56
vino tinto	1.68	0.0515	15	1.57	1.78

confidence level used: 0.95

\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
control - etanol	0.575	0.0728	15	7.895	<.0001
control - vino tinto	0.350	0.0728	15	4.806	0.0006
etanol - vino tinto	-0.225	0.0728	15	-3.089	0.0193



Métodos de estimación de parámetros

Mínimos cuadrados ordinarios (MCO u OLS por sus siglas en inglés *ordinary least squares*)

- Consiste en estimar los parámetros de manera tal de minimizar

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i))^2$$

- Método “clásico” para estimar parámetros de modelos lineales generales
- Cálculos sencillos
- Supuestos: independencia, normalidad y homocedasticidad. Sensible al desbalanceo
- ¿Y si no se cumplen los supuestos?

Las estimaciones de los parámetros siguen siendo insesgadas, pero los errores estándar no, por lo que la inferencia no es confiable

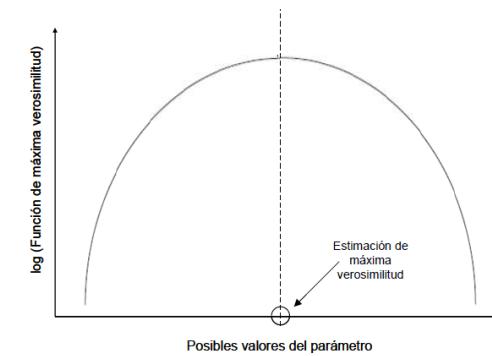
Métodos de estimación de parámetros

Máxima verosimilitud (MV o ML por sus siglas en inglés *maximum likelihood*)

- Consiste hallar los valores de los parámetros – asumiendo una distribución de probabilidades – que maximicen la función de verosimilitud
- Es una medida de la “creencia racional” de que los parámetros tomen cierto valor dados los datos y una distribución de probabilidades

$$\max L(\theta / \text{datos, distr de prob}) = \prod_{i=1}^n f(y_i / \theta)$$

- Cálculos complejos
- Para la estimación es fundamental asumir una distribución de probabilidades de la variable
- Si se cumplen los supuestos de independencia, normalidad y homocedasticidad y el diseño es balanceado, las estimaciones son las de MCO



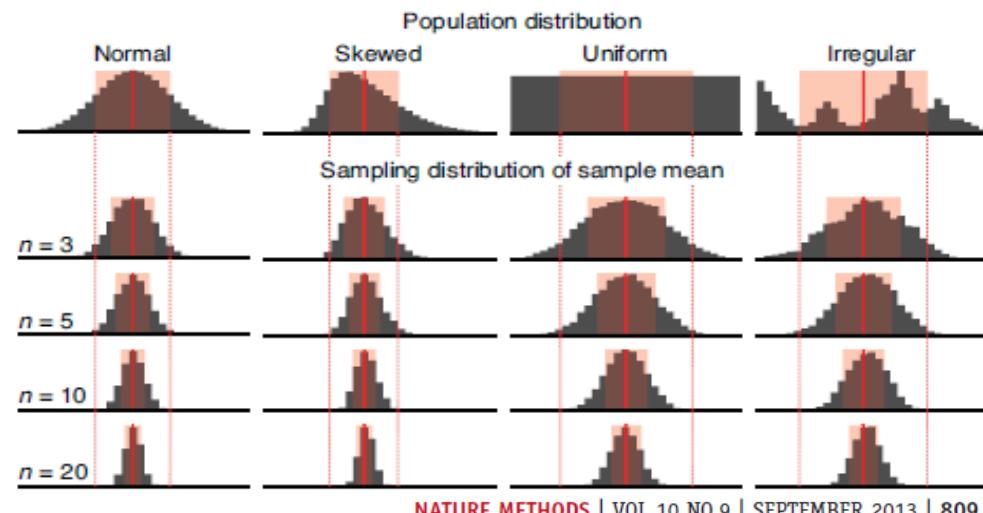
Propiedades de los estimadores

35

- Un estimador es una **variable aleatoria** (ya que varía al variar aleatoriamente la muestra) y por lo tanto tiene una distribución de probabilidades.
- Si el n es grande se demuestra (Teorema central del límite) que la media muestral \bar{y} y los coeficientes de regresión $\hat{\beta}_i$ tienden a la distribución **normal**

Dos propiedades deseables de los estimadores son:

- **Insegadez**: La esperanza del estimador es igual al parámetro que estima
- **Consistencia**: a medida que aumenta n la distancia entre el estimador puntual y el parámetro tiende a cero



Desvío estándar y error estándar

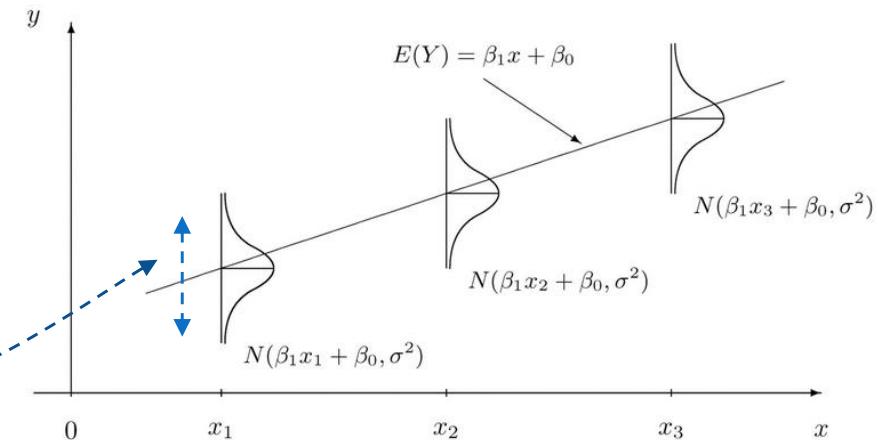
36

desvío estándar: Es la raíz de la esperanza de las distancias entre las observaciones y su media, elevadas al cuadrado. Mide la dispersión de la variable para cada nivel de X

$$\sigma = \sqrt{E(Y_i - E[Y])^2}$$

$$S = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n - 1}}$$

error estándar: Es el DE de un estimador. Da idea de la **precisión en la estimación del parámetro** (cuanto mayor el EE menor la precisión)



$$S_{\bar{Y}} = \sqrt{\frac{S^2}{n}} \quad S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}$$
$$S_{\hat{\beta}_1} = \sqrt{\frac{S^2}{\sum(x_i - \bar{x})^2}}$$

¿Y si el estimador no sigue una distribución de probabilidades conocida y/o no existe una fórmula exacta para su error estándar?

Método Bootstrap

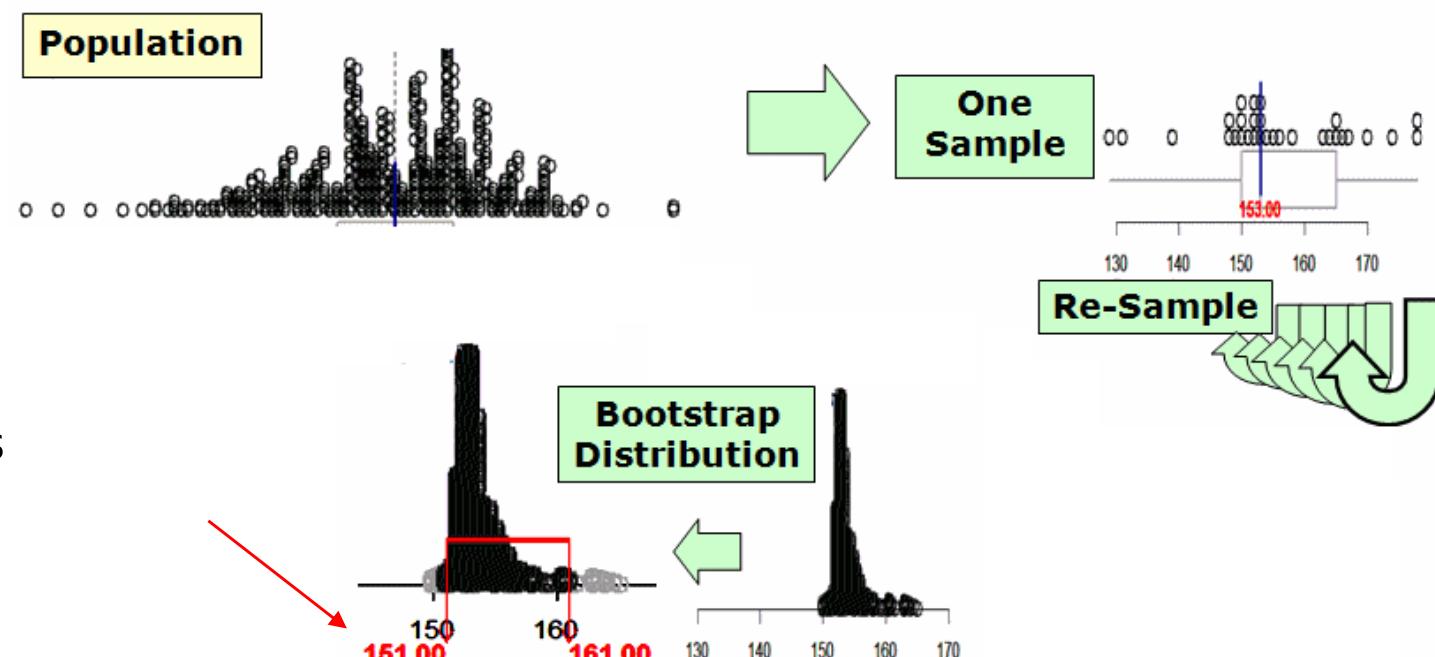
37

- Es un método de remuestreo que se utiliza para aproximar la distribución de probabilidades de un estimador, ya que por ejemplo se desconoce su distribución teórica
- Es el caso de la mediana, de muchos índices en biología que surgen de funciones complejas (diversidad, árboles filogenéticos, etc)
- Solo se cuenta con datos muestrales (n). Consideramos que su distribución constituye una buena aproximación a la distribución real de la variable
- Entonces aproximamos la distribución muestral mediante la simulación de experimentos repetidos sobre nuestros datos muestrales
- Mediante la simulación podemos obtener EE, predecir sesgo e incluso comparar varias formas de estimar el mismo parámetro
- El único requisito es que los datos hayan sido independientemente muestreados de una única distribución

Bootstrap: Procedimiento

38

- Se extraen muchas muestras con reposición (i.e. 1000) de tamaño n de la muestra original (se “re-muestrea”)
- En cada muestra se calcula el estimador de interés
- Con esos valores se construye la distribución muestral
- El estimador por bootstrap del parámetro será la media de dicha distribución y su error estándar sería el desvío estándar de la distribución



Estimación mediante intervalos de confianza paramétricos

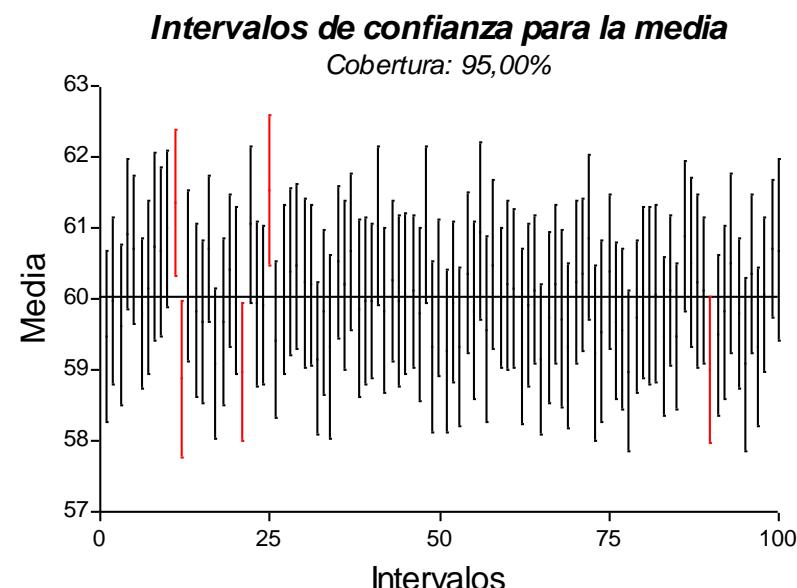
39

Consiste en obtener un rango de valores, utilizando la distribución de probabilidades del estimador y su EE, que se espera contenga al parámetro con una cierta probabilidad (a priori) o nivel de confianza (a posteriori)

$$P(\text{Límite inf} < \text{parámetro} < \text{Límite sup}) = \text{nivel de confianza}(1 - \alpha)$$

$$\text{estimador} \pm \text{percentil}_{GL,\alpha/2} \text{ EE}_{\text{estimador}}$$

El nivel de confianza es el porcentaje de intervalos que se espera contengan al parámetro (para ese tamaño de muestra) [VER](#)



Nuestro IC es un rango de valores plausibles para el parámetro. Los valores fuera del IC son relativamente inverosímiles

Nuestros datos son compatibles con cualquier valor del parámetro dentro del IC pero relativamente incompatibles con cualquier valor fuera de él.

Intervalos de confianza

$$\bar{Y} \pm t_{n-1;1-\alpha/2} \sqrt{\frac{s^2}{n}}$$

40

IC para la magnitud del efecto de una VE:

- IC para el coeficiente de regresión / pendiente en un modelo de regresión
- IC para la diferencia de medias en un modelo de comparación de medias

Permiten:

$$\hat{\beta}_1 \pm t_{GLE;\alpha/2} \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}$$

$$\Delta Y \pm t_{GLE,\alpha/2} \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$$

- detectar si el efecto es significativo (si el cero no pertenece al IC)
- cuantificar la relevancia de una VE según los valores que toma

```
> summary(m1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.96250  0.06999 28.038 4.96e-15
etanol      -0.43750  0.05422 -8.069 4.96e-07
---
```

```
> confint(m1)
              2.5 %      97.5 %
(Intercept) 1.8141204 2.1108796
etanol      -0.5524343 -0.3225657
```

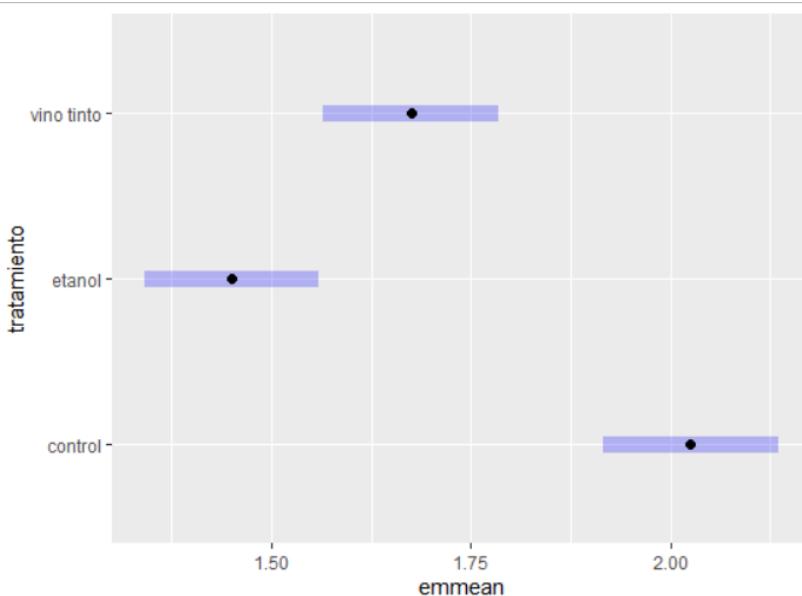
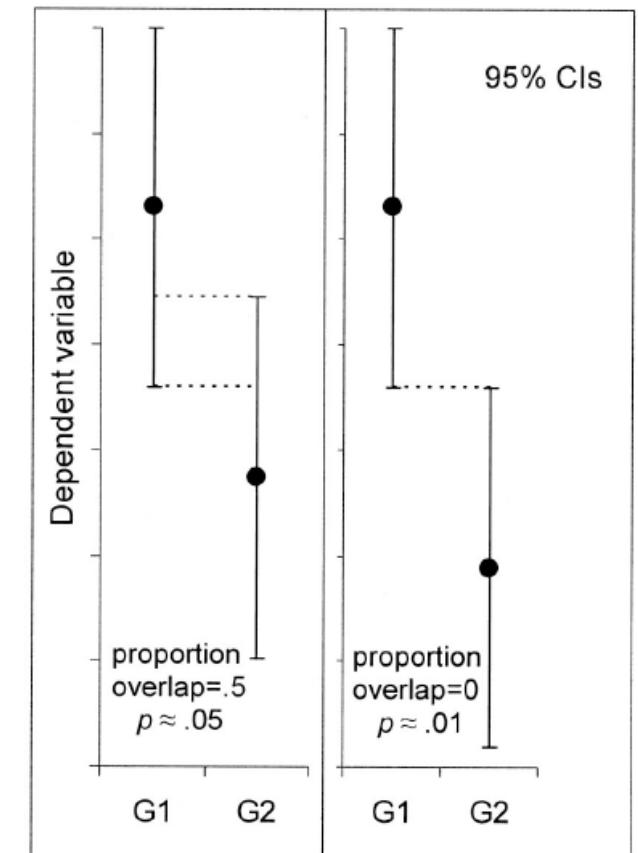
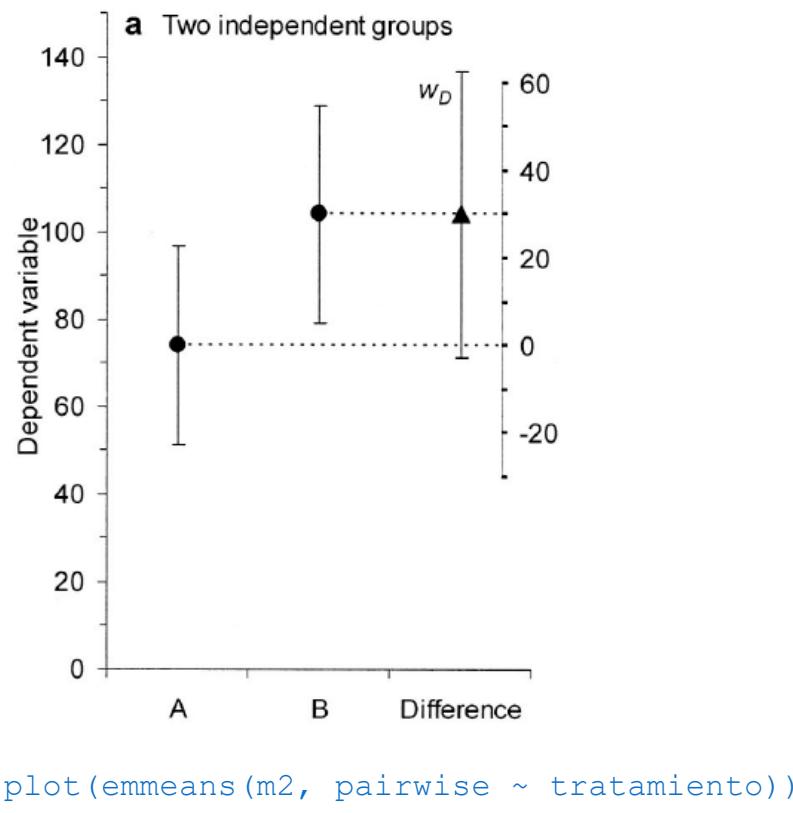
```
> confint(emmeans(m2, pairwise ~ tratamiento))
$emmeans
  tratamiento emmean    SE df lower.CL upper.CL
control        2.02 0.0515 15    1.92    2.13
etanol         1.45 0.0515 15    1.34    1.56
vino tinto     1.68 0.0515 15    1.57    1.78
Confidence level used: 0.95

$contrasts
  contrast      estimate    SE df lower.CL upper.CL
control - etanol  0.575 0.0728 15   0.386  0.7642
control - vino tinto 0.350 0.0728 15   0.161  0.5392
etanol - vino tinto -0.225 0.0728 15  -0.414 -0.0358
```

Nuestros datos son compatibles con cualquier valor del parámetro dentro del IC pero relativamente incompatibles con cualquier valor fuera de él.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American psychologist*, 60(2), 170.

Two Independent Groups, Both of Size 50 and With Equal Margins of Error



Rule of Eye 4: For a comparison of two independent means, $p \leq .05$ when the overlap of the 95% CIs is no more than about half the average margin of error, that is, when proportion overlap is about .50 or less (see Figure 4; Figure 5, left panel). In addition, $p \leq .01$ when the two CIs do not overlap, that is, when proportion overlap is about 0 or there is a positive gap (see Figure 5, right panel). These relationships are sufficiently accurate when both sample sizes are at least 10, and the margins of error do not differ by more than a factor of 2.

El valor p

42

Es la probabilidad de obtener una diferencia entre los grupos tan o más extrema que la obtenida en nuestro ensayo, si los grupos en realidad no difiriesen (si no hubiese efecto en la población)

- ✓ Si el valor p es **bajo** se concluye que **sí existe efecto** en la población (la prueba es **significativa**)
- ✓ Si el valor p no es bajo se concluye que **no hay evidencias de que exista efecto** en la población (la prueba no es significativa)

A menor valor de “p”, menor es la credibilidad sobre la hipótesis de “no efecto”

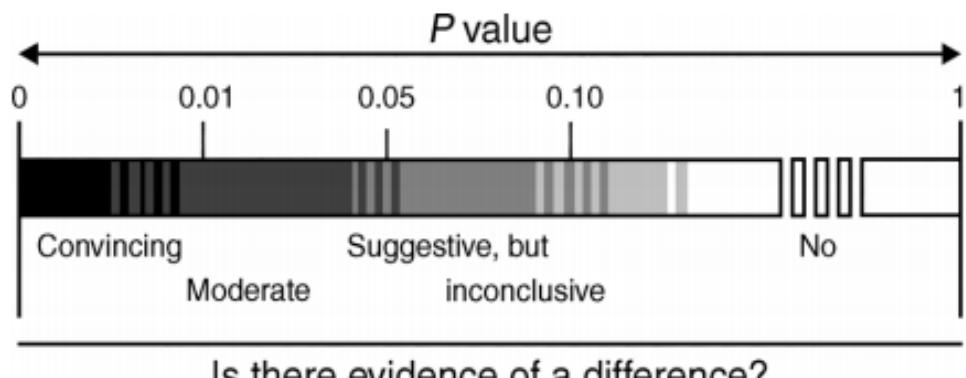


FIG. 1. Interpretation of the P value. Reprinted with permission from Ramsey and Schafer (2002).

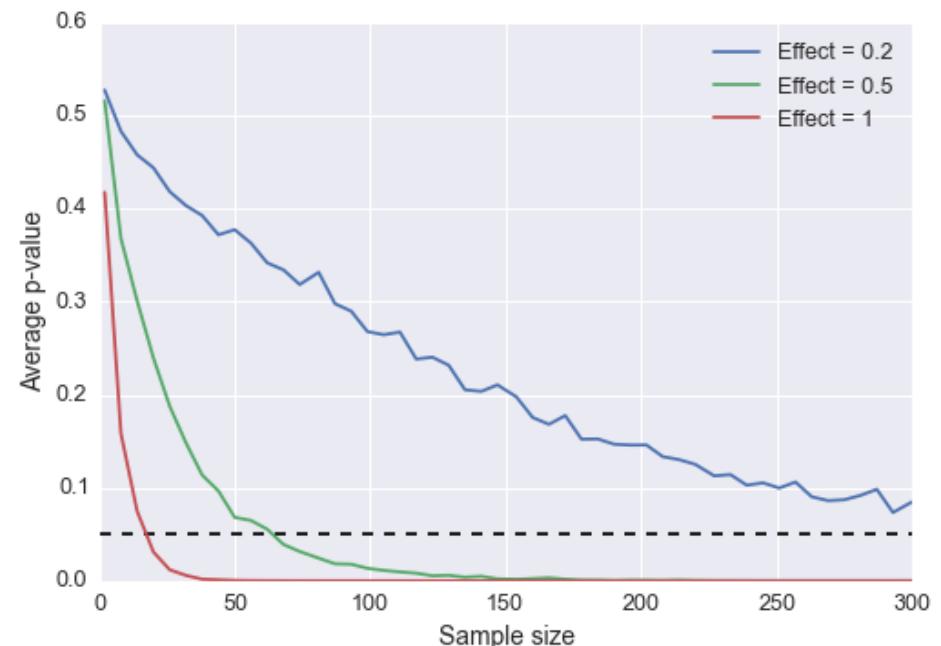
¿De qué depende el valor p ?

43

- ✓ De la magnitud del efecto observado
- ✓ Del tamaño de la muestra
- ✓ De la variabilidad no explicada en la VR (ruido)
- ✓ Del tipo de prueba aplicada (paramétrica vs no paramétrica; una cola vs dos colas)

Atenti p-hacking!

Los valores p raramente son comparables entre ensayos



The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein  & Nicole A. Lazar

44

- Los valores *p* pueden indicar qué tan incompatibles son los datos con una hipótesis o un modelo estadístico
- Los valores *p* no miden la probabilidad de que la hipótesis estudiada sea verdadera, o la probabilidad de que los datos se hayan producido solo por azar.
- Las conclusiones científicas y las decisiones políticas no deberían basarse únicamente en si un valor *p* supera un umbral específico.
- Una inferencia adecuada requiere información completa y transparente
- Un valor *p* no mide la magnitud del efecto o la importancia de un resultado.
- Por sí mismo, un valor *p* no proporciona una buena medida de evidencia con respecto a un modelo o hipótesis.

Check-list

45

- ¿Se determinó a priori la cantidad de réplicas para una dada potencia y un efecto de tratamiento?
- ¿Se exploraron los datos? (estadística descriptiva: tendencia central, dispersión, forma de la distribución, datos atípicos, datos faltantes)
- ¿Se chequearon los supuestos de la prueba?
- ¿Se estimó la magnitud del efecto mediante intervalos de confianza? Recordar que la significación estadística no necesariamente asegura significación biológica
- ¿Qué error se puede estar cometiendo?
¿Con qué máxima probabilidad?
- ¿Se puede establecer una relación causal?
¿Estudio observacional o experimental?
- ¿Sobre qué población se aplican las conclusiones?



BIOMETRÍA II

CLASE 3

SUPUESTOS DE LOS MODELOS LINEALES

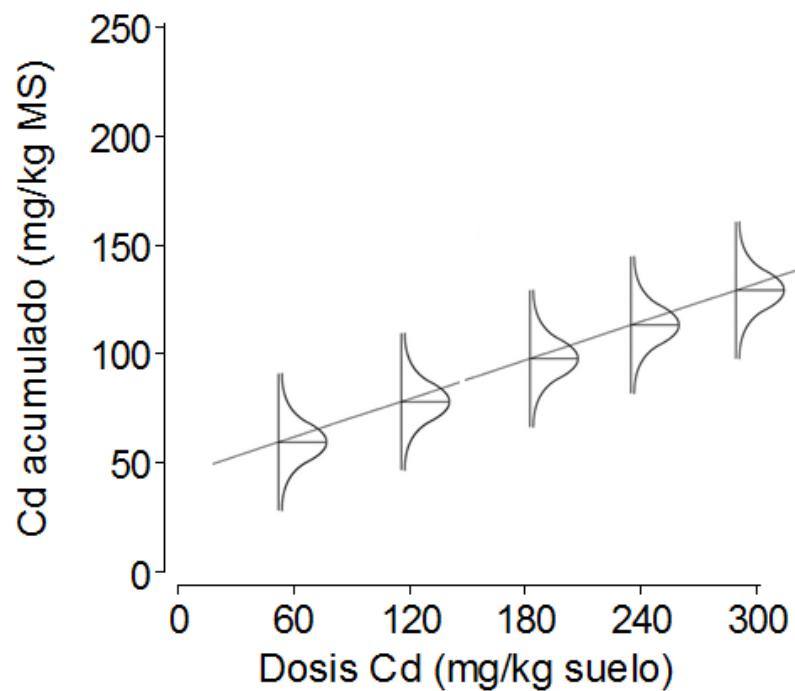
Adriana Pérez
Depto de Ecología, Genética y Evolución
FCEN, UBA

Restauración con césped de suelos contaminados con cadmio



2

- Interesa estudiar la capacidad detoxificadora del césped *Eremochloa ophiuroides* en suelos contaminados con Cd
 - UE
 - VR (Y)
 - VE (X)
 - Réplicas
 - Modelo
- A 20 macetas con césped se les asignará una de **5 dosis de Cd** diferente ($60, 120, 180, 240$ y $300 \text{ mg Cd kg}^{-1}$); 4 macetas por dosis
- Luego de 36 días en invernadero se medirá el **Cd acumulado** por la planta (expresado como mg Cd kg^{-1} materia seca)
- Se sospecha una relación lineal



Modelo de regresión lineal simple

3

equivalentes

$$\left\{ \begin{array}{l} Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i=1 \dots n \\ E_{Y/X} = \beta_0 + \beta_1 X_i \end{array} \right.$$

Valor esperado de $Y = \mu_{Y/x}$

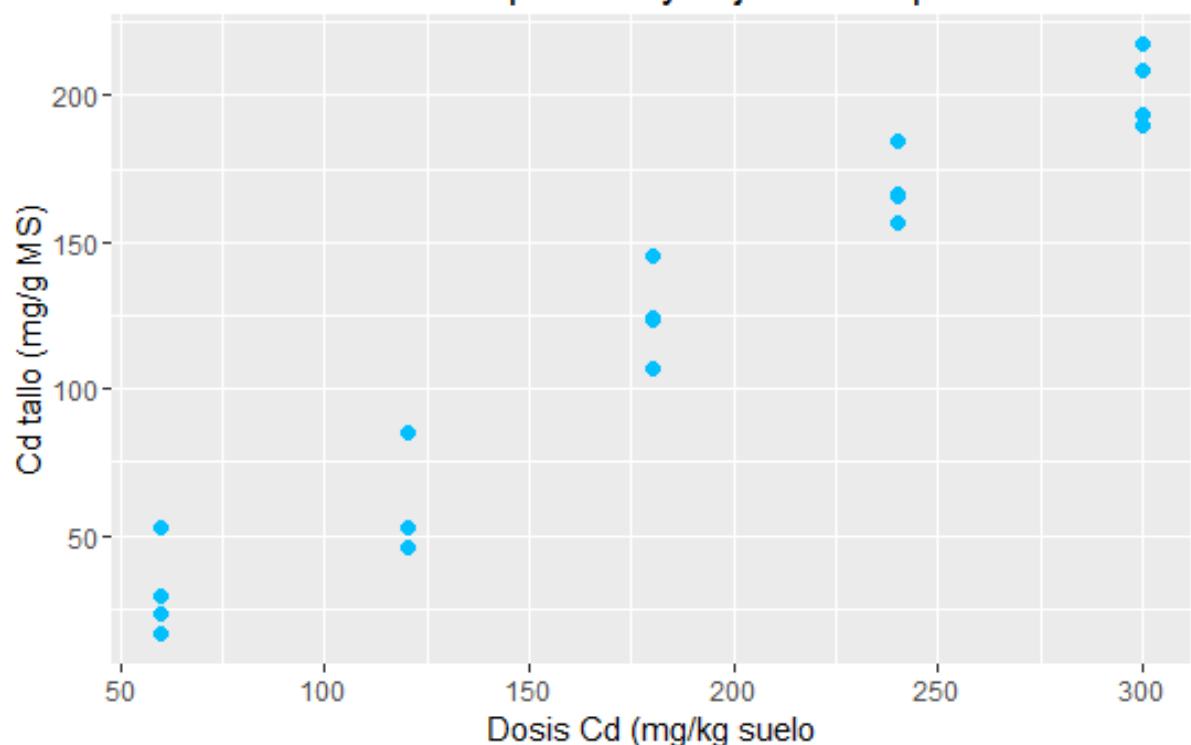
- Y_i es la i -ésima observación de la variable dependiente Y
- X_i es el i -ésimo valor de la variable predictora X
- β_0 y β_1 son los **parámetros** ordenada al origen y pendiente (o coeficiente de regresión)
 - Si el alcance del modelo incluye a $X=0$, β_0 es el valor esperado de Y cuando $X=0$
 - β_1 indica el cambio esperado en Y por cada aumento unitario de X
- ε_i es el error aleatorio, variación de Y no explicada por X ;

$$\varepsilon_i \sim NID(0, \sigma^2)$$

Dosis Cd (mg Cd/kg)	Concentración Cd (mg Cd/kg MS)	
	Tallo y hojas	Raíz
60	23,2	104,6
	16,2	156,0
	52,7	114,9
	29,1	176,9
120	52,5	258,4
	45,7	340,9
	52,9	205,3
	84,9	366,8
180	123,5	457,5
	106,9	540,8
	123,9	472,5
	145,7	294,3
240	166,8	612,7
	165,9	789,6
	184,3	622,9
	157,0	562,6
300	208,4	988,9
	189,9	1067,1
	217,7	959,6
	193,2	962,9

```
> summary(cadmio)
dosis_cd          cd_tallo          cd_raiz
Min.   : 60   Min.   :16.20   Min.   :104.6
1st Qu.:120  1st Qu.:52.65  1st Qu.:245.1
Median :180  Median :123.70 Median :465.0
Mean   :180  Mean   :117.02  Mean   :502.8
3rd Qu.:240  3rd Qu.:171.18 3rd Qu.:664.6
Max.   :300  Max.   :217.70 Max.   :1067.1
```

Absorción de Cd por tallo y hojas de *E.ophiuroides*



```
modelo1<-lm(cd_tallo~dosis_cd, data=cadmio)
```

Estimación de los parámetros del modelo

5

```
> summary(modelo1)
```

Call:

```
lm(formula = cadmio$cd_tallo ~ cadmio$dosis_cd)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.970	-11.220	1.730	7.655	28.680

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-19.03000	8.35547	-2.278	0.0352 *
cadmio\$dosis_cd	0.75583	0.04199	18.001	5.88e-13 ***

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

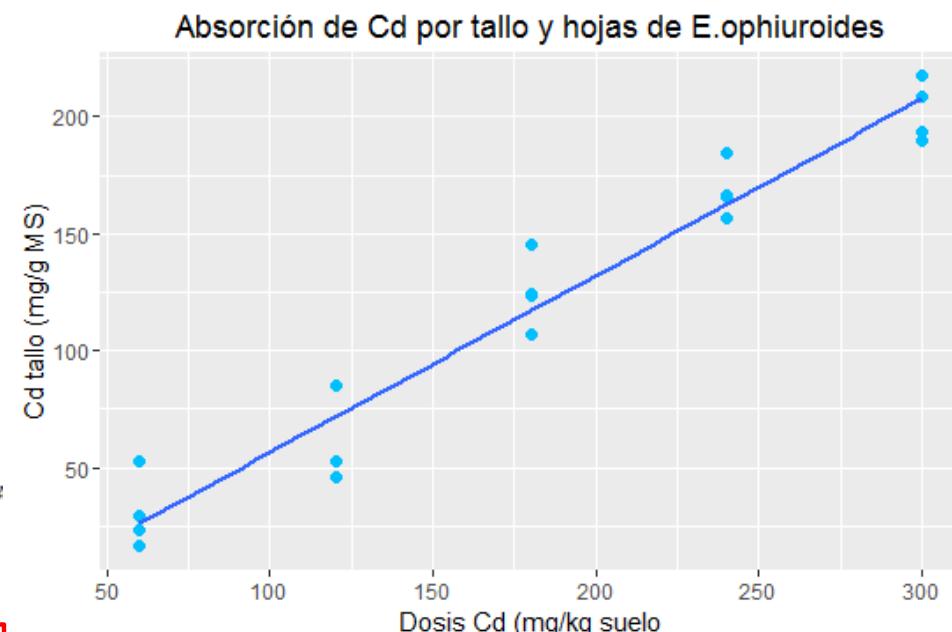
Residual standard error: 15.93 on 18 degrees of freedom

Multiple R-squared: 0.9474, Adjusted R-squared: 0.9445

F-statistic: 324 on 1 and 18 DF, p-value: 5.882e-13

$$\hat{y} = -19,03 + 0,76x$$

$$Cd\ tallo = -19,03 + 0,76 \cdot Cd\ suelo$$



Si los errores son independientes y su distribución es normal, los estimadores por mínimos cuadrados son los estimadores por máxima verosimilitud

Calculando varianzas

6

	dosis	Cd	Cd	tallo	Predichos	Residuos
1		60		23.2	26.32	-3.12
2		60		16.2	26.32	-10.12
3		60		52.7	26.32	26.38
4		60		29.1	26.32	2.78
5		120		52.5	71.67	-19.17
6		120		45.7	71.67	-25.97
7		120		52.9	71.67	-18.77
8		120		84.9	71.67	13.23
9		180		123.5	117.02	6.48
10		180		106.9	117.02	-10.12
11		180		123.9	117.02	6.88
12		180		145.7	117.02	28.68
13		240		166.8	162.37	4.43
14		240		165.9	162.37	3.53
15		240		184.3	162.37	21.93
16		240		157.0	162.37	-5.37
17		300		208.4	207.72	0.68
18		300		189.9	207.72	-17.82
19		300		217.7	207.72	9.98
20		300		193.2	207.72	-14.52

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad e_i = y_i - \hat{y}_i$$

$$S^2_Y = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{86834.53}{19} = 4570.23 \text{ (mg/g MS)}^2$$

$$S^2_e = S^2_{Y/X} = CM_{error} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \\ = \frac{\sum_{i=1}^n (y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i))^2}{n-2} = \frac{4569.63}{18} = 253.87 \text{ (mg/g MS)}^2$$

Residuos
estandarizados

$$RE = \frac{e_i}{\sqrt{S^2_e}}$$

Variación total de VR =
variación explicada por el modelo + Variación no explicada (error o residual)

Grados de libertad: Piezas de información independiente = n – cantidad de parámetros que se debieron estimar previamente. Dividir por GL en vez de por n asegura que el estimador de σ^2 sea insesgado

Supuestos del modelo

7

- X medida sin error, no es V.A., sus valores son determinados por el investigador

Muchas veces no se cumple; pero es grave sólo si la magnitud del error de X es grande en relación a la magnitud de X (ie, $> 10\%$). En ese caso, **Modelo II**

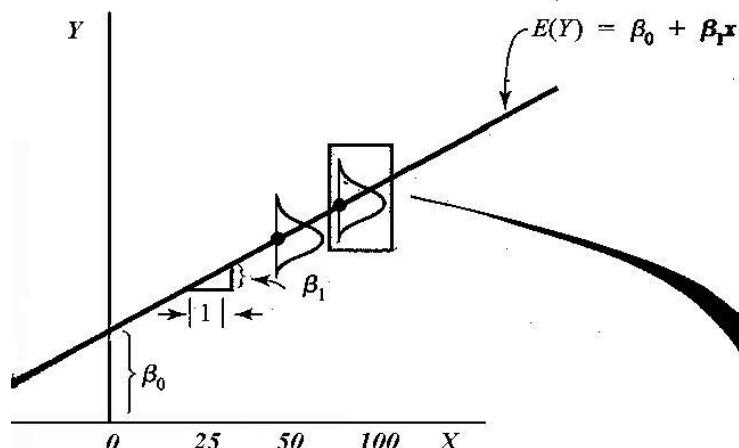
- **Independencia** entre las observaciones (no se cumple para medidas repetidas o datos estructurados espacialmente). No debe existir correlación entre observaciones. $cov(i;j) = 0$ para $i \neq j$.

Supuestos del modelo

8

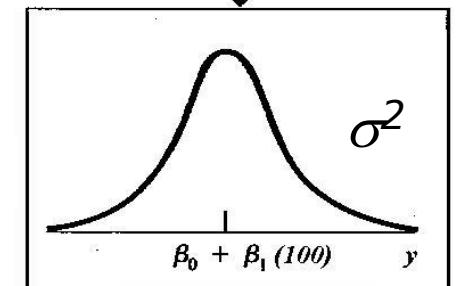
No es necesarios para estimar β_0 y β_1 , pero sí para hacer inferencia

- Para cada valor de X existe una subpoblación de Y
 - La media de cada una de estas subpoblaciones es
$$E_{Y/X} = \beta_0 + \beta_1 X_i \text{ (linealidad)}$$
 - La distribución de cada subpoblación es **normal**
$$Y_{i/X} \approx NID(\mu_{Y/X}, \sigma^2)$$
 - las varianzas de las subpoblaciones son iguales, es decir que el modelo asume una **varianza constante** σ^2 , sin importar el nivel de X
$$\text{Var}[Y/X] = \sigma^2$$



Estos supuestos se pueden resumir en: $\varepsilon_i \approx NID(0, \sigma^2)$

Los residuos constituyen el insumo básico para estudiar los supuestos del modelo



Linealidad y homocedasticidad

9

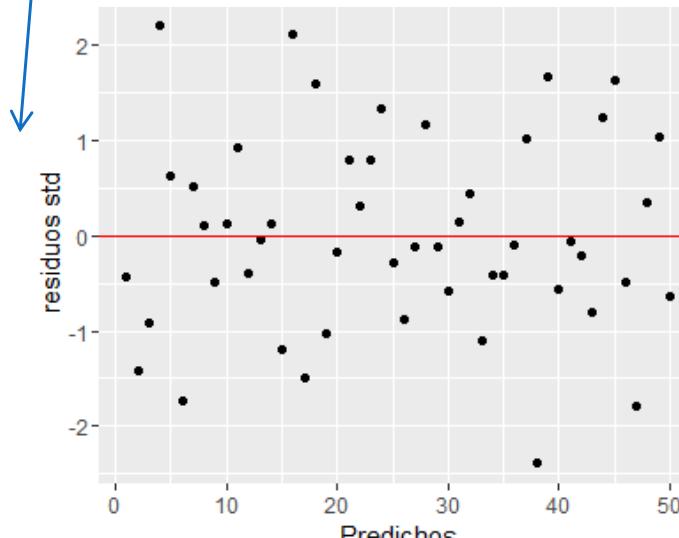
Gráficamente: **gráfico de dispersión de residuos vs predichos**

- Determinar si el modelo lineal está bien especificado (los residuos deberían distribuirse aleatoriamente, sin patrones)
- Determinar si la variabilidad es constante (homocedasticidad)
- Detectar **outliers o datos atípicos en Y** (con residuo muy grande)

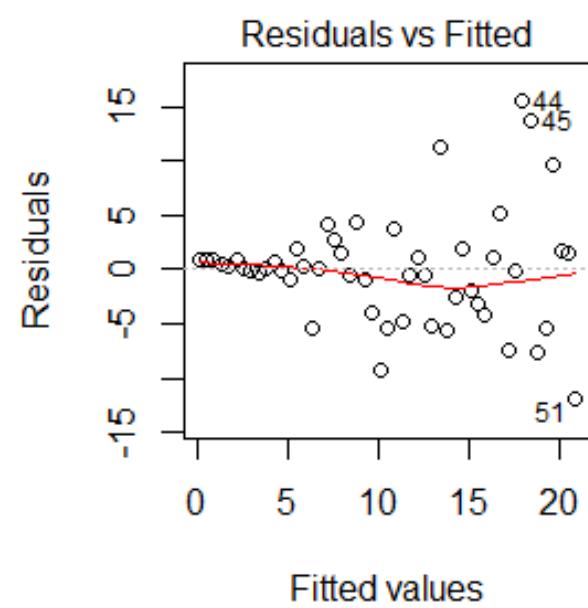
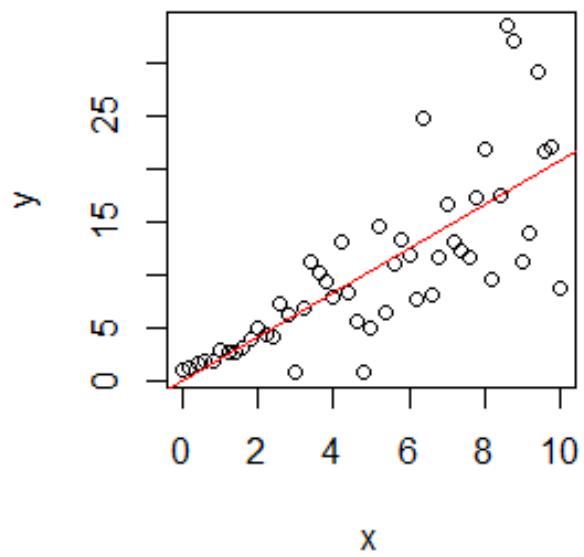
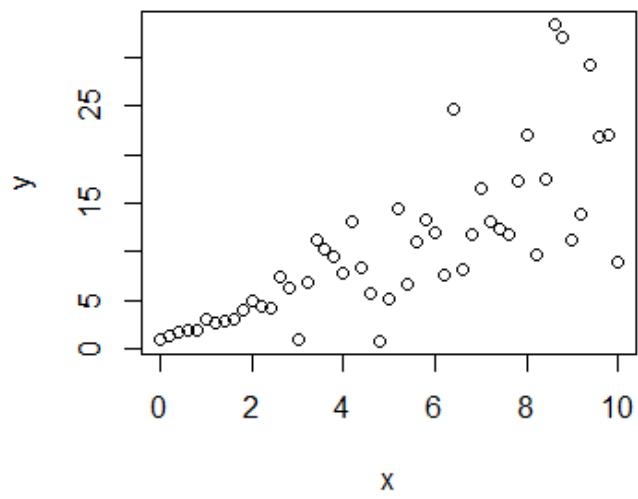
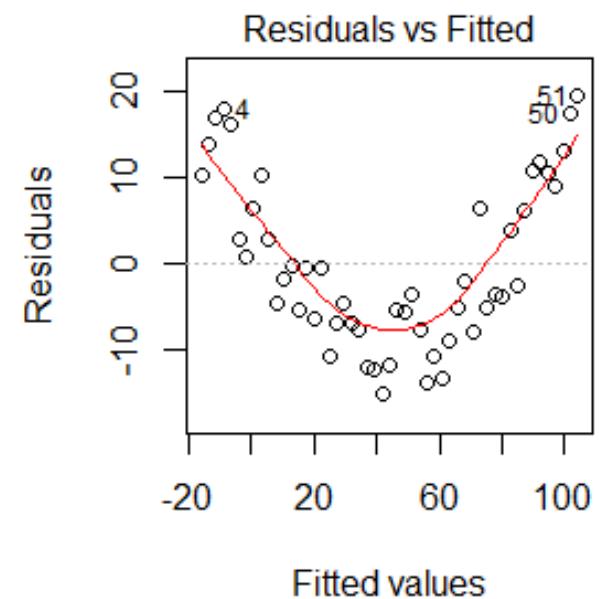
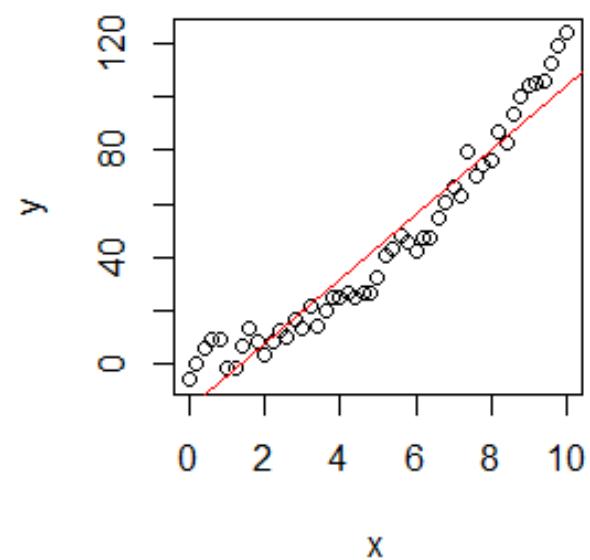
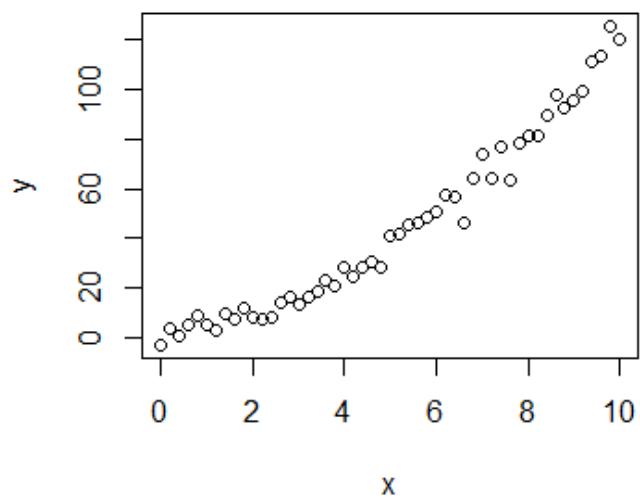
Se espera encontrar una distribución al azar (sin patrones) y con variabilidad constante

Conviene utilizar residuos estandarizados, ya que permite detectar outliers ($RE>2$ o $RE<2$)

Gráfico de dispersión de residuos vs predichos



El valor predicho para cada observación es la respuesta obtenida a partir de la ecuación estimada



Homocedasticidad

11

- Analíticamente: Prueba de Levene
- Es un análisis de la varianza de un factor utilizando como VR el valor absoluto de los residuos
- H_0 : todas las varianzas poblacionales son iguales
- Solo se puede efectuar cuando existen réplicas para cada nivel de X (raro en estudios observacionales)

```
library(car)
levene.test(modelo2, data=cadmio)
```

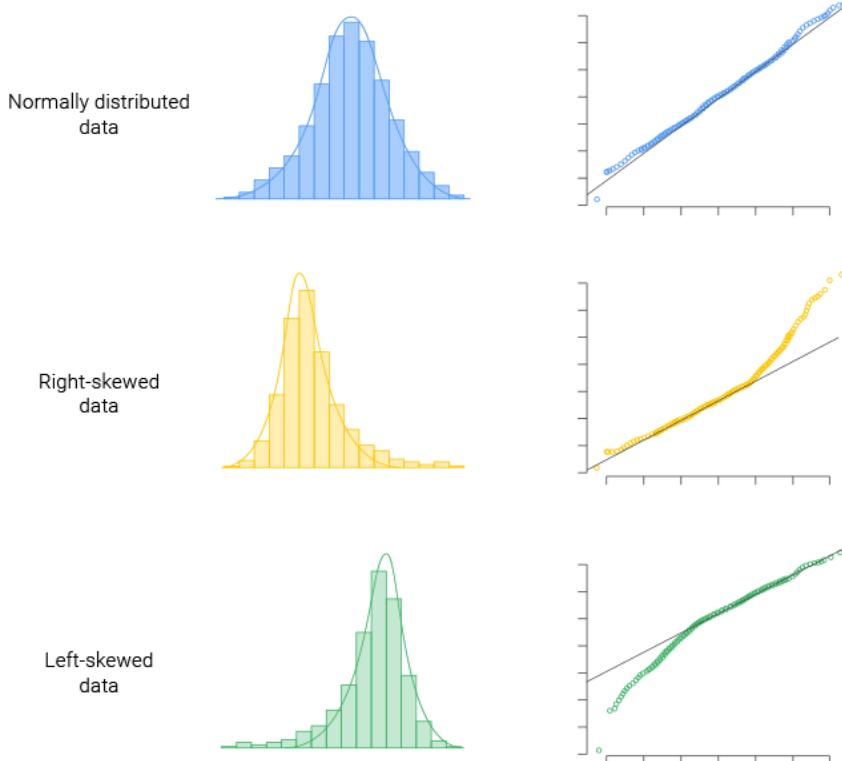
Normalidad

12

- Gráficamente: [QQ plot](#)

Es un gráfico de dispersión de los percentiles (quantiles) de las observaciones vs los percentiles (cuantiles) de una distribución normal con media y DE estimados a partir de la muestra

La normalidad se estudia utilizando los residuos del modelo



Normalidad

13

- Analíticamente: Prueba de Shapiro-Wilk

$$H_0: \varepsilon_i \approx \text{normal}$$

El estadístico W^* de la prueba de Shapiro-Wilk oscila entre 0 y 1. Cuanto más cercano a 1 mayor evidencia de normalidad. Básicamente, mide cuan cerca de una recta está la curva que describen los puntos graficados en el QQ-plot

Los errores del modelo no son observables; para probar el supuesto se utilizan sus correlatos empíricos, los residuos

Causas del incumplimiento de los supuestos

14

- la presencia de outliers puede generar heterocedasticidad
- Si la distribución de la variable no es normal (lognormal, gamma, etc) puede detectarse tanto falta de normalidad como de homocedasticidad
- La falta de linealidad implica que la relación de la VR con la VE no es lineal. Puede solucionarse agregando más términos al modelo (cuadrático, cúbico, interacciones, etc) o tratando a las VE cuantitativas como cualitativas

Consecuencias del incumplimiento de los supuestos

15

Heterocedasticidad

Las estimaciones de los parámetros son insesgadas y consistentes, pero los errores estándares de los estimadores no ⇒ la inferencia (Pruebas de hipótesis e IC) no es confiable; provoca un aumento en la probabilidad de cometer error tipo I; el nivel de confianza no es el propuesto

- Los efectos son más graves si el diseño es desbalanceado
- Más grave si una varianza es mucho mayor que el resto
- Menos grave si una varianza es mucho menor que el resto

No normalidad

Menos grave. Si el apartamiento de la normalidad no es severo y no hay heterocedasticidad, las estimaciones e inferencia son razonables

No linealidad

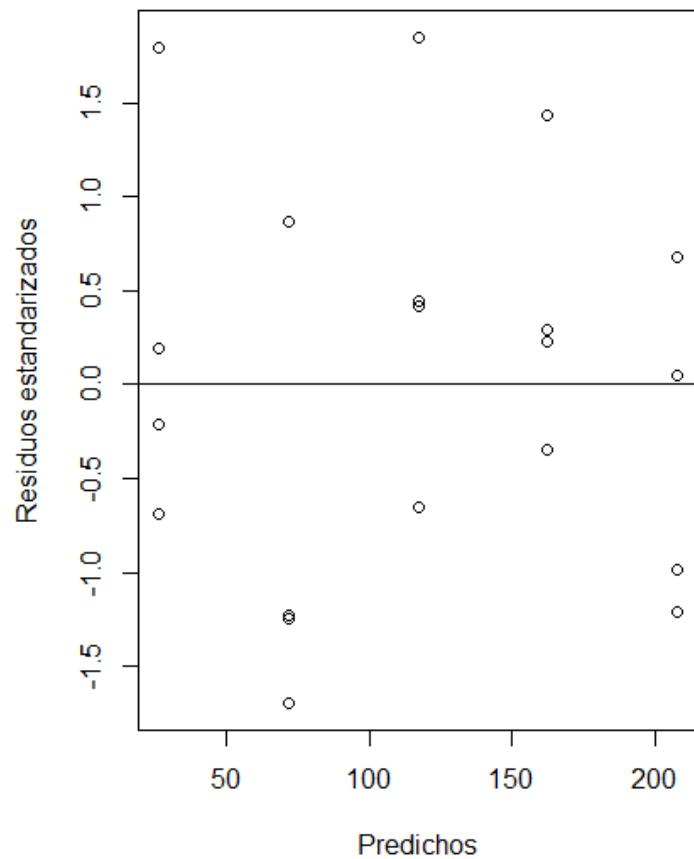
Los coeficientes de regresión no miden la verdadera relación con la VE

¿Cómo corregimos la heterocedasticidad o la falta de normalidad?

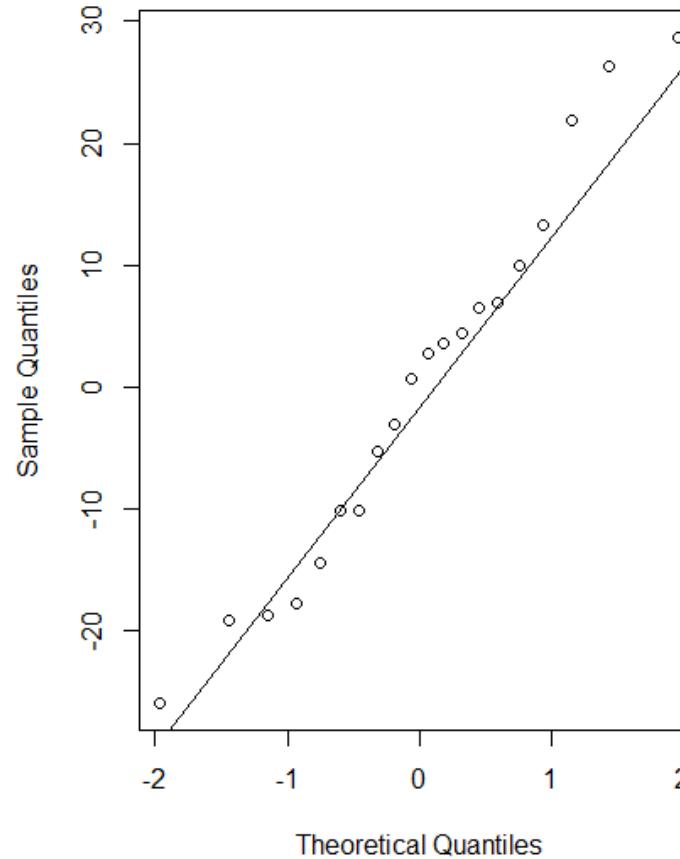
16

1. Aplicando modelos que permitan modelar la heterocedasticidad
2. Aplicando modelos que permitan otra distribución de probabilidad de la VR (**modelos lineales generalizados**)
3. Aplicando transformaciones monotónicas a los datos (i.e. aplicando logaritmo), pero que implican cambiar la escala de la VR
4. Regresión ponderada (para heterocedasticidad): menor peso a los datos con mayor dispersión
5. Métodos robustos

Gráfico de dispersión de RE vs PRED



Normal Q-Q Plot



Levene's Test for Homogeneity of Variance (center = median)

Df	F value	Pr(>F)
group	4	0.0782
		0.9878

15

> shapiro.test(e)

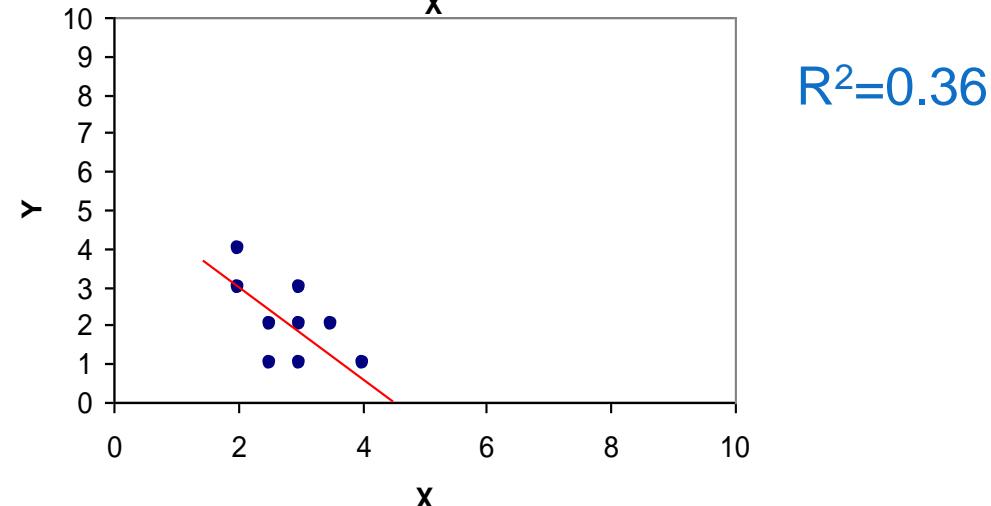
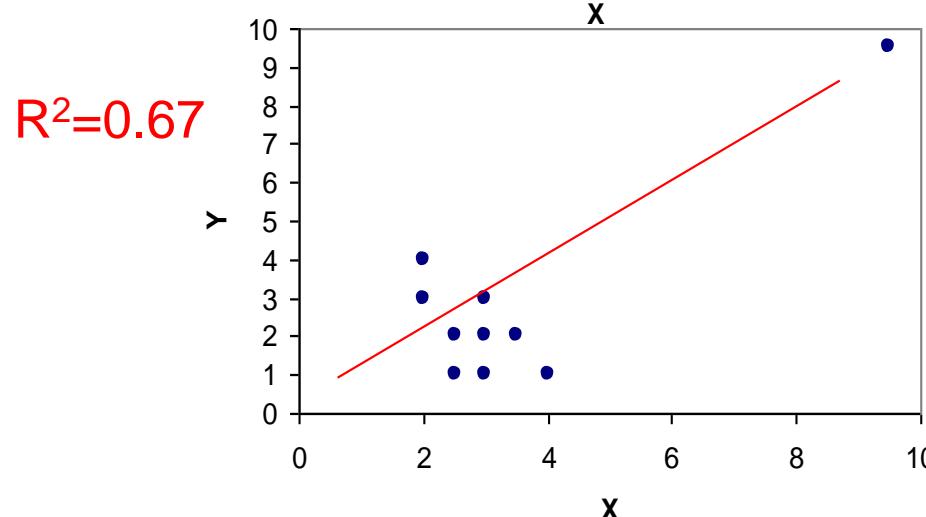
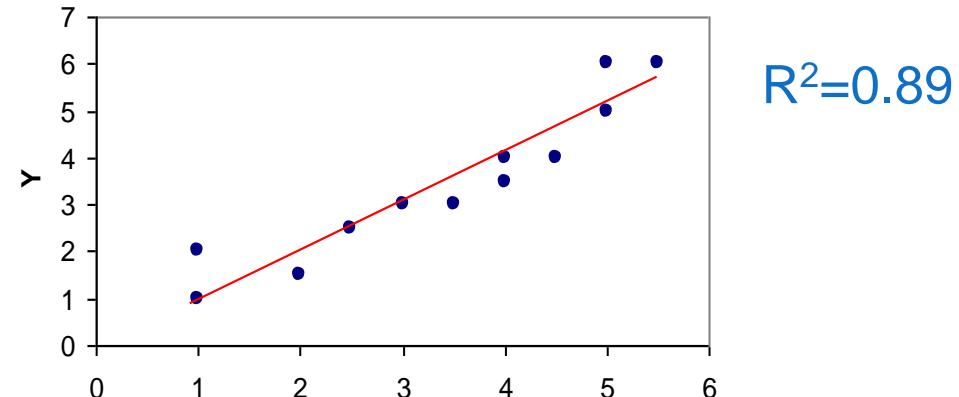
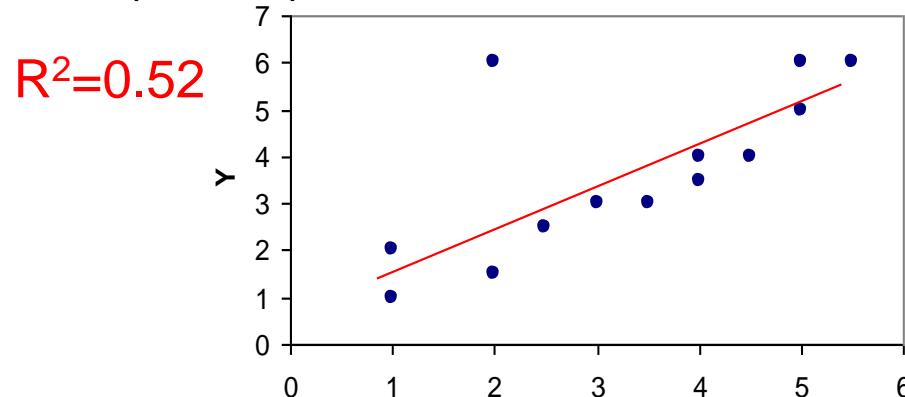
Shapiro-Wilk normality test

data: e
W = 0.9676, p-value = 0.7035

Observaciones atípicas y observaciones influyentes

18

- **Atípicas (outliers en Y):** Observaciones con un patrón distinto al resto de los datos, que producen un residuo grande
- **Influyentes (con alta palanca):** Observaciones cuyo valor de X se encuentra alejado del promedio y que tienen mucho peso en las estimaciones de los parámetros. Al ser eliminadas pueden provocar cambios sustanciales en las estimaciones



Cómo detectar observaciones atípicas

19

□ Residuos estandarizados

- Permite detectar outliers en Y
- Se identifican valores con RE <-2 o >2

$$RE = \frac{e_i}{\sqrt{S_e^2}}$$

□ Residuos studentizados

- Permite detectar outliers en Y
- Se calculan como:
- Se identifican valores con RS <-2 o >2
- h_{ii} es el Leverage o palanca

$$RS = \frac{e_i}{\sqrt{S_e^2(1 - h_{ii})}}$$

Cómo detectar observaciones influyentes

20

- **Leverage o palanca** h_{ii}
 - Es una medida que mide cuán lejos cae el valor de X_i de la media muestral de las X (outlier en X)
 - Mide, de alguna manera, cuánto es el aporte de la observación i-ésima a la varianza muestral de las X
 - Puede tomar valores entre $1/n$ y 1
 - Valores altos (alta palanca) indican que esa observación contribuye más en la predicción de Y, es decir que fuerza a la recta a pasar por el valor observado de Y
 - Se consideran outliers en X las observaciones con $h_{ii} > 2k/n$, donde k es la cantidad de variables predictoras del modelo

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

Cómo detectar observaciones influyentes

21

□ Distancia de Cook

- Mide el efecto global de una observación sobre las estimaciones de los parámetros del modelo y sobre los valores predichos
- Grandes valores indican observaciones cuya eliminación tiene gran influencia sobre las estimaciones y sobre los valores predichos (dato influyente)

$$D_{Cook} = \left(\frac{e_i}{\sqrt{CMerror(1-h_{ii})}} \right)^2 \left(\frac{h_{ii}}{1-h_{ii}} \right) \left(\frac{1}{p} \right)$$

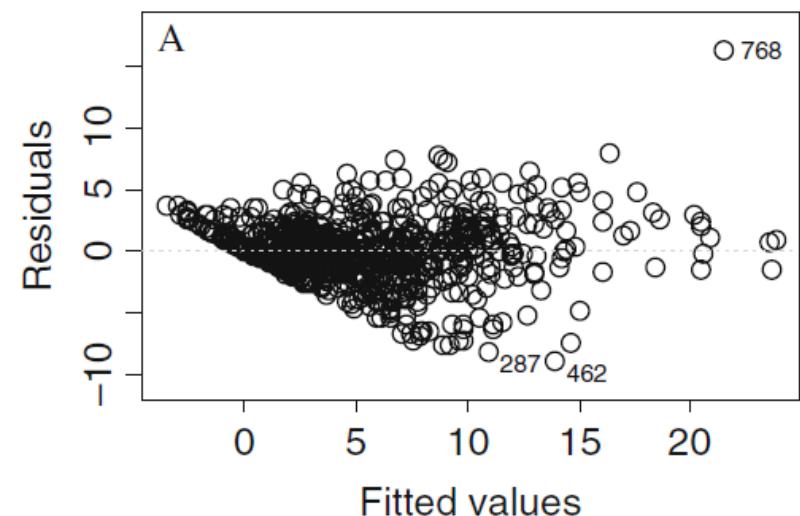
$$D_{Cook} = \frac{\sum (\hat{y}_j - \hat{y}_{j(i)})^2}{pCMerror}$$

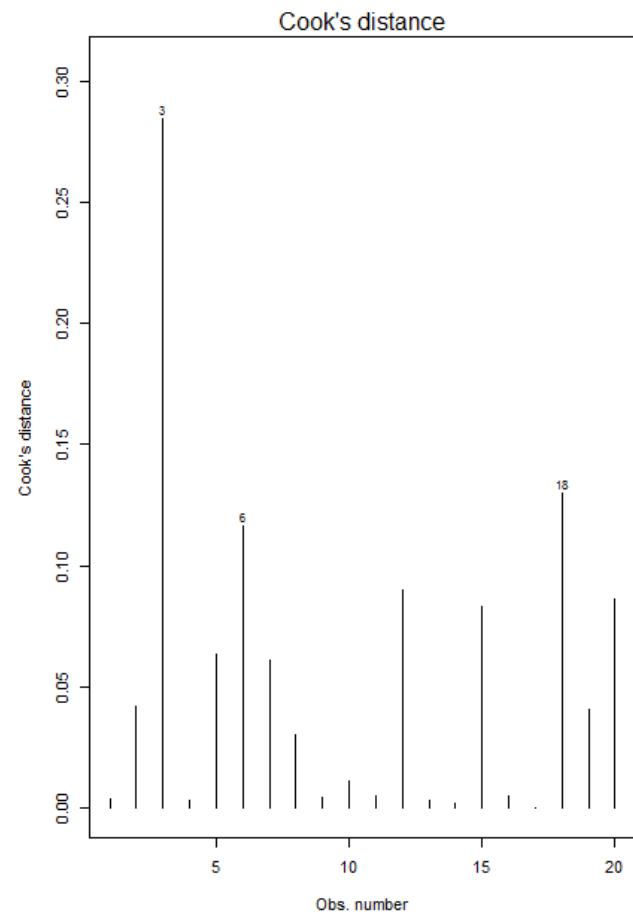
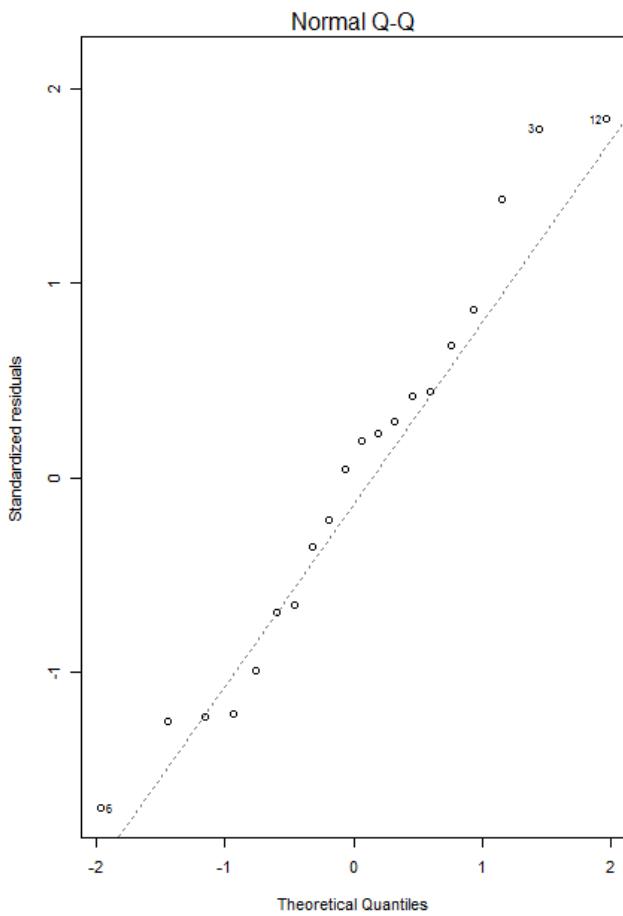
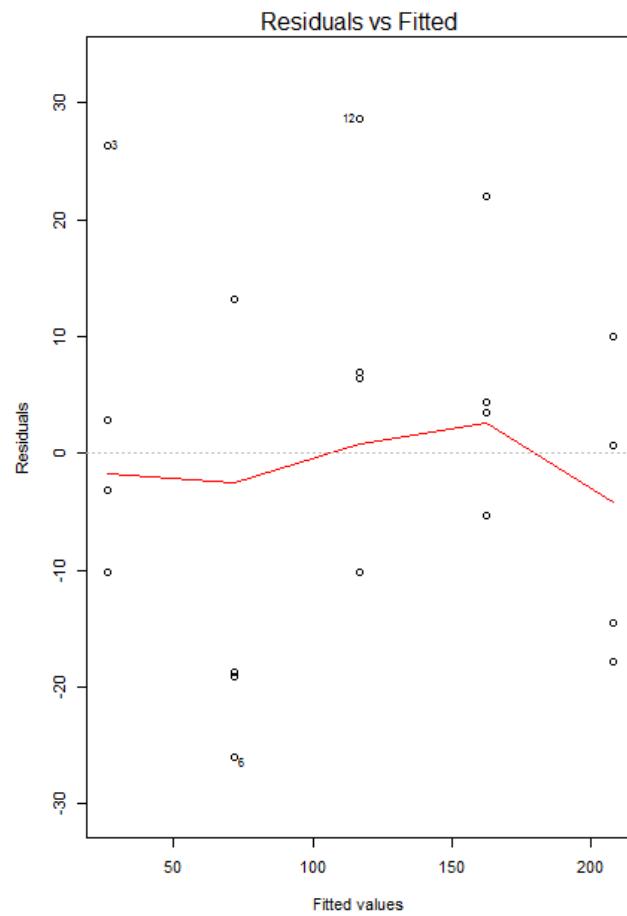
- Se consideran influyentes las observaciones con $D > 1$

VR limitada: Datos censurados o truncados

22

- Truncamiento: cuando por el proceso de recopilación de los datos solo se obtiene datos de un subconjunto de una población de interés más grande.
 - Por ejemplo: VR: Glucosa en plasma, pero solo participaron individuos con $\text{Glu} > 110$
- Censura: cuando todos los valores de un cierto rango se transforman (o se informan como) un solo valor.
 - Por ejemplo: VR: tiempo de respuesta (hasta 60 seg)
- Se detectan patrones en los residuos
- Exigen un modelado especial





No se detectan violaciones a los supuestos -> podemos hacer inferencia

```
> summary(model1)
```

Call:

```
lm(formula = cadmio$cd_tallo ~ cadmio$dosis_cd)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.970	-11.220	1.730	7.655	28.680

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-19.03000	8.35547	-2.278	0.0352 *
cadmio\$dosis_cd	0.75583	0.04199	18.001	5.88e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 15.93 on 18 degrees of freedom

Multiple R-squared: 0.9474, Adjusted R-squared: 0.9445

F-statistic: 324 on 1 and 18 DF, p-value: 5.882e-13

Ojo: diferencias
"significativas" no quiere decir "importantes", sino
"poco probables de obtener sólo por azar"

Alternativamente puede construirse un IC para para β_1
y determinar si cero pertenece o no a dicho intervalo

$$\hat{\beta}_1 \pm t_{n-2;1-\alpha/2} \sqrt{\frac{s^2_{error}}{\sum(x_i - \bar{x})^2}}$$

```
> round(confint(modelo1), 2) (en paquete ISwR)
```

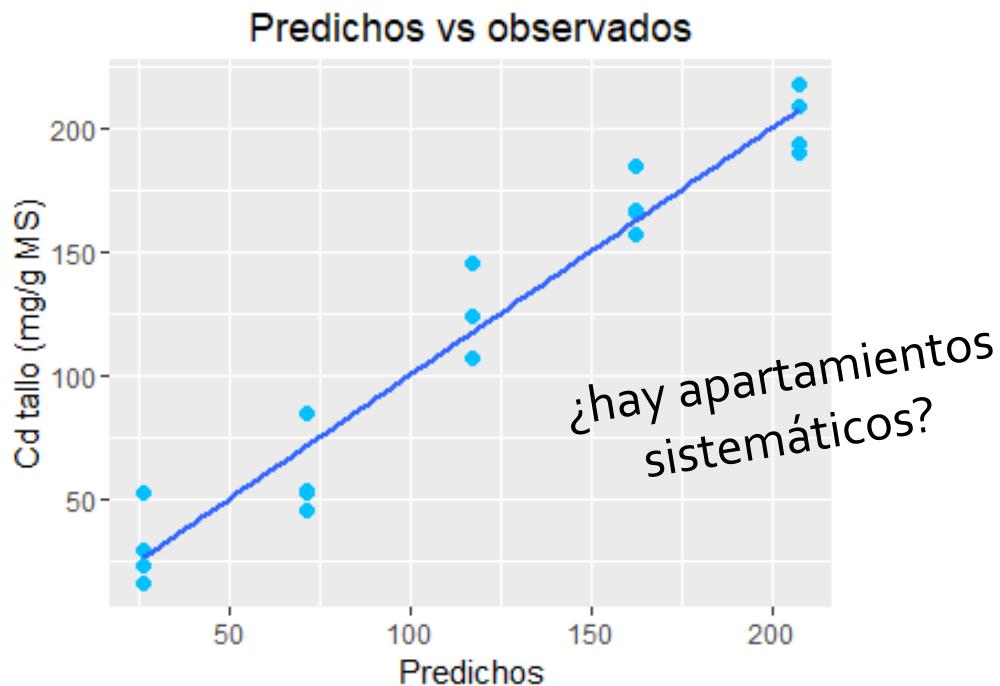
	2.5 %	97.5 %
(Intercept)	-36.58	-1.48
cadmio\$dosis_cd	0.67	0.84



β_1 mide la magnitud del efecto

Validación del modelo

25



$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

`cor(pre, cd_tallo)`
0.9733321

- Coeficiente de determinación R^2 para evaluar cuánto de la variabilidad de Y está explicada por el modelo
- Correlación entre predichos y observados para evaluar la capacidad predictiva del modelo. Pero está sobrevaluada!
- La validación cruzada es un conjunto de métodos para medir el desempeño de un modelo evaluando su capacidad para predecir **un nuevo conjunto de datos** (próximamente)

Importancia de visualizar los datos

Intervalos de confianza para las predicciones

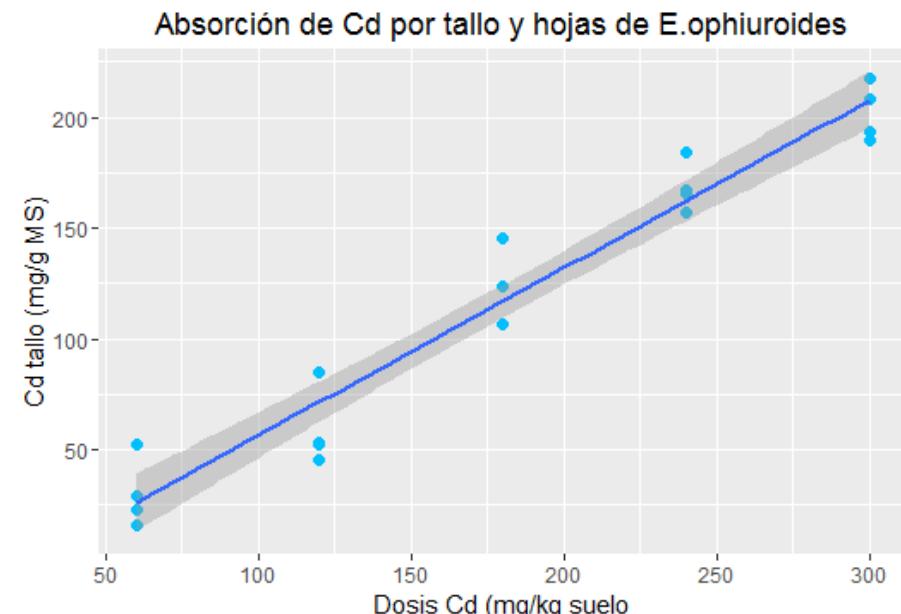
26

- Dos aplicaciones de los modelos de regresión: explicación y predicción
- Una vez **estimados los parámetros y validado el modelo**, es posible realizar **predicciones** acerca del valor que tomaría la VR para una unidad extramuestral.
- Se pueden construir **intervalos de confianza** sobre dichos valores
- Los pronósticos son válidos en el rango estudiado

IC para $\mu_{Y/X}$

$$\hat{y}_0 \pm t_{n-2;1-\alpha/2} \sqrt{S_e^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right]}$$

Penaliza el alejamiento del centro





PHYSIOLOGICAL RESPONSES AND TOLERANCE THRESHOLD TO CADMIUM CONTAMINATION IN *EREMOCHLOA OPHIUROIDES*

Yiming Liu, Kai Wang, Peixian Xu, and Zhaolong Wang

School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai, China

Table 2 The capacity of Cd phytoextraction of centipedegrass under various Cd concentration treatments

Treatments (mg Cd/kg)	Biomass (g/pot)		Cd concentration (mg/kg DW)		Cd accumulation (mg/pot)			Phytoextraction (%)
	Shoot	Root	Shoot	Root	Shoot	Root	Total	
0	58.9 a	5.1 a	—	—	—	—	—	—
60	58.5a	5.2 a	30.3 e	104.6 e	1.8 bc	0.5 d	2.3	0.60
120	57.6 ab	4.7 a	59.0 d	258.4 d	3.4 b	1.2 cd	4.6	0.57
180	57.9 a	4.4 a	135.1 c	457.5 c	7.8 a	2.0 bc	9.8	0.87
240	53.9 ab	4.3 a	168.5 b	612.7 b	9.1 a	2.6 ab	11.7	0.76
300	46.4 b	3.4 b	202.3 a	988.9 a	9.4 a	3.4 a	12.8	0.63

Dosis Cd (mg Cd/kg)	Concentración Cd (mg Cd/kg MS)	
	Tallo y hojas	Raíz
60	23,2	104,6
	16,2	156,0
	52,7	114,9
	29,1	176,9
120	52,5	258,4
	45,7	340,9
	52,9	205,3
	84,9	366,8
180	123,5	457,5
	106,9	540,8
	123,9	472,5
	145,7	294,3
240	166,8	612,7
	165,9	789,6
	184,3	622,9
	157,0	562,6
300	208,4	988,9
	189,9	1067,1
	217,7	959,6
	193,2	962,9

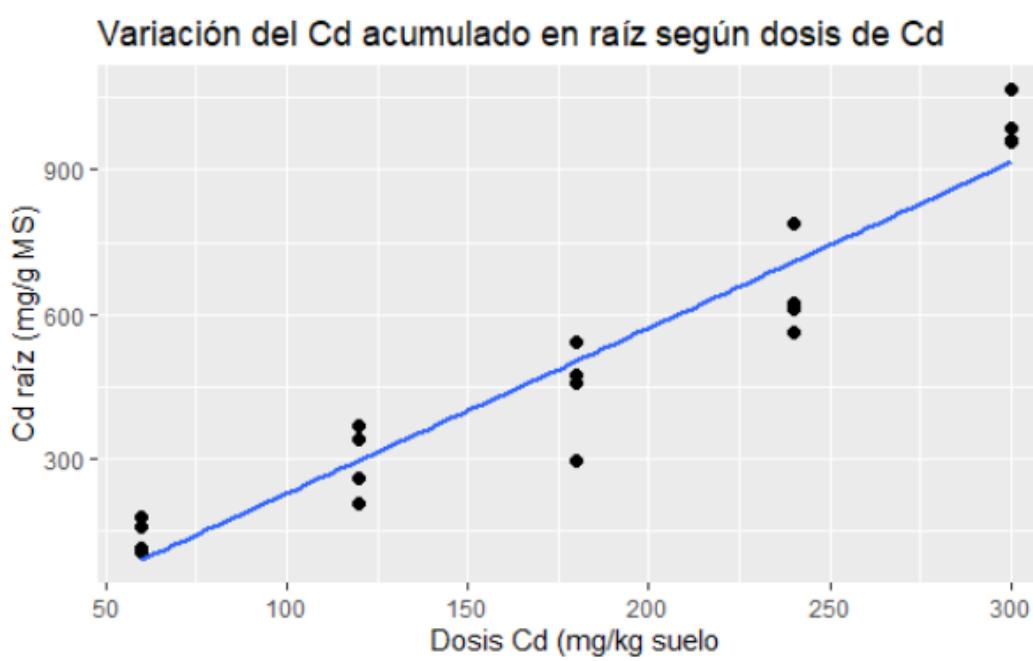
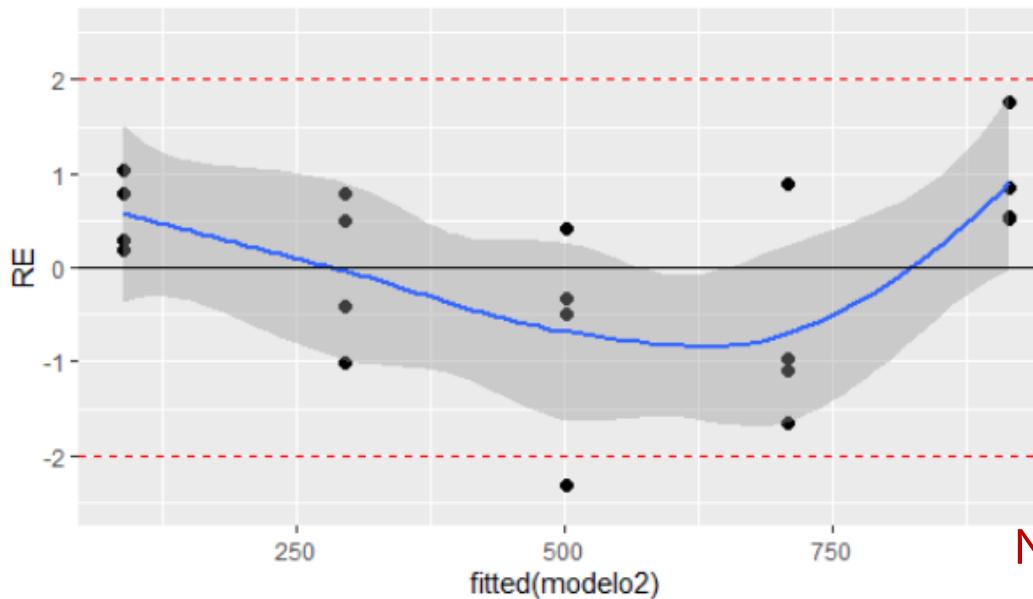


Gráfico de RE vs predichos



No linealidad

Residual standard error: 92.65 on 18 degrees of freedom
 Multiple R-squared: 0.9171, Adjusted R-squared: 0.9125

Regresión polinomial

29

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \varepsilon_i$$

- El modelo incluye términos de potencias sucesivas de la VE cuantitativa X
- Es un caso particular de **regresión múltiple**: las distintas potencias de X actúan como distintas v. explicatorias
- p es el **grado** del polinomio (máxima potencia)
- Si $p = 1$ entonces la regresión polinomial se reduce a regresión lineal simple
- El grado máximo al que se puede ajustar un conjunto de n datos es $n-1$. Si se desea hacer inferencia, $n-2$
- Y como siempre:

$$\varepsilon_i \approx NID(0, \sigma^2)$$

Variación del Cd acumulado en raíz según dosis de Cd

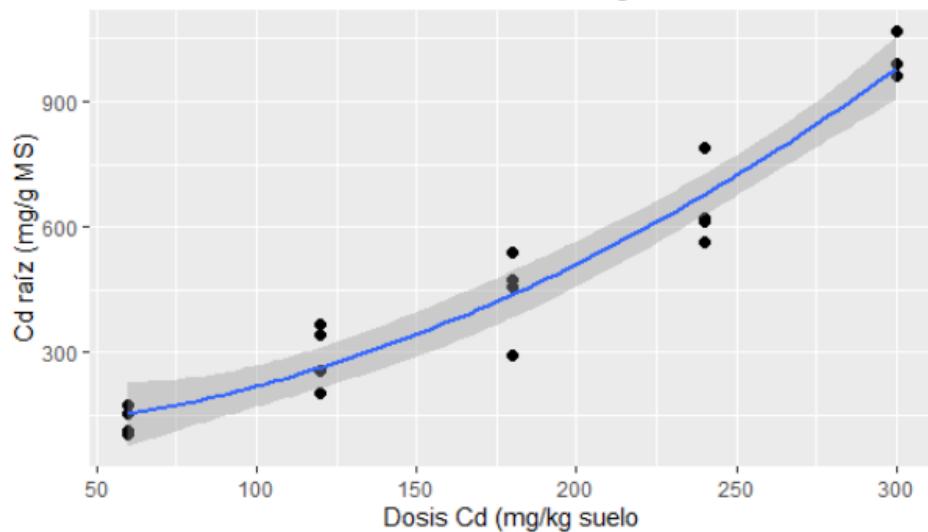
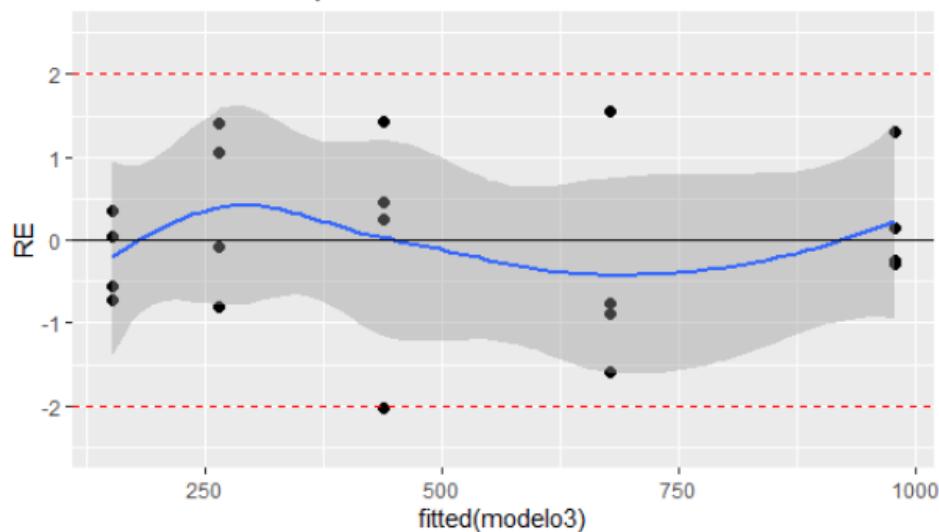


Gráfico de RE vs predichos



$$E(y_i) = \beta_0 + \beta_1 \text{dosis}_i + \beta_2 \text{dosis}_i^2$$

```
modelo3 <- lm(cd_raiz ~ dosis_cd
+ I(dosis_cd^2), bd)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.042e+02	8.160e+01	1.277	0.21891
dosis_cd	2.802e-01	1.036e+00	0.270	0.79009
dosis_cd_cuad	8.792e-03	2.824e-03	3.113	0.00633 **

Residual standard error: 76.09 on 17 degrees of freedom
Multiple R-squared: 0.9472, Adjusted R-squared: 0.941
F-statistic: 152.5 on 2 and 17 DF, p-value: 1.39e-11

Bibliografía

31

Quinn, G., & Keough, M. (2002). Cap 5: Correlation and regression.
In *Experimental Design and Data Analysis for Biologists*. Cambridge:
Cambridge University Press.

- 5.3.8 Assumptions of regression analysis
- 5.3.9 Regression diagnostics
- 5.3.10 Diagnostic graphics
(pp 92 -98)

BIOMETRÍA II

CLASE 4

ANALISIS DE LA VARIANZA MODELADO DE VARIANZAS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FCEN, UBA

Marcadores biológicos de contaminación ambiental

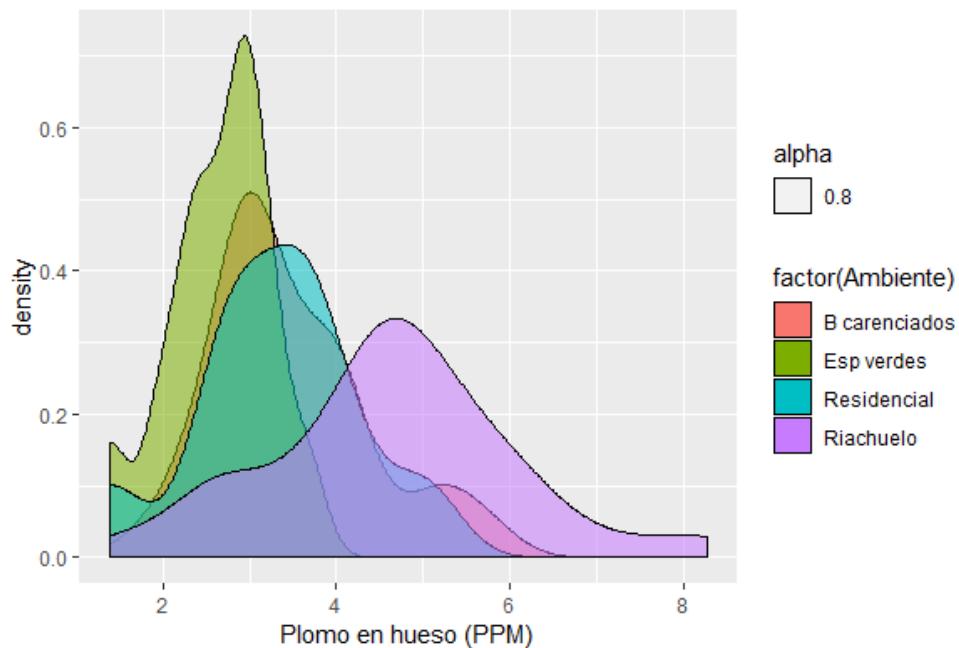
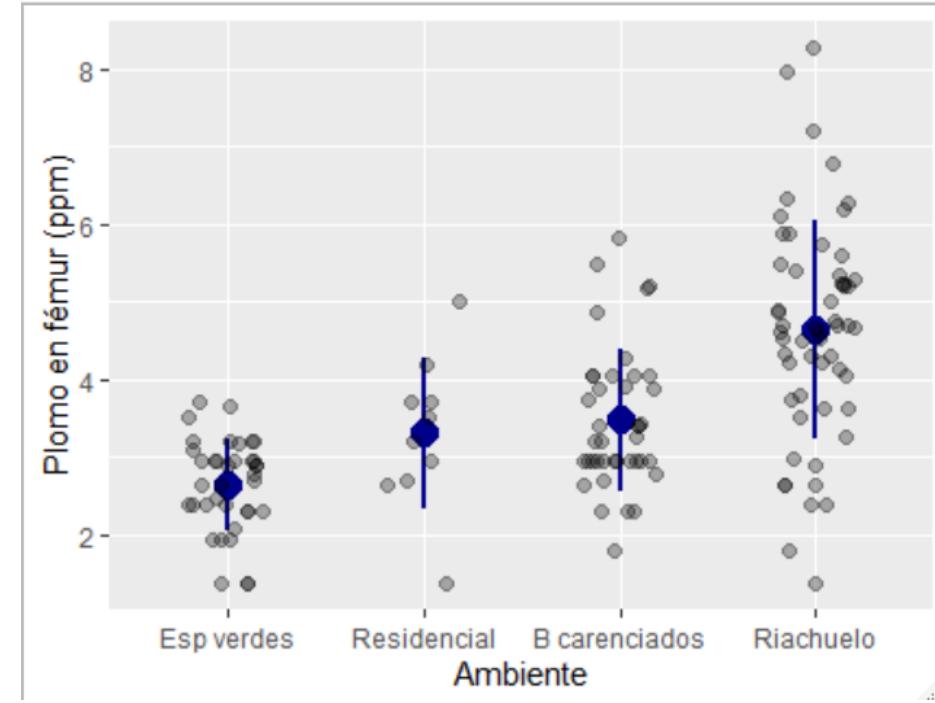
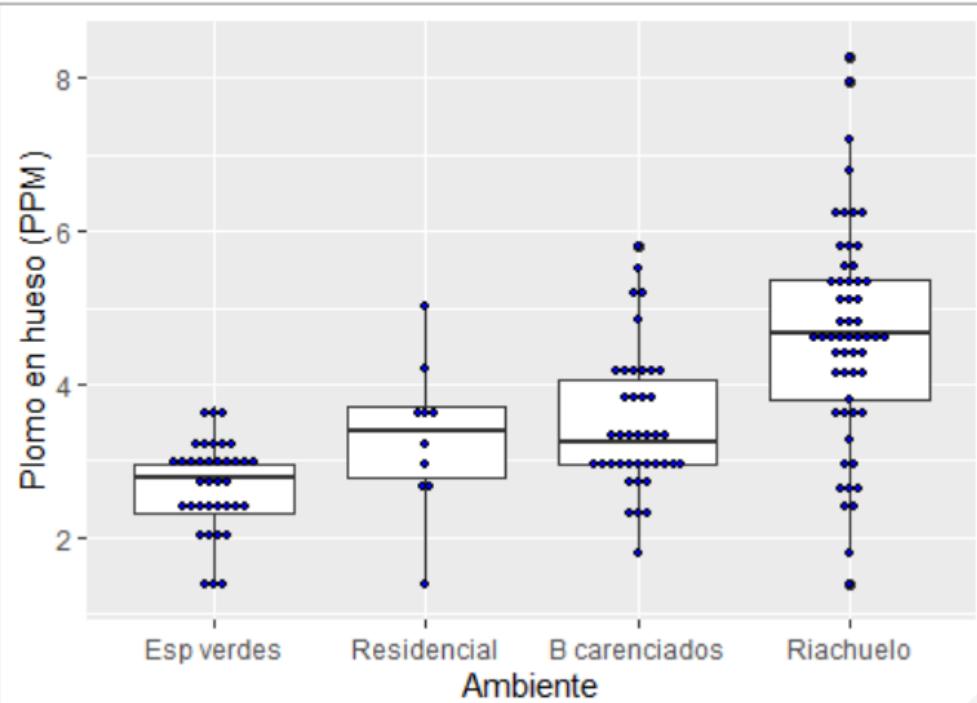


- Las ratas (genero *Rattus*) presentan áreas de actividad relativamente pequeñas, por lo que se ha sugerido que podrían ser usadas para la detección de contaminación ambiental por metales pesados del ambiente en el que viven
- Se desea comparar el nivel de acumulación media de plomo en ratas provenientes de distintos ambientes de CABA
- Para ello se efectuó un muestreo aleatorio de *Rattus norvegicus* en 4 ambientes contrastantes: *Espacios verdes*; *Barrios residenciales*; *Barrios carenciados* y *Costa del Riachuelo*.
- En cada ejemplar se registró el nivel de acumulación de plomo en fémur (ppm) (n=143)

- Experimento o estudio observacional?
- Unidad muestral / experimental? Independencia?
- variable respuesta? Tipo y potencial distribución de probabilidades?
- variable explicativa? Tipo y niveles?
- Modelo?

Plomo.txt

	Ambiente	Pb
1	Esp verdes	2.484907
2	Esp verdes	2.944439
3	Esp verdes	2.772589
4	Esp verdes	2.397895
5	Esp verdes	2.890372
6	Esp verdes	2.397895
7	Esp verdes	2.079442
8	Esp verdes	2.890372
9	Esp verdes	2.944439
10	Esp verdes	2.639057
11	Esp verdes	2.639057
12	Esp verdes	1.945910
13	Fen verdes	3.178054

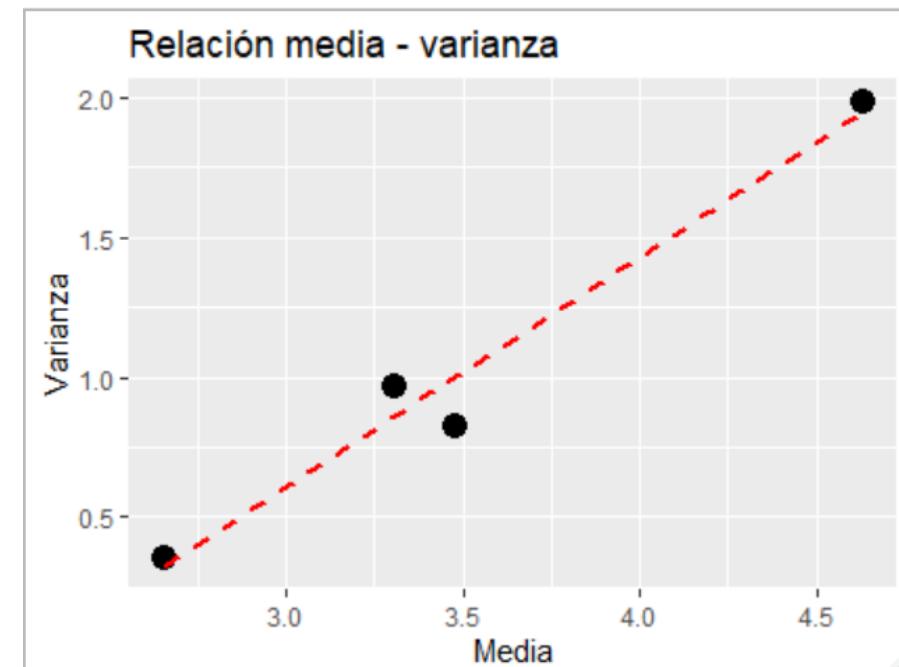


Estadística descriptiva

Variable respuesta: Pb

VE o factor: Ambiente con 4 niveles

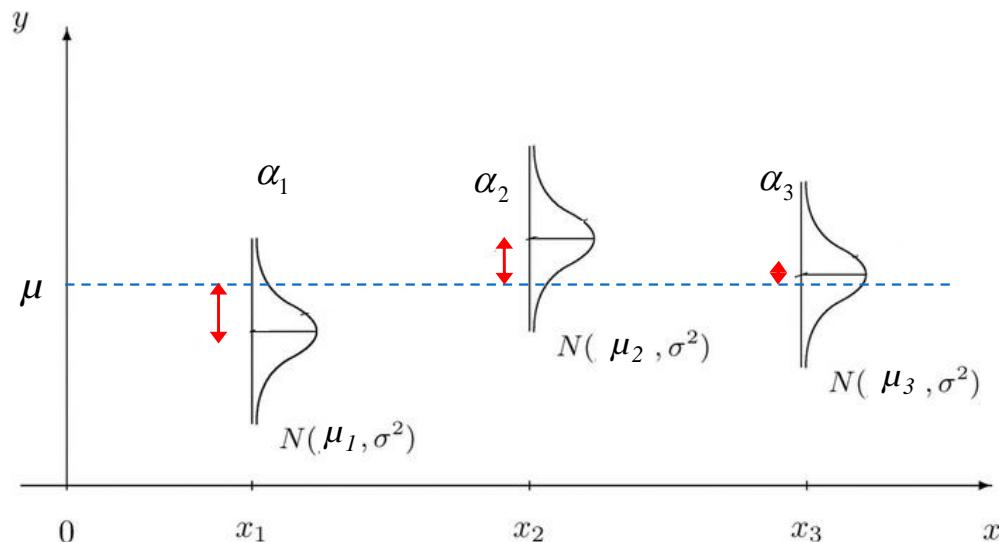
Ambiente <i><fct></i>	n	media	DE	var	min	max
	<i><int></i>	<i><dbl></i>	<i><dbl></i>	<i><dbl></i>	<i><dbl></i>	<i><dbl></i>
1 Esp verdes	37	2.65	0.593	0.351	1.39	3.71
2 Residencial	10	3.31	0.984	0.968	1.39	5.02
3 B carenciados	40	3.48	0.910	0.828	1.79	5.81
4 Riachuelo	56	4.63	1.41	1.98	1.39	8.27



Modelo de comparación de medias

5

$$E(Y_i) = \mu + \alpha_i \quad \varepsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$



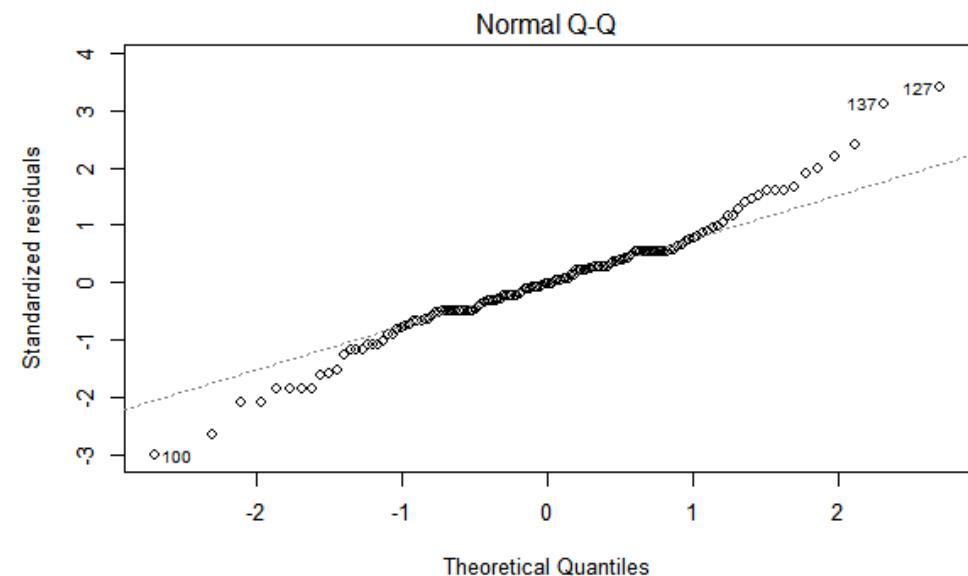
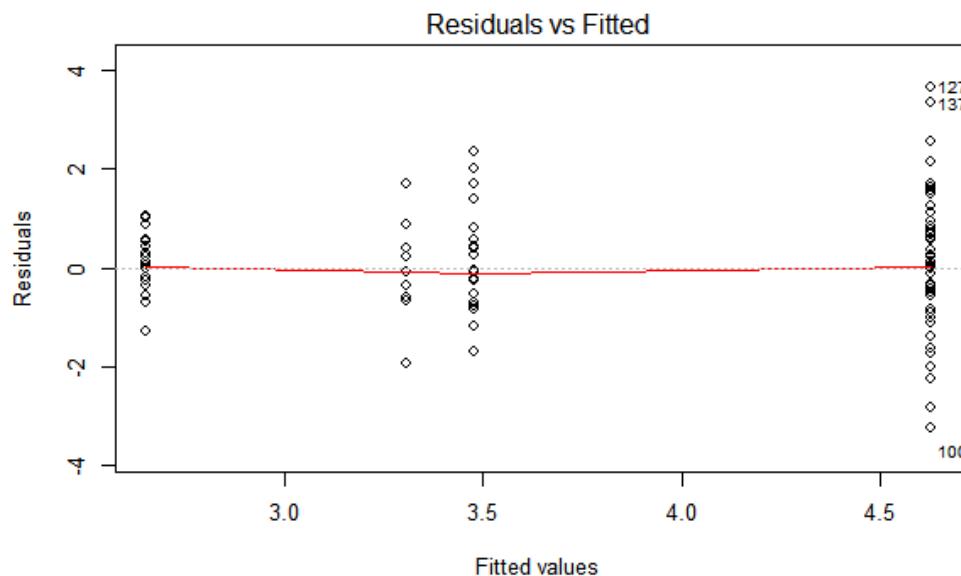
```
modelo1<-lm(Pb ~ Ambiente, data=bd)
```

Supuestos del modelo

(mismos supuestos menos linealidad)

6

```
> modelo1<-lm(Pb ~ Ambiente, data=bd)
```



```
> LeveneTest(modelo1)
Levene's Test for Homogeneity of variance
  Df F value    Pr(>F)
group  3  5.0704 0.002313 **
139
```

```
> anova(modelo1)
Analysis of Variance Table
```

Response: Pb

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Ambiente	3	92.348	30.7828	26.308	1.541e-13
Residuals	139	162.643	1.1701		

```
> shapiro.test(residuals(modelo1))
```

Shapiro-Wilk normality test

```
data: residuals(modelo1)
W = 0.97456, p-value = 0.009099
```

El p-valor obtenido no es confiable

La buena noticia: podemos modelar la estructura de varianzas

$$\text{var}(\varepsilon_i) = \sigma^2 \cdot \text{función de varianza}$$

$$\text{var}(\varepsilon_i) = \sigma^2 \cdot f(\mu_i, X, \delta)$$

Se incorpora al modelo una función de varianza que puede depender de:

- μ_i = media o esperanza de la variable respuesta
- X = **covariable** para la varianza. Cualquier variable utilizada para modelar la estructura de varianzas de los errores
- δ = parámetro; es estimado en función de la estructura de varianzas propuesta

Las estimaciones son por mínimos cuadrados generalizados

Funciones de varianza

Identidad (`varIdent`): una varianza distinta para cada grupo

$$\varepsilon_i \sim N(0, \sigma^2_i)$$

Exponencial (`varExp`): la varianza como función exponencial de alguna covariable

$$\varepsilon_i \sim N(0, \sigma^2 \cdot e^{2\delta \cdot X_i})$$

Potencia (`varPower`): la varianza como función de potencia de alguna covariable

$$\varepsilon_i \sim N(0, \sigma^2 \cdot |X_i|^{2\delta})$$

Fija (`varFixed`): la varianza como función lineal de alguna covariable

$$\varepsilon_i \sim N(0, \sigma^2 \cdot X_i)$$

```
library("nlme")
gls(Y ~ X, weights="xx", data)
```

```
varIdent(form=~1|A)
varPower()
varExp()
VarFixed(~X)
```

¿Cuál función utilizar?

varIdent

- Es la única que admite variables **cualitativas** como covariable
- Estima diferentes varianzas para cada nivel de la covariable (σ^2_i). Se estiman tantas varianzas como niveles -1

varPower

- No se puede usar cuando la covariable toma valores iguales a 0
- Requiere estimar un parámetro (δ)

varConstantPower

- Simil varPower pero se le suma una constante a la covariable:
- Requiere estimar dos parámetro (δ y la cte)
- Para covariables que toman valores iguales a 0

varExp

- Se puede usar cuando la covariable toma valores iguales a 0
- Puede tener problemas de estimación cuando los valores de la covariable son altos (i.e. > 100); en esos casos conviene reescalar
- Requiere estimar un parámetro (δ)

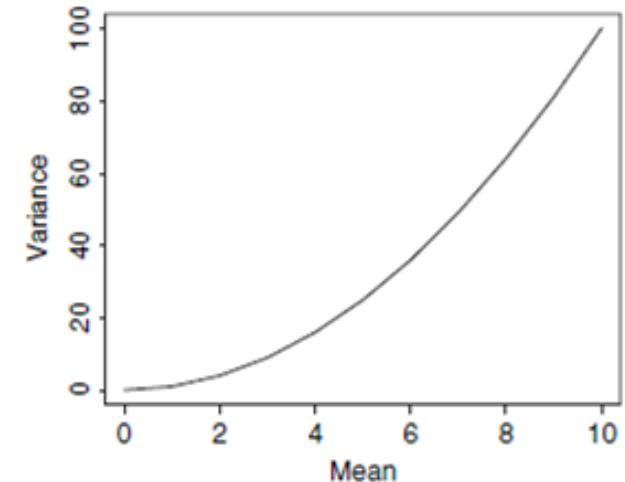
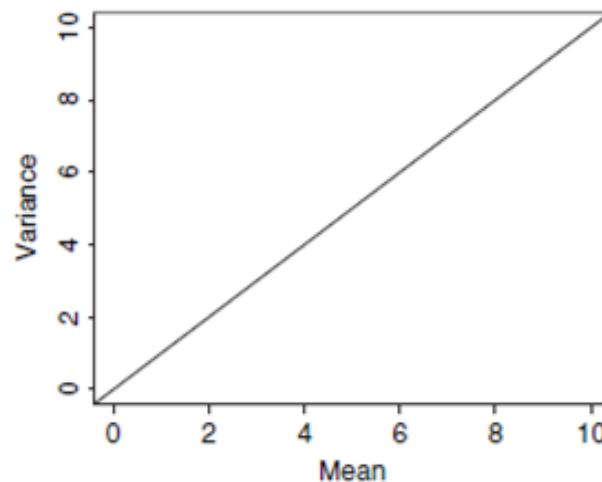
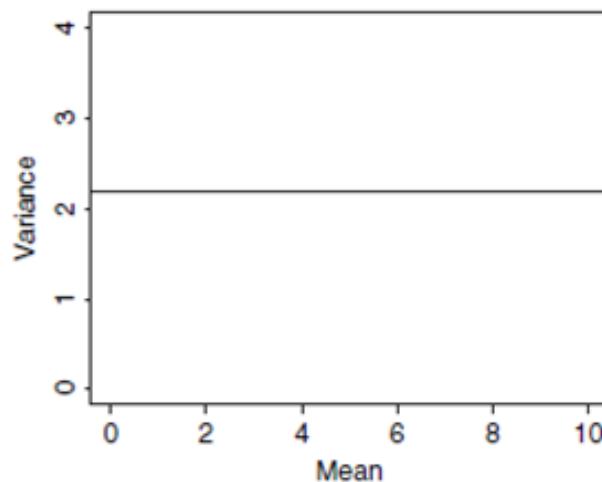
varFixed

- No requiere estimar ningún parámetro

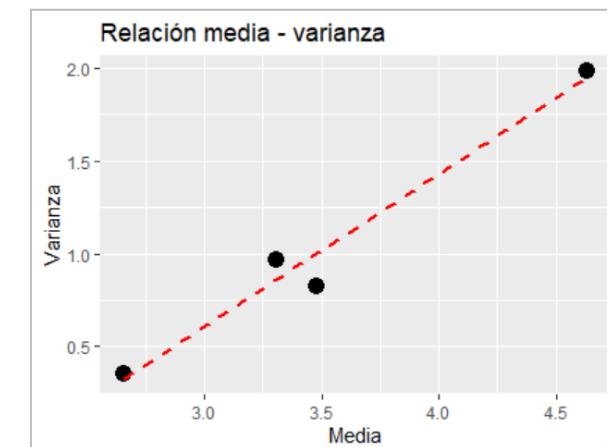
Explorar la relación
varianza-media

Relación entre esperanza y varianza

10



Crawley 2007



Cuadrados mínimos generalizados (gls)

Modelando varianzas: varIdent(Ambiente)

11

```
> #modelo 3 modelado varianzas. varIdent(ambiente)
> modelo3<-gls(Pb~Ambiente, weights=varIdent(form=~1|Ambiente), data=Plomo)
> anova(modelo3)
Denom. DF: 139
      numDF   F-value p-value
(Intercept)     1 1951.0503 <.0001
Ambiente         3   30.9453 <.0001
> modelo3
Generalized least squares fit by REML
  Model: Pb ~ Ambiente
  Data: Plomo
 Log-restricted-likelihood: -200.2043
```

Estimación por MV

Coefficients:

(Intercept)	AmbienteEsp verdes	AmbienteResidencial	AmbienteRiachuelo
3.4763281	-0.8221418	-0.1706125	1.1526315

Variance function:

Structure: Different standard deviations per stratum

Formula: ~1 | Ambiente

Parameter estimates:

Esp verdes B carenciados	Residencial	Riachuelo
1.000000	1.534740	1.659941
Degrees of freedom: 143 total; 139 residual		
Residual standard error: 0.5928515		

Desvío estándar de Esp verdes

D.E. de Riachuelo/D.E. Esp verdes

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

↑

</

Cuadrados mínimos generalizados (gls)

Modelando varianzas: varIdent(Ambiente)

12

```
modelo3<-gls(Pb~Ambiente, weights=varIdent(form=~1|Ambiente), method="ML", data=bd)
```

```
> modelo3
```

Generalized least squares fit by maximum likelihood

Model: Pb ~ Ambiente

Data: bd

Log-likelihood: -196.7303

coefficients:

(Intercept)	AmbienteResidencial	AmbienteB carenciados	AmbienteRiachuelo
2.6541863	0.6515293	0.8221418	1.9747733

Variance function:

Structure: Different standard deviations per stratum

Formula: ~1 | Ambiente

Parameter estimates:

Esp verdes B carenciados	Residencial	Riachuelo
1.000000	1.536338	1.596478
2.385593		

Degrees of freedom: 143 total; 139 residual

Residual standard error: 0.5847852

Cambian las estimaciones!

Cuando n es grande los estimadores de varianza por MV son normalmente asintóticos y consistentes. Cuando n es chico la estimación es **sesgada** ya que no divide por los GL sino por n . Eso se corrige empleando MV restringida (REML) que genera una función de verosimilitud sin considerar los efectos fijos

Modelando varianzas: varIdent(Ambiente)

Análisis de residuos (modelo3)

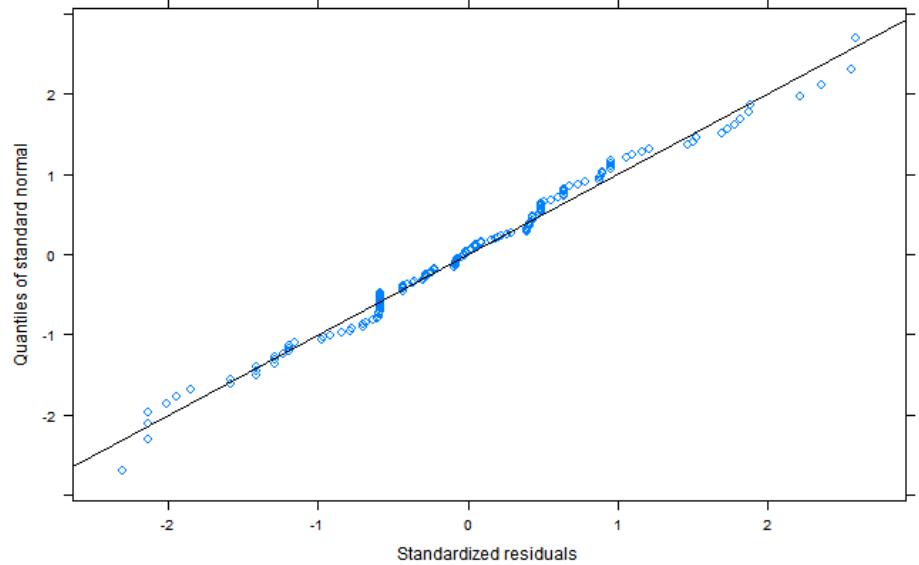
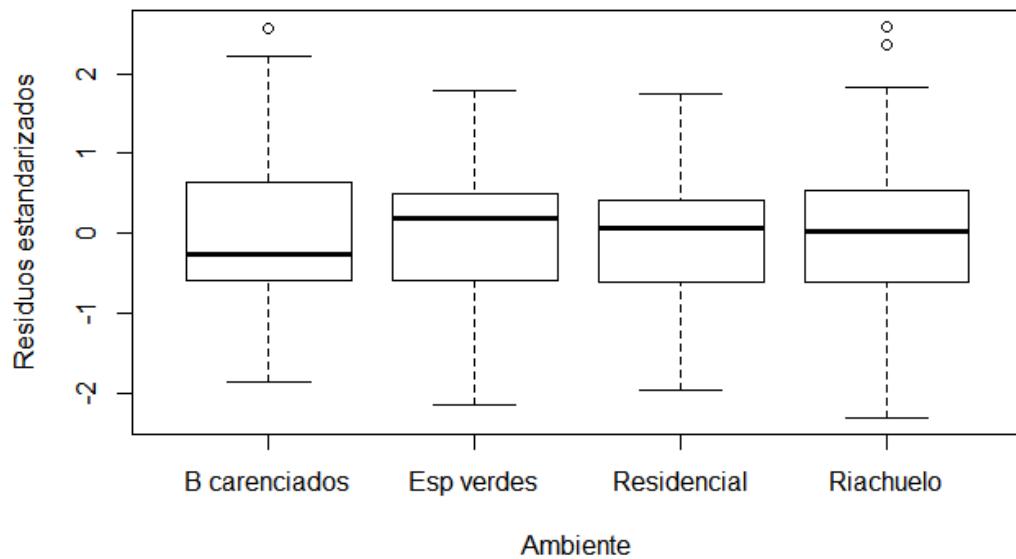
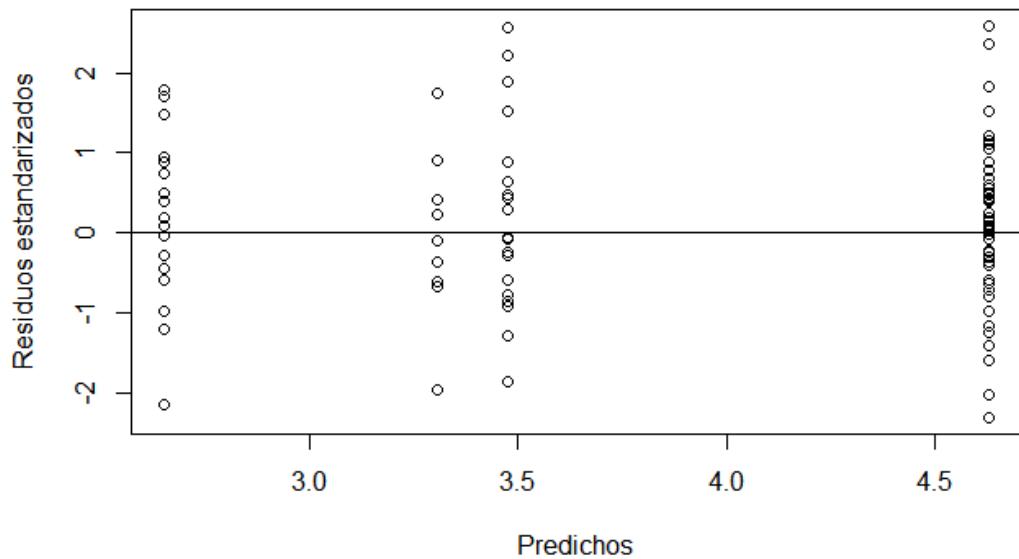


Gráfico de dispersión de RE vs PRED



$$RE_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\sigma}_i^2}}$$

O la función de varianza que corresponda

Cuadrados mínimos generalizados (gls)

Modelando varianzas: varPower

14

```
modelo4<-gls(Pb~Ambiente, weights=varPower(), data=Plomo)
```

```
> anova(modelo4)
```

Denom. DF: 139

	numDF	F-value	p-value
(Intercept)	1	1940.6731	<.0001
Ambiente	3	30.1536	<.0001

```
> modelo4
```

Generalized least squares fit by REML

Model: Pb ~ Ambiente

Data: Plomo

Log-restricted-likelihood: -200.4478

Coefficients:

(Intercept)	AmbienteEsp verdes	AmbienteResidencial	AmbienteRiachuelo
3.4763281	-0.8221418	-0.1706125	1.1526315

Combination of variance functions:

Structure: Power of variance covariate

Formula: ~fitted(.)

Parameter estimates:

power 1.529801 $\hat{\delta}$

Degrees of freedom: 143 total; 139 residual

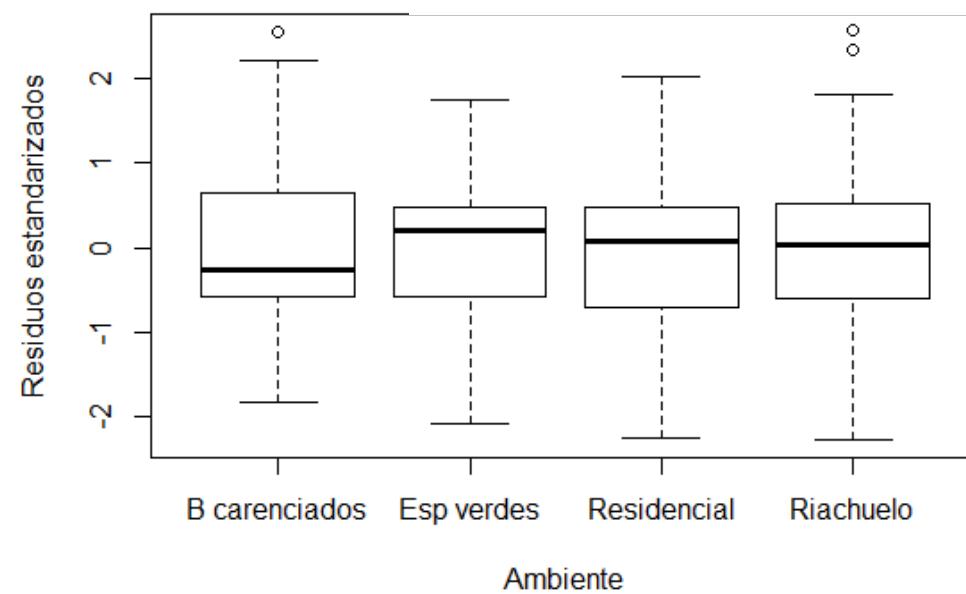
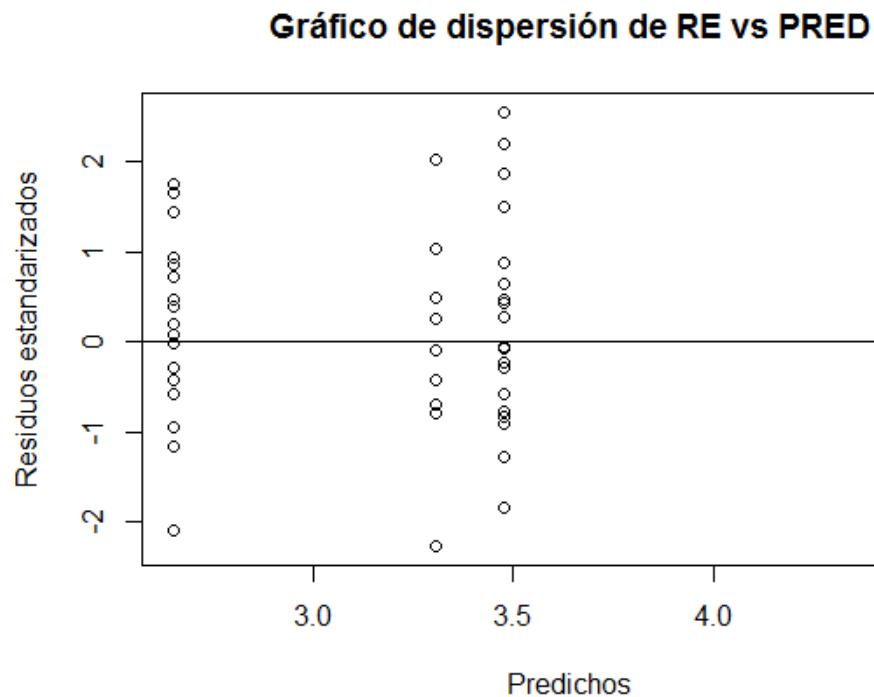
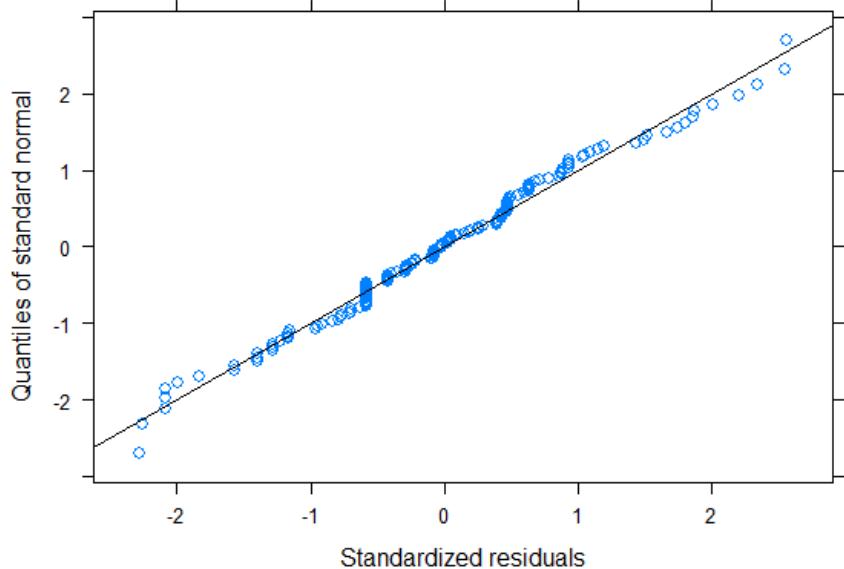
Residual standard error: 0.1362282

$$RE_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\sigma}^2 * |X_i|^{2*\hat{\delta}}}}$$

$$RE_i = \frac{y_i - \hat{y}_i}{\sqrt{0,136^2 * |\hat{y}_i|^{2*1,53}}}$$

Modelando varianzas: VarPower

Análisis de residuos (modelo4)



Varios modelos posibles

16

	Residuos
Modelo 2: gls sin modelar varianzas. Se descarta por los residuos	x
Modelo 3: gls modelando varianzas por varIdent(ambiente)	ok
Modelo 4: gls modelando varianzas por varPower	ok
Modelo 5: gls modelando varianzas por varExp	ok

¿Cuál elegir?

Selección de modelos

17

Criterios de información:

- Resumen la información de un modelo, teniendo en cuenta la función de verosimilitud $L(\theta)$ (cuanto mayor, mejor) y el número de parámetros a estimar del modelo (p) (cuanto mayor, peor)
- Estiman la distancia relativa entre el modelo ajustado y el mecanismo verdadero pero desconocido (de tal vez infinitos parámetros) que generó los datos observados
- El valor individual no es interpretable, solo sirve con fines comparativos:
cuanto menor, mejor el modelo

Comparación de modelos

18

Criterios de información:

- Akaike (AIC)
- Bayesiano de Schwartz (BIC)

- AIC y BIC menores, implican mejor ajuste

L	Log(L)	-2xLog(L)
0	--	--
0.1	-1	2
0.2	-0.70	1.40
0.3	-0.52	1.05
0.4	-0.40	0.80
0.5	-0.30	0.60
0.6	-0.22	0.44
0.7	-0.15	0.31
0.8	-0.10	0.19
0.9	-0.05	0.09
1	0	0

- BIC penaliza más que AIC los modelos con más parámetros a estimar

$$AIC = -2 \log L(\theta) + 2p$$

$$BIC = -2 \log L(\theta) + p \ln(n)$$



Medida del Penalización
ajuste por la
complejidad
del modelo

Varios modelos

19

Modelo 2: gls sin modelar varianzas. Se descarta por los residuos

Modelo 3: gls modelando varianzas por varIdent(ambiente)

Modelo 4: gls modelando varianzas por varPower

Modelo 5: gls modelando varianzas por varExp

```
> AIC(modelo2,modelo3,modelo4,modelo5)
      df      AIC
```

```
modelo3  8 416.4087
```

```
modelo4  6 412.8956
```

```
modelo5  6 413.3685
```

- 1- Seleccionamos los modelo con residuos adecuados (modelos 3 a 5)
- 2- Seleccionamos el que presente menor AIC (modelo 4)

$$E(Y_i) = \mu + \alpha_i$$

H1: Algun $\alpha_i \neq 0$

Anova

20

```
modelo4<-gls(Pb~Ambiente, weights=varPower(), data=P1omo)
```

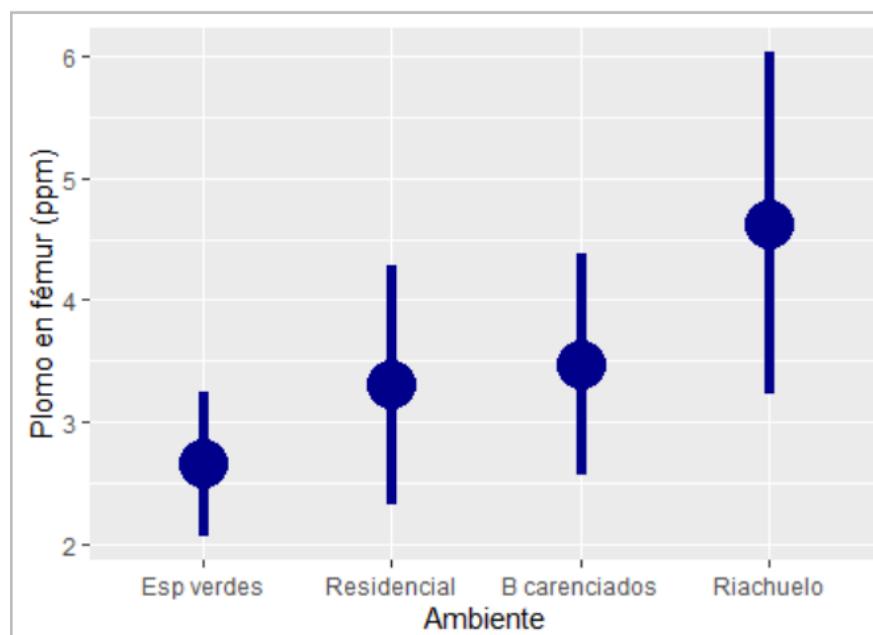
```
> anova(modelo4)
```

Denom. DF: 139

	numDF	F-value	p-value
(Intercept)	1	1940.6731	<.0001
Ambiente	3	30.1536	<.0001
.	.	.	.

P-valor < 0,05

La concentración media de plomo en los fémures de las ratas de **alguno** de los ambientes difiere significativamente de la media general



Comparaciones múltiples

Anova

Comparaciones múltiples

21

- Sirven para detectar diferencias entre las medias de los tratamientos como complemento a la prueba global
- Controlan de alguna manera el error global, es decir la probabilidad de cometer al menos un error tipo I.
- Equivalen a hacer múltiples test t con algún tipo de ajuste
- Existen distintos métodos de comparación, según como controlen el error global

Comparaciones múltiples

22

- Compara subgrupos de medias. Por ejemplo de a pares:
 - $H_0: \mu_i = \mu_j$ (o lo que es lo mismo, $\mu_i - \mu_j = 0$)
 - $H_1: \mu_i \neq \mu_j$ (o lo que es lo mismo, $\mu_i - \mu_j \neq 0$)
- La lógica es la del test de t para la comparación de dos medias. Los distintos métodos difieren en cómo se calcula el EE, en la distribución de probabilidades, en ajustes en nivel de significación, etc

$$t_{GL} = \frac{\Delta \bar{Y} - \Delta \mu}{EE_{\bar{Y}_1 - \bar{Y}_2}} = \frac{\Delta \bar{Y} - \Delta \mu}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}} = \frac{\Delta \bar{Y} - \Delta \mu}{\sqrt{\frac{CMerror}{n_1} + \frac{CMerror}{n_2}}}$$

EE para la diferencia de medias

$$\Delta \bar{Y} \pm P_{GL,\alpha} EE_{\Delta \bar{Y}}$$

Se puede expresar como un IC para la diferencia de dos medias y ver si cero pertenece al IC. Además permite estimar la magnitud del efecto

Clasificación de los métodos

23

A priori: planeados

- Las hipótesis se plantean antes del muestreo
- Se basan en información independiente del experimento
- pueden detectar diferencias aunque el Anova no lo haya hecho, ya que incorporan información que ni la hipótesis global ni las comparaciones a posteriori poseen
- Mayor sensibilidad, mayor potencia
 - Contrastes ortogonales
 - Método de Bonferroni
 - Método de Dunnet

A posteriori: no planeados

- se aplican sólo si el Anova dio significativo
- “búsqueda de significación”, exploratorios
- Más conservativos, menos potentes
 - Método de Tukey

Comparaciones no planeadas

Método de Tukey



24

- Compara todos los pares posibles de medias
- La cantidad total de comparaciones de a pares posibles con a grupos es:

$$q_{GL\ dentro,a} = \frac{\Delta \bar{Y} - \Delta \mu}{\sqrt{\frac{CMerror}{n_i}}}$$

$\frac{a(a-1)}{2}$

Distribución de TUKEY



- Procedimiento: Se calculan todas las diferencias entre las medias de los tratamientos y se comparan con una diferencia crítica o **diferencia mínima significativa** (DMS):

$$\Delta \bar{Y} \ vs \ DMS = q_{GL\ dentro,\alpha,k} \sqrt{\frac{CMdentre}{n_i}}$$

- Alternativamente se calcula un IC para la diferencia de dos medias

Usando paquete emmeans. Tukey por default emmeans(modelo4, pairwise ~ Ambiente)

\$emmeans

Ambiente	emmean	SE	df	lower.CL	upper.CL
Esp verdes	2.65	0.0993	43.1	2.45	2.85
Residencial	3.31	0.2680	109.3	2.77	3.84
B carenciados	3.48	0.1448	128.7	3.19	3.76
Riachuelo	4.63	0.1904	76.8	4.25	5.01

d.f. method: satterthwaite
Confidence level used: 0.95

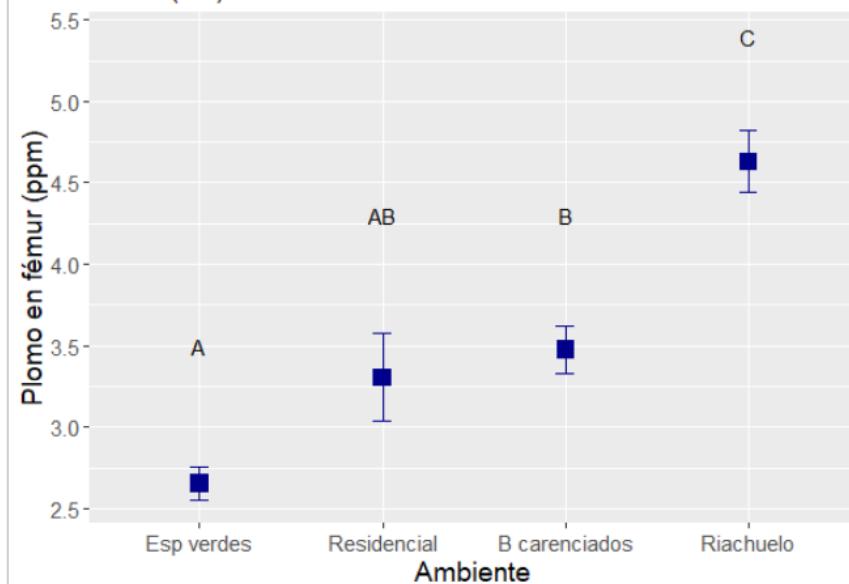
\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
Esp verdes - Residencial	-0.652	0.286	98.5	-2.280	0.1098
Esp verdes - B carenciados	-0.822	0.176	94.7	-4.683	0.0001
Esp verdes - Riachuelo	-1.975	0.215	122.9	-9.196	<.0001
Residencial - B carenciados	-0.171	0.305	113.9	-0.560	0.9436
Residencial - Riachuelo	-1.323	0.329	142.5	-4.025	0.0005
B carenciados - Riachuelo	-1.153	0.239	117.5	-4.818	<.0001

P value adjustment: tukey method for comparing a family of 4 estimates

No usar los IC para la media para determinar si hay diferencias.
Lo correcto es analizar IC para la diferencia de medias

Acumulación de plomo en ratas de distintos ambientes e
Media (EE)



¿Las medias difieren significativamente?
¿En qué magnitud?

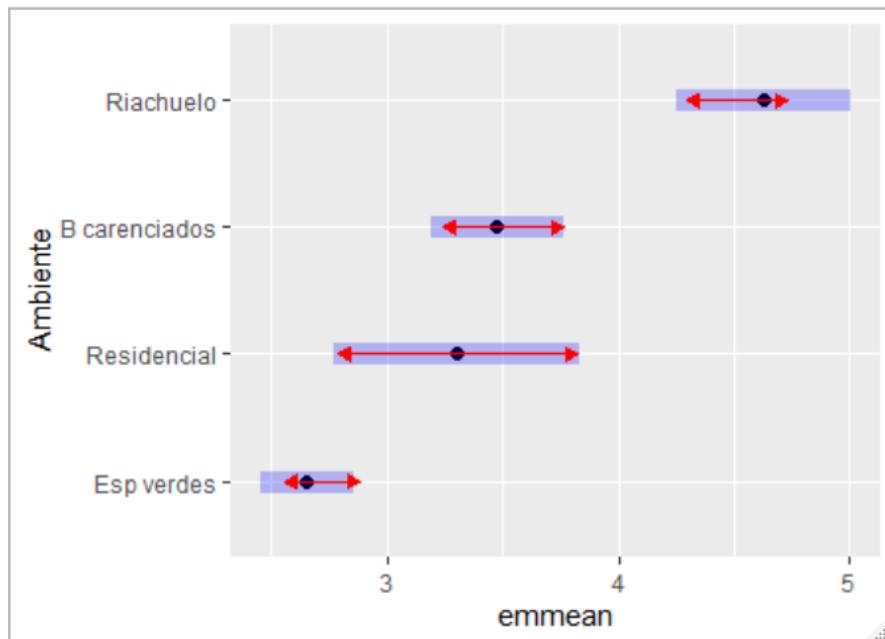
Graficar siempre las estimaciones que surgen del modelo

contrast	estimate	SE	df	lower.CL	upper.CL
Esp verdes - Residencial	-0.652	0.286	98.5	-1.398	0.0953
Esp verdes - B carenciados	-0.822	0.176	94.7	-1.281	-0.3630
Esp verdes - Riachuelo	-1.975	0.215	122.9	-2.534	-1.4155
Residencial - B carenciados	-0.171	0.305	113.9	-0.965	0.6236
Residencial - Riachuelo	-1.323	0.329	142.5	-2.178	-0.4686
B carenciados - Riachuelo	-1.153	0.239	117.5	-1.776	-0.5291

IC95% para la diferencia de medias:

Contiene al cero? Si lo contiene, cuál es su magnitud?

Las barras azules son IC para la media. Las flechas rojas indican comparaciones entre medias. Si las flechas de dos grupos se solapan, las diferencias no son estadísticamente significativas



Con una confianza del 95% se estima que la concentración de plomo en fémur de ratas de Riachuelo es, en promedio, entre 1,43 y 2,52 ppm mayor a la de ratas de espacios verdes

Método de Bonferroni

27

- Consiste en ajustar los valores p de las pruebas de hipótesis individuales de manera de controlar el error global ([corrección por múltiples tests](#))

$$\text{valor } p \text{ corregido} = \text{valor } p \cdot m$$

- O lo que es lo mismo, $\alpha \text{ corregido} = \alpha^* = \frac{\alpha}{m}$
donde m es la cantidad de comparaciones que se efectúan
- Si solo se hace una comparación, equivale a un test t. Pero la prueba va perdiendo potencia si m es muy grande
- Existen variantes no tan conservadoras **Bonferroni secuencial (Holm, 1979)**

\$contrasts	contrast	estimate	SE	df	t.ratio	p.value
	Esp verdes - Residencial	-0.652	0.286	98.5	-2.280	0.1486
	Esp verdes - B carenciados	-0.822	0.176	94.7	-4.683	0.0001
	Esp verdes - Riachuelo	-1.975	0.215	122.9	-9.196	<.0001
	Residencial - B carenciados	-0.171	0.305	113.9	-0.560	1.0000
	Residencial - Riachuelo	-1.323	0.329	142.5	-4.025	0.0006
	B carenciados - Riachuelo	-1.153	0.239	117.5	-4.818	<.0001

P value adjustment: bonferroni method for 6 tests

Contrastes ortogonales



28

- Otra lógica: preguntas específicas planteadas a priori basadas en información por fuera del ensayo (pocas, bien dirigidas, máxima potencia)
- Los contrastes deben ser independientes (ortogonales) entre sí, ya que consisten en una descomposición de la SC de los tratamientos
- Como máximo $\# \text{grupos} - 1$ contrastes
- Solo para diseños balanceados
- Equivale a un test t

Contrastes ortogonales

29

- Máximo 3 contrastes
- ¿Zonas con mayor cobertura vegetal (espacios verdes y residenciales) difieren de zonas con menor cobertura (barrios carenciados y Riachuelo)?
- ¿espacios verdes y residenciales difieren entre sí?

$$Ho^1 : \frac{\mu_{esp.verd} + \mu_{res}}{2} = \frac{\mu_{b.car} + \mu_{riach}}{2} \Rightarrow \frac{\mu_{esp.verd} + \mu_{res}}{2} - \frac{\mu_{b.car} + \mu_{riach}}{2} = 0$$

$$Ho^2 : \mu_{esp.verd} = \mu_{res} \Rightarrow \mu_{esp.verd} - \mu_{res} = 0$$

Contraste	1	2	$c_i c_j$
Esp verdes			
Residenciales			
Carenciados			
Riachuelo			

$$\sum c_i = 0$$

$$\sum c_i c_j = 0 \text{ para todos } (i, j)$$

Contrastes ortogonales

30

1. Creamos la matriz de coeficientes

```
c1<-c(1/2,1/2, -1/2,-1/2)
c2<-c(1,-1,0,0)
matriz_contrastes<-cbind(c1,c2)
```

```
> matriz_contrastes
      c1   c2
[1,] 0.5  1
[2,] 0.5 -1
[3,] -0.5 0
[4,] -0.5 0
```

2. Ajustamos el modelo

```
modelo4b <- lm(Pb ~ Ambiente, data=bd,contrasts =
list(dosis_cd_cuali=matriz_contrastes))
summary(modelo4b)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.5643	0.1208	29.516	< 2e-16 ***
Ambientec1	-1.1687	0.2415	-4.839	3.4e-06 ***
Ambientec2	-0.3258	0.2096	-1.554	0.122

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



¿Cuál método de comparación elegir?

31

Tukey: todos contra todos, de a pares. "Búsqueda de significación", exploratorio. El más recomendable para ese objetivo

Bonferroni: Cuantas más comparaciones menos potente

Dunnet: todos contra el control

Ortogonal: el más potente de todos, pero con serias restricciones acerca de las comparaciones que se pueden efectuar (deben ser matemáticamente ortogonales)

⇒ Compromiso entre cantidad de preguntas que se desean responder y potencia (a fin de mantener el error global)

value

AmbienteEsp verdes	2.654186
AmbienteResidencial	3.305716
AmbienteB carenciados	3.476328
AmbienteRiachuelo	4.628960

Modelo4 como regresión lineal

32

```
modelo4<-gls(Pb~Ambiente, weights=varPower(), data=P1omo)
```

```
> summary(modelo4)
```

Generalized least squares fit by REML

Variance function:

Structure: Power of variance covariate

Formula: ~fitted(.)

Parameter estimates:

power	1.529801
-------	----------

Coefficients:

	value	Std. Error	t-value	p-value
(Intercept)	2.6541863	0.09970036	26.621633	0.0000
AmbienteResidencial	0.6515293	0.28623657	2.276192	0.0244
AmbienteB carenciados	0.8221418	0.17588010	4.674445	0.0000
AmbienteRiachuelo	1.9747733	0.21436703	9.212113	0.0000

Magnitud del efecto
(diferencia de medias)

EE para la
diferencia de
medias

- Salvo cuando hay solo dos niveles, no hay una prueba “global” sobre el efecto de la VE
- Los coeficientes son diferencias de medias con respecto al nivel de referencia; no se informan otras comparaciones
- No son comparaciones ortogonales
- No controlan el error global

Resumiendo

33

- El análisis de la varianza puede usarse para comparar cualquier cantidad de grupos con respecto a su media
- Si la VE es cuantitativa y afecta a la VR según una función modelable, es más eficiente utilizar regresión que anova (más parsimoniosa, permite interpolar)
- Si la VE es cuali o si es cuanti pero sin una relación clara, se recomienda Anova
- Luego del ANOVA se debe aplicar algún método de comparaciones entre grupos. La elección del método depende de los objetivos
- Para leer sobre IC en comparaciones: Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170.

BIOMETRÍA II

CLASE 5

MODELOS CON MÁS DE UNA PREDICTORA

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

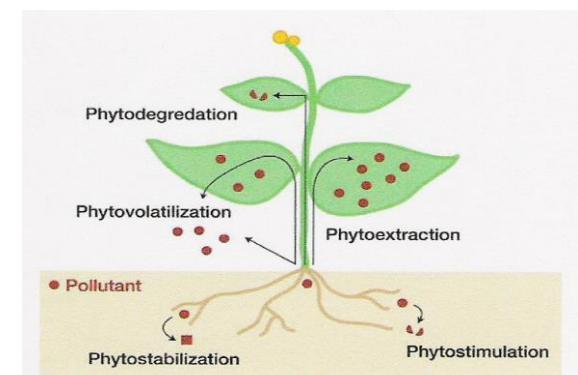
Recuperación de suelos empetrados de Comodoro Rivadavia

Luque, JL - Estación Experimental Chubut del INTA Trelew - 2009

2

- La fitorremediación es una técnica que emplea vegetales en combinación con microorganismos asociados a la rizósfera para remover, degradar o inmovilizar contaminantes contenidos en suelos, sedimentos y aguas
- Se desea estudiar el desempeño de dos especies vegetales perennes: charcao (*Senecio filaginoides*), nativa, y agropiro alargado (*Thynopiron ponticum*), exótica, para fitorremediar suelos empetrados de Comodoro Rivadavia.
- La bioestimulación es la adición de nutrientes al suelo para estimular la actividad de microorganismos degradadores del contaminante. Se desea estudiar el efecto de la adición al suelo de un fertilizante (fósforo + nitrógeno).

¿Podemos, en un mismo ensayo, estudiar la efectividad de las plantas y la de los microorganismos en remover los HC del suelo?



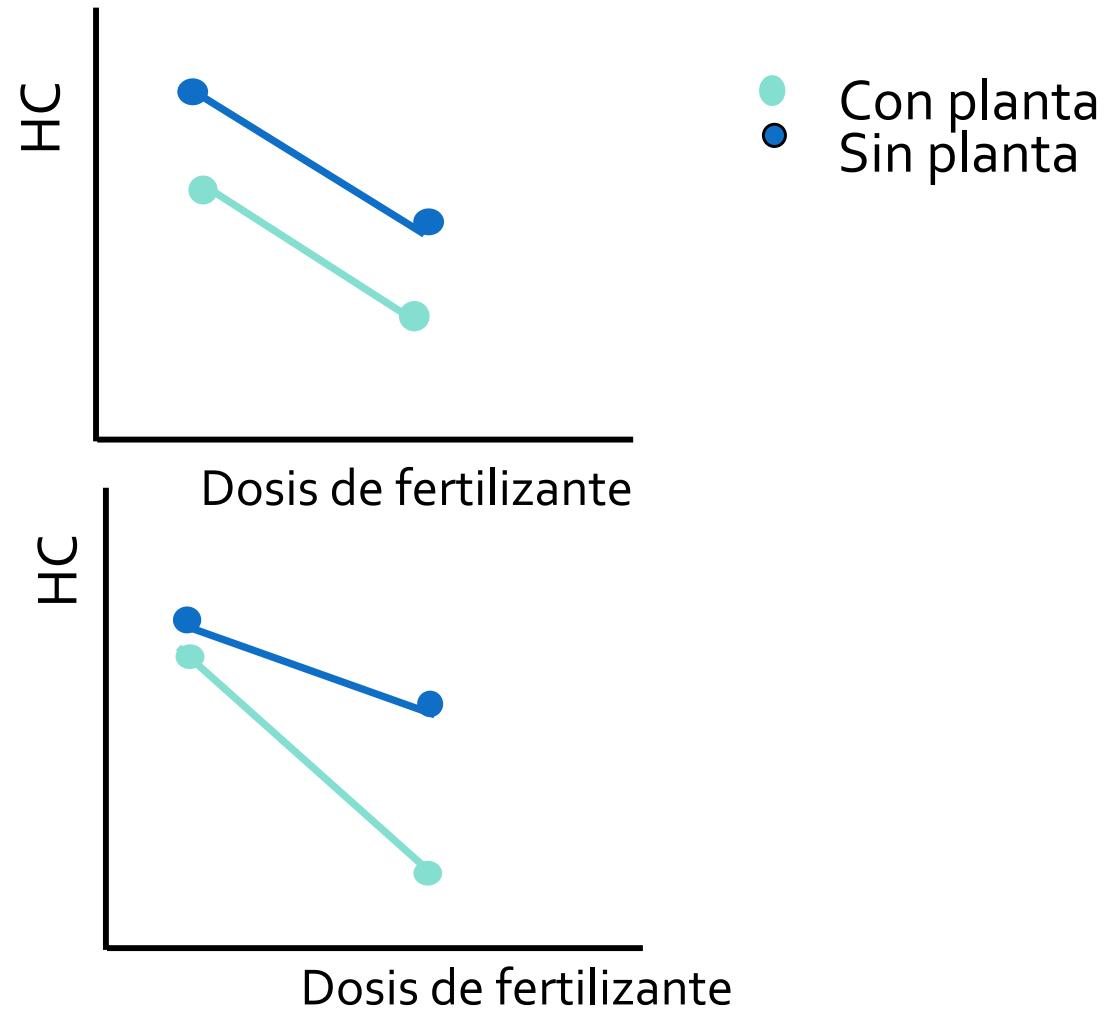
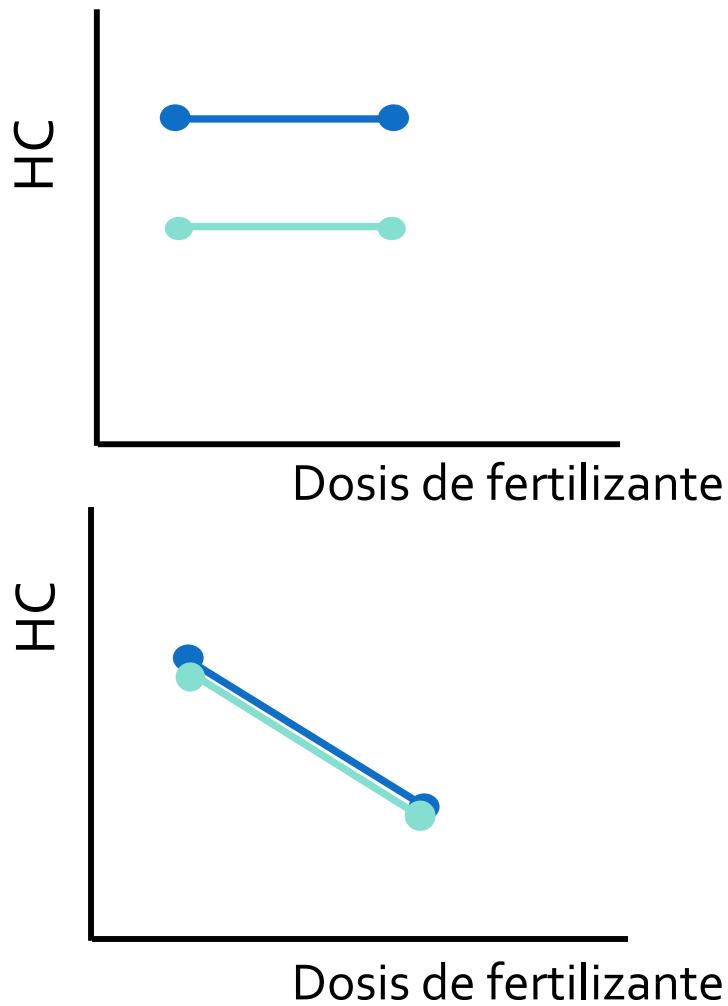
Recuperación de suelos empetrificados de C. Rivadavia

Luque, JL - Estación Experimental Chubut del INTA Trelew - 2009



3

- Se planea un ensayo utilizando macetas conteniendo suelo extraído de la zona petrolera de Cdro Rivadavia conteniendo 4.1% de hidrocarburos (HC) (4,1 g/100 g suelo seco)
- Se desea contestar las siguientes preguntas:
 - ¿Las especies son efectivas en remover los HC? (independientemente del agregado de fertilizante)
 - ¿Cuánto de la remoción de HC es por microorganismos del suelo y no por acción de las plantas?
 - ¿La fertilización es efectiva en promover la remoción de los HC? (independientemente del agregado de plantas)
 - La capacidad de remoción de HC de las especies ¿cambia según el agregado de fertilizante?



- ✓ Efectos nulos
- ✓ Efectos aditivos
- ✓ Interacción

Interacción entre variables explicatorias

5

- El efecto de una VE sobre la VR cambia según los valores que tome otra VP
- Es decir que el efecto de una VE **depende de / se asocia con** el valor que tome otra VE (y viceversa) (**modificación de efectos**)
- Si hay interacción entre VE, pierde relevancia estimar los efectos de una dada VE independientemente de los valores que tome la otra VE con la que interactúa (**principio de marginalidad**)
- Las interacciones pueden ser entre cualquier tipo de variables (categóricas con categóricas, cuantitativas con categóricas, cuanti con cuanti...)



Recuperación de suelos empetrificados de C. Rivadavia

Luque, JL - Estación Experimental Chubut del INTA Trelew - 2009

6

- 30 Macetas conteniendo suelo extraído de la zona petrolera de Cdro Rivadavia conteniendo 4.1% de hidrocarburos (HC) (4,1 g/100 g suelo seco)
- Cada maceta fue asignada al azar a una combinación de Planta (Charcao, Agropiro o Testigo sin vegetación) y Fertilización (con o sin)
- A los 350 días se midió contenido en suelo de HC totales de petróleo (% P/P, g/100 g suelo seco)
- Experimento o estudio observacional?
- UE:
- VR (tipo y potencial distribución de probabilidades):
- VE (tipo):
- Factores y niveles:
- Tratamientos:
- Diseño factorial de 3x2
- Replicación, aleatorización, control del error

Diseños factoriales

7

Son aquellos que incluyen más de una VE categórica (factor). Modelos de comparación de medias

- Mayor eficiencia en el uso de los recursos, menor error global y mayor potencia que varios unifactoriales
- Permiten evaluar la **interacción** entre factores

- Los **gráficos de perfiles** son muy útiles para describir el comportamiento de la VR
- Cuando un experimento tiene dos o más factores, la cantidad de medias que pueden compararse surge de las combinaciones de los niveles de los factores

El modelo es

$$Y \sim A^* B$$

8

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

$i=1,2\dots a$
 $j=1,2\dots b$
 $k=1,2\dots n_{ij}$

- donde Y_{ijk} es la concentración de HC en suelo de cada UE
- μ es la media general o media de la población
- α_i es el efecto del factor i (especie)
- β_j es el efecto del factor j (fertilización)
- $\alpha\beta_{ij}$ es el efecto de la interacción ij
- ε_{ijk} es el error aleatorio de cada UE

$$\varepsilon_{ijk} \sim NID(0, \sigma^2)$$

tienen que ser cuales

En R: `m1<-lm(HC~veg*fert, bd)`

Hipótesis en Diseño factorial

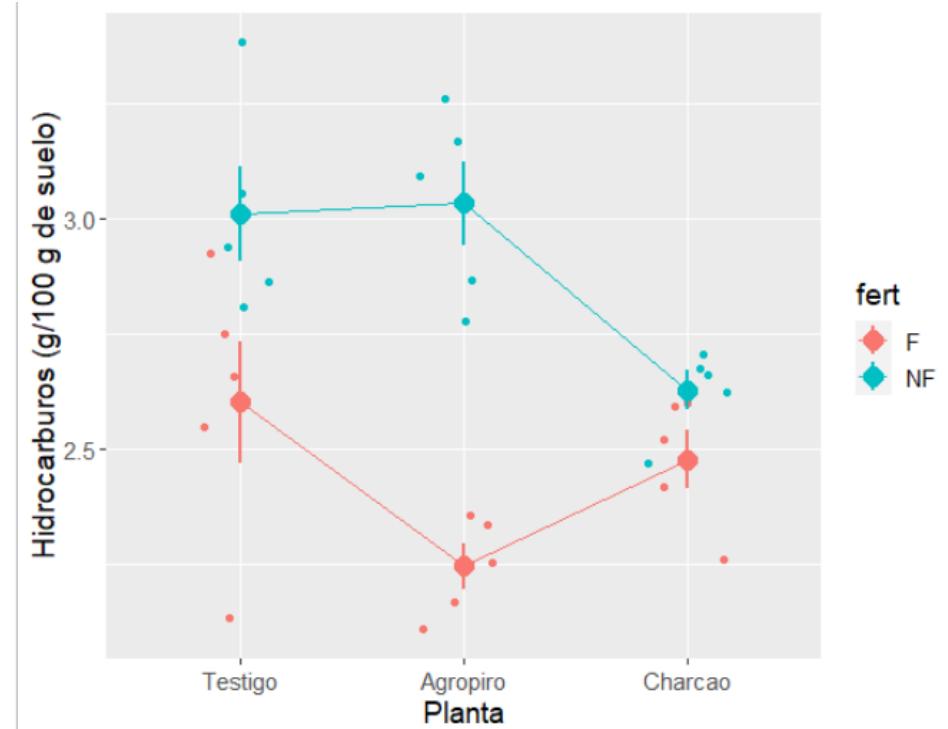
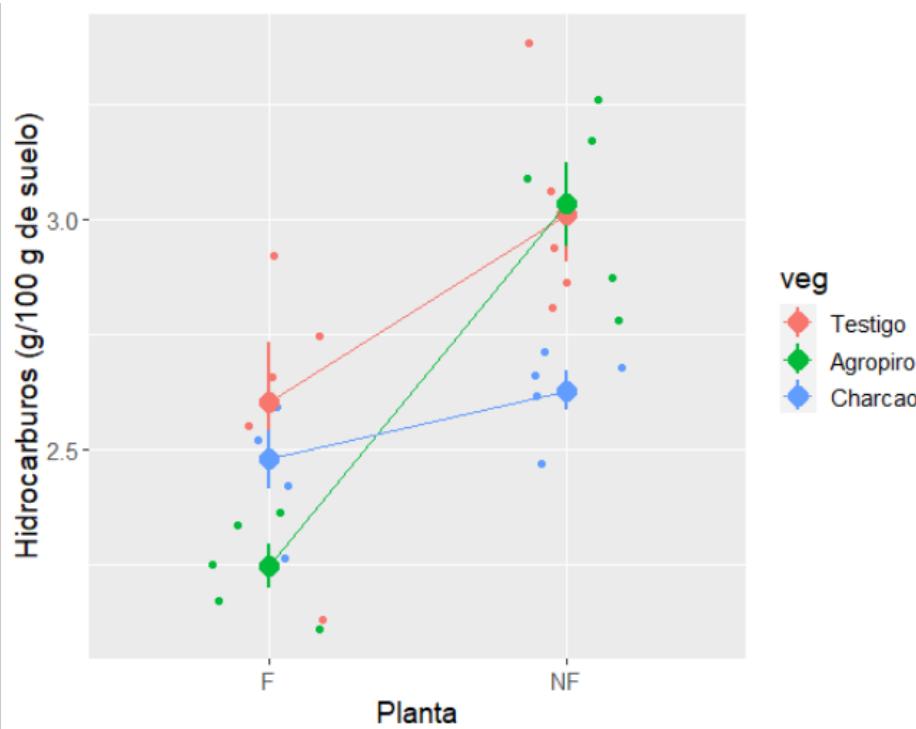
9

- $H_0: \gamma_{ij} = 0$
es decir no existe interacción entre especies y fertilización \Rightarrow El efecto de la fertilización es independiente de la especie (y viceversa)

Es la que debe analizarse primero, ya que si existe interacción no hay independencia entre los efectos de A y B y no es correcto analizar los factores por separado

- $H_0: \alpha_i = 0$
es decir no existe efecto sobre el contenido de HC del suelo debido a la especie (suponiendo independencia entre A y B)
- $H_0: \beta_j = 0$
es decir no existe efecto sobre el contenido de HC del suelo debido a la fertilización (suponiendo independencia entre A y B)

Gráficos de perfiles



Se pueden reordenar los niveles



```
factor(bd$veg, levels=c("Testigo", "Agropiro", "Charcao"))
```

Calculando residuos

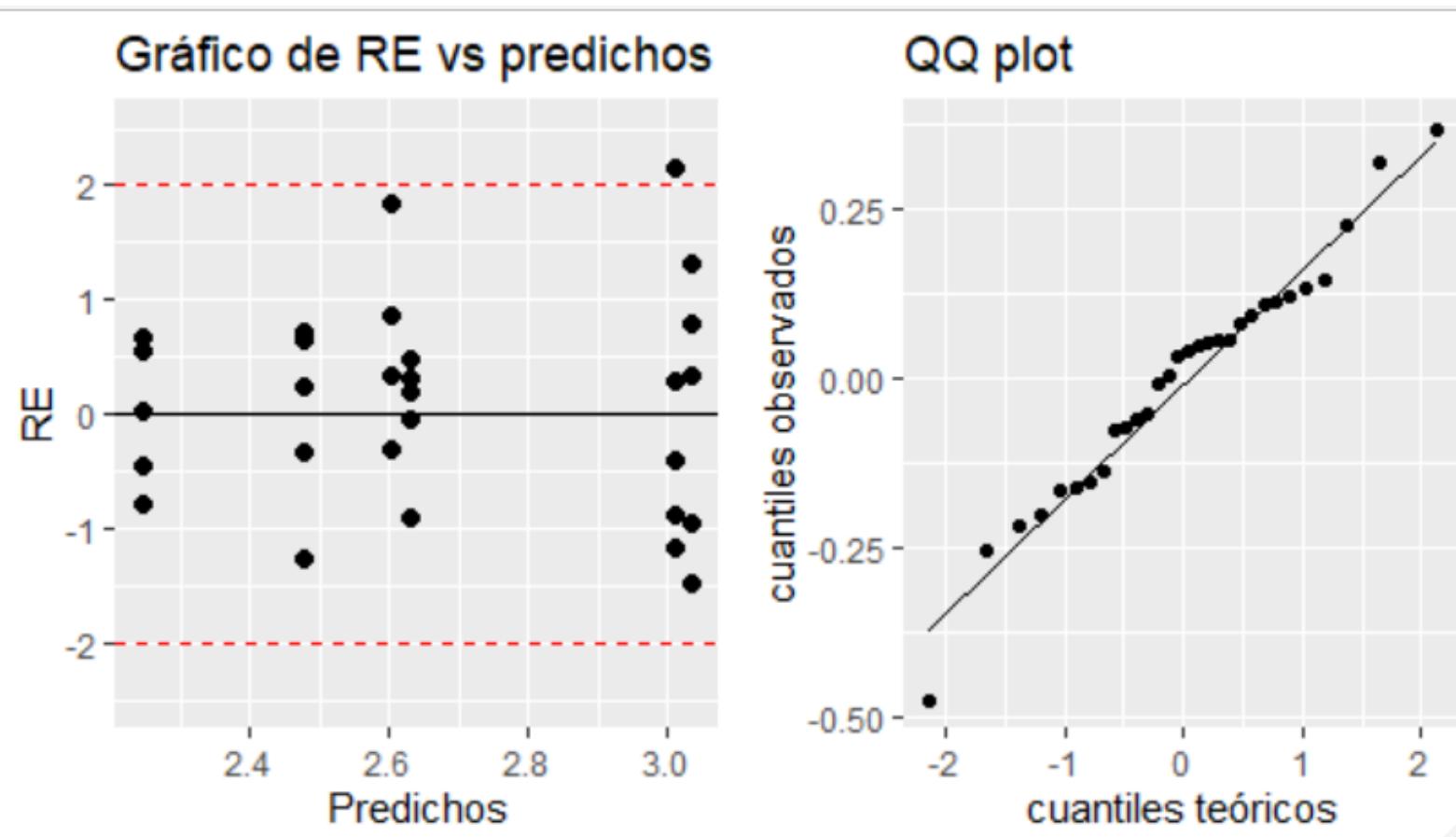
11

	Especie	Fertil	HC	Predichos	Residuos	residuos std
1		3	2	3.06	3.010	0.050
2		3	2	2.86	3.010	-0.150
3		3	2	2.81	3.010	-0.200
4		3	2	2.94	3.010	-0.070
5		3	2	3.38	3.010	0.370
6		3	1	2.66	2.602	0.058
7		3	1	2.75	2.602	0.148
8		3	1	2.55	2.602	-0.052
9		3	1	2.92	2.602	0.318
10		3	1	2.13	2.602	-0.472
11		2	2	2.66	2.628	0.032
12		2	2	2.71	2.628	0.082
13		2	2	2.62	2.628	-0.008
14		2	2	2.68	2.628	0.052
15		2	2	2.47	2.628	-0.158
16		2	1	2.59	2.478	0.112
17		2	1	2.42	2.478	-0.058
18		2	1	2.60	2.478	0.122
19		2	1	2.52	2.478	0.042
20		2	1	2.26	2.478	-0.218

$$e_{ijk} = y_{ijk} - \bar{y}_{ij}$$

Estudiando los supuestos

12



```
> leveneTest(HC~veg*fert, FITOR, center=mean)
```

Levene's Test for Homogeneity of Variance (center = mean)

Df	F value	Pr(>F)
----	---------	--------

group	5	1.4057	0.2578
-------	---	--------	--------

24

```
> shapiro.test(e)
```

Shapiro-Wilk normality test

data: e

w = 0.97461, p-value = 0.6713



Tabla de ANOVA

13

FdV	SC	GL	CM	F
Entre niveles factor A	$\sum bn_{ij} (\bar{y}_{i\cdot} - \bar{\bar{y}})^2$	$a-1$	\underline{SC}_A Gl_A	\underline{CM}_A CM_{error}
Entre niveles factor B	$\sum an_{ij} (\bar{y}_{\cdot j} - \bar{\bar{y}})^2$	$b-1$	\underline{SC}_B Gl_B	\underline{CM}_B CM_{error}
AxB (interacción)	$\sum n_{ij} (\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{\bar{y}})^2$	$(a-1)$ $(b-1)$	\underline{SC}_{AB} Gl_{AB}	\underline{CM}_{AB} CM_{error}
Error o dentro	$\sum (y_{ijk} - \bar{y}_{ij})^2$	$n-ab$	\underline{SC}_{error} GL_{error}	
Total	$\sum (y_{ijk} - \bar{\bar{y}})^2$	$n-1$		

Anova

```
m1<-lm(HC~veg*fert, bd)  
anova(m1)
```

Analysis of Variance Table

Response: HC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
veg	2	0.33045	0.16522	4.4833	0.022164	*
fert	1	1.50976	1.50976	40.9668	1.283e-06	***
veg:fert	2	0.51501	0.25750	6.9872	0.004061	**
Residuals	24	0.88448	0.03685			

signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'.'
					0.1	''
					1	

Pruebas globales

No es correcto analizar
efectos principales

Efectos principales y simples

15

Efectos principales o marginales de un factor son las comparaciones entre los niveles de un factor promediados para todos los niveles del otro factor. Es decir, *independientemente* del otro factor.

Efectos simples o comparaciones de interacción (de celdas) son comparaciones entre distintos niveles de un factor fijando los niveles del otro factor.

Medias HC	Especie			$\bar{y}_{\cdot j}$
	Testigo	Charcao	Agropiro	
Fertilización	3,01	2,63	3,03	2,890
no	3,01	2,63	3,03	2,890
sí	2,60	2,48	2,25	2,443
\bar{y}_i	2,805	2,555	2,640	2,667

Efectos simples
(celdas; dentro de la tabla)

Efectos principales
(marginales)

Principio de Marginalidad

- No deben interpretarse los efectos principales de las VE que interactúan
- No deben plantearse modelos con términos de interacción sin incluir los efectos principales

Comparaciones múltiples

16

- Los métodos disponibles y el procedimiento es el mismo que para anova de un factor
- Si la interacción fue no significativa se comparan los efectos principales y se efectúan comparaciones entre niveles de cada factor
- Si la interacción fue significativa no se comparan efectos principales, sino comparaciones entre celdas

Comparaciones de interacción

(entre todas las combinaciones)

```
library(emmeans)
emmeans(m1, pairwise ~ fert*veg)
```

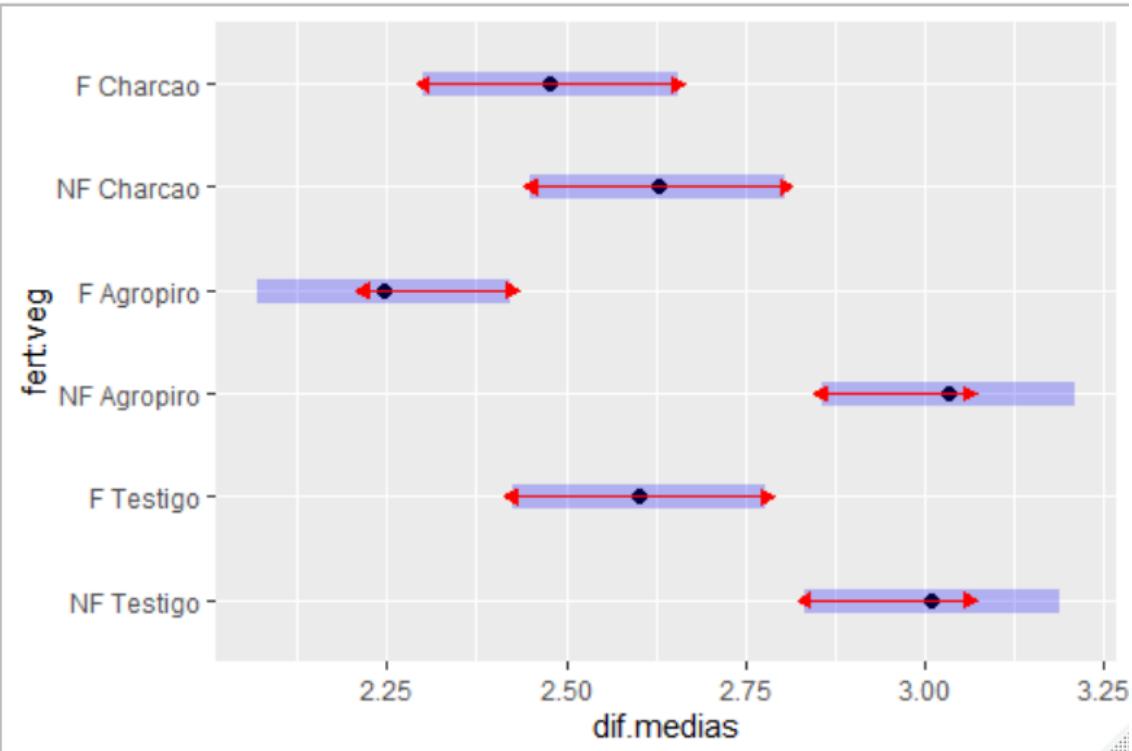
confidence level used: 0.95

\$contrasts	contrast	estimate	SE	df	t.ratio	p.value	lower.CL	upper.CL
	NF,Testigo - F,Testigo	0.408	0.121	24	3.360	0.0277	0.0326	0.7834
	NF,Testigo - NF,Agropiro	-0.024	0.121	24	-0.198	1.0000	-0.3994	0.3514
	NF,Testigo - F,Agropiro	0.764	0.121	24	6.293	<.0001	0.3886	1.1394
	NF,Testigo - NF,charcao	0.382	0.121	24	3.146	0.0445	0.0066	0.7574
	NF,Testigo - F,Charcao	0.532	0.121	24	4.382	0.0024	0.1566	0.9074
	F,Testigo - NF,Agropiro	-0.432	0.121	24	-3.558	0.0176	-0.8074	-0.0566
	F,Testigo - F,Agropiro	0.356	0.121	24	2.932	0.0701	-0.0194	0.7314
	F,Testigo - NF,Charcao	-0.026	0.121	24	-0.214	0.9999	-0.4014	0.3494
	F,Testigo - F,charcao	0.124	0.121	24	1.021	0.9062	-0.2514	0.4994
	NF,Agropiro - F,Agropiro	0.788	0.121	24	6.490	<.0001	0.4126	1.1634
	NF,Agropiro - NF,charcao	0.406	0.121	24	3.344	0.0287	0.0306	0.7814
	NF,Agropiro - F,charcao	0.556	0.121	24	4.579	0.0015	0.1806	0.9314
	F,Agropiro - NF,charcao	-0.382	0.121	24	-3.146	0.0445	-0.7574	-0.0066
	F,Agropiro - F,Charcao	-0.232	0.121	24	-1.911	0.4200	-0.6074	0.1434
	NF,charcao - F,charcao	0.150	0.121	24	1.235	0.8153	-0.2254	0.5254

P value adjustment: tukey method for comparing a family of 6 estimates

IC para la diferencia de medias:
¿Cero pertenece al IC?
¿Cuál es la magnitud del efecto?

```
emmeans(m1, pairwise ~ fert*veg)
```



- Las barras grises son los IC para las medias
- las flechas rojas son para las comparaciones entre grupos. Si una flecha de un grupo se superpone a una flecha de otro grupo, la diferencia no es significativa
- Nota: Nunca usar IC para una media para realizar comparaciones, pueden ser muy engañosos

```
> CLD(comp1)
   fert veg      emmean        SE  df lower.CL upper.CL .group
   F   Agropiro  2.246 0.08585259 24  2.068809 2.423191    1
   F   Charcao   2.478 0.08585259 24  2.300809 2.655191   12
   F   Testigo   2.602 0.08585259 24  2.424809 2.779191   12
   NF  Charcao   2.628 0.08585259 24  2.450809 2.805191    2
   NF  Testigo   3.010 0.08585259 24  2.832809 3.187191    3
   NF  Agropiro  3.034 0.08585259 24  2.856809 3.211191    3
```

Confidence level used: 0.95

P value adjustment: tukey method for comparing a family of 6 estimates
significance level used: alpha = 0.05

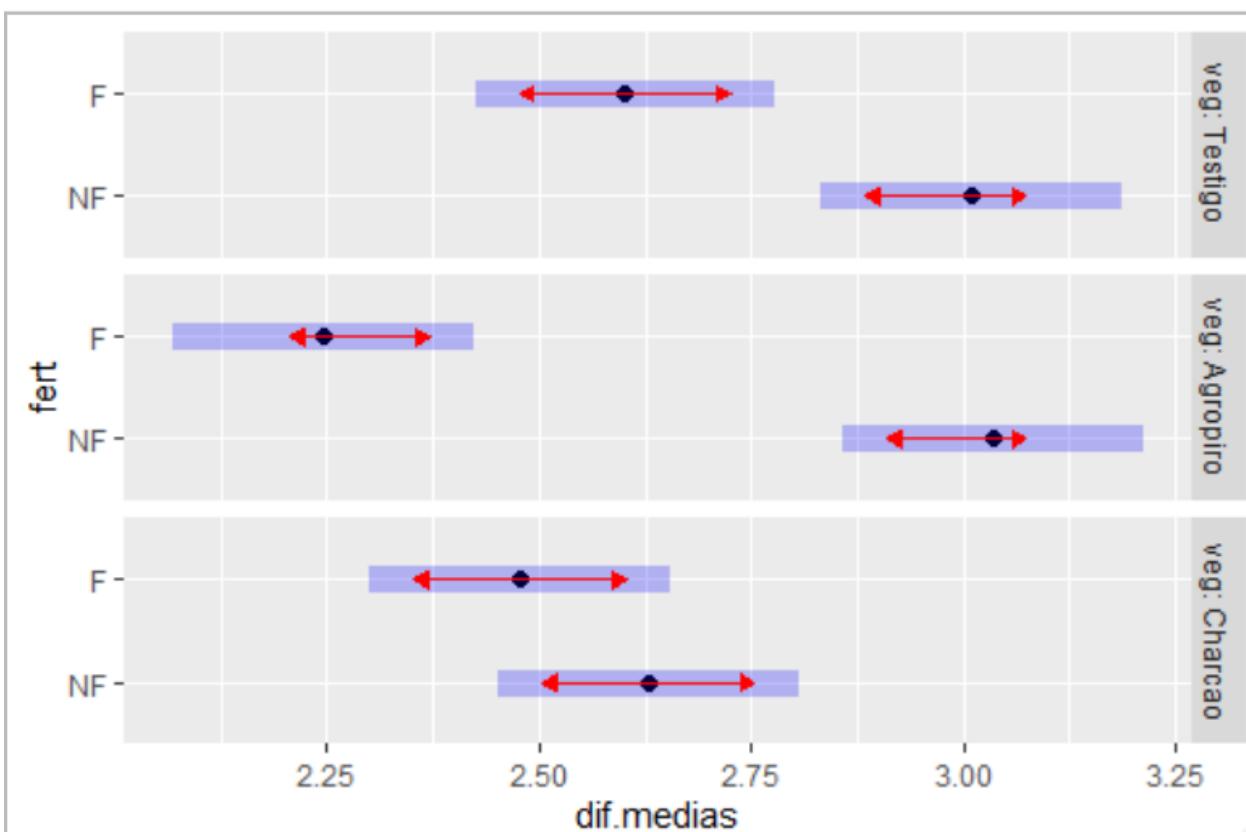
2- Otra posibilidad: efectos simples

emmeans(m1, ~ fert | veg)

```
$contrasts
veg = Agropiro:
contrast estimate      SE df t.ratio p.value
F - NF      -0.788 0.1214139 24  -6.490 <.0001

veg = Charcao:
contrast estimate      SE df t.ratio p.value
F - NF      -0.150 0.1214139 24  -1.235 0.2286

veg = Testigo:
contrast estimate      SE df t.ratio p.value
F - NF      -0.408 0.1214139 24  -3.360 0.0026
```



- Mayor potencia, ya que son menos comparaciones
- La elección depende de los objetivos del ensayo

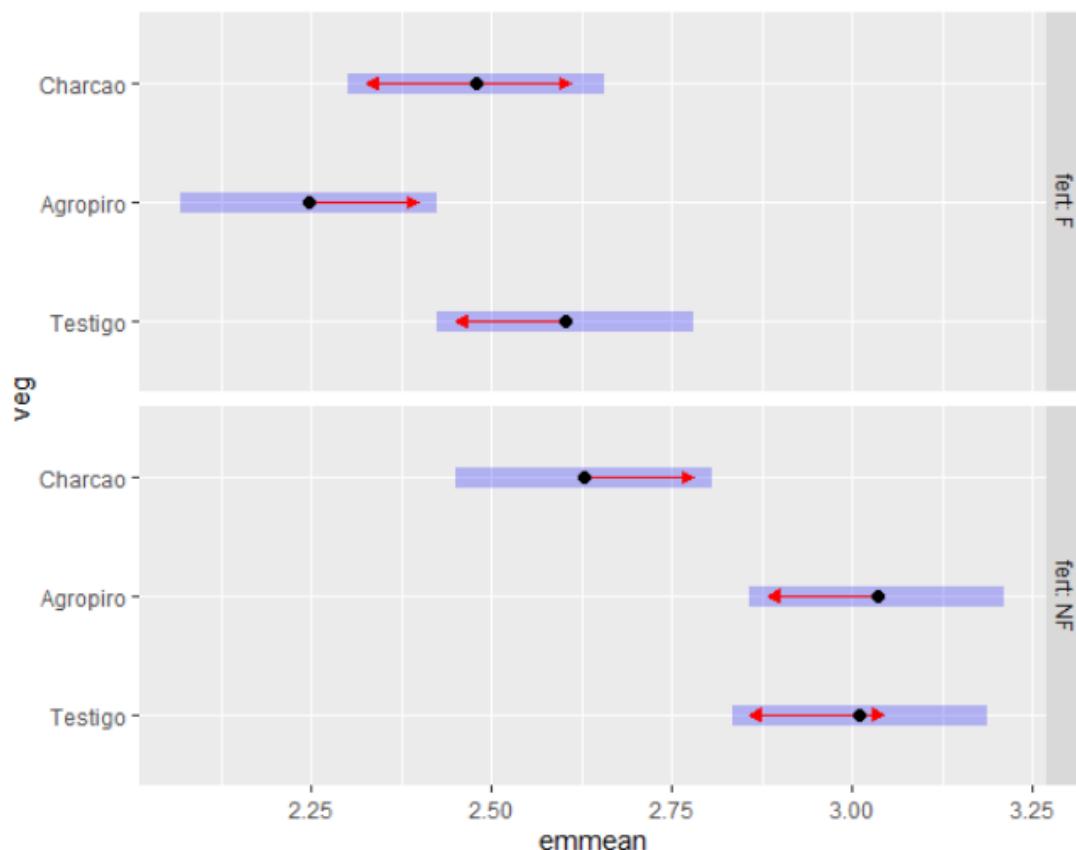
- ✓ Para cada tipo de planta ¿es efectiva la fertilización?
- ✓ O podría interesar: Para cada nivel de fertilización ¿cuál es la especie más efectiva?

2- efectos simples en el otro sentido

emmeans(m1, ~ veg | fert)

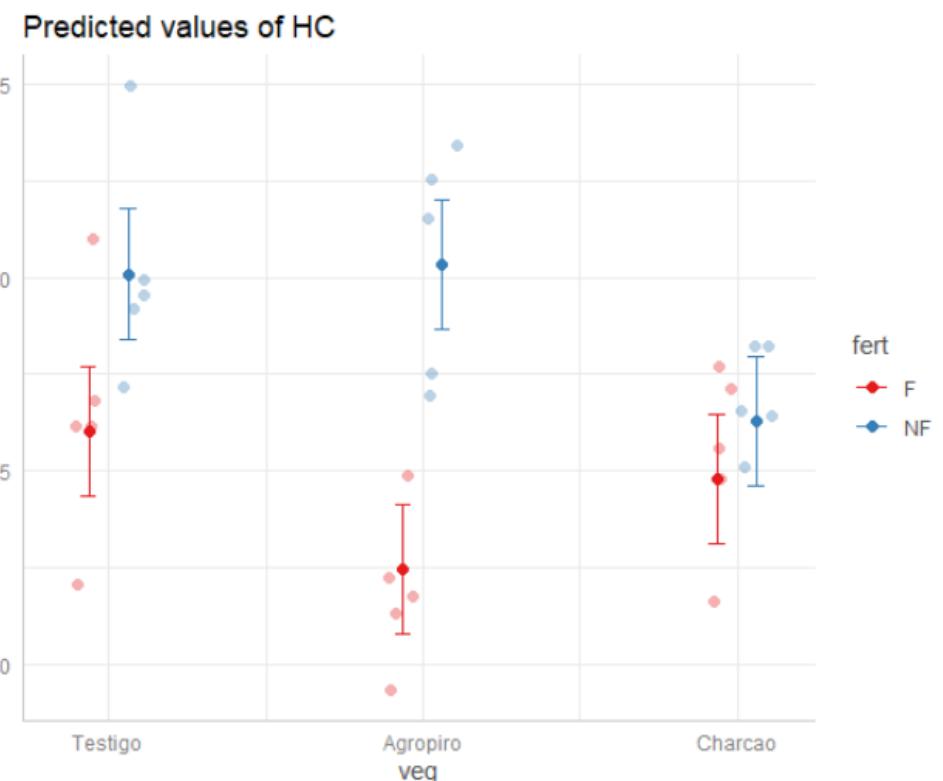
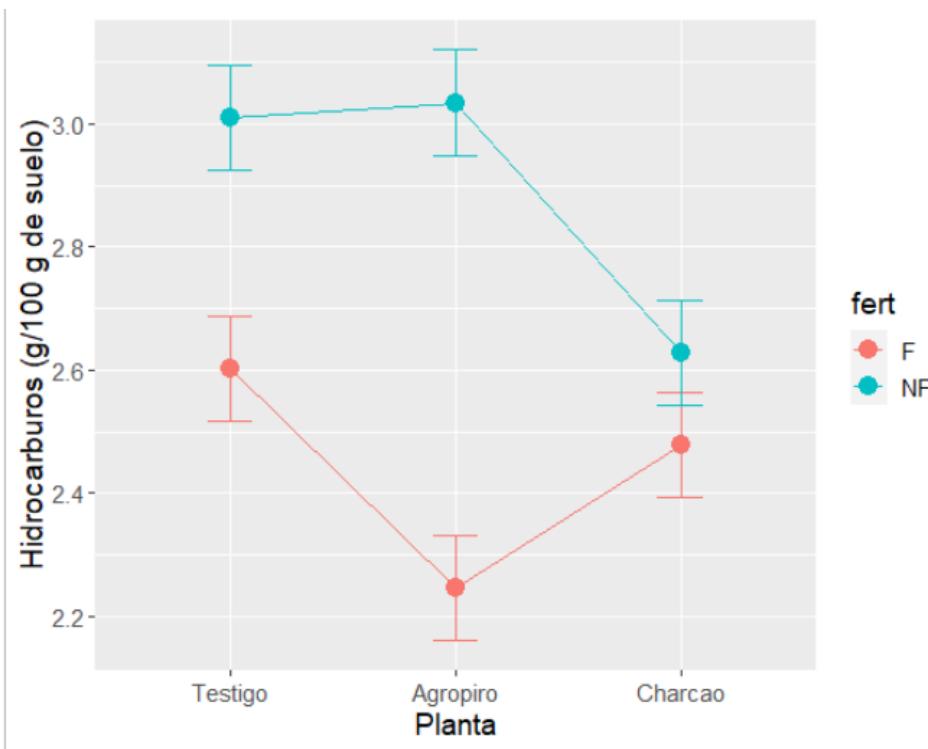
```
$contrasts
fert = F:
contrast          estimate    SE df t.ratio p.value
Testigo - Agropiro   0.356 0.121 24  2.932  0.0192
Testigo - Charcao     0.124 0.121 24  1.021  0.5709
Agropiro - charcao   -0.232 0.121 24 -1.911  0.1574

fert = NF:
contrast          estimate    SE df t.ratio p.value
Testigo - Agropiro   -0.024 0.121 24 -0.198  0.9787
Testigo - Charcao     0.382 0.121 24  3.146  0.0117
Agropiro - charcao    0.406 0.121 24  3.344  0.0073
```

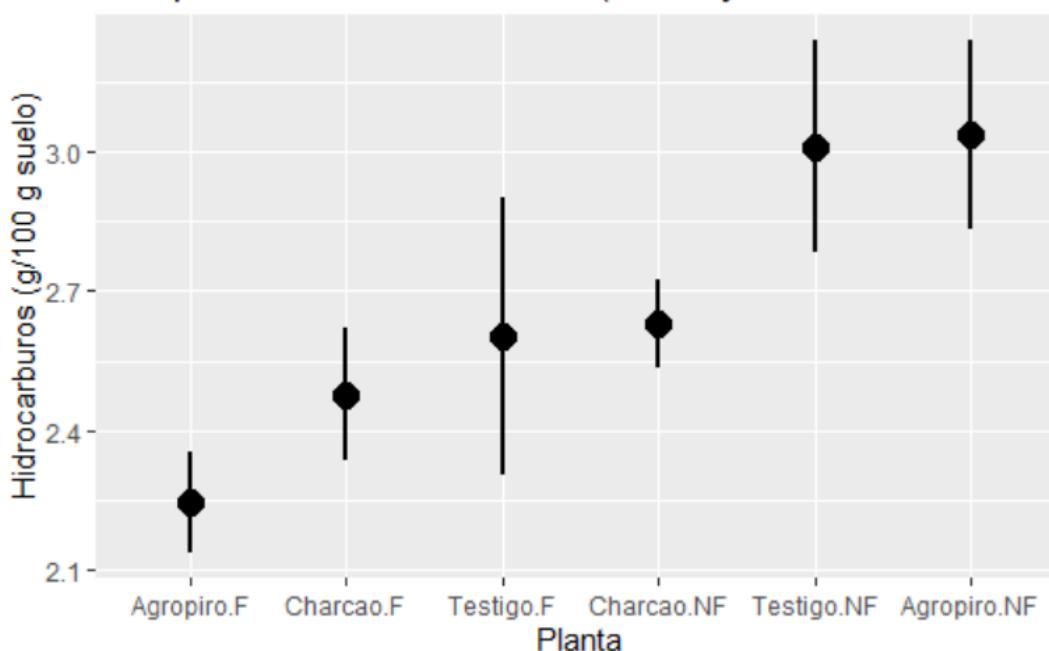


- Mayor potencia, ya que son menos comparaciones
- La elección depende de los objetivos del ensayo

- ✓ Para cada nivel de fertilización ¿cuál es la especie más efectiva?
- ✓ O uno u otro, pero no ambos



Comparación de tratamientos (media y DE)



¿Cuál es la pregunta a responder?

- ✓ Para cada nivel de vegetación ¿es efectiva la fertilización?
- ✓ Para cada nivel de fertilización ¿cuál es la especie más efectiva?
- ✓ ¿Cuál es el mejor de estos tratamientos para restaurar el suelo?
- ✓ ¿Cuál es la magnitud de los efectos?

> CLD(comp1)		
fert	veg	.group
F	Agropiro	1
F	Charcao	12
F	Testigo	12
NF	Charcao	2
NF	Testigo	3
NF	Agropiro	3

Test de Tukey si la interacción no es significativa

Efectos principales:

Comparaciones entre niveles del factor fert

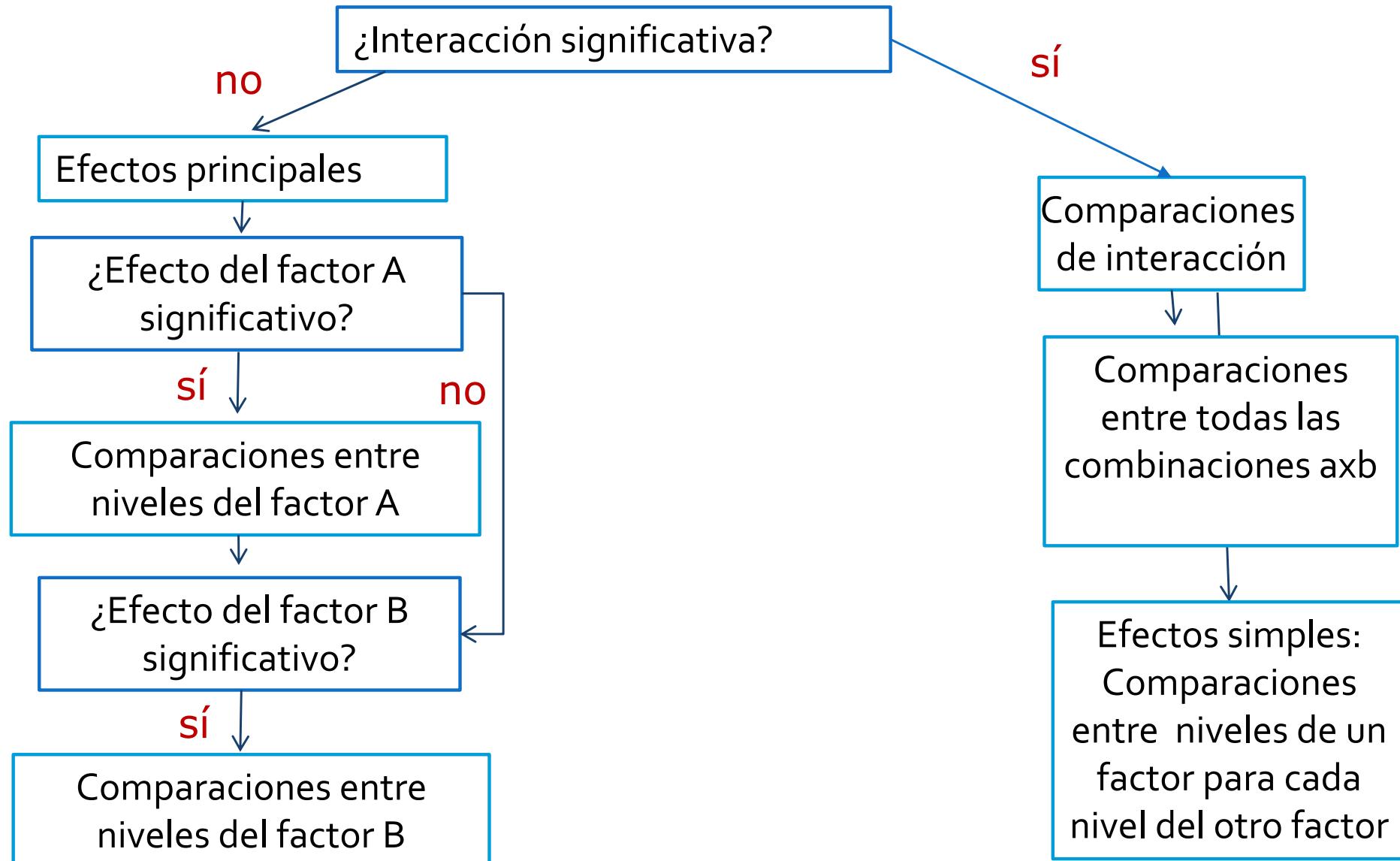
`emmeans(modelo2, pairwise ~ fert)`

Comparaciones entre niveles del factor veg

`emmeans(modelo2, pairwise ~ veg)`

Análisis en diseños factoriales

45



Diseños más complejos

24

- Si se tienen 3 VE (A, B, C) hay una interacción triple y 3 dobles, además de los efectos principales (A, B, C, A*B, A*C, B*C, A*B*C)
- El modelo debe respetar el principio de marginalidad: si se incluye una interacción, se deben incluir los efectos principales de las VE que la constituyen

Fuente de variación	Escenario 1	Escenario 2	Escenario 3	Escenario 4
A				
B				
C				
A*B		S	S	NS
A*C		S	NS	NS
B*C		S	NS	NS
A*B*C	S	NS	NS	NS

Ej de efectos simples en escenario 1 y 2:

`emmeans(m1, pairwise~ A*B|C, adjust= "tukey")`

Selección de modelos con más de una v. explicatoria

25

- Experimentos diseñados:
 - Las VE son en general cualitativas (factores)
 - el modelo viene dado por el diseño experimental, no se lo debería simplificar
 - VE generalmente ortogonales

- Estudios observacionales:
 - VE cuantitativas y/o cualitativas (regresión múltiple)
 - es necesario simplificar el modelo => métodos de selección de modelos
 - VE casi nunca ortogonales

VE ortogonales: La variabilidad explicada por un factor es la misma, independientemente de si el otro factor es tenido en cuenta o no.
Verdadera partición de la variabilidad

Diseños desbalanceados

26

- Cuando los diseños factoriales están **desbalanceados** (distinta cantidad de réplicas en las combinaciones) las SC dejan de ser ortogonales y los resultados pueden diferir según cómo se calculen
- La pérdida de ortogonalidad puede darse también por asociación entre las VE (infrecuente en experimentos pero muy frecuente en estudios observacionales)
- Existen distintos métodos para calcularlas (tipo I, tipo III), que difieren en cómo se calculan las medias marginales

- Simulo desbalanceo en el diseño

```
FITOR[sample(1:nrow(FITOR), 20,
replace=FALSE),]
```

- Calculo SC tipo I y tipo III

```
> anova(modelo3) #solo tipo I
Analysis of Variance Table
```

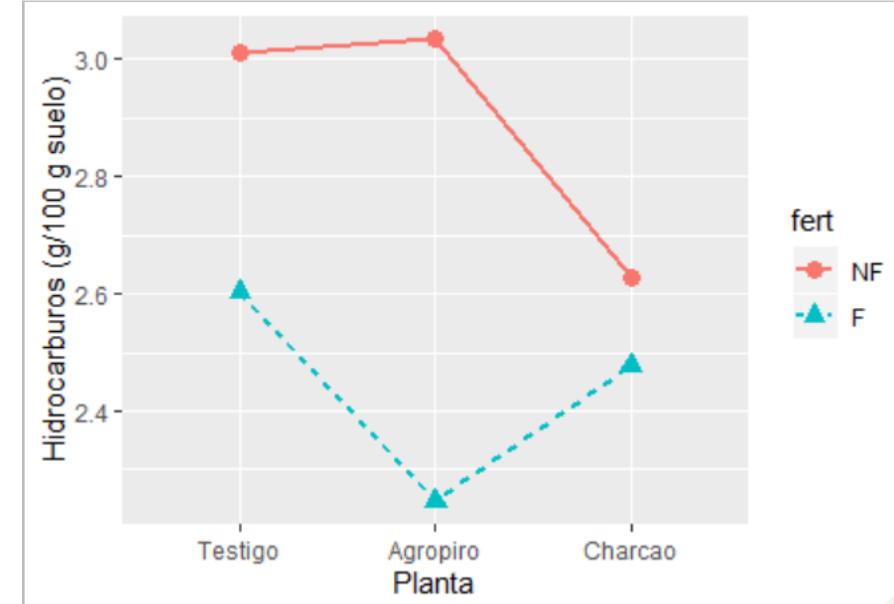
Response: HC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
veg	2	1603.5	801.8	1.8792	0.1893
fert	1	13041.8	13041.8	30.5677	7.431e-05 ***
veg:fert	2	3297.3	1648.7	3.8642	0.0461 *
Residuals	14	5973.1	426.7		

```
> Anova(modelo3, type="III") #paquete car
Anova Table (Type III tests)
```

Response: HC

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	154587	1	362.3254	2.1e-11 ***
veg	980	2	1.1487	0.3451918
fert	11109	1	26.0376	0.0001609 ***
veg:fert	3297	2	3.8642	0.0461000 *
Residuals	5973	14		



Sumas de cuadrados

28

□ SC secuenciales o Tipo I

- Particionan la SC del modelo según la secuencia de incorporación de términos => el orden importa
- Miden la contribución de una VE siendo que las VE que **la preceden** en el modelo ya están incluidas en el mismo
- Como pesa las medias marginales por la cantidad de observaciones, el tamaño de las celdas importa
- Es una verdadera partición de la SC total

$$SC_{X_1}$$

$$SC_{X_2/X_1}$$

$$SC_{X_3/X_1, X_2}$$

□ SC parciales, ajustadas o Tipo III

- Miden la contribución de una VE siendo que **todas las otras** VE ya están incluidas en el modelo => el orden no importa
- se basa en medias marginales sin ponderar (les da el mismo peso independientemente de la cantidad de observaciones)
- No es una verdadera partición de la SC total

$$SC_{X_1/X_2, X_3}$$

$$SC_{X_2/X_1, X_3}$$

$$SC_{X_3/X_1, X_2}$$

Resumiendo:

29

Cuando hay ortogonalidad entre las VE:

- La variabilidad explicada por un factor es la misma, independientemente si la otra VE es tenida en cuenta o no
- SC total puede descomponerse en fuentes de variación independientes y aditivas
- SC secuencial = SC parcial o ajustada
- En experimentos diseñados y balanceados, las SC son ortogonales. En regresión múltiple sólo se daría si las VE fueran estrictamente independientes entre sí (correlación nula). Imposible en estudios observacionales...

Cuando no hay ortogonalidad entre las VE:

- La significación de la interacción es la misma en ambos métodos
- La SC tipo III respeta el principio de marginalidad, más recomendable
- Alternativamente los datos pueden analizarse utilizando un **diseño de celdas** (comparando a x b tratamientos)

BIOMETRÍA II

CLASE 6

REGRESIÓN CON VARIABLES CATEGÓRICAS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Valores de referencia para pruebas de función pulmonar

2



- La ventilación voluntaria máxima (VVM) es el máximo volumen que puede ser ventilado dentro y fuera de los pulmones en un intervalo de 10 a 15 seg mediante esfuerzo voluntario (en litros)
- Se desea establecer valores de referencia de VVM en función de la edad para la población sana brasileña
- Participaron 100 individuos sanos, no fumadores (50 hombres y 50 mujeres), de entre 20 y 80 años de edad

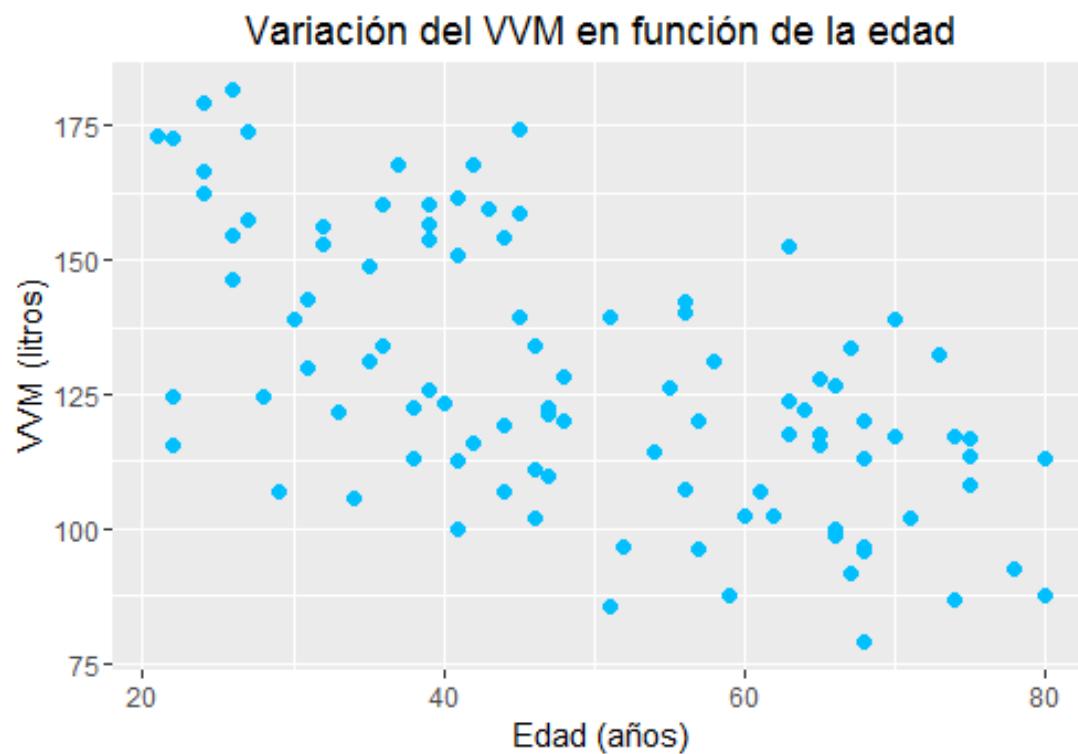
Datos

3

	sexo	edad	vvm
1	varón	55	126.2
2	varón	24	162.5
3	varón	65	117.4
4	varón	36	160.5
5	varón	43	159.6
6	varón	63	117.5
7	varón	37	167.6
8	varón	74	117.3
9	varón	70	138.8

Showing 1 to 9 of 100 entries

sexos
mujer:50 edad
varón:50 Min. :21.00 vvm
 1st Qu.:36.75 1st Qu.:107.9
 Median :46.50 Median :122.5
 Mean :49.27 Mean :127.1
 3rd Qu.:65.00 3rd Qu.:147.0
 Max. :80.00 Max. :181.7



Modelo nulo

modelo0<-lm(vvm ~ 1, VVM)

$$Y_i = \beta_0 + \varepsilon_i$$

4

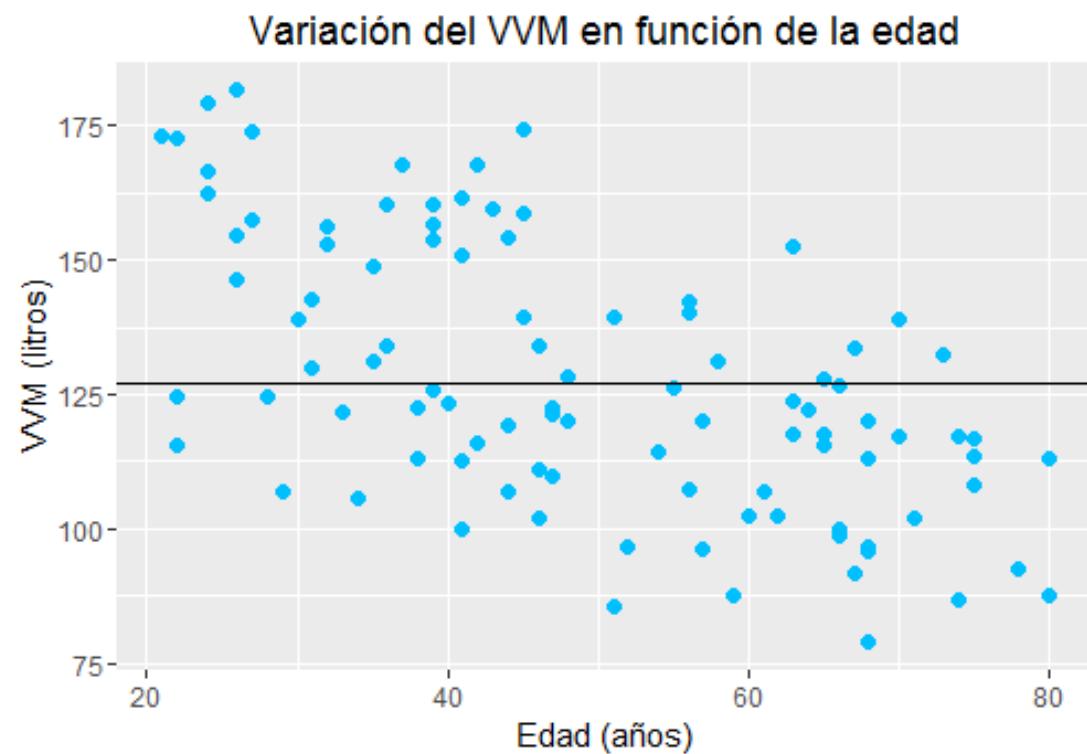
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	127.118	2.495	50.95	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 24.95 on 99 degrees of freedom

```
> round(confint(modelo0),2)
      2.5 % 97.5 %
(Intercept) 122.17 132.07
> summary(modelo0)$r.squared
[1] 0
> AIC(modelo0)
[1] 930.1611
```



Modelo con una VE

modelo1<-lm(vvm ~ edad, VVM)

$$Y_i = \beta_0 + \beta_1 edad + \varepsilon_i$$

5

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	172.3187	6.3671	27.064	< 2e-16 ***
edad	-0.9174	0.1227	-7.478	3.24e-11 ***

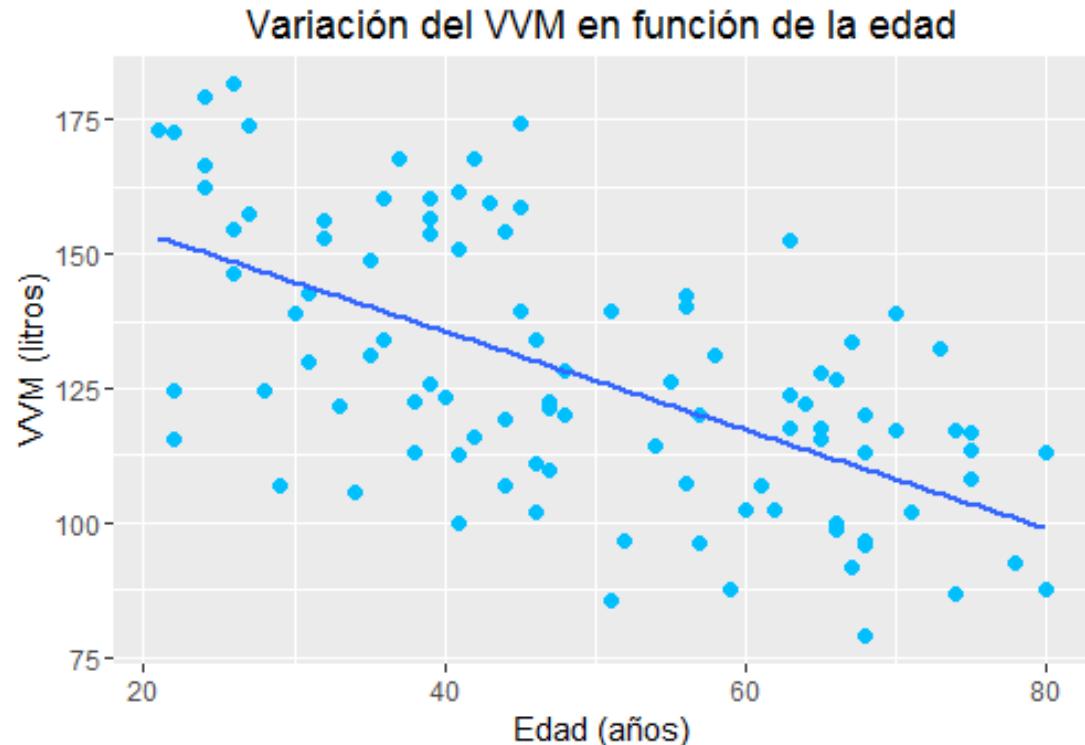
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 20.01 on 98 degrees of freedom

Multiple R-squared: 0.3633, Adjusted R-squared: 0.3568

F-statistic: 55.92 on 1 and 98 DF, p-value: 3.238e-11

```
> round(confint(modelo1),2)
      2.5 % 97.5 %
(Intercept) 159.68 184.95
edad        -1.16 -0.67
> summary(modelo1)$r.squared
[1] 0.3633089
> AIC(modelo1)
[1] 887.014
```



Modelo de regresión múltiple con dos v. explicatorias, una continua y otra categórica con dos categorías

6

- Las v. cualitativas deben ser codificadas para poder ser incluidas en la regresión ([v. auxiliares, indicadoras o dummy](#))
- Si la variable cualitativa tiene sólo dos categorías se la puede codificar utilizando una única variable cuantitativa que tome valores 0 o 1 – presencia/ausencia (aunque puede ser cualquier valor numérico). La categoría que toma el valor 0 es la de [referencia](#)
- En nuestro ejemplo, creamos la variable auxiliar varón: 0: mujer 1:varón

$$E(VVM) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$E(VVM) = \beta_0 + \beta_1 Edad + \beta_2 Varón$$

Para Mujeres($Varón = 0$): $E(VVM) = \beta_0 + \beta_1 Edad$

Para Varones($Varón = 1$): $E(VVM) = (\beta_0 + \beta_2) + \beta_1 Edad$

- β_0 es el valor esperado de Y cuando X_1 y X_2 valen 0
- β_1 es el cambio esperado en Y por cada aumento unitario en X_1
- β_2 es el cambio esperado en β_0 cuando $X_2=1$

Modelo con 2 VE sin interacción

7

Coefficients:

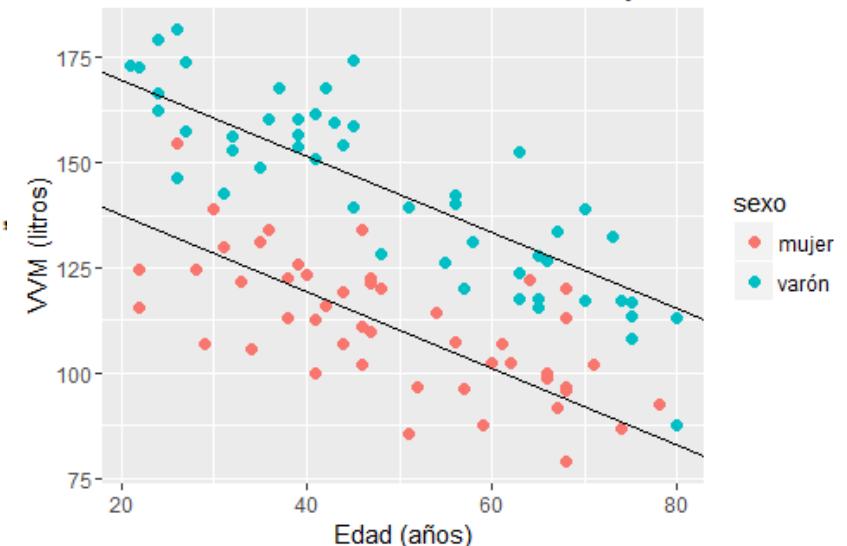
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	155.8252	3.9143	39.81	<2e-16	***
edad	-0.9095	0.0718	-12.67	<2e-16	***
sexovarón	32.2115	2.3421	13.75	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 11.71 on 97 degrees of freedom
 Multiple R-squared: 0.7842, Adjusted R-squared: 0.7797
 F-statistic: 176.2 on 2 and 97 DF, p-value: < 2.2e-16

```
> summary(modelo2)$r.squared
[1] 0.7841738
> AIC(modelo2)
[1] 780.8329
```

Variación del VVM en función de la edad y el sexo



- H₀₁: $\beta_0 = 0$
- H₀₂: $\beta_1 = 0$
- H₀₃: $\beta_2 = 0$ Prueba de igualdad de ordenada al origen

Ecuaciones para hombres y mujeres?

Interacción entre variables explicatorias

8

- El **efecto** de una VE sobre la VR cambia según los valores que tome otra VE
- Es decir que el efecto de una VE **depende de / se asocia con** el valor que tome otra VE (y viceversa) (**modificación de efectos**)
- Si hay interacción entre VE, pierde relevancia estimar los efectos de una dada VE independientemente de los valores que tome la otra VE con la que interactúa (**principio de marginalidad**)
- Las interacciones pueden ser entre cualquier tipo de variables (categóricas con categóricas, cuantitativas con categóricas, cuanti con cuanti...)

Modelo de regresión múltiple con dos VE, una continua y otra categórica con dos categorías e interacción

9

$$E(VVM) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$E(VVM) = \beta_0 + \beta_1 Edad + \beta_2 Varón + \beta_3 Edad \cdot Varón$$

Para Mujeres($Varón = 0$): $E(VVM) = \beta_0 + \beta_1 Edad$

Para Varones($Varón = 1$): $E(VVM) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Edad$

- β_0 es el valor esperado de Y cuando X_1 y X_2 valen 0
- β_1 es el cambio esperado en Y por cada aumento unitario en X_1
- β_2 es el cambio esperado en β_0 cuando $X_2=1$
- β_3 es el cambio esperado en β_1 cuando $X_2=1$

Modelo con 2 VE e interacción (máximo)

10

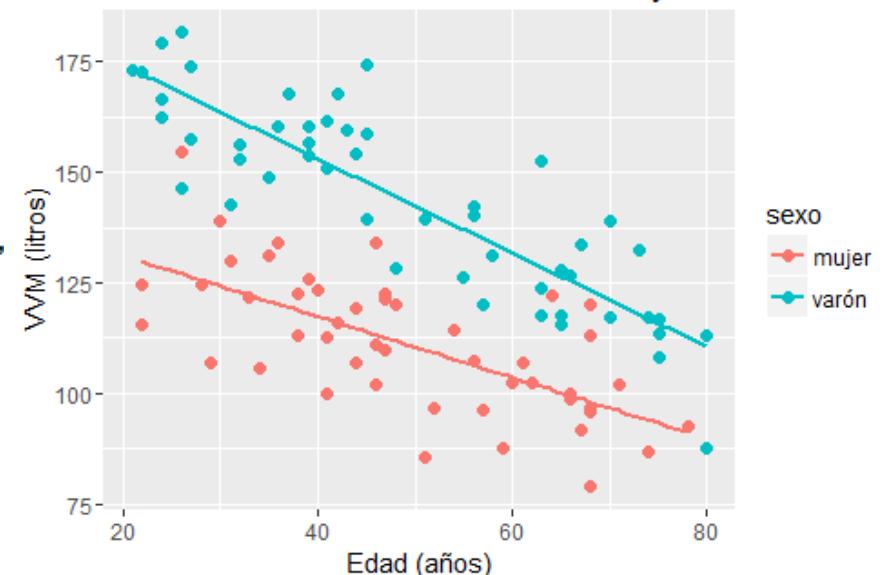
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	144.9467	5.6124	25.826	< 2e-16 ***
edad	-0.6893	0.1089	-6.333	7.73e-09 ***
sexovarón	50.6090	7.3459	6.889	5.84e-10 ***
edad:sexovarón	-0.3732	0.1417	-2.634	0.00984 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 'Residual standard error: 11.37 on 96 degrees of freedom
Multiple R-squared: 0.7987, Adjusted R-squared: 0.7924
F-statistic: 127 on 3 and 96 DF, p-value: < 2.2e-16

```
> summary(modelo3)$r.squared
[1] 0.7987179
> AIC(modelo3)
[1] 775.8563
```

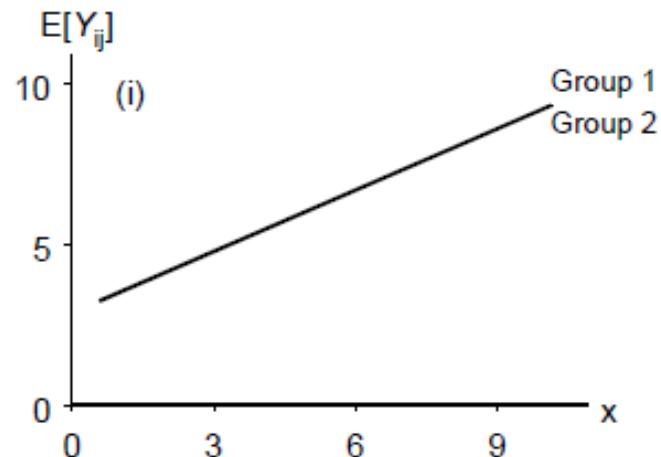
Variación del VVM en función de la edad y el sexo



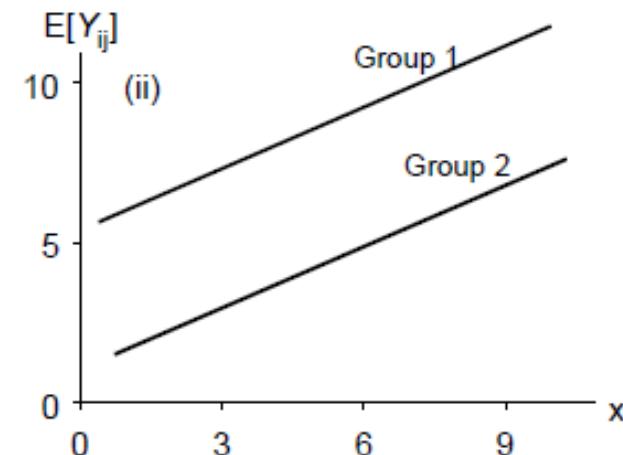
- H₀₁: $\beta_0 = 0$
- H₀₂: $\beta_1 = 0$
- H₀₃: $\beta_2 = 0$ Prueba de igualdad de ordenada al origen
- H₀₃: $\beta_3 = 0$ Prueba de igualdad de pendientes (paralelismo)

Ecuaciones para hombres y mujeres?

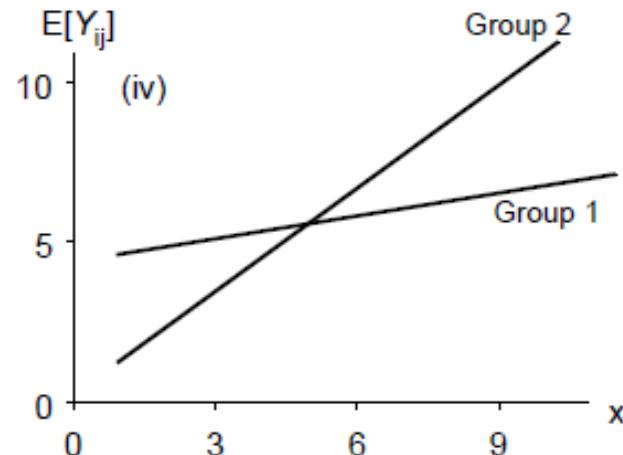
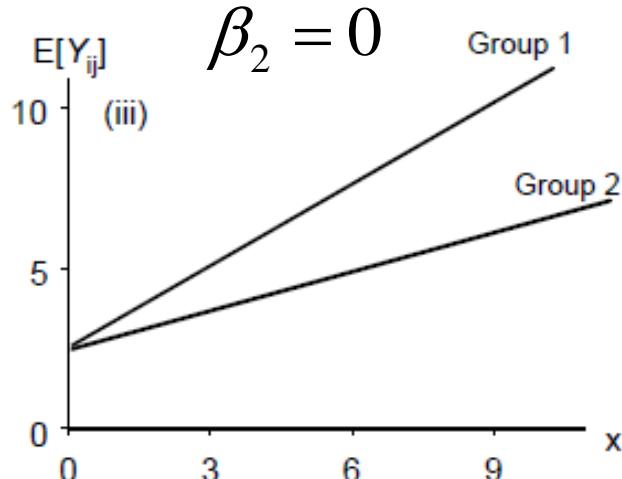
$$\beta_2 = \beta_3 = 0$$



$$\beta_3 = 0$$



$$\beta_2 = 0$$

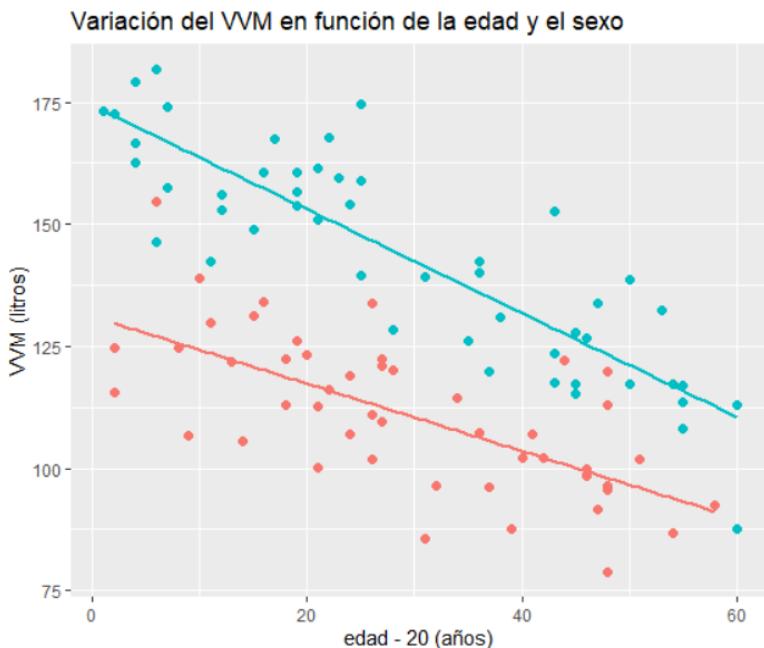


$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

¿Y si a cada valor de edad le restamos 20 años (la edad mínima para la que se desean efectuar predicciones)?

	sexo	edad	vvm	edad_c
1	varón	55	126.2	35
2	varón	24	162.5	4
3	varón	65	117.4	45
4	varón	36	160.5	16
5	varón	43	159.6	23
6	varón	63	117.5	43
7	varón	37	167.6	17
8	varón	74	117.3	54
9	varón	70	138.8	50
10	varón	57	119.9	37

Showing 1 to 11 of 100 entries



modelo4<-lm(vvm ~ edad_c*sexo, VVM)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	131.1602	3.5813	36.624	< 2e-16 ***
edad_c	-0.6893	0.1089	-6.333	7.73e-09 ***
sexovarón	43.1445	4.7329	9.116	1.18e-14 ***
edad_c:sexovarón	-0.3732	0.1417	-2.634	0.00984 **

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	1	1	1	1

Residual standard error: 11.37 on 96 degrees of freedom
Multiple R-squared: 0.7987, Adjusted R-squared: 0.7924
F-statistic: 127 on 3 and 96 DF, p-value: < 2.2e-16

¿Qué cambia?

modelo3<-lm(vvm ~ edad*sexo, VVM)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	144.9467	5.6124	25.826	< 2e-16 ***
edad	-0.6893	0.1089	-6.333	7.73e-09 ***
sexovarón	50.6090	7.3459	6.889	5.84e-10 ***
edad:sexovarón	-0.3732	0.1417	-2.634	0.00984 **

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	1	1	1	1

Residual standard error: 11.37 on 96 degrees of freedom
Multiple R-squared: 0.7987, Adjusted R-squared: 0.7924
F-statistic: 127 on 3 and 96 DF, p-value: < 2.2e-16

Centrado de X

13

- Cuando cero está fuera del rango de X, la ordenada al origen no tiene interpretación en contexto
- El centrado de X consiste en restar a los valores de X una constante (promedio, mínimo o cualquier otro valor con sentido para X)
- La ordenada al origen β_0 se interpreta después del centrado como el valor esperado de Y cuando X es igual a la constante
- Si hay interacción significativa, β_2 se interpreta como la diferencia en el valor esperado de Y con respecto a la categoría de referencia cuando X es igual a la constante
- Además evita problemas de colinealidad cuando se incluyen interacciones (lo veremos en la próxima clase)
- Para leer más: Afshartous, D., & Preston, R. A. (2011). Key results of interaction models with centering. *Journal of Statistics Education*, 19(3), 1-24.

Selección de modelos

14

Para p v. explicatorias, existen $2^p - 1$ modelos posibles. Por ejemplo, si hay 4 v. explicatorias, existen 15 modelos posibles:

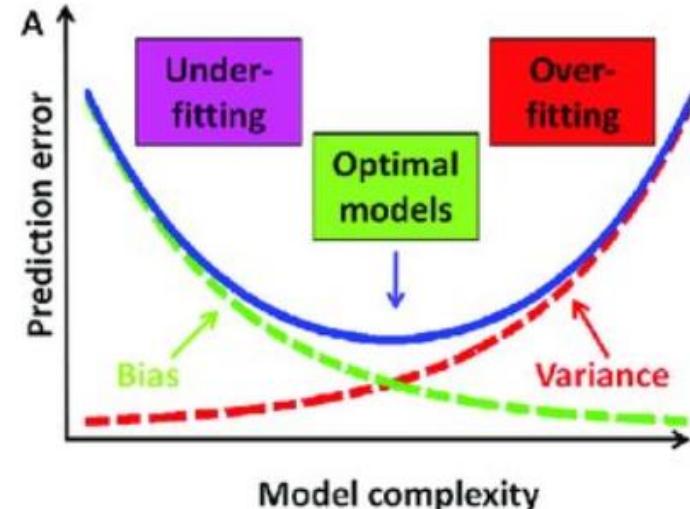
1	X1	11	X1 X2 X3
2	X2	12	X1 X2 X4
3	X3	13	X1 X3 X4
4	X4	14	X2 X3 X4
5	X1 X2	15	X1 X2 X3 X4
6	X1 X3		
7	X1 X4		
8	X2 X3		
9	X2 X4		
10	X3 X4		

Si hay 10 v. explicatorias, 1023 modelos posibles!

Selección de modelos

15

Compromiso entre parsimonia
(la menor cantidad posible de parámetros)
y ajuste (el menor error)



Principio de parsimonia: dado un conjunto de explicaciones igualmente buenas para un fenómeno, la explicación más simple es la correcta. Este principio aplicado a selección de modelos implica:

- Los modelos deben tener la menor cantidad posible de parámetros
- Los modelos con relaciones más simples (por ej lineales) son preferibles a los más complejos (por ej no lineales)
- Los modelos deben ser reducidos hasta encontrar el mínimo adecuado

Criterios para seleccionar el mejor modelo

16

- Mínima varianza residual (S^2_e , CM error)
- Máximo R^2 ajustado
- Mínimo Criterio de información de Akaike (AIC)/Bayesiano
- Retener variables con coeficientes significativos
- Retener variables que provoquen una reducción significativa de la SC residual
- Mínimo Error cuadrático medio de predicción ECMP

Varianza residual

17

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n - p} = \frac{\sum_{i=1}^n e_i^2}{n - p}$$

- En la tabla de anova es el CMerror o residual
- Es el cuadrado del error estándar residual
- Mide la variabilidad en la VR no explicada por las predictoras
- Cuanto más elevada, peor el ajuste del modelo

modelo	CMe
0	622.525
1	400.401
2	137.128
3	129.219
4	129.219

Coeficiente de determinación ajustado

18

- Cuantas más VE se agreguen al modelo, mayor será R^2 (se explica más variabilidad de Y) => no sirve para comparar modelos con distinta cantidad de VE => $R^2_{ajustado}$

$$R^2_{aj} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-p)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)} = 1 - \frac{\hat{\sigma}_{modelo}^2}{\hat{\sigma}_{nulo}^2}$$

- El R^2_{ajust} penaliza la incorporación de VE
- Se utiliza para comparar modelos con distinta cantidad de VE

modelo	R2	R2 ajust
0	0.000	0.000
1	0.363	0.357
2	0.784	0.780
3	0.799	0.792
4	0.799	0.792

Criterio de información de Akaike

19

- Resumen la información de un modelo, teniendo en cuenta la falta de ajuste (verosimilitud) y la cantidad de parámetros (parsimonia)

$$AIC = -2 \log L(\theta) + 2p$$

- Como la verosimilitud \mathcal{L} es un producto de probabilidades, depende de la cantidad de datos. Por lo tanto AIC puede utilizarse para comparar cualquier par de modelos siempre y cuando se estimen sobre los mismos datos
- Cuanto menor, mejor el modelo
- Idem BIC

	df	AIC
modelo0	2	930.1611
modelo1	3	887.0140
modelo2	4	780.8329
modelo3	5	775.8563
modelo4	5	775.8563

Error cuadrático medio de predicción (ECMP)

20

- se estiman los coeficientes del modelo excluyendo a un subconjunto de observaciones
- Se calcula el valor esperado de dichas observaciones \hat{y}_{i-1}
- Se definen los residuos de validación cruzada como

$$e_{i-1} = y_i - \hat{y}_{i-1}$$

- Se define ECMP como

$$ECMP = \frac{\sum e_{i-1}^2}{n}$$

Pruebas de hipótesis

21

- $H_0: \beta_i = 0$
- Equivale a comparar modelos anidados:

$$\text{Modelo 2 } y_i = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{ijk} \quad a \text{ parámetros}$$

$$\text{Modelo 1 } y_i = \beta_o + \beta_1 x_1 + \varepsilon_{ijk} \quad b \text{ parámetros}$$

El modelo 1 está anidado en el modelo 2 si todas las VE que se encuentran en el modelo 1 se incluyen en el modelo 2, es decir, el conjunto de VE en el modelo 1 es un subconjunto del conjunto de VE en el modelo 2

$b < a$, el modelo 1 (más simple, reducido) está anidado en el modelo 2

El criterio para establecer si una o un conjunto de VE deben ser retenida en un modelo con k VE es determinar la significación de la reducción en la SC residual

$$F = \frac{(SCres_1 - SCres_2)/(GL_1 - GL_2)}{SCres_2/GL_2}$$

`anova(modelo1, modelo 2)`

o

`drop1(modelo2, test="F")`

Selección de modelos

22

Table I. Commonly used model selection methods

Model selection method	Calculation ^a	Elements	Refs
Adjusted R^2	$R_{adj}^2 = 1 - \frac{RSS/n - p - 1}{\sum(y_i - \bar{y})^2/n - 1}$	Fit	[7]
Likelihood ratio test	$LRT = -2\{\ln[L(\hat{\theta}_p y)] - \ln[L(\hat{\theta}_{p+q} y)]\} \sim \chi_q^2$	Fit and complexity	[7]
Akaike information criterion (AIC)	$AIC = -2\ln[L(\hat{\theta}_p y)] + 2p$	Fit and complexity	[3]
Small sample unbiased AIC (AIC_c)	$AIC_c = -2\ln[L(\hat{\theta}_p y)] + 2p\left(\frac{n}{n - p - 1}\right)$	Fit and complexity (with bias correction term for small sample size)	[3]
Schwarz criterion	$SC = -2\ln[L(\hat{\theta}_p y)] + p \cdot \ln(n)$	Fit, complexity, and sample size	[10]

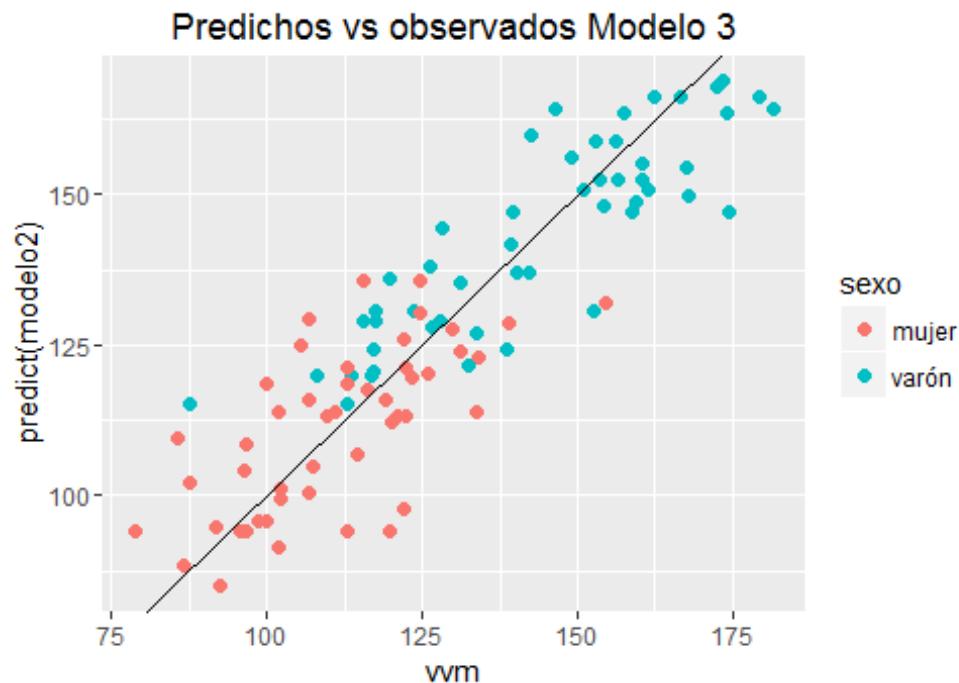
^aRSS, residual sum of squares for a linear model; n , sample size; p , count of free parameters (σ^2 must be included if it is estimated from the data); q , additional parameters of a fuller model; y : data; $L(\hat{\theta}|y)$: likelihood of the model parameters (more precisely, their maximum likelihood estimates, $\hat{\theta}_p$) given the data, y ; for a model fitted by least squares with the usual assumptions, $\ln[L(\hat{\theta}_p|y)] = -n/2\ln(RSS/n)$, enabling computation of LRTs, AIC, AIC_c , and SC from standard regression output.

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2), 101-108.

Validación del modelo

23

□ Predichos vs observados



Coeficiente de correlación:
 $r = 0.8937$

Coeficiente de determinación:
 $R^2 = 0.8937^2 = 0.799$

Validación cruzada

24

- Conjunto de métodos para medir el desempeño de un modelo evaluando su capacidad para predecir **un nuevo conjunto de datos independientes**
- La idea básica consiste en dividir los datos en dos conjuntos:
 - el **conjunto de entrenamiento** (training set) utilizado para entrenar (es decir, construir) el modelo
 - el **conjunto de prueba** (validation set) utilizado para probar (es decir, validar) el modelo mediante la estimación del error de predicción
- Se compara el desempeño predictivo de los modelos usando distintos estadísticos:
 - Raíz del error cuadrático medio (RMSE)
 - Error absoluto medio (MAE)
 - R² entre predichos y observados

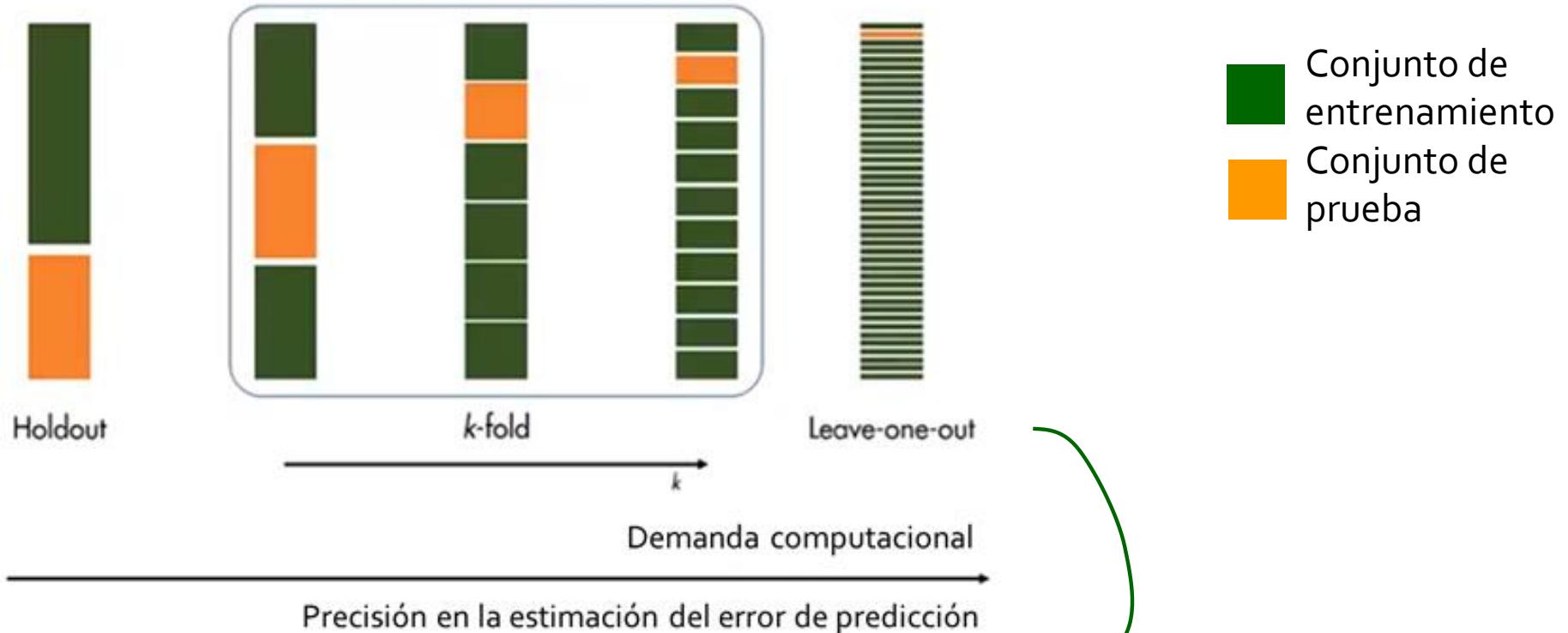
Según modelo estimado con conjunto de entrenamiento

$$RMSE = \sqrt{ECM} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n}} = \sqrt{\frac{\sum e_i^2}{n}}$$
$$MAE = \sqrt{\frac{\sum|y_i - \hat{y}|}{n}} = \sqrt{\frac{\sum|e_i|}{n}}$$

Métodos de validación cruzada

25

- ▣ **Método de retención (holdout):** Consiste simplemente en dividir la base de datos aleatoriamente en dos partes (por ej 70:30). Con una se estima el modelo y con la otra se estima el error de predicción. Requiere n grande. Puede presentar sesgos
- ▣ **Validación cruzada de K-iteraciones (K-fold cross-validation):** se divide la muestra en K submuestras, de forma que se utilizan $K-1$ para estimar el modelo y la restante como conjunto de prueba, este proceso se repite K veces, de forma que cada submuestra es utilizada una vez para evaluar el modelo y $K-1$ veces para estimarlo. Una vez finalizadas las iteraciones, se calcula el error de predicción para cada uno de los modelos producidos, y para obtener el error final se calcula el promedio de los K modelos entrenados
- ▣ **Leave-one-out:** Idem anterior, salvo que la muestra de validación está formada por un único caso. Computacionalmente demandante, pero máxima precisión 
- ▣ Se estima el error de predicción de todos los modelos que se están evaluando y se selecciona aquel que produzca el **menor** error promedio de estimación



Indicamos la función para el entrenamiento

```
trainControl(method = "LOOCV")
```

Entrenamos (estimamos) el modelo 1 (n modelos con n-1 observaciones)

```
m1loo <- train(VVM~ edad*sexo, data=bd, method ="lm", trControl= train.control(method = "LOOCV"))
```

Indicadores de desempeño

```
RMSE(pred = predict(m1loo, bd), obs = bd$vvm)
```

$$\text{error relativo} = \frac{RMSE}{\bar{Y}}$$

	RMSE	Rsquared	MAE	ER
1	20.20068	0.3383328	17.372683	15.89128
2	11.90436	0.7701267	9.562724	9.36481
3	11.61890	0.7810703	9.074065	9.14025

Mejorando la precisión en la estimación

27

- Si $n \gg p$, poco error en las estimaciones
- Si n no es mucho mayor que p , el error es mayor, generalmente hay sobreajuste (el modelo describa a la muestra particular) y las predicciones de futuros casos serán pobres
- Si $n < p$, no puede aplicarse cuadrados mínimos, ya que no existe una única estimación de los coeficientes del modelo y la varianza es infinita

Reducción de la cantidad de VE
(y por lo tanto los coeficientes a estimar)

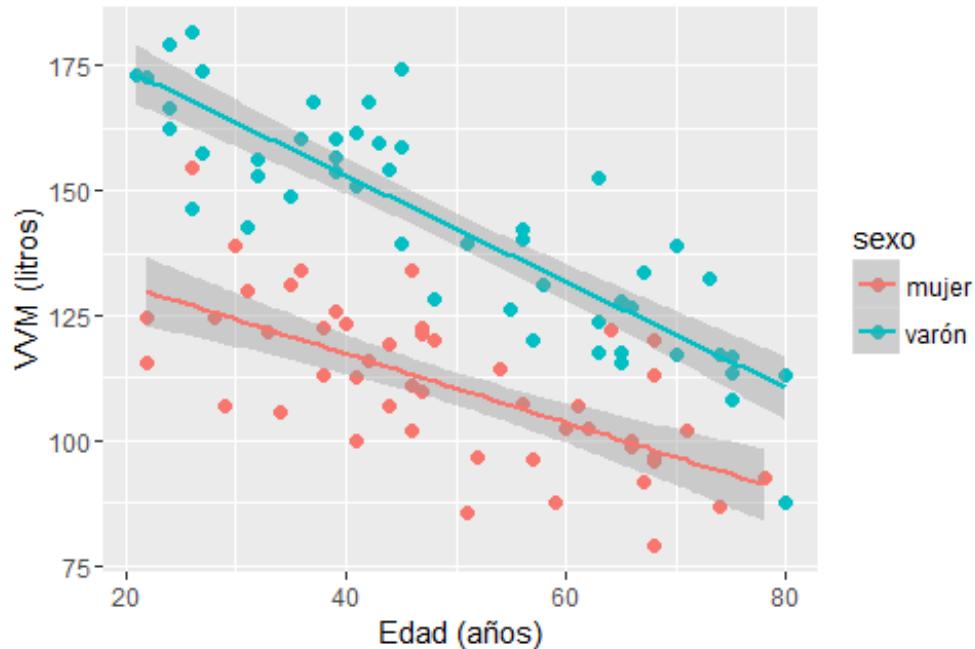
- Se busca mejora la precisión en las estimaciones, con poco aumento del sesgo
- Métodos:
 - Subconjunto de VE (criterios de selección de modelos ya vistos)
 - Shrinkage o encogimiento (regresión penalizada) - LASSO
 - Reducción de la dimensionalidad (análisis de componentes principales, técnica multivariada)

Predicciones de VVM



28

Variación del VVM en función de la edad y el sexo



Para Mujeres(Varón = 0):

$$VVM = 144,95 - 0,69 \text{ Edad}$$

Para Varones(Varón = 1):

$$\begin{aligned} VVM &= (144,95 + 50,61) + (-0,69 - 0,37) \text{ Edad} = \\ &= 195,56 - 1,06 \text{ Edad} \end{aligned}$$

¿Cuál es el VVM esperado para un hombre de 50 años?

```
nuevo = data.frame(sexo="varón", edad=50)
predict(modelo3, nuevo, interval="predict")
```

fit	lwr	upr
142.4282	119.6389	165.2175

Ojo, si usamos el modelo
centrado, edad_c = 50-20

¿Por qué no efectuar dos regresiones simples en vez de una múltiple?

29

- Tanto para RLS como para RLM las estimaciones de los parámetros son las mismas!
 - Para Mujeres : $VVM = 144,95 - 0,69 \text{Edad}$*
 - Para Varones : $VVM = 195,56 - 1,06 \text{Edad}$*
- Pero la RLM:
 - Permite comparar estadísticamente los parámetros de ambas regresiones
 - Mejor estimación de la varianza del modelo σ^2 , más GL
 - Si no hay interacción, mejor estimación de β_1
 - Menor error global

	RLS Mujeres	RLS Varones	RLM
n	50	50	100
R ²	0,45	0,75	0,80
CMerror	133,83	124,61	129,22
GLerror	48	48	96

Reference values for lung function tests. II. Maximal respiratory pressures and voluntary ventilation

Brazilian Journal of Medical and Biological Research (1999) 32: 719-727

3.

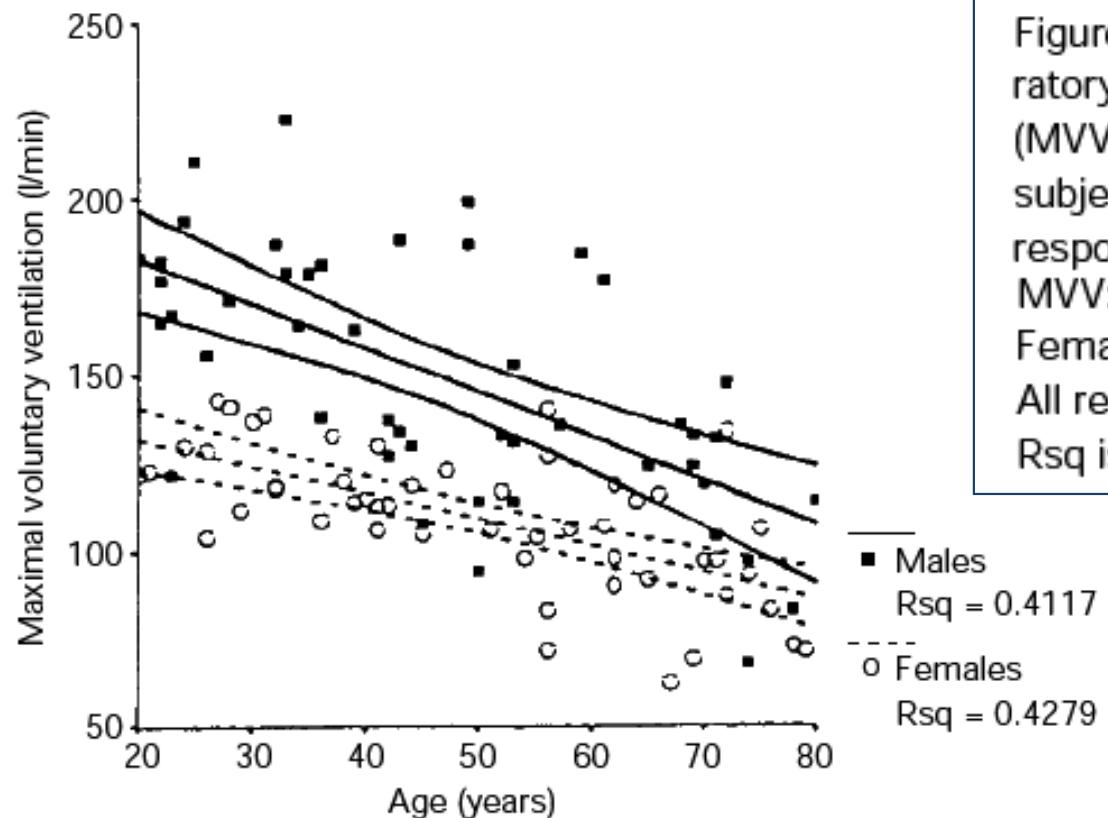


Figure 1 - Maximal inspiratory pressure (MIP) (A), expiratory pressure (MEP) (B) and voluntary ventilation (MVV) (C) as a function of age in 100 healthy sedentary subjects. Regression lines are presented with the corresponding 95% confidence limits (CL).
MVV: Males: $y = -1.12 \text{ (age)} + 199.1$, SEE = 27.5;
Females: $y = -0.76 \text{ (age)} + 147.4$, SEE = 15.3.
All regressions were statistically significant at $P < 0.01$.
Rsq is the coefficient of determination.

Algunos comentarios

31

- Si existen más de dos categorías se deben generar tantas v. dummy como categorías menos 1 (todas las dummy tomarán el valor 0 para la categoría de referencia)
- Por ejemplo, si hubiese tres categorías de nivel de actividad física:

- Baja (**referencia**)
- Moderada
- Alta

	D ₁ moderada	D ₂ alta
baja	0	0
moderada	1	0
alta	0	1

- No es correcto asignar valores crecientes (por ejemplo 1, 2 y 3) ya que la escala de la variable es ordinal y se la convierte en cuantitativa, asignándole una métrica que no posee
- Como ya se vio, los coeficientes miden diferencias con respecto a la categoría de referencia. Pero las comparaciones no están corregidas por múltiples test. Deben aplicarse métodos de **comparaciones**

BIOMETRÍA II

CLASE 7

REGRESIÓN MÚLTIPLE

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Valores de referencia para pruebas de función pulmonar

2



- Continuando con el estudio, en 50 hombres se registró la edad, altura (en cm) y peso (en kg), con el objetivo de estimar la ventilación voluntaria máxima (VVM) (en litros/min)
- VR
- VE
- Modelo

resp.csv

Regresión lineal múltiple

3

- Una única variable respuesta o dependiente (Y) cuantitativa y más de una variable VE, explicativas o independientes (varias X), que pueden ser cuantitativas o cualitativas
- Sin interacción (efectos aditivos): el efecto de X_1 sobre la respuesta media no depende del nivel de X_2 y viceversa

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad \varepsilon_i \approx NID(0, \sigma^2)$$

$$E(Y / X_1, \dots, X_k) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k X_{ki}$$

- Pueden agregarse al modelo anterior términos cuadráticos, cúbicos, etc:

$$E(Y / X_1, \dots, X_k) = \beta_0 + \beta_1 x_{1i} + \beta_2 {x_{1i}}^2 + \dots + \beta_k X_{ki}$$

- β_i son los coeficientes de regresión parcial, ya que indican la influencia (parcial) de cada VE sobre Y , cuando se mantiene constante la influencia de las otras VE
- k es la cantidad de VE, por lo tanto la ecuación del modelo posee p parámetros

Múltiples VE

4

- Lo ideal es que cada una proporcione información “independiente” sobre el comportamiento de Y => modelos sin información redundante, más parsimoniosos
- Si ello ocurre, se dice que las VE son **ortogonales**

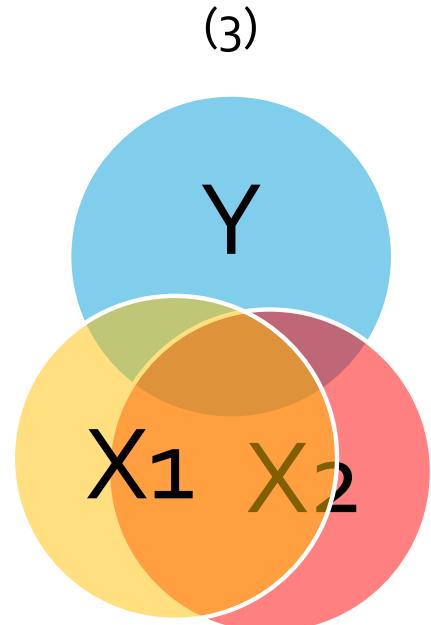
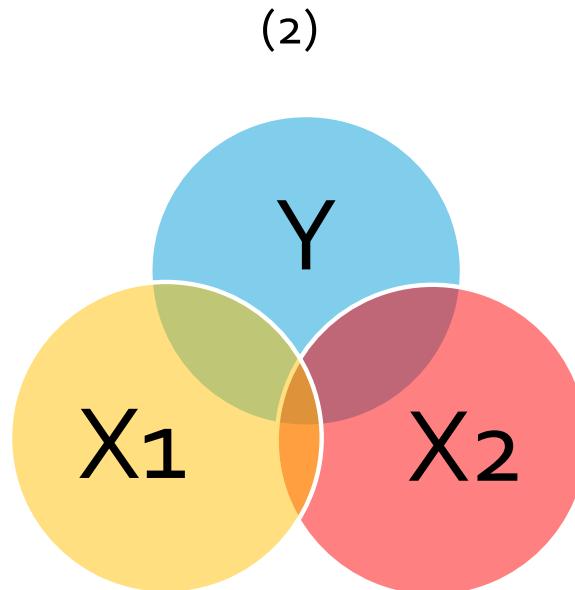
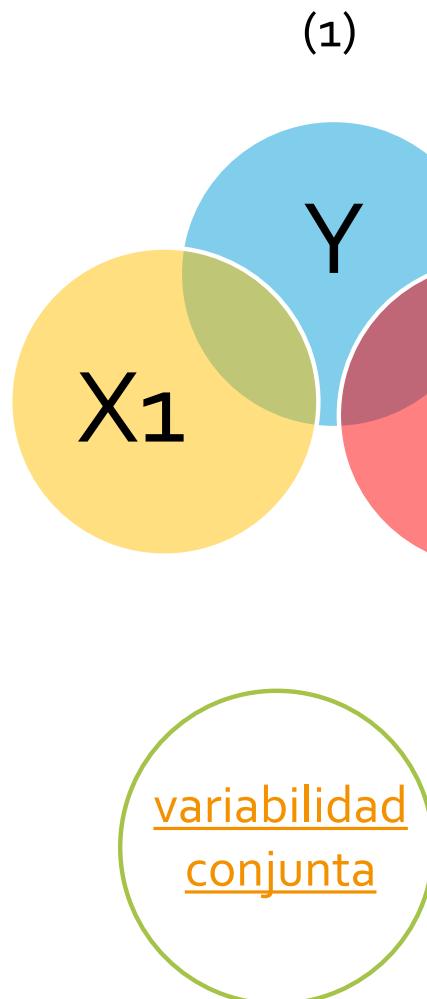
Dos variables VE son ortogonales (independientes) cuando el conocimiento de una no proporciona información sobre la otra, es decir que no están asociadas

- Es habitual en experimentos diseñados pero casi imposible en estudios observacionales
- Si las VE están asociadas linealmente se habla de **colinealidad**

Dos variables VE son colineales cuando están asociadas linealmente. Si una VE es combinación lineal exacta de otra, la colinealidad es perfecta y la estimación por cuadrados mínimos de los coeficientes β no tiene una solución única

La regresión múltiple busca estimar la contribución independiente de X_1 y X_2 a la variación de Y , es decir, estimar β_1 y β_2 .

- (1) X_1 y X_2 son independientes, ortogonales. La asociación entre ellas es nula
- (2) X_1 y X_2 están asociadas débilmente
- (3) X_1 y X_2 están asociadas fuertemente



Aunque la variación de Y explicada por X_1 y X_2 es similar a (1), cuanto mayor es la asociación entre X_1 y X_2 menor es la contribución independiente de cada variable y eventualmente se convierte no significativa

¿Cómo estudiamos asociación entre variables cuantitativas?

6

- Gráficamente: matrices de diagramas de dispersión
- Analíticamente: Coeficiente de correlación lineal de Pearson ρ
- Mide el **grado de asociación lineal** entre dos variables **aleatorias**
- No depende de las unidades de medida de las variables originales
- Toma valores entre $[-1,1]$. Cuanto más cerca esté de $+1$ o -1 más fuerte será el grado de relación lineal (siempre que no existan datos anómalos)
- Su **signo** nos indica si la posible relación es directa o inversa
- Su estimador muestral es:

$$r = \frac{S_{Y_1 Y_2}}{S_{Y_1} S_{Y_2}} = \frac{\sum_{i=1}^n (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sqrt{\sum_{i=1}^n (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^n (y_{i2} - \bar{y}_2)^2}}$$

En un experimento diseñado

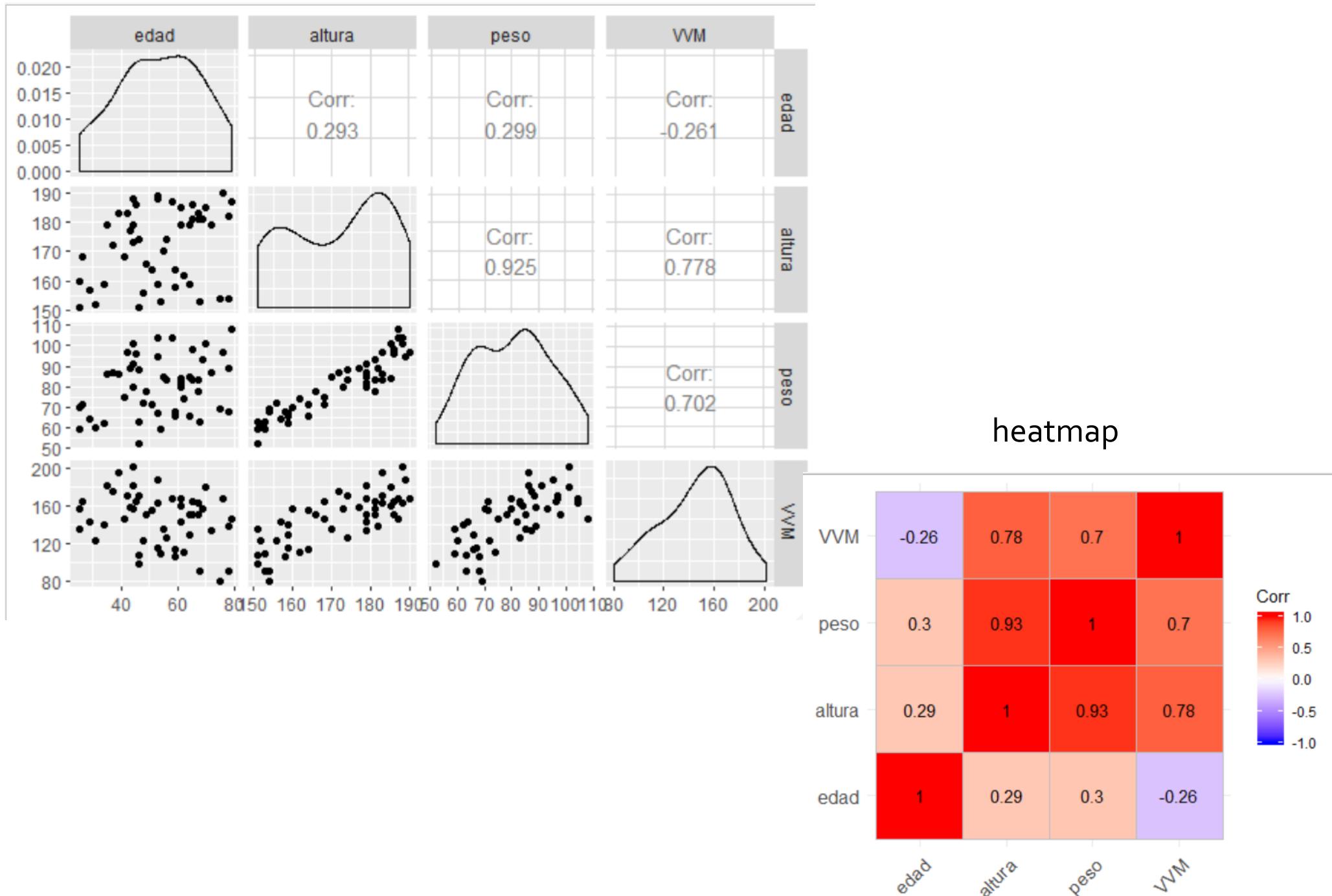
7

- Se prueban 4 dosis de un nitrógeno (0, 10, 20 y 30 mg) y 2 temperaturas (20 y 30 C) en plántulas, y se mide la longitud de la plántula (en cm) al mes de iniciado el tratamiento

Dosis	Temperatura	Longitud
0	20	89
0	20	88
0	30	66
0	30	59
10	20	93
10	20	73
10	30	82
10	30	77
20	20	100
20	20	67
20	30	57
20	30	68
40	20	69
40	20	59
40	30	62
40	30	59

Coeficiente de correlación lineal entre dosis y temperatura $r = 0!$
Las VE son ortogonales

En nuestro estudio observacional



Colinealidad

9

¿Qué provoca?

- Las estimaciones de los coeficientes tendrán varianzas muy altas (alto EE), es decir que tendrán poca precisión. Eso puede provocar que las PH individuales sean no significativas aunque el modelo global sea significativo o el R^2 sea alto
- Los coeficientes de regresión pueden presentar signos contrarios a los esperados
- Sin embargo, los estimadores de los valores esperados de la VR seguirán siendo insesgados

$$EE_{\hat{\beta}_j} = \sqrt{\frac{S_e^2}{(1 - R_j^2) \cdot \sum(x_i - \bar{x})^2}}$$

Menor EE cuanto mayor
es la dispersión de X

Se efectúa una regresión de X_i
en función de las restantes VE y
se calcula el R^2 . Si las VE son
ortogonales, $R_j^2 = 0$

```
m1<-lm(vvm ~ edad + altura + peso)
```

```
m2<-lm(vvm ~ edad + peso + altura)
```

SC secuencial o Tipo I

```
> anova(m1)
```

Analysis of Variance Table

Response: VVM

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
edad	1	2694	2694	23.5461	1.44e-05 ***
altura	1	31678	31678	276.8207	< 2.2e-16 ***
peso	1	3	3	0.0287	0.8662
Residuals	46	5264	114		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

```
> anova(m2)
```

Analysis of Variance Table

Response: VVM

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
edad	1	2694.5	2694.5	23.546	1.440e-05 ***
peso	1	26470.4	26470.4	231.317	< 2.2e-16 ***
altura	1	5210.4	5210.4	45.532	2.179e-08 ***
Residuals	46	5263.9	114.4		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

SC_{X_1}

SC_{X_2/X_1}

$SC_{X_3/X_1,X_2}$

$SC_{X_1/X_2,X_3}$

$SC_{X_2/X_1,X_3}$

$SC_{X_3/X_1,X_2}$

SC marginal o Tipo III

Anova Table (Type III tests)

Response: VVM

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	2570.5	1	22.4628	2.095e-05 ***
edad	10263.9	1	89.6928	2.229e-12 ***
altura	5210.4	1	45.5323	2.179e-08 ***
peso	3.3	1	0.0287	0.8662
Residuals	5263.9	46		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

```
> Anova(m2, type="III")
```

Anova Table (Type III tests)

Response: VVM

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	2570.5	1	22.4628	2.095e-05 ***
edad	10263.9	1	89.6928	2.229e-12 ***
peso	3.3	1	0.0287	0.8662
altura	5210.4	1	45.5323	2.179e-08 ***
Residuals	5263.9	46		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

```
> |
```



Obviamente, si las VE son ortogonales,
ambas SC coinciden

Es la que calcula lm

Colinealidad

11

¿Cómo se detecta? Estudiando la asociación entre VE

- Gráficos de dispersión y coeficientes de correlación para todos los pares posibles de X (pero solo detecta colinealidad de a pares)
- Efectuando una regresión de X_i en función de las restantes VE y calculando R^2 . El proceso se efectúa para todas las VE. Valores cercanos a 1 indican problemas de colinealidad que pueden involucrar a más de una VE
- VIF (factor de inflación de la varianza): mide para cada X el aumento de la varianza del coeficiente de regresión debido a la correlación entre VE

$$VIF_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el R^2 que se obtiene al efectuar una RLM con X_j como VD vs las demás X

- Toma valores entre 1 e infinito. Valores superiores a 5 indican colinealidad importante

```
> library(car)
> vif(m1)
      edad    altura     peso
1.100223 6.967441 6.994053
```

Colinealidad

12

¿Cómo se resuelve?

- Eliminando variables: aquellas que proporcionan una información que se obtiene de otras variables ya incluidas en el modelo. Pero ojo, eso no significa que no estén asociadas con la VR
- Combinando las VE asociadas en una nueva variable (técnicas multivariadas) o en índices

¿Mejor modelo?

13

```
m1<-lm(VVM~ edad+altura+peso)
m2<-lm(VVM~ edad+peso)
m3<-lm(VVM~ edad+altura)
m4<-lm(VVM~ edad)
m5<-lm(VVM~ altura)
```

	sigma	R2	R2	ajust	df	AIC
m1	10.70	0.867		0.859	5	384.724
m2	14.93	0.736		0.725	4	417.127
m3	10.59	0.867		0.861	4	382.756
m4	27.74	0.068		0.049	3	478.152
m5	18.04	0.606		0.598	3	435.108

Selección de modelos:

- Error estándar residual: por ajuste
- R2 ajustado: por ajuste y parsimonia
- AIC: por ajuste y parsimonia
- ECMP: por predicción
- Pruebas de hipótesis: para VE y modelos

	RMSE	Rsquared	MAE	ER
1	11.026	0.847	9.067	7.57
2	15.347	0.703	12.671	10.54
3	10.813	0.853	8.867	7.42
4	28.301	0.013	23.898	19.43
5	18.416	0.573	15.825	12.64

Pruebas de hipótesis para comparar modelos

14

- Para modelos anidados:

El modelo 2 está anidado en el modelo 1 si todas las VE que se encuentran en el modelo 2 se incluyen en el modelo 1, es decir, el conjunto de VE en el modelo 2 es un subconjunto del conjunto de VE en el modelo 1. Si dos modelos están anidados, el más complejo puede convertirse en el más simple si algunos coeficientes se hacen nulos

$$\text{Modelo 1 } Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_{ijk} \quad a \text{ parámetros}$$

$$\text{Modelo 2 } Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_{ijk} \quad b \text{ parámetros}$$

$b < a$, el modelo 1 (más simple, reducido) está anidado en el modelo 2

El criterio para establecer si una o un conjunto de VE deben ser retenidas en un modelo es determinar la significación de la reducción en la SC residual

$$F = \frac{(SCres_2 - SCres_1) / (GL_2 - GL_1)}{SCres_1 / GL_1}$$

`anova(modelo1, modelo 2)`

o

`drop1(modelo1, test="F")`

¿Mejor modelo?

m1<-lm(VVM~ edad+altura+peso)
m2<-lm(VVM~ edad+peso)
m3<-lm(VVM~ edad+altura)
m4<-lm(VVM~ edad)
m5<-lm(VVM~ altura)

15

Comparando modelos por anova (extra SC)
`anova(m1,m2)`

Analysis of Variance Table

Model 1: VVM ~ edad + altura + peso

Model 2: VVM ~ edad + peso

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	46	5263.9				
2	47	10474.4	-1	-5210.4	45.532	2.179e-08 ***

m2 es significativamente peor que m1

RSS: SCresidual

Diferencia en la cant.de parámetros
y en la SCres

`anova(m1,m5)`

Analysis of Variance Table

Model 1: VVM ~ edad + altura + peso

Model 2: VVM ~ altura

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	46	5263.9			
2	48	15619.9	-2	-10356	45.249

1.366e-11 ***

`anova(m2,m3)`

¿Mejor modelo?

m1<-lm(VVM~ edad+altura+peso)
m2<-lm(VVM~ edad+peso)
m3<-lm(VVM~ edad+altura)
m4<-lm(VVM~ edad)
m5<-lm(VVM~ altura)

16

drop1(m1)

Compara por anova (extra SC) el modelo completo (m1) con modelos anidados, eliminando una variable a la vez. Además informa AIC

single term deletions

Model:

VVM ~ edad + altura + peso							
		Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
m1	<none>			5263.9	240.83		
	edad	1	10263.9	15527.8	292.92	89.6928	2.229e-12 ***
m2	altura	1	5210.4	10474.4	273.23	45.5323	2.179e-08 ***
m3	peso	1	3.3	5267.2	238.86	0.0287	0.8662

m2 significativamente
peor que m1; equivale a
las pruebas t del
summary

Inferencia multimodelo

17

- Burnham et al (2011). AIC model selection and multimodel inference in behavioral ecology. *Behavioral Ecology and Sociobiology*, 65(1), 23-35
- Se estiman todos los modelos posibles (anidados o no anidados)
- Se rankean según la teoría de la información (AIC)
- No utiliza PH
- Los modelos tienen distinto “peso” basado en la evidencia muestral

Todos los modelos posibles

AIC corregido para
muestras pequeñas

18

```
library(MuMIN)
```

```
dredge(lm(VVM~ edad+peso+altura, na.action = "na.fail"))
```

Model selection table

	(Intrc)	altur	edad	peso	df	logLik	AICc	delta	weight
4	-155.30	2.075	-1.0280		4	-187.378	383.6	0.00	0.772
8	-159.90	2.124	-1.0260	-0.04902	5	-187.362	386.1	2.44	0.228
7	58.70		-0.9930	1.74300	4	-204.563	418.0	34.37	0.000
2	-150.70	1.727			3	-214.554	435.6	51.98	0.000
6	-174.70	1.988		-0.25870	4	-214.406	437.7	54.06	0.000
5	30.86			1.42900	3	-220.864	448.3	64.61	0.000
3	172.40		-0.5012		3	-236.076	478.7	95.03	0.000
1	145.70				2	-237.836	479.9	96.28	0.000

Models ranked by AICc(x)

```
dredge(lm(VVM~ edad*peso*altura, na.action = "na.fail"))
```

Model selection table

	(Int)	alt	edd	pes	alt:edd	alt:pes	edd:pes	alt:edd:pes
4	-155.30	2.075	-1.0280					
8	-159.90	2.124	-1.0260	-0.04902				
12	-152.60	2.058	-1.0780		0.0003013			
24	-299.60	2.907	-1.0240	1.96700			-0.01124	
40	-164.70	2.123	-0.9351	0.01390				-0.001166
16	-158.70	2.117	-1.0460	-0.04843	0.0001177			

Tabla 1. Algunas características de la inferencia clásica y multimodelo.

	Inferencia clásica	Inferencia multimodelo
Contraste	Contrasta un estadístico obtenido de los datos muestrales con respecto a una hipótesis “nula” del valor de un parámetro en la población. Informa probabilidad de cometer un error de tipo 1 y de tipo 2.	Contrasta varios modelos anidados y no anidados de manera simultánea. Se intenta evitar el sesgo de “enamorarnos” de una sola hipótesis y ver en los datos información que la sustente, cuando en realidad en muchas situaciones hay más de una hipótesis factible de explicar los patrones observados en los datos.
Valor <i>P</i> y error de tipo 1	Indica la probabilidad de obtener un valor igual o más extremo al estadístico muestral si la hipótesis “nula” es verdadera. El valor <i>P</i> se compara con una probabilidad fijada a priori de cometer un error de tipo 1 (nivel de significancia).	No aplica.
Criterio de información de Akaike (AIC)	No aplica.	Se utiliza para comparar la parsimonia de los distintos modelos planteados. El contraste es relativo; nos indica cuál o cuáles modelos son los que tienen mayor parsimonia. Es deseable, por lo tanto, incorporar en el contraste un modelo nulo (sin predictores).
<i>r</i> ²	Se utiliza como índice de bondad de ajuste en aquellos casos en los que puede ser calculado, como los modelos lineales generales.	Otros índices como el AICc, QAIC, cAIC, BIC y DIC se utilizan con fines similares. Se utiliza como índice de bondad de ajuste en aquellos casos en los que puede ser calculado, como los modelos lineales generales. El <i>r</i> ² “ajustado” (el cual penaliza por complejidad del modelo) puede ser usado con fines similares a los del AIC.
Potencia = 1- <i>P</i> (error de tipo 2)	Se puede calcular dicha probabilidad al fijar una hipótesis alternativa de interés.	No aplica.
Tamaño muestral	Al aumentar el tamaño muestral la potencia aumenta, pero no influye sobre el nivel de significancia (y por lo tanto la probabilidad de cometer un error de tipo 1). Para una diferencia dada, a mayor tamaño muestral más probabilidad de rechazar la hipótesis nula.	Siempre es deseable tener mayor tamaño muestral, pero ello no influye de manera previsible sobre el orden (ranking) de los modelos a ser comparados. Cuando se emplea el AIC, todos los modelos a comparar deben tener el mismo tamaño muestral.

[Garibaldi, L. A. et al. \(2017\). Inferencia multimodelo en ciencias sociales y ambientales. Ecología Austral 27:348-363](#)

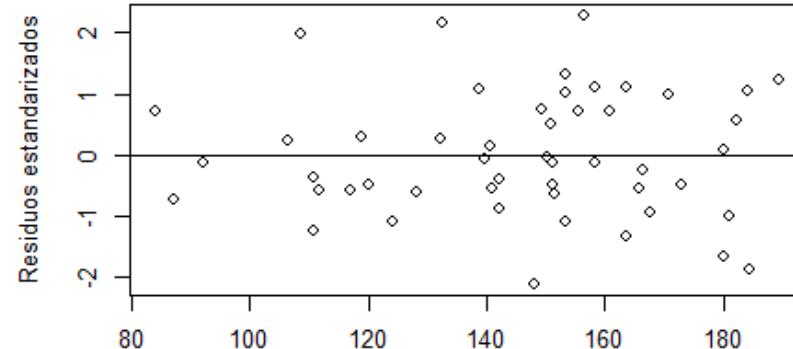
Supuestos y validación m3

vif(m3)
edad altura
1.093778 1.093778

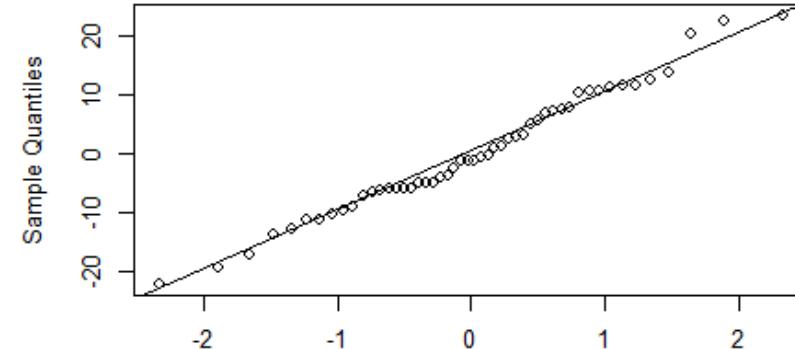
20

m3<-lm(vvm~ edad+altura)

Gráfico de dispersión de RE vs PRED



Normal Q-Q Plot



Cook's distance

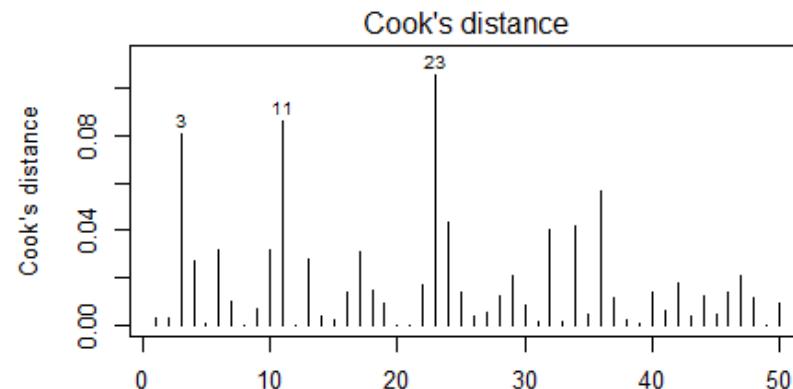
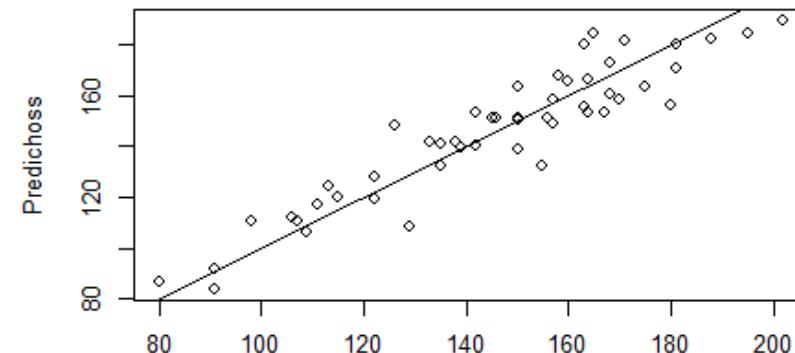


Gráfico de dispersión de PRED vs OBS



Pruebas de hipótesis

21

`m3<-lm(VVM~ edad+altura)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-155.3453	20.3030	-7.651	8.48e-10	***
edad	-1.0276	0.1069	-9.611	1.13e-12	***
altura	2.0746	0.1234	16.813	< 2e-16	***

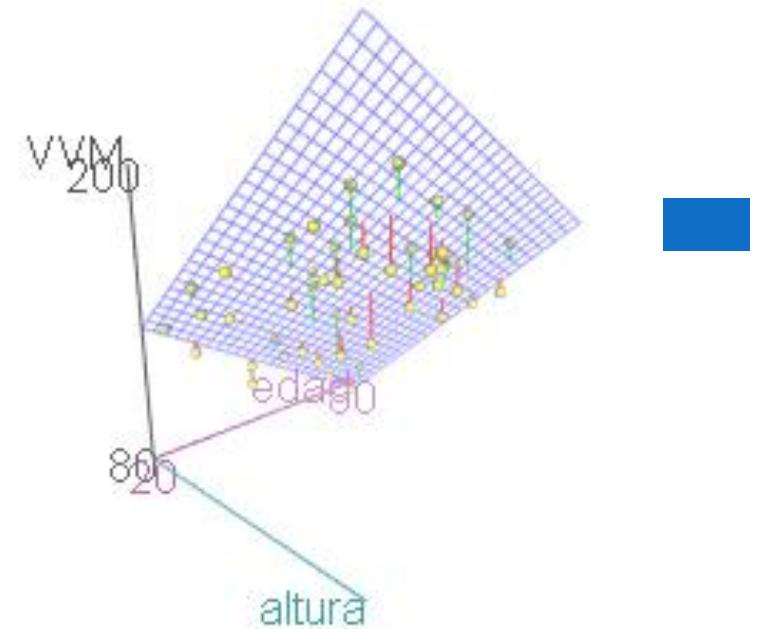
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.59 on 47 degrees of freedom

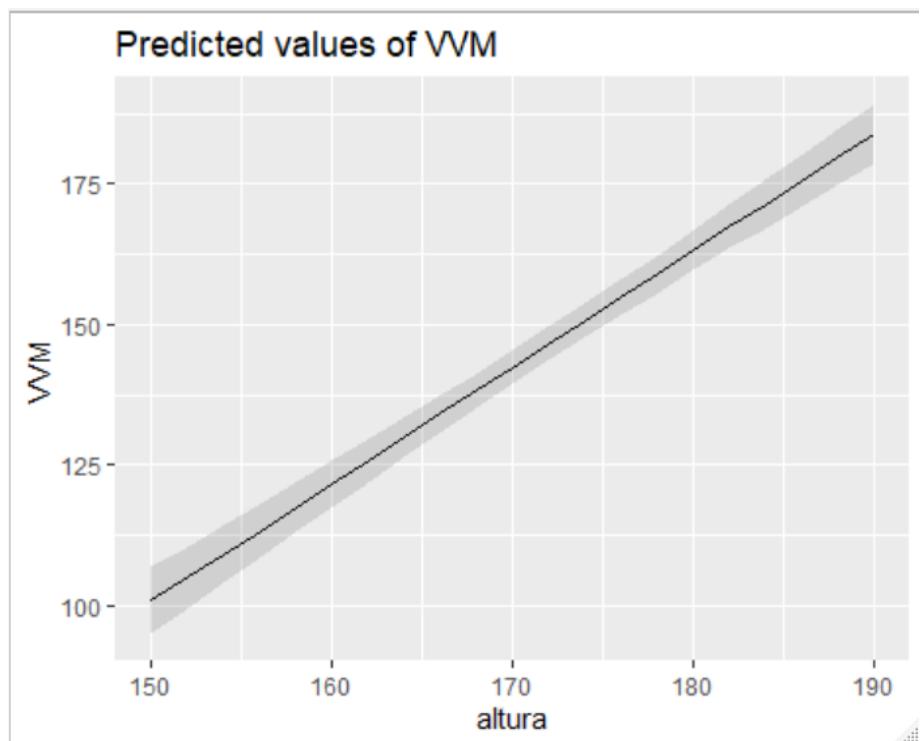
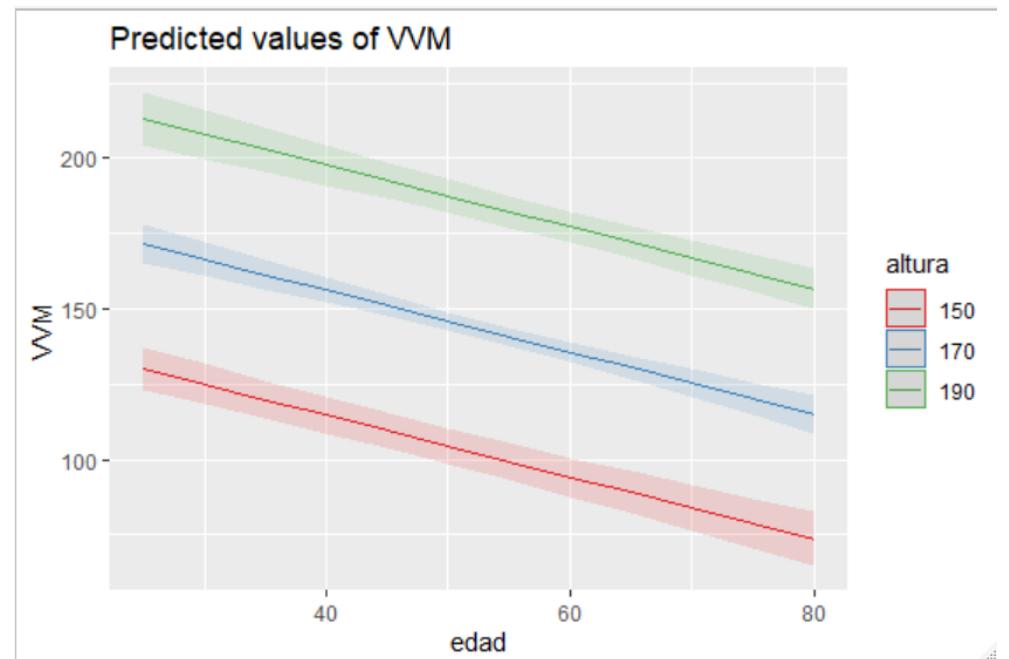
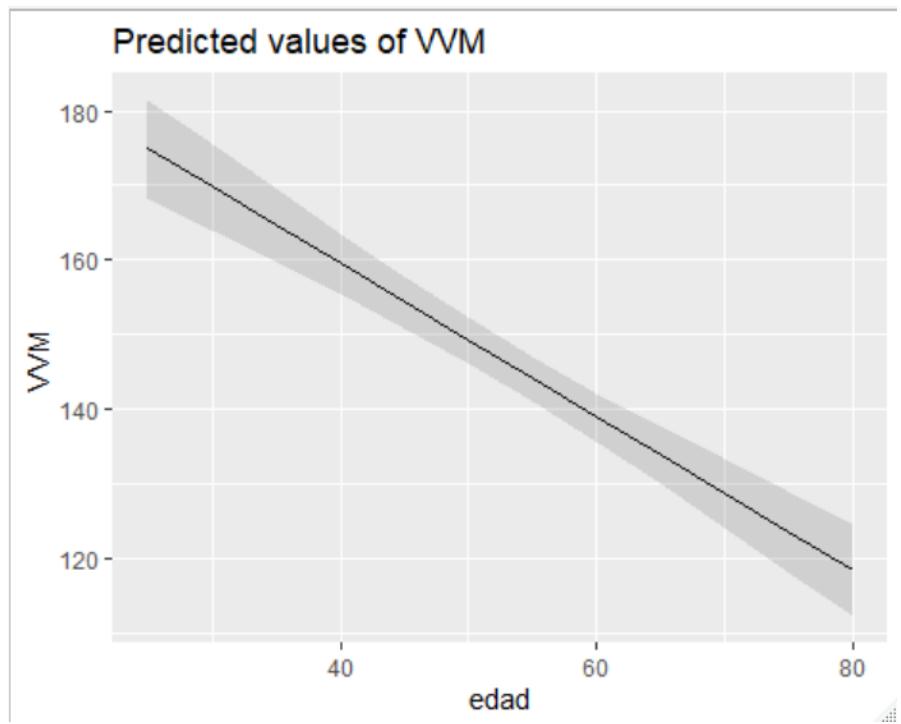
Multiple R-squared: 0.8671, Adjusted R-squared: 0.8615

F-statistic: 153.4 on 2 and 47 DF, p-value: < 2.2e-16

- ✓ ¿Ecuación estimada?
- ✓ ¿Interpretación de los coeficientes?
- ✓ ¿Cómo darle sentido a la ordenada al origen?
- ✓ ¿Porcentaje de la variabilidad en VVM explicada por la edad y la altura?



La representación gráfica de modelos múltiples sólo es posible para 2 VE (plano). Para k VE (siendo $k > 2$) \Rightarrow espacio k -dimensional (hiperplano)

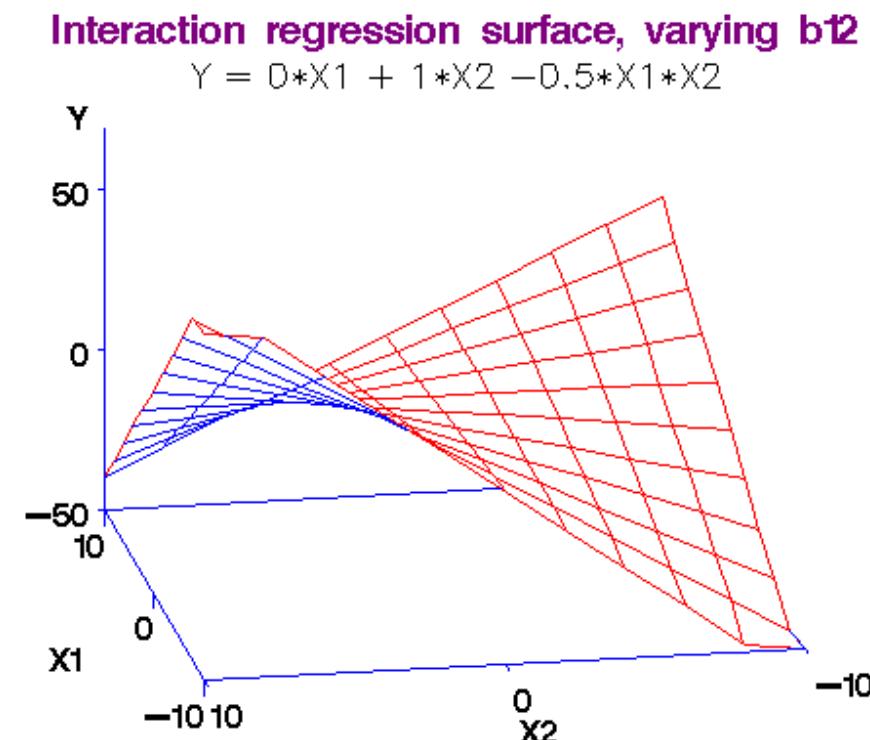
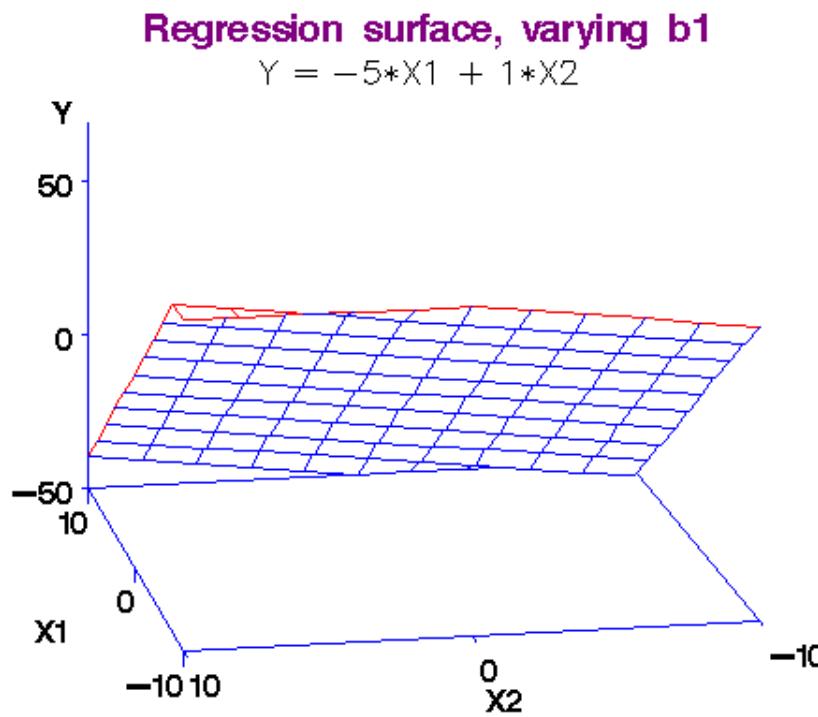


Modelo con interacción

23

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \dots + \varepsilon_i$$

- Con interacción: tanto el efecto de X_1 para un dado nivel de X_2 como el efecto de X_2 para un dado nivel de X_1 dependen del valor de la otra VE



¿Interacción significativa?

24

```
m6<-lm(VVM~ edad*altura)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.526e+02	7.625e+01	-2.001	0.0513 .
edad	-1.078e+00	1.348e+00	-0.800	0.4279
altura	2.058e+00	4.550e-01	4.523	4.27e-05 ***
edad:altura	3.013e-04	7.965e-03	0.038	0.9700

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘			

Residual standard error: 10.7 on 46 degrees of freedom

Multiple R-squared: 0.8671, Adjusted R-squared: 0.8585

F-statistic: 100.1 on 3 and 46 DF, p-value: < 2.2e-16

	sigma	R2	R2	ajust	df	AIC
m1	10.70	0.867		0.859	5	384.724
m2	14.93	0.736		0.725	4	417.127
m3	10.59	0.867		0.861	4	382.756
m4	27.74	0.068		0.049	3	478.152
m5	18.04	0.606		0.598	3	435.108
m6	10.70	0.867		0.867	5	384.754

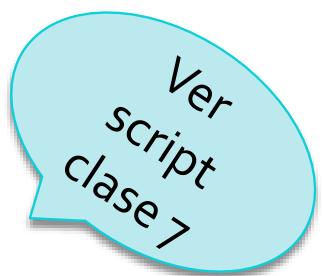
```
> anova(m4,m7)
Analysis of Variance Table

Model 1: VVM ~ edad + altura
Model 2: VVM ~ edad * altura
Res.Df   RSS Df Sum of Sq    F Pr(>F)
1       47 5267.2
2       46 5267.1  1   0.16388 0.0014   0.97
```

Interacción entre variables cuantitativas

25

- La inclusión de un término de interacción $X_1 \times X_2$ provoca colinealidad entre las variables involucradas y $X_1 \times X_2$, dando lugar a valores altos de VIF. Las estimaciones de los coeficientes son insesgadas pero los EE de X_1 y de X_2 (pero no de $X_1 \times X_2$) están inflados. Puede solucionarse centrando las variables antes de generar la interacción
- La inclusión de potencias (X^2 , X^3 , etc) genera el mismo efecto



Interpretación de la interacción

26

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

Se puede reescribir como:

$$Y_i = \beta_0 + \beta_2 X_{2i} + (\beta_1 + \beta_3 X_{2i}) X_{1i} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + (\beta_2 + \beta_3 X_{1i}) X_{2i} + \varepsilon_i$$

El coeficiente de X_1 cambia según el valor de X_2

El cambio en la respuesta media por el incremento de una unidad en X_1 cuando X_2 se mantiene constante es $\beta_1 + \beta_3 X_2$

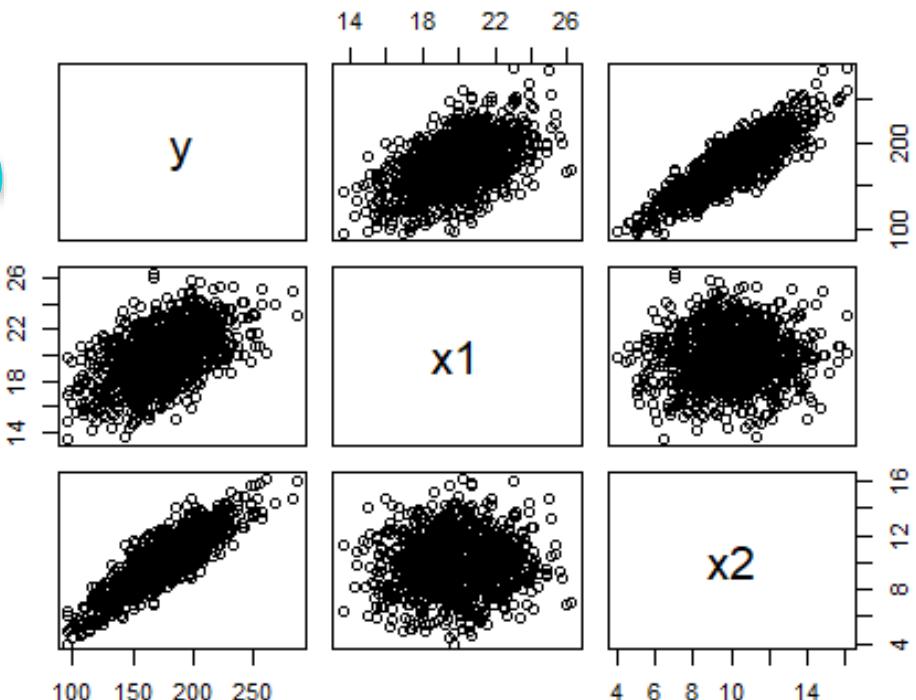
Se pueden elegir valores de X_2 (por ej: $\bar{X}, \bar{X} - S, \bar{X} + S$ y calcular el coeficiente para X_1

Similarmente, el cambio en la respuesta media con un incremento de una unidad en X_2 cuando X_1 se mantiene constante es $\beta_2 + \beta_3 X_1$ y se pueden elegir valores de X_1 y estimar el coeficiente de X_2

Simulamos modelo con interacción

```
x1 = rnorm(1000,20,2)
x2 = rnorm(1000,10,2)
beta0 <- 5
beta1 <- 2
beta2<-3
beta3<-0.5
e = rnorm(1000,mean=0,sd=2)
y=
beta0+beta1*x1+beta2*x2+beta3*x1*x2+e
bd1<-cbind.data.frame(y,x1,x2)
```

Ver
script



Ajustamos un modelo aditivo

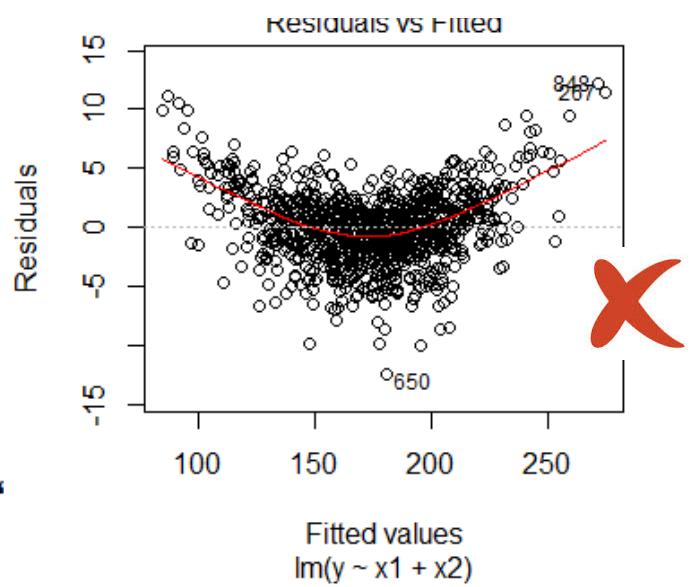
```
m1= lm(y ~ x1+x2, data=bd1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-92.6117	0.9925	-93.3	<2e-16	***
x1	6.9572	0.0448	155.4	<2e-16	***
x2	12.8564	0.0463	278.0	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 3.02 on 997 degrees of freedom
Multiple R-squared: 0.991, Adjusted R-squared: 0.991
F-statistic: 5.23e+04 on 2 and 997 DF, p-value: <2e-16



vif(m1)
x1 x2 1 1

Ajustamos un modelo multiplicativo

`m1= lm(y ~ x1*x2, data=bd1)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.630	3.003	2.21	0.027	*
x1	1.925	0.151	12.71	<2e-16	***
x2	2.780	0.299	9.31	<2e-16	***
x1:x2	0.510	0.015	33.92	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

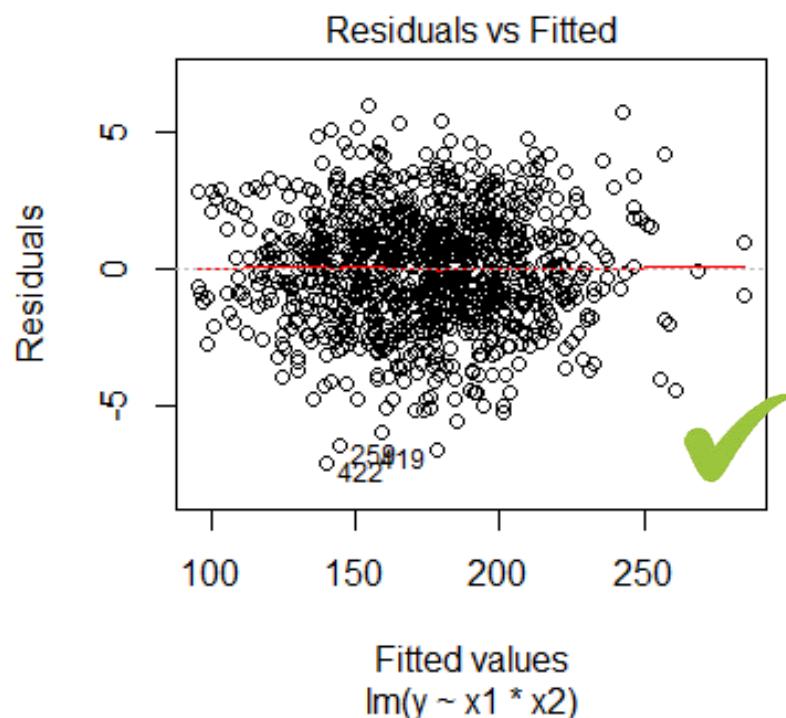
Residual standard error: 2.06 on 996 degrees of freedom

Multiple R-squared: 0.996, Adjusted R-squared: 0.996

F-statistic: 7.55e+04 on 3 and 996 DF, p-value: <2e-16

`vif(m2)`

x1	x2	x1:x2
24.7	89.9	116.9

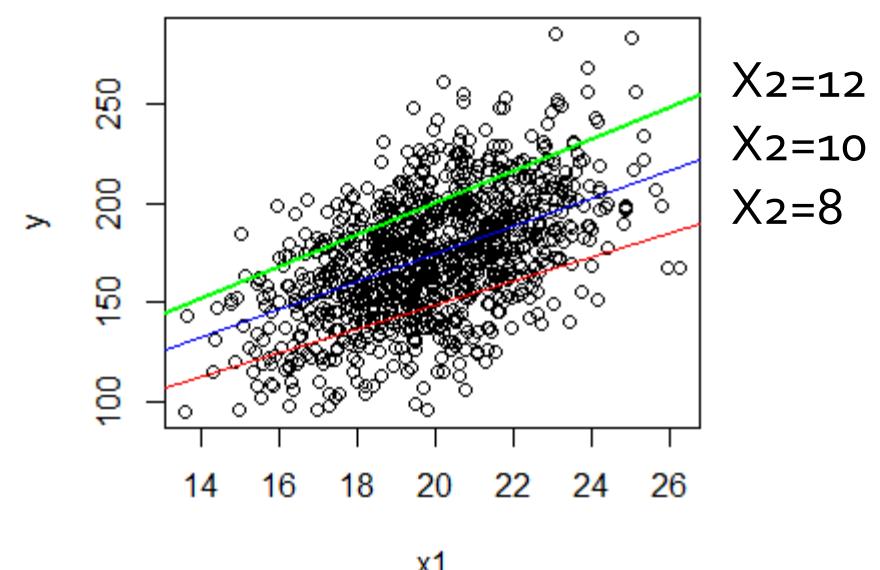


$$Y_i = \beta_0 + \beta_2 X_{2i} + (\beta_1 + \beta_3 X_{2i}) X_{1i} + \varepsilon_i$$

$$\hat{Y}_i = 6,63 + 2,78 X_{2i} + (1,93 + 0,51 X_{2i}) X_{1i}$$

X2	Y
$\bar{X} - S = 10 - 2 = 8$	$28,87 + 6X_1$
$\bar{X} = 10$	$34,4 + 7,03X_1$
$\bar{X} + S = 10 + 2 = 12$	$40 + 8,05X_1$

Idem para X1



Estrategia de análisis

29

1. Estadística descriptiva
2. Estudiar supuestos: linealidad, igualdad de varianzas, normalidad, outliers, observaciones influyentes, colinealidad
3. Estimación y selección de modelos; eventualmente incluir interacciones (siempre en los experimentos diseñados / según las hipótesis en estudios observacionales)
4. Estudiar el desempeño del modelo final: observados vs predichos; R^2 ; validación cruzada
5. Magnitud del efecto: a través de los coeficientes de regresión. Algunos sugieren utilizar los coeficientes estandarizados (llamados beta), sin unidades:

$$\text{beta}_i = b_i \frac{S_x}{S_y} = r$$

Construcción de modelos

30

- Asegurarse de incluir a todas las VE relevantes, basándose en la pregunta de investigación, teoría y conocimiento de la temática
- Ojo con la colinealidad si el objetivo es explicativo. Se pueden combinar las VE que tienden a medir la misma dimensión del fenómeno (por ejemplo mediante un índice o técnicas multivariadas) o seleccionar la más relevante del conjunto
- Considerar la posibilidad de incluir interacciones (principalmente entre variables con mayores efectos)
- Estrategias para retener o eliminar VEs:
 - VE NS pero con el signo esperado: mantener
 - VE NS y sin el signo esperado: eliminar
 - VE S y con el signo esperado: mantener
 - VE S y sin el signo esperado: revisar

Gelman, Andrew, Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2007

Recomendaciones

31

- La inclusión de una interacción o un término cuadrático suele inducir colinealidad, que solo afecta los EE de los términos de menor orden
- La no inclusión de un término de interacción relevante puede dar residuos con patrón
- Centrar las X si se desea ganar interpretación en la ordenada al origen
- Se puede ajustar el modelo máximo, registrar el porcentaje de variabilidad explicado e ir descartando términos , o al revés
- También pueden estimarse todos los modelos y rankearlos por algún criterio, por ej AIC (MuMin)
- Considerar el objetivo: ¿predicción o explicación? Ver Shmueli, G. (2010). To explain or to predict?. Statistical science, 25(3), 289-310.
- Evitar ajustar modelos complejos con pocos datos. Algunos sugieren 10 observaciones por cada VE
- En la selección de modelos se debe respetar el principio de marginalidad (ver diapo siguiente)

Principio de marginalidad

32

Implica que los términos de menor orden no deberían ser removidos antes que los de mayor orden.

- Polinomios: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$ 
 $E(Y) = \beta_0 + \beta_2 X_1^2$  aunque β_1 sea NS

- Modelos con interacciones: si el modelo incluye la interacción de $X_1 * X_2$, el efecto principal de cada variable debe ser incluido en el modelo

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$E(Y) = \beta_0 + \beta_3 X_1 X_2$$

$$E(Y) = \beta_0 + \beta_1 X_1$$


Aunque si la interacción es significativa no tiene sentido evaluar los coeficientes ni la significación de los efectos principales de las VE involucradas, se sobreentiende que ambas variables afectan a la VR.

BIOMETRÍA II

CLASE 10

DISEÑOS ANIDADOS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Efecto del pastoreo sobre el banco de semillas de un pastizal



2

- Se seleccionaron diez potreros de 1ha cada uno. Cada potrero fue asignado al azar a uno de dos tratamientos: régimen de pastoreo por ganado bovino o régimen de clausura de ganado durante 5 años, en un diseño balanceado
- En cada parcela se eligieron 10 puntos al azar y en cada uno se extrajo con un barreno de 5 cm de diámetro por 5 cm de altura una muestra de suelo y se determinó en cada muestra la biomasa de semillas (en gramos/m²)

Experimento o estudio observacional?

VR:

Tipo? Potencial distribución de probabilidades?

VE:

Tipo? De efectos fijos o aleatorios?

Datos (gramos semillas/m²)

3

	tratamiento	potrero	biomasa
1	pastoreo	1	8.2
2	pastoreo	1	8.8
3	pastoreo	1	9.5
4	pastoreo	1	12.7
5	pastoreo	1	15.2
6	pastoreo	1	13.0
7	pastoreo	1	8.5
8	pastoreo	1	6.7
9	pastoreo	1	9.9
10	pastoreo	1	8.5
11	pastoreo	2	4.9
12	pastoreo	2	10.5
13	pastoreo	2	7.5
14	pastoreo	2	9.0
15	pastoreo	2	6.4
16	pastoreo	2	8.5
17	pastoreo	2	4.9
18	pastoreo	2	3.7
19	pastoreo	2	6.6
20	pastoreo	2	7.5

Showing 1 to 20 of 100 entries

	tratamiento	potrero	biomasa
oz	clausura	9	9.5
83	clausura	9	7.9
84	clausura	9	10.4
85	clausura	9	6.2
86	clausura	9	5.3
87	clausura	9	6.3
88	clausura	9	4.7
89	clausura	9	8.4
90	clausura	9	6.6
91	clausura	10	8.6
92	clausura	10	12.2
93	clausura	10	9.3
94	clausura	10	8.6
95	clausura	10	8.1
96	clausura	10	7.2
97	clausura	10	10.3
98	clausura	10	8.0
99	clausura	10	7.7
100	clausura	10	7.8

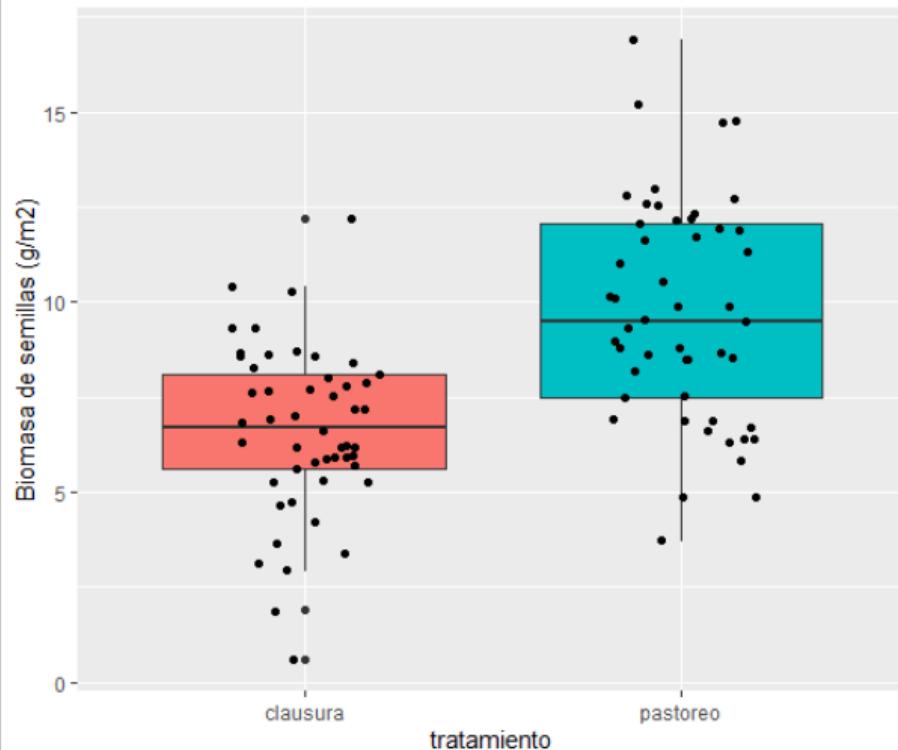
Showing 81 to 100 of 100 entries

semillas.csv

Opción 1

Ignorando los potreros

4



tratamiento	n	media	DE	EE
<fct>	<int>	<dbl>	<dbl>	<dbl>
clausura	50	6.64	2.21	0.313
pastoreo	50	9.74	2.93	0.414

```
m1<-lm(biomasa~tratamiento, semillas)
anova(m1)
```

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tratamiento	1	240.3	240.25	35.69	3.73e-08 ***
Residuals	98	659.7	6.73		

Seudoreplicación!
Los EE están subestimados, p menores a lo correcto, mayor probabilidad de error tipo I



Opción 2

Promediando la VR por potrero

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$i=1,2$$

$$j=1 \text{ a } 5$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

5



interaction(trata)	tratamiento	potrero	biomasa
1	pastoreo	1	10.10
2	pastoreo	2	6.95
3	pastoreo	3	8.54
4	pastoreo	4	10.38
5	pastoreo	5	12.75
6	clausura	6	5.78
7	clausura	7	5.47
8	clausura	8	5.81
9	clausura	9	7.38
10	clausura	10	8.78

```
m2<-lm(biomasa~tratamiento, medias.potrero)
anova(m2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tratamiento	1	24.02	24.025	7.186	0.0279 *
Residuals	8	26.75	3.343		

tratamiento	n	media	DE	EE
<fct>	<int>	<dbl>	<dbl>	<dbl>
clausura	5	6.64	1.41	0.629
pastoreo	5	9.74	2.17	0.970



Pero se pierde información

Opción 3. Modelo condicional (mixto)

Incorporando la variable potrero al modelo

6

$$Y_{ijk} = \mu + \alpha_i + B_j + \varepsilon_{ijk}$$

$$i=1,2$$

$$j=1 \text{ a } 5$$

$$K=1 \text{ a } 10$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$$B_j \sim N(0, \sigma_{potreros}^2)$$

$$\varepsilon_{ij}, B_j \text{ indep}$$

las observaciones son condicionalmente independientes, pero marginalmente estarán correlacionadas debido al efecto aleatorio.

```
library(lme4)
m3<- lmer(biomasa ~ tratamiento + (1|potrero), semillas)
```

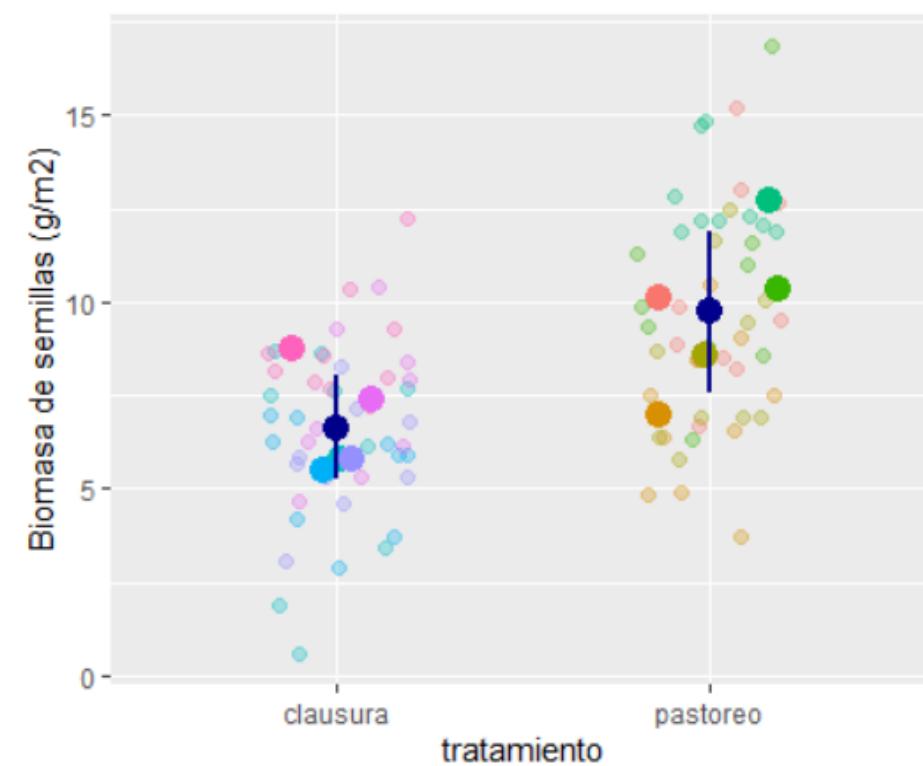
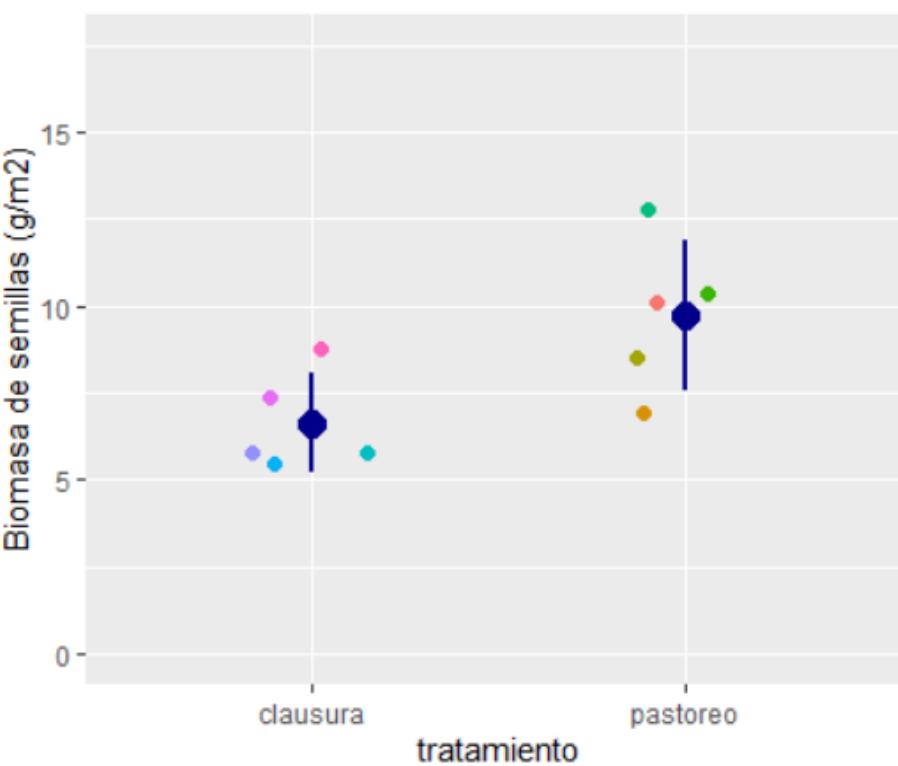
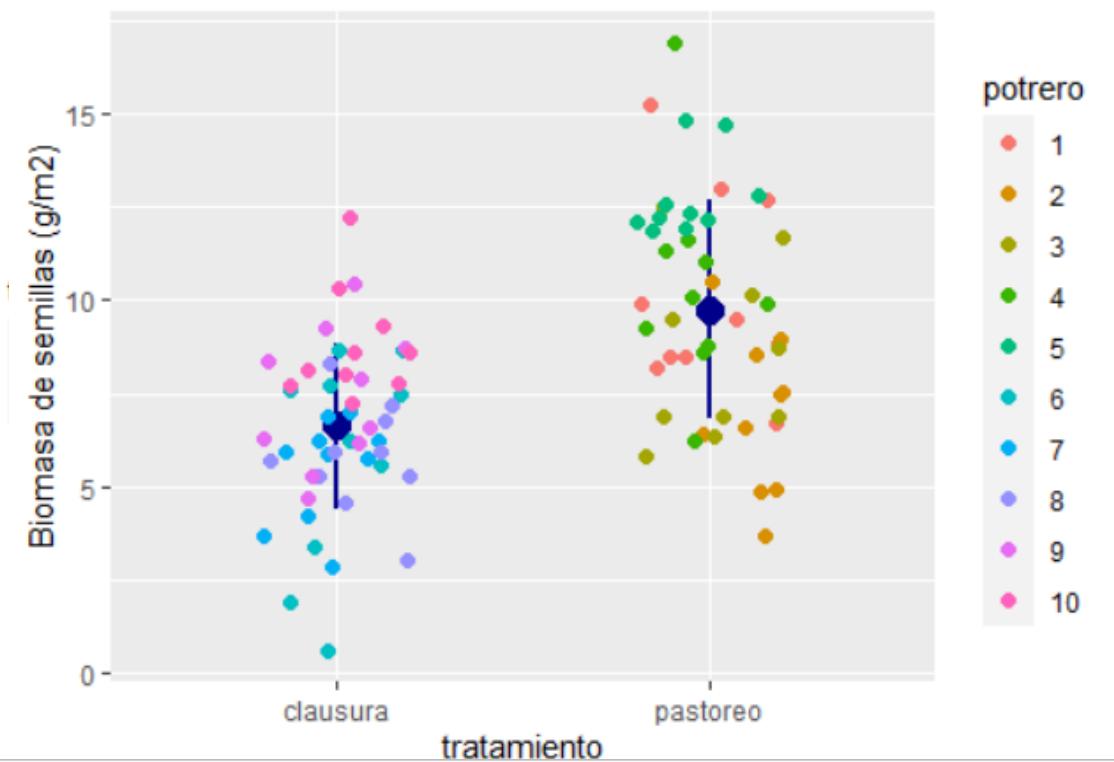
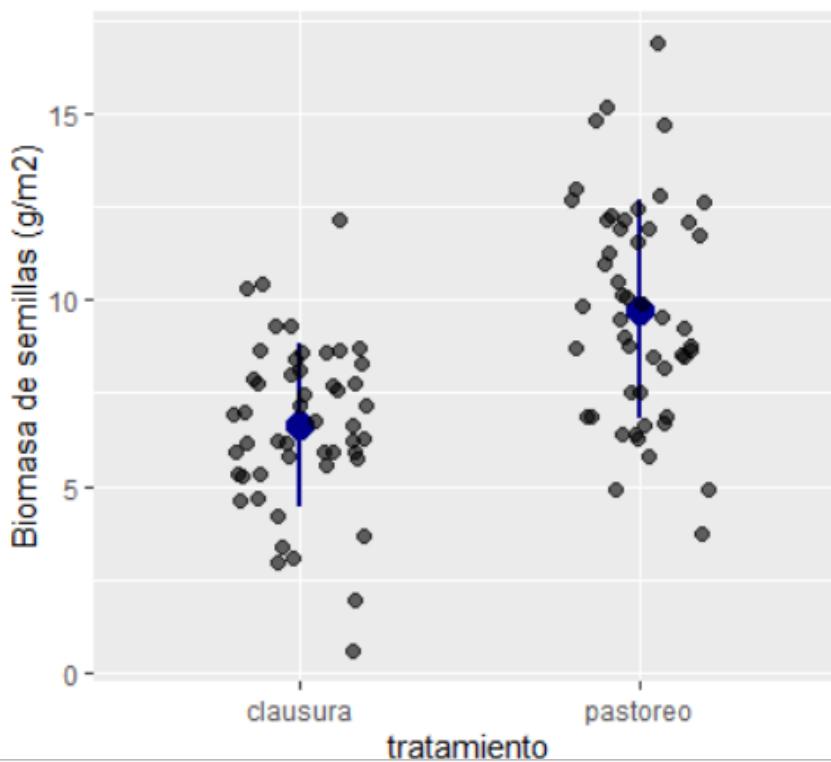
```
> anova(m3)
```

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
tratamiento	31.314	31.314	1	8	7.1856	0.0279 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Parámetros? Efectos aleatorios?



Opción 3. Modelo condicional (mixto)

Incorporando la variable potrero al modelo

8

```
library(lme4)
m3<- lmer(biomasa ~ tratamiento + (1|potrero), semillas)
Summary(m3)

Linear mixed model fit by REML t-tests use Satterthwaite approximations to
degrees of freedom [lmerMod]
Formula: biomasa ~ tratamiento + (1 | potrero)
Data: semillas

REML criterion at convergence: 446.5

scaled residuals:
    Min     1Q   Median     3Q     Max 
-2.5353 -0.6256 -0.0507  0.6122  3.1630 

Random effects:
Groups   Name        Variance Std.Dev.
potrero  (Intercept) 2.908    1.705  
Residual           4.358    2.088  
Number of obs: 100, groups: potrero, 10

Fixed effects:
              Estimate Std. Error    df t value Pr(>|t|)    
(Intercept)  6.6440    0.8177 8.0000  8.125 3.91e-05 ***
tratamientopastoreo 3.1000    1.1565 8.0000  2.681  0.0279 *  
---

```

Mismos resultados
usando la función
lme del paquete
nlme

Opción 4. Modelo marginal

9

```
m4 <- gls(biomasa ~ tratamiento, correlation=corCompSymm(form = ~  
1 | potrero), data = semillas)
```

```
Generalized least squares fit by REML  
Model: biomasa ~ tratamiento  
Data: semillas  
AIC      BIC      logLik  
454.4912 464.8311 -223.2456
```

Correlation Structure: Compound symmetry

```
Formula: ~1 | potrero  
Parameter estimate(s):  
Rho  
0.4002022
```

Coefficients:

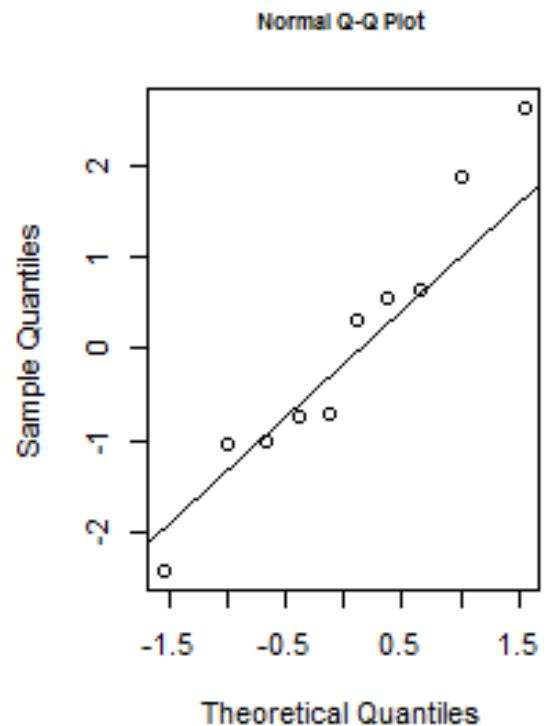
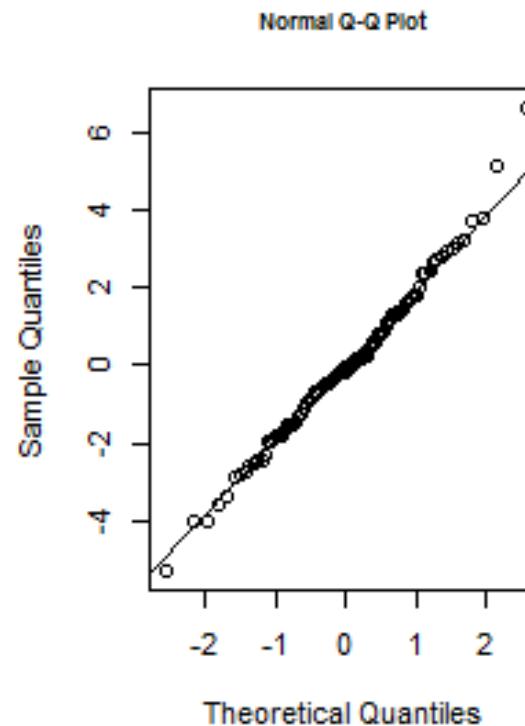
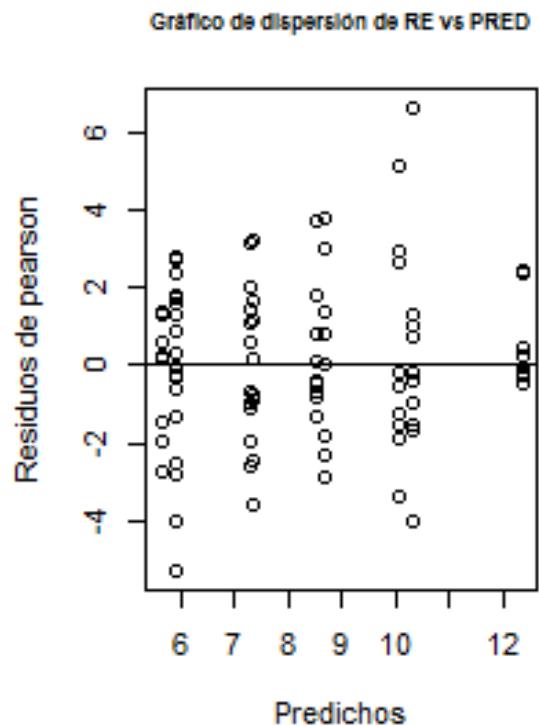
	value	Std. Error	t-value	p-value
(Intercept)	6.644	0.8177383	8.124848	0.0000
tratamientopastoreo	3.100	1.1564566	2.680602	0.0086

Mismos resultados
para la parte fija
que lmer y lme

Parámetros? Efectos ~~X~~ aleatorios?

Supuestos

10



Shapiro-wilk normality test data:
e w = 0.98998, p-value = 0.6632

Shapiro-wilk normality test data:
alfa1 w = 0.9601, p-value = 0.787

Parte fija Significación?

Problemas:

- No hay consenso sobre los GL
- Las distribuciones de los estadísticos son asintóticas

11

□ Prueba de Wald

Fixed effects:

	Estimate	std. Error	df	t value	Pr(> t)
(Intercept)	6.6440	0.8177	8.0000	8.125	3.91e-05 ***
tratamientopastoreo	3.1000	1.1565	8.0000	2.681	0.0279 *

Comparar con
opción 2

□ Prueba de cociente de verosimilitud (Likelihood ratio test LRT):

Permite comparar modelos anidados con distinta estructura fija. Ojo, la estimación debe ser por máxima verosimilitud (no por REML). Se puede usar drop1 o anova

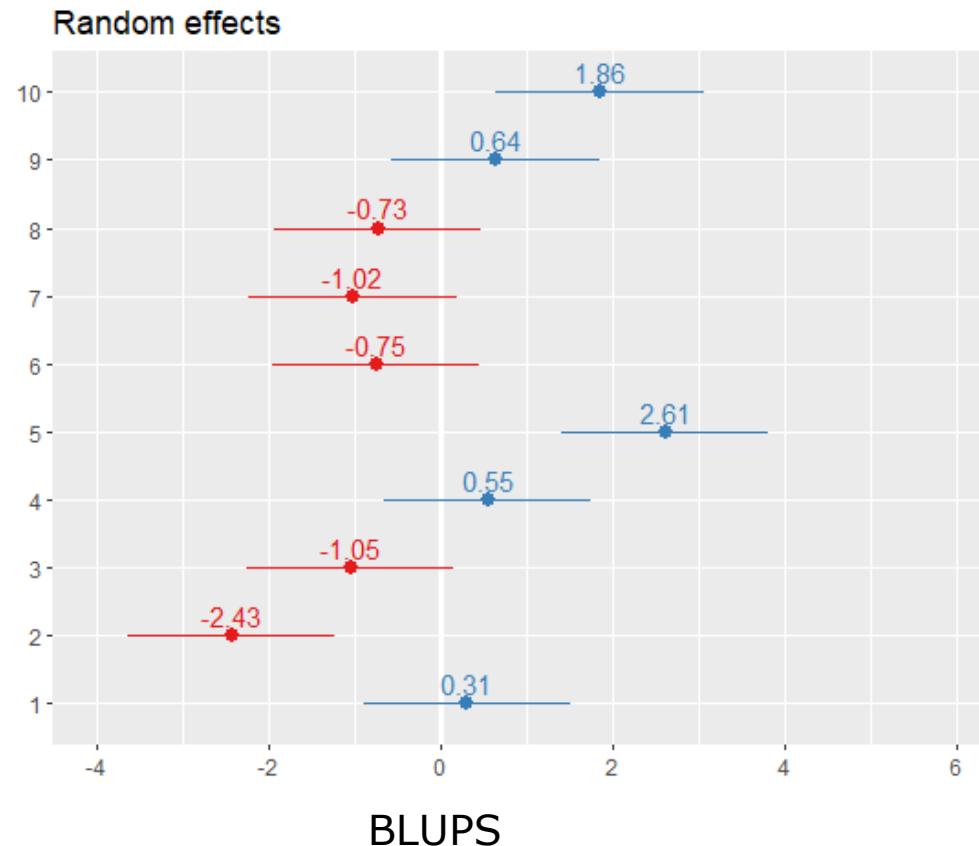
```
> m0<- lmer(biomasa ~ (1|potrero), semillas)
> anova(m0,m3)
refitting model(s) with ML (instead of REML)
Data: semillas
Models:
m0: biomasa ~ (1 | potrero)
m3: biomasa ~ tratamiento + (1 | potrero)
      Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>chisq)
m0   3 461.54 469.36 -227.77    455.54
m3   4 457.13 467.55 -224.56    449.13 6.4091     1 0.01135 ,
```

Parte aleatoria

Efectos aleatorios

$$BLUP_j = BLUE_j \left(\frac{\sigma_{potreros}^2}{\sigma_{potreros}^2 + \sigma^2 / n_i} \right)$$

```
$potrero
  (Intercept)
1  0.3095992
2 -2.4298321
3 -1.0470715
4  0.5531042
5  2.6142002
6 -0.7513869
7 -1.0209817
8 -0.7252971
9  0.6400703
10 1.8575954
```



Parte aleatoria

Componentes de varianza

Random effects:

Groups	Name	Variance	Std.Dev.
potrero	(Intercept)	2.908	1.705
Residual		4.358	2.088
Number of obs: 100, groups: potrero, 10			

$$\hat{\sigma}_{Y_{ij}}^2 = 2,908 + 4,358 = 7,266$$

$$\hat{\sigma}_{Y_{ij}} = 2,7 \text{ g / m}^2$$

$$CCI = \frac{\sigma_{potrero}^2}{\sigma_{potrero}^2 + \sigma^2} = 0,4$$

El 40% de la variación en la biomasa de semillas está aportada por la variación entre potreros sometidos a un determinado tratamiento; el 60% restante está aportado por la variación entre muestras dentro de un mismo potrero

El coeficiente de correlación intraclass CCI mide la correlación entre puntos de un mismo lote; cuanto más alta sea indica que las mediciones dentro de un mismo lote son muy similares y por lo tanto la variación viene dada por los potreros

Parte aleatoria

Significación?

14

- Según algunos autores, si el efecto aleatorio está dado por diseño, debería permanecer en el modelo
- Puede usarse la prueba de cociente de verosimilitud, tanto con ML o REML, ya que estamos comparando modelos con la misma parte fija. La prueba es conservativa.

> `ranova(m3)`

ANOVA-like table for random-effects: single term deletions

Model:

```
biomasa ~ tratamiento + (1 | potrero)
          npar   LogLik    AIC     LRT  DF Pr(>chisq)
<none>           4 -223.25 454.49
(1 | potrero)     3 -236.40 478.80 26.311  1 2.906e-07 ***
---
---
```

- Al probar si una o más varianzas son cero estamos en la frontera del espacio de parámetros (ya que el mínimo de una varianza es cero), y por lo tanto la distribución asintótica del estadístico de la prueba es aproximada
- Se pueden construir intervalos de confianza para los componentes de varianza del método usando el método de verosimilitud perfilada (profile maximum likelihood)

`confint(mod.mix4, level = 0.95, method = c("profile"))`

2.5 % 97.5 %

.sig01	0.8935747	2.650294
.sigma	1.8160876	2.434040

Presentación de resultados

15

```
confint(emmeans(m3, pairwise ~ tratamiento))
```

\$emmeans

tratamiento	emmmean	SE	df	lower.CL	upper.CL
clausura	6.64	0.818	8	4.76	8.53
pastoreo	9.74	0.818	8	7.86	11.63

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

\$contrasts

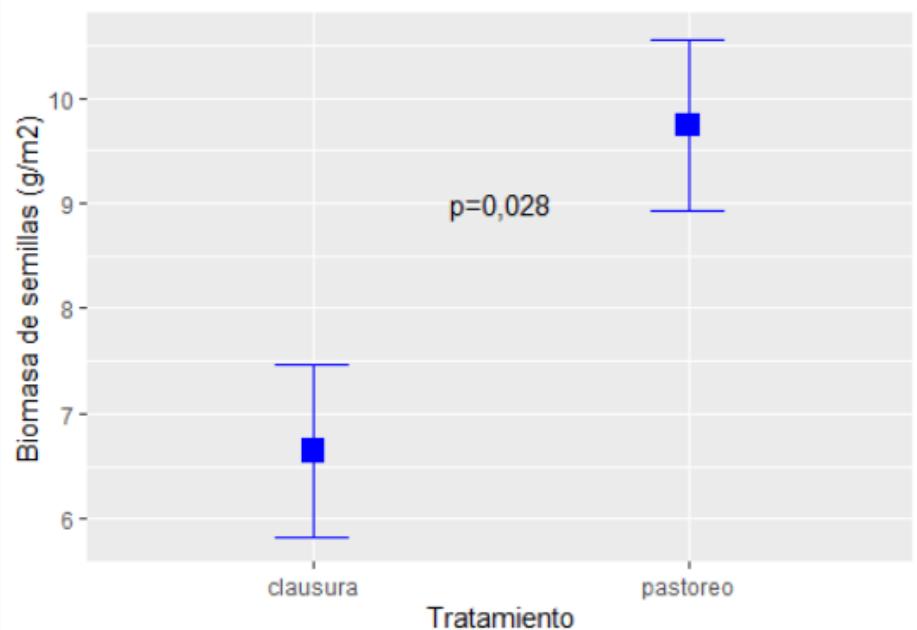
contrast	estimate	SE	df	lower.CL	upper.CL
clausura - pastoreo	-3.1	1.16	8	-5.77	-0.433

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

Comparación de la biomasa de semillas según pastoreo

Media ± error estándar



¿Para qué sirvió anidar?

16

- Obtener submuestras es usualmente menos costoso que incrementar la cantidad de ue
- Pero ojo: el submuestreo no incrementa el número de réplicas. El número de réplicas sigue siendo el número de UE a las que se le aplicó el tratamiento en forma aleatoria, y no el número de observaciones por tratamiento
- Por lo tanto el aumento en la cantidad de submuestras no incrementa directamente la potencia de la prueba (ésta depende de la cantidad de réplicas)
- Sin embargo, si existe mucha variación a pequeña escala (elevado σ^2), si se aumenta la cantidad de submuestras la estimación de la variación entre UE será más precisa, lo que indirectamente aumentará la potencia de la prueba
- Además, si existe desbalanceo en las submuestras, este análisis provee estimaciones que lo toman en cuenta
- Los BLUP, es decir los efectos aleatorios, se encogen (se parecen más a la media general) si:
 - el componente de varianza para el término en cuestión es pequeño
 - la varianza residual es grande
 - el número de repeticiones del nivel de factor considerado es pequeño
- Las submuestras son en muchos casos réplicas técnicas

Estimación por MV restringida vs MV

17

```
m2 <- lmer(biomasa ~ tratamiento  
+ (1|potrero), data = semillas,  
REML=TRUE)
```

Random effects:

Groups	Name	Variance	Std.Dev.
potrero	(Intercept)	2.908	1.705
Residual		4.358	2.088

Number of obs: 100, groups: potrero, 10

Fixed effects:

	Estimate	Std. Error
(Intercept)	6.6440	0.8177
tratamientopastoreo	3.1000	1.1565

```
m2ML <- lmer(biomasa ~ tratamiento  
+ (1 | potrero), data = semillas,  
REML=FALSE)
```

Random effects:

Groups	Name	Variance	Std.Dev.
potrero	(Intercept)	2.239	1.496
Residual		4.358	2.088

Number of obs: 100, groups: potrero, 10

Fixed effects:

	Estimate	Std. Error	1
(Intercept)	6.6440	0.7314	
tratamientopastoreo	3.1000	1.0344	

Es el método por defecto

Si se utiliza estimación por MV:
Las varianzas y EE están subestimados,
pero no los estimadores de los
coeficientes para VE de efectos fijos
Solo es indicada para comparar
modelos

Factores cruzados vs anidados

18

- Dos factores (VE cualitativas) están **cruzados** cuando cada nivel de un factor está observado en todos los niveles del otro (y viceversa). Corresponde a un diseño **factorial**. No hay jerarquía
- El factor B está **anidado** en A cuando cada nivel del factor B está observado en un solo nivel de A (hay jerarquía). Como cada nivel de B no se cruza con cada nivel de A, no es posible que exista interacción entre A y B
- Bloques? Potreros?
- Para que R detecte que los potreros están anidados en los tratamientos y no cruzados, se los debe identificar unívocamente:
 - Potreros 1 a 10 => anidados en tratamiento
 - Potreros 1 a 5 en Control y 1 a 5 en Clausura => cruzados con tratamiento

$$\begin{aligned}Y_{ij} &= \mu + \alpha_i + B_j + \varepsilon_{ij} \\ \varepsilon_{ij} &\approx NID(0, \sigma^2) \\ B_j &\approx NID(0, \sigma^2_{potreros})\end{aligned}$$

Variaciones en rasgos del cedro amargo



19

- Se llevó a cabo un estudio en el NOA a fin de caracterizar la variabilidad fenotípica en el cedro americano (*Cedrela odorata*) , una especie vulnerable.
- Se estudiaron 7 poblaciones elegidas al azar en el área de estudio. De cada población se eligieron entre 12 y 20 familias y de cada familia se estudiaron al menos dos ejemplares.
- Se registró el largo de cada ejemplar

Experimento o estudio observacional?

VR:

Tipo? Potencial distribución de probabilidades?

VE:

Tipo? De efectos fijos o aleatorios?

Agrupamiento?

cedro.csv

Diseño totalmente anidado

20

Totalmente anidado: Factor A (aleatorio), Factor B anidado en A, Factor C anidado en B

```
lmer(largo ~ 1 + (1 | poblacion/familia), BD)
lmer(largo ~ 1 + (1 | poblacion)+(1 | familia), BD)
lme(largo ~ 1, random = ~ 1|poblacion/familia, BD)
```

```
> BD
      poblacion familia largo
1       Charagre   Ch_71    6.0
2       Charagre   Ch_71    6.0
3       Charagre   Ch_710   6.0
4       Charagre   Ch_710   13.0
5       Charagre   Ch_711   14.0
6       Charagre   Ch_711   8.0
7       Charagre   Ch_712  12.5
8       Charagre   Ch_712  10.0
9       Charagre   Ch_713   6.5
10      Charagre   Ch_713   6.0
```

```
> summary(m4)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: largo ~ 1 + (1 | poblacion/familia)
Data: BD
```

REML criterion at convergence: 2008.5

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.24033	-0.42502	-0.05879	0.55051	2.43795

Random effects:

Groups	Name	Variance	Std.Dev
familia:poblacion	(Intercept)	219.0	14.80
poblacion	(Intercept)	737.5	27.16
Residual		463.7	21.53

Number of obs: 214, groups: familia:poblacion, 115; poblacion, 7

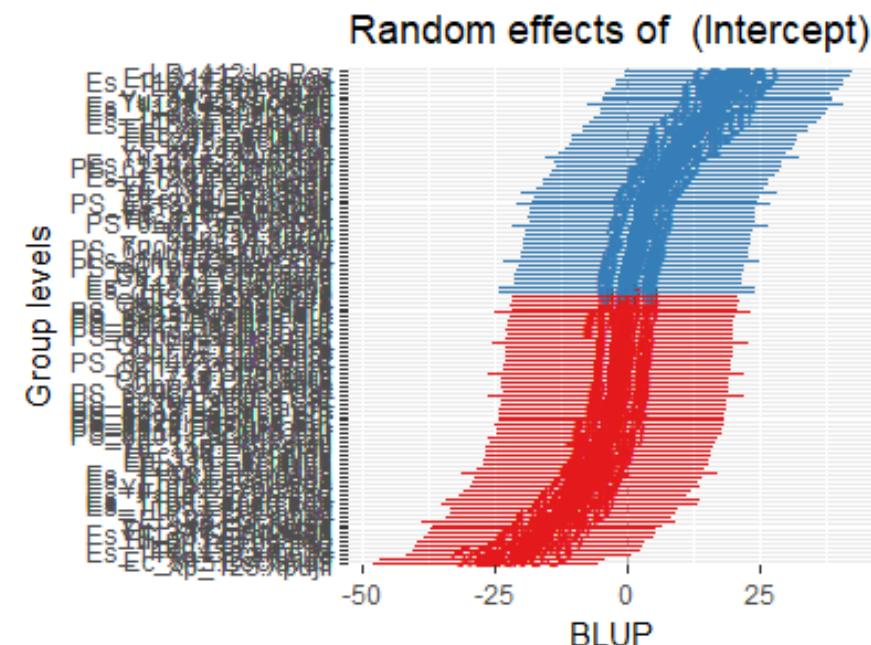
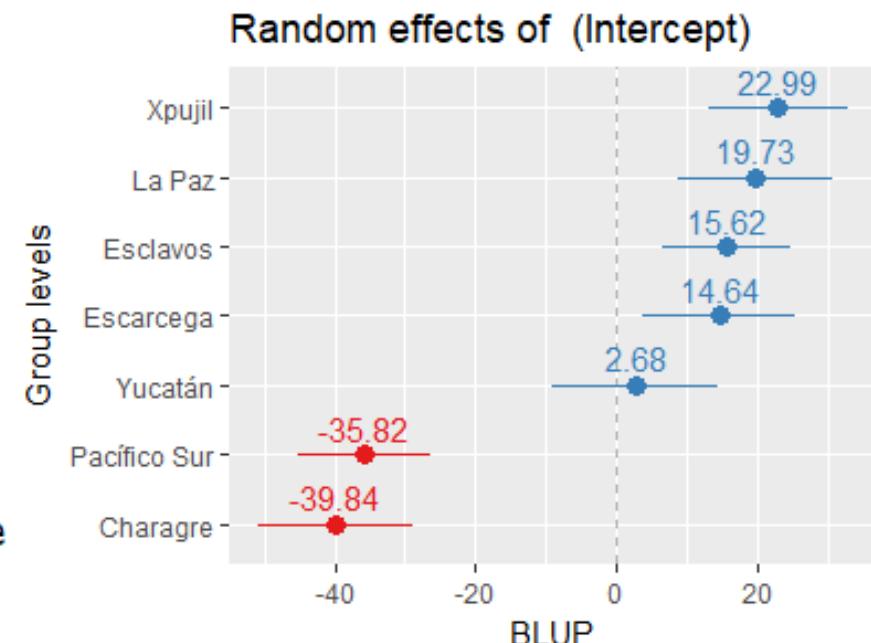
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	49.85	10.47	4.762

¿Qué miden?

¿Cuánto aportan?

¿Cuáles son sus unidades?





Complicando el modelo

22

- ¿Y si de cada ejemplar se eligieron 10 semillas al azar y se registró el peso de cada una?
- ¿Y si de cada ejemplar se registró el pH del suelo e interesa saber si el largo del ejemplar se asocia con el pH?
- ¿Y si se sospecha que el “efecto” del pH sobre el largo del ejemplar cambia entre poblaciones?



Complicando el modelo

23

- ¿Y si de cada ejemplar se eligieron 10 semillas al azar y se registró el peso de cada una?

`lmer(peso ~ 1 + (1 | poblacion/familia/ejemplar))`

`lme(peso ~ 1, random = ~ 1|poblacion/familia/ejemplar)`

- ¿Y si de cada ejemplar se registró el pH del suelo e interesa saber si el largo del ejemplar se asocia con el pH?

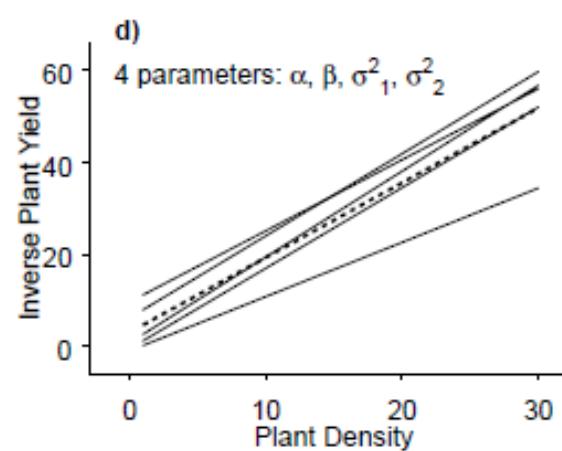
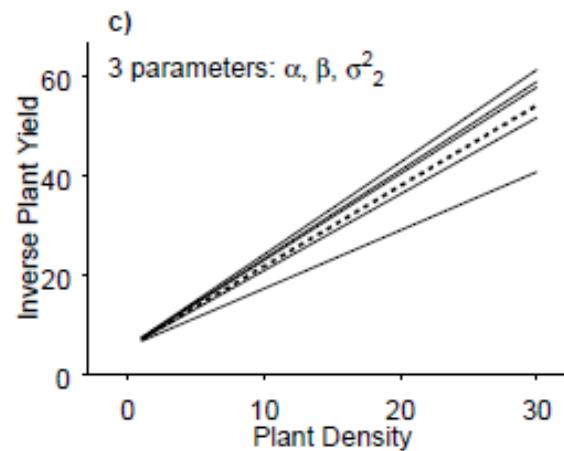
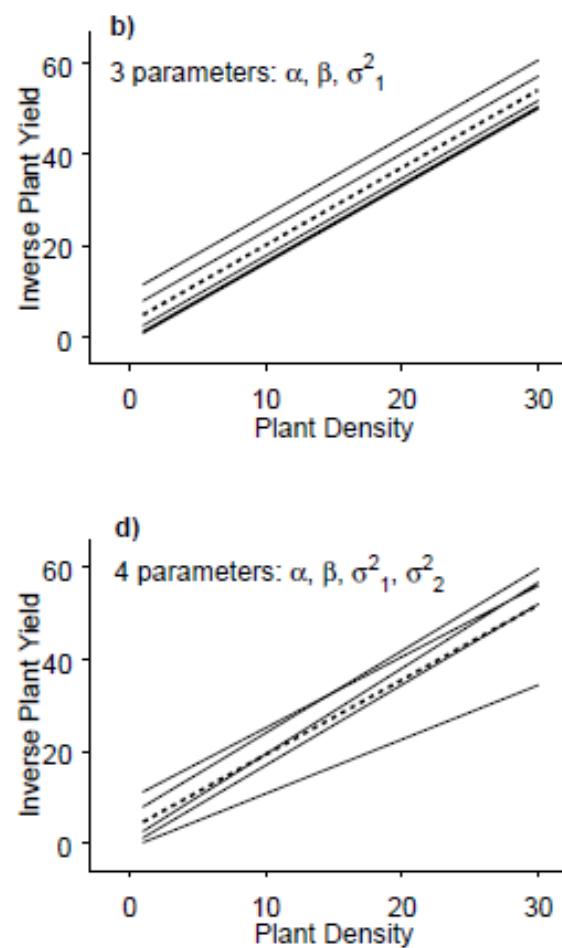
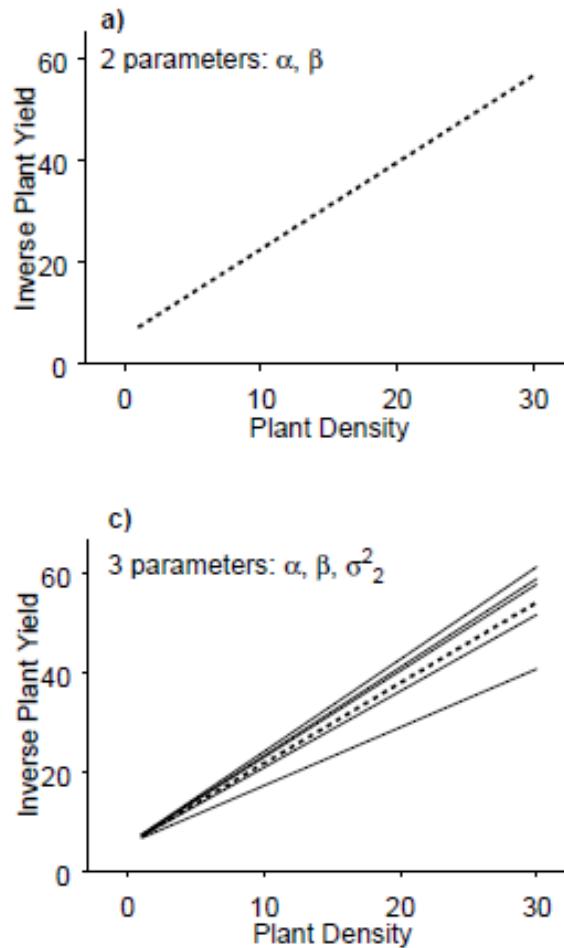
`lmer(largo ~ pH + (1 | poblacion/familia))`

`lme(largo ~ pH , random = ~ 1|poblacion/familia)`

- ¿Y si se sospecha que el “efecto” del pH sobre el largo del ejemplar cambia entre poblaciones?

Modelos lineales mixtos

24



- a) Modelo sin efectos aleatorios
- b) Modelo con intercepto aleatorio
- c) Modelo con pendiente aleatoria
- d) Modelo con intercepto y pendiente aleatoria

```
a <- lm(Y ~ X, data)
b <- lmer (Y ~ X + (1|Factor_aleatorio), data)
c <- lmer (Y ~ X + (0+X|Factor_aleatorio), data)
d <- lmer (Y ~ X + (1+X|Factor_aleatorio), data)
```

Modelos con intercepto y pendiente aleatorios

25

- ¿Y si se sospecha que el “efecto” del pH sobre el largo del ejemplar entre poblaciones?

`lmer(largo ~ pH + (1 + pH | poblacion))`

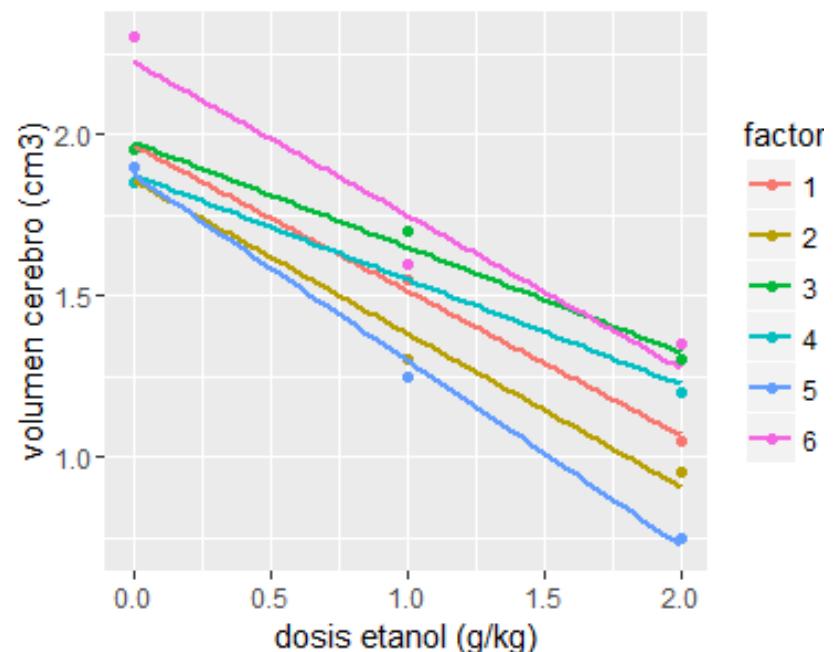
`lme(largo ~ pH , random = ~ 1 + pH | poblacion)`

En el ej de DBA:

Implica interacción
tratamiento x bloque

`lmer(vol ~ etanol + (etanol| camada) , bd)`

Implica una interacción trans-nivel



Modelos con VE de efectos aleatorios cruzados

26

En un ensayo de comparabilidad interlaboratorios, se suministraron 5 muestras de suero de pacientes a 10 laboratorios. Cada laboratorio midió la concentración de un anticuerpo por triplicado

Se desea estudiar la variabilidad intra e interlaboratorios

`lmer(Y ~ X + (1 | A) + (1 | B), data)`

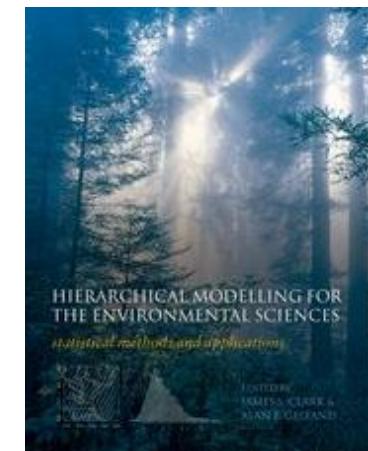
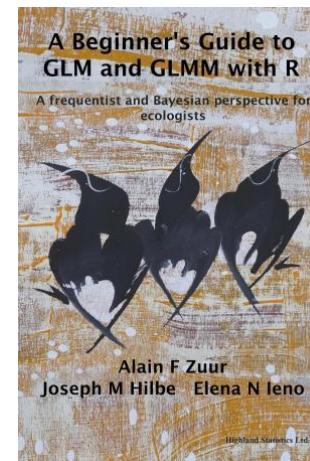
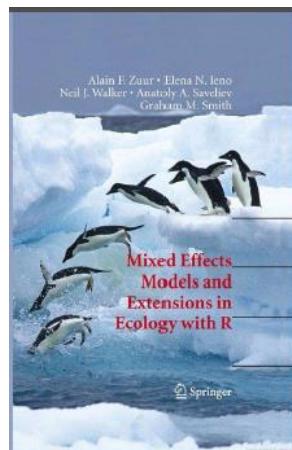
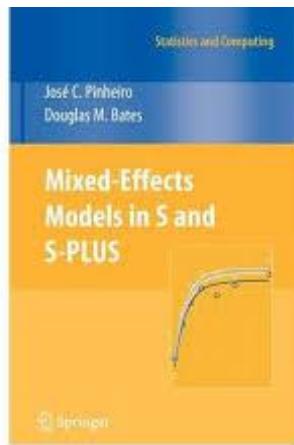
La clave está en el armado de la base de datos

Laboratorio	Paciente	Concentr
L1	1	12
L1	1	19
L1	1	23
L1	5	
L10	5	
L10	5	

Bibliografía

27

- Pinheiro J.C., Bates D.M. 2004. Mixed-Effects Models in S and S-PLUS. Springer, New York
- Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, New York
- Zuur AF, Hilbe JM and Ieno EN. 2013. Beginner's Guide to GLM and GLMM with R . Highland Statistics Ltd
- Clark JS. 2006. Hierarchical modelling for the environmental sciences. Oxford University Press



BIOMETRÍA II

CLASE 11

DISEÑO DE MEDIDAS REPETIDAS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Efecto del tratamiento con broncodilatadores para el tratamiento del asma

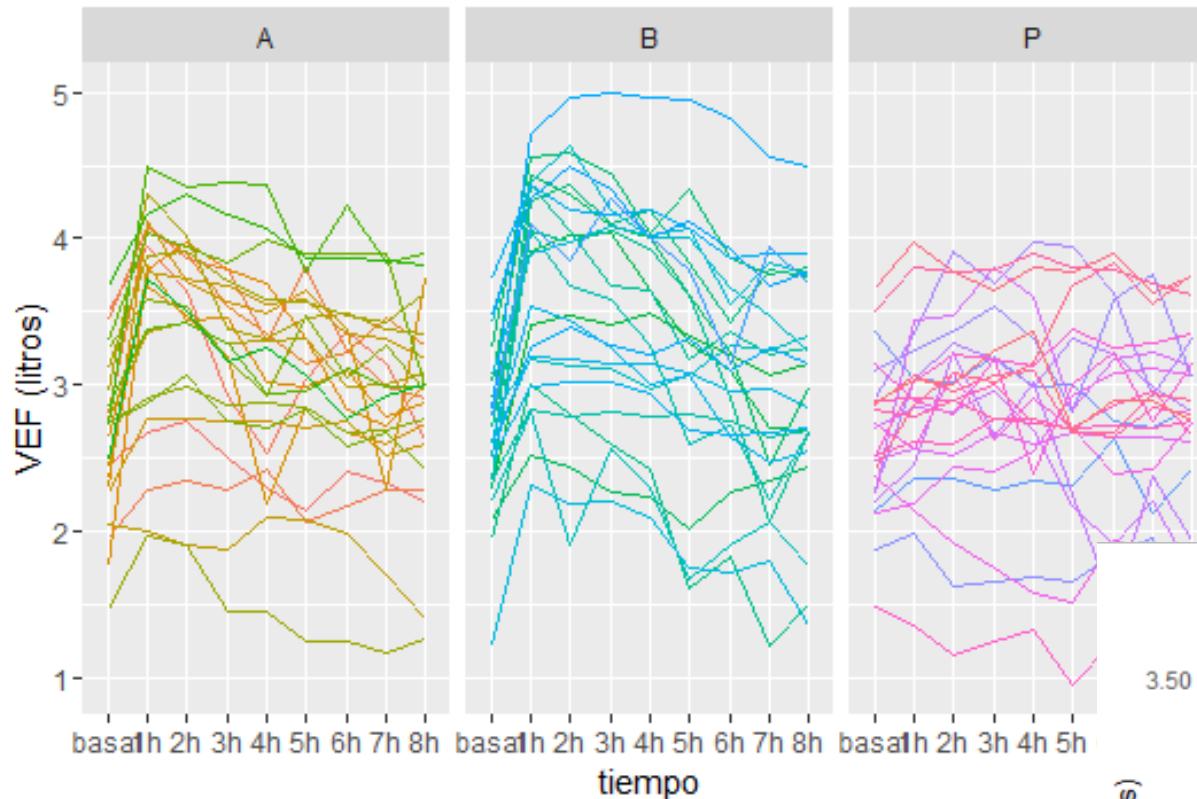
2

- 72 pacientes asmáticos fueron asignados en forma aleatoria y balanceada a uno de los siguientes tratamientos: droga A, droga B o placebo (P)
- Se midió el volumen espiratorio forzado (VEF) antes de la administración del tratamiento (basal) y a intervalos de 1 hora luego de esta, hasta las 8 hs
- UE para el tratamiento
- VR
- Tipo, potencial distribución de probabilidades
- VE
- Tipo, de efectos fijos o aleatorios

paciente	droga	tiempo	vef
1	A	basal	2.46
1	A	1h	2.68
1	A	2h	2.76
1	A	3h	2.50
1	A	4h	2.30
1	A	5h	2.14
1	A	6h	2.40
1	A	7h	2.33
1	A	8h	2.20
2	A	basal	3.50
2	A	1h	3.95
2	A	2h	3.65
2	A	3h	2.93

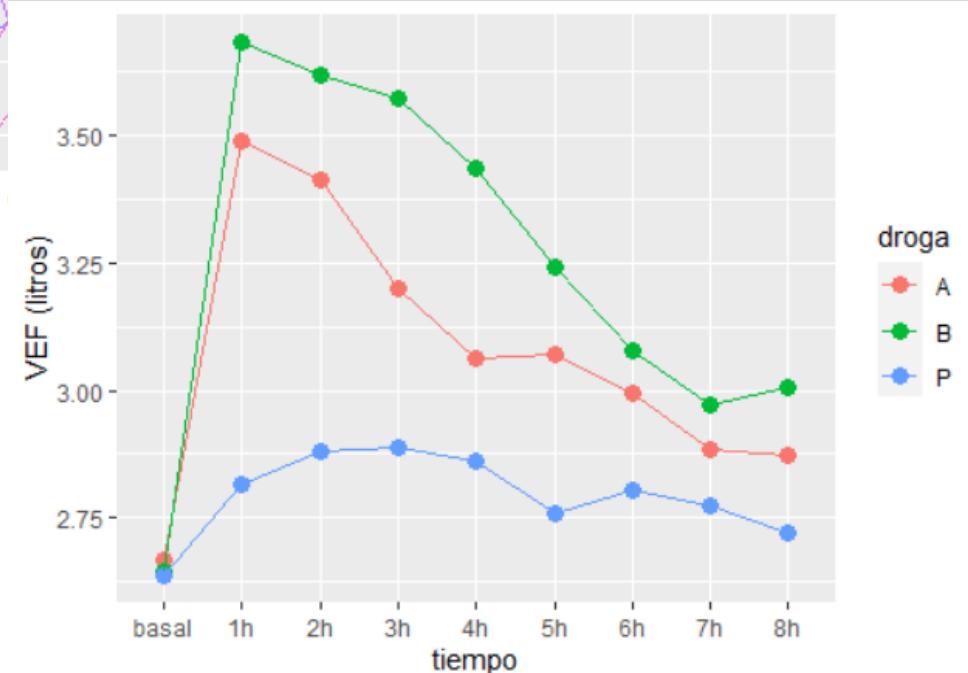
Spaghetti plots /Gráficos de perfiles

3



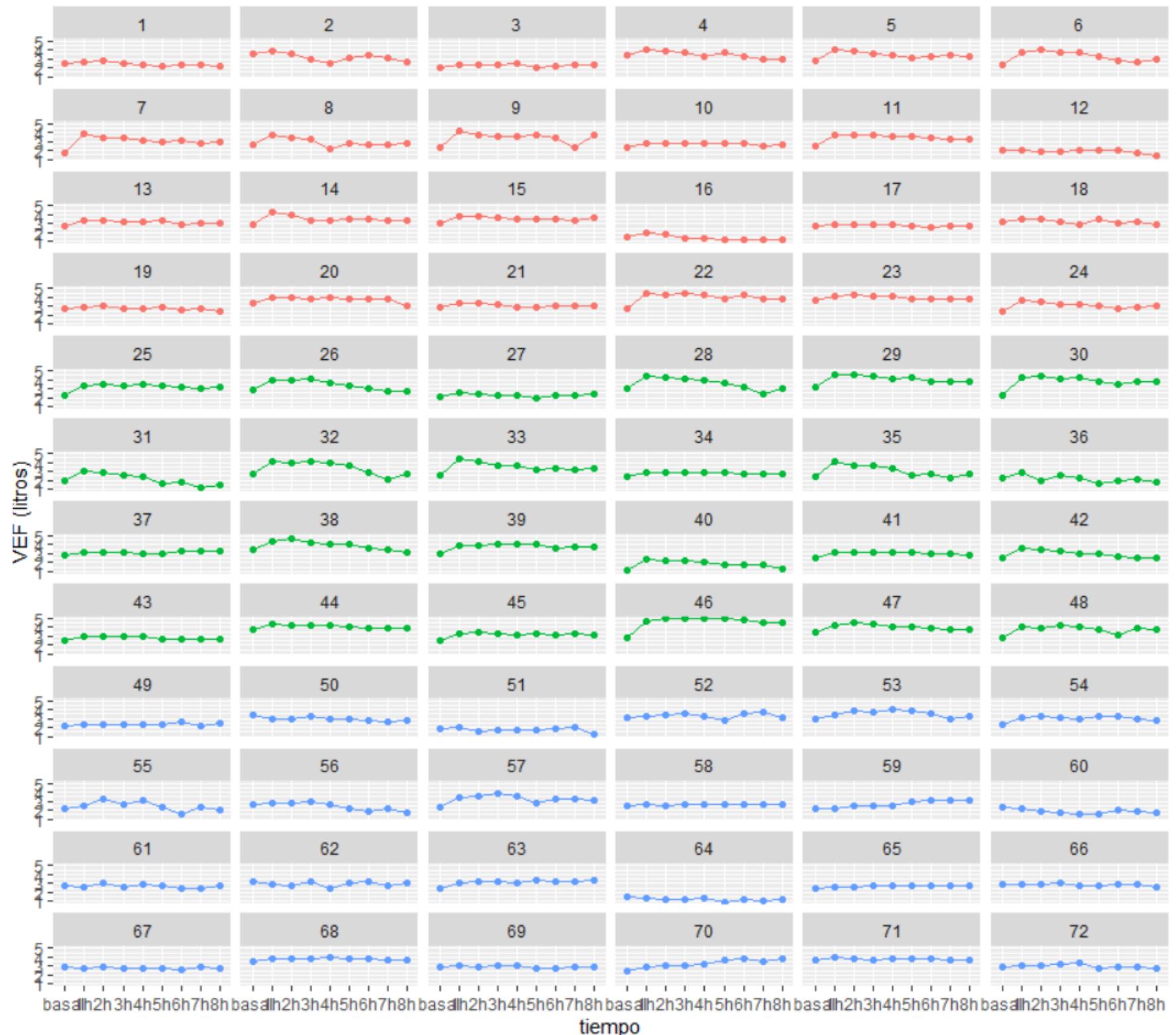
asma.csv

Un Id único
por paciente



droga

- A
- B
- P



Diseño de medidas repetidas

5

- Se utiliza cuando una misma unidad experimental es sometida a mediciones sucesivas a lo largo del tiempo o en cierto orden
- Proporcionan información sobre **tendencias en el tiempo** de la variable respuesta bajo distintas condiciones (tratamientos)
- Se los denominan también **datos longitudinales**
- Las observaciones efectuadas sobre la misma ue están **correlacionadas** – acarrean un mismo efecto de ue - y no pueden por tanto considerarse como observaciones independientes
- Debemos modelar esa estructura de correlación. Y como ya vimos hay dos opciones:
- Eso se hace mediante distintos modelos para la matriz de covarianza

¿Cómo modelamos datos correlacionados?

6

- **Modelos Condicionales**, sujeto-específicos (efectos fijos + efectos aleatorios). El modelo incluye VE de efecto aleatorio. Esto induce la correlación entre observaciones

Las mediciones de VEF a lo largo del tiempo para cada individuo se modelan individualmente para cada individuo. Ordenada al origen: efecto aleatorio (distinto para cada sujeto). Esto resulta en residuos independientes. Permite estimar componentes de varianza. [lme](#), [lmer](#)

- **Modelos Marginales** (efectos fijos + estructura de correlación residual). El modelo incluye una estructura de correlación explícita entre las observaciones (matriz de covarianzas).

Las mediciones de VEF para cada individuo se modelan con un modelo de regresión lineal múltiple de efectos fijos y se explicita la estructura de correlación de los residuos dentro de cada individuo con una matriz de covarianza. [gls](#)

Varianza, covarianza, correlación?

7

- Varianza: es el promedio de las desviaciones a la media, elevadas al cuadrado. Es una medida de la variabilidad de UNA variable. [Y]

$$\sigma_Y^2 = E(Y - E_Y)^2 = \frac{\sum (y_i - \mu_Y)^2}{N}$$

- Covarianza: es una medida de la variación conjunta de DOS variables aleatorias cuantitativas. [Y₁Y₂]

$$Cov_{Y_1 Y_2} = \sigma_{Y_1 Y_2} = E(Y_1 - E_{Y_1})(Y_2 - E_{Y_2}) = \\ = \frac{\sum (y_{i1} - \mu_{Y_1})(y_{i2} - \mu_{Y_2})}{N}$$

- Estimador insesgado de la varianza:

$$S_Y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

- Estimador insesgado de la covarianza:

$$S_{Y_1 Y_2} = \frac{\sum (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{n-1}$$

Covarianza

8

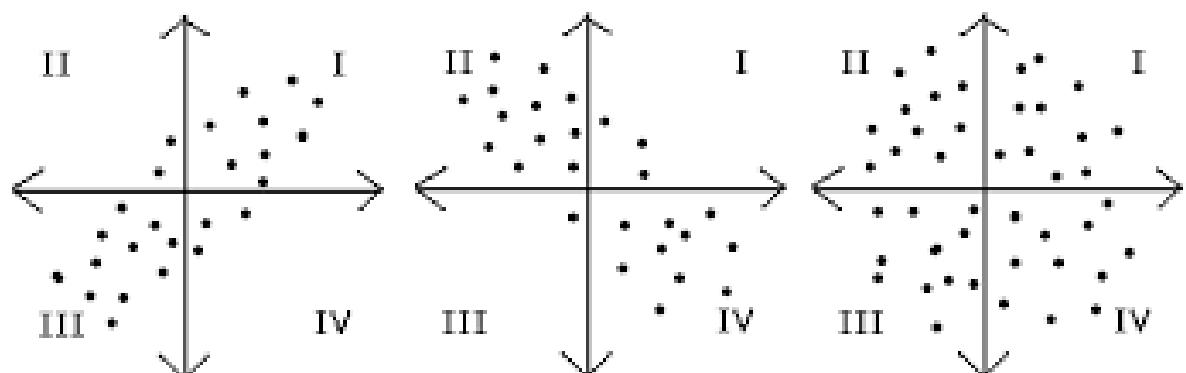
- Las variables asociadas en forma directa tienen covarianza positiva, mientras que las asociadas en forma indirecta, la tienen negativa
- Además

$$\sigma_{YY} = \sigma_Y^2$$

$$\sigma_{Y_1 Y_2} = \sigma_{Y_2 Y_1}$$

$$\sigma_{aY} = 0$$

$$-\infty < \sigma_{Y_1 Y_2} < +\infty$$



$$\sigma_{Y_1 Y_2} = E(Y_1 - E_{Y_1})(Y_2 - E_{Y_2})$$

- Si dos variables son independientes, su covarianza vale 0 (la recíproca no es necesariamente cierta)

Correlación

9

- Coeficiente de correlación lineal de Pearson: es una estandarización de la covarianza. Sin unidades

$$\rho_{Y_1 Y_2} = \frac{\sigma_{Y_1 Y_2}}{\sigma_{Y_1} \sigma_{Y_2}}$$

- Además $\rho_{YY} = 1$

$$\rho_{Y_1 Y_2} = \rho_{Y_2 Y_1}$$

$$\rho_{aY} = 0$$

$$-1 < \rho_{Y_1 Y_2} < +1$$

- Si dos variables son independientes el coeficiente vale 0
- Equivale a calcular la covarianza con las variables estandarizadas

- Estimador insesgado del coeficiente de correlación de Pearson:

$$r_{Y_1 Y_2} = \frac{S_{Y_1 Y_2}}{S_{Y_1} S_{Y_2}}$$

- Se deduce que la covarianza puede expresarse como:

$$\sigma_{Y_1 Y_2} = \rho_{Y_1 Y_2} \sigma_{Y_1} \sigma_{Y_2}$$

Asociación en medidas repetidas

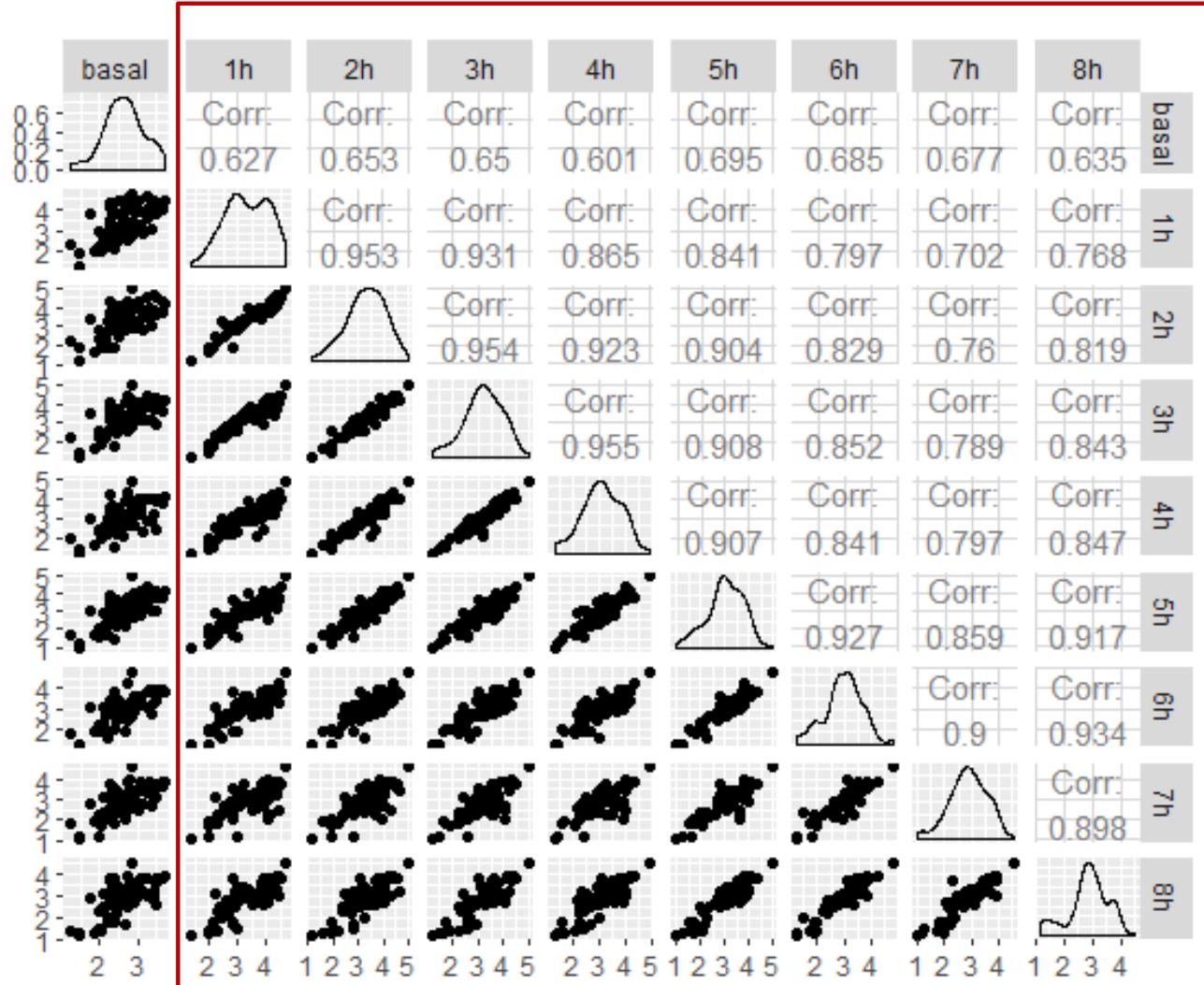
10

Podemos pensarlos como
distintas variables

paciente	droga	basal	1h	2h	3h	4h	5h	6h	7h	8h
1	A	2.46	2.68	2.76	2.50	2.30	2.14	2.40	2.33	2.20
2	A	3.50	3.95	3.65	2.93	2.53	3.04	3.37	3.14	2.62
3	A	1.96	2.28	2.34	2.29	2.43	2.06	2.18	2.28	2.29
4	A	3.44	4.08	3.87	3.79	3.30	3.80	3.24	2.98	2.91
5	A	2.80	4.09	3.90	3.54	3.35	3.15	3.23	3.46	3.27
6	A	2.36	3.79	3.97	3.78	3.69	3.31	2.83	2.72	3.00
7	A	1.77	3.82	3.44	3.46	3.02	2.98	3.10	2.79	2.88
8	A	2.64	3.67	3.47	3.19	2.19	2.85	2.68	2.60	2.73
9	A	2.30	4.12	3.71	3.57	3.49	3.64	3.38	2.28	3.72
10	A	2.27	2.77	2.77	2.75	2.75	2.71	2.75	2.52	2.60

Formato “wide” vs “long”

Matriz de diagramas de dispersión



Las
observaciones
están
asociadas

Matriz de covarianza Σ

12

paciente	droga	basal	1h	2h	3h	4h	5h	6h	7h	8h
1	A	2.46	2.68	2.76	2.50	2.30	2.14	2.40	2.33	2.20
2	A	3.50	3.95	3.65	2.93	2.53	3.04	3.37	3.14	2.62
3	A	1.96	2.28	2.34	2.29	2.43	2.06	2.18	2.28	2.29
4	A	3.44	4.08	3.87	3.79	3.30	3.80	3.24	2.98	2.91
5	A	2.80	4.09	3.90	3.54	3.35	3.15	3.23	3.46	3.27
6	A	2.36	3.79	3.97	3.78	3.69	3.31	2.83	2.72	3.00
7	A	1.77	3.82	3.44	3.46	3.02	2.98	3.10	2.79	2.88
8	A	2.64	3.67	3.47	3.19	2.19	2.85	2.68	2.60	2.73
9	A	2.30	4.12	3.71	3.57	3.49	3.64	3.38	2.28	3.72
10	A	2.27	2.77	2.77	2.75	2.75	2.71	2.75	2.52	2.60

$$\Sigma = \begin{matrix} T1 & T2 & T3 & T4 \\ T1 & \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ T2 & \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ T3 & \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} \\ T4 & \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{matrix}$$

Solo 4 tiempos para que las matrices sean chicas

- ✓ en la diagonal principal, las varianzas de cada variable σ_i^2 . En el resto, las covarianzas σ_{ij} entre pares de variables
- ✓ Matriz cuadrada y simétrica ($\sigma_{12} = \sigma_{21}$)
- ✓ Matriz no estandarizada, tiene unidades
- ✓ Σ : matriz poblacional, S : matriz muestral

Estructura de la matriz de covarianza para medidas repetidas

13

- Simple
 - Simetría compuesta **corCompSymm**
 - Autoregresiva de orden 1 (AR1) **corAR1**
 - Toeplitz o autoregresivo general
 - Desestructurada **corSymm**
 - Autoregresiva continua de orden 1 o Interdependencia de primer orden **corCAR1**

□ Se pueden combinar con varianzas heterogéneas



Requieren tiempos igualmente espaciados

Estructura simple de la matriz de covarianza Σ

14

- Si las observaciones fuesen independientes (i.e. suponiendo que en cada tiempo se midió a un individuo distinto, o todos los diseños vistos antes de mixtos) las covarianzas son nulas.

$$\begin{matrix} & T1 & T2 & T3 & T4 \\ T1 & \sigma^2 & 0 & 0 & 0 \\ T2 & 0 & \sigma^2 & 0 & 0 \\ T3 & 0 & 0 & \sigma^2 & 0 \\ T4 & 0 & 0 & 0 & \sigma^2 \end{matrix}$$

$$\begin{matrix} & T1 & T2 & T3 & T4 \\ T1 & \sigma_1^2 & 0 & 0 & 0 \\ T2 & 0 & \sigma_2^2 & 0 & 0 \\ T3 & 0 & 0 & \sigma_3^2 & 0 \\ T4 & 0 & 0 & 0 & \sigma_4^2 \end{matrix}$$

Suponiendo homocedasticidad

Más parsimoniosa; más restringida

No suponiendo homocedasticidad
(varident)

de parámetros?

Estructura de simetría compuesta

15

- Si los datos provienen de la misma UE no son independientes y por lo tanto la covarianza entre mediciones sucesivas no es nula
- Suponiendo misma varianza en cada tiempo y misma covarianza entre tiempos:

$$\sigma_{Y_1 Y_2} = \rho_{Y_1 Y_2} \sigma_{Y_1} \sigma_{Y_2} = \rho \sigma^2$$

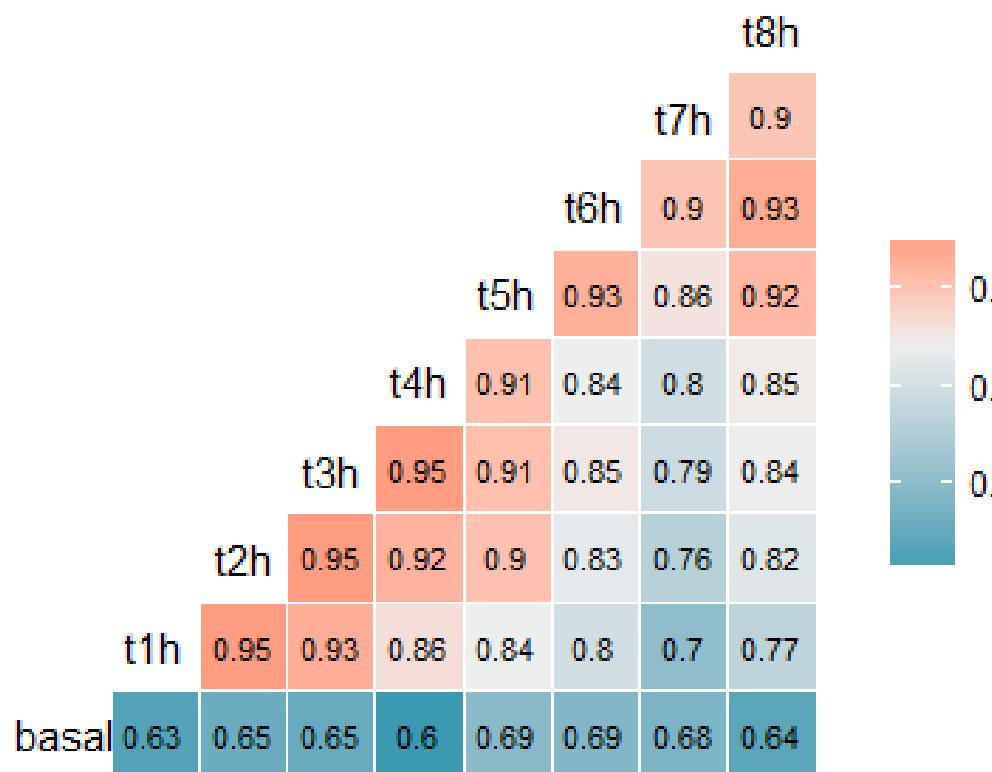
$$\begin{array}{cccc} T1 & T2 & T3 & T4 \\ \hline T1 & \left[\begin{matrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{matrix} \right] & \rightarrow & \left[\begin{matrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{matrix} \right] \\ T2 & & & = \\ T3 & & & \\ T4 & & & \end{array}$$

Matriz de simetría compuesta
corCompSymm

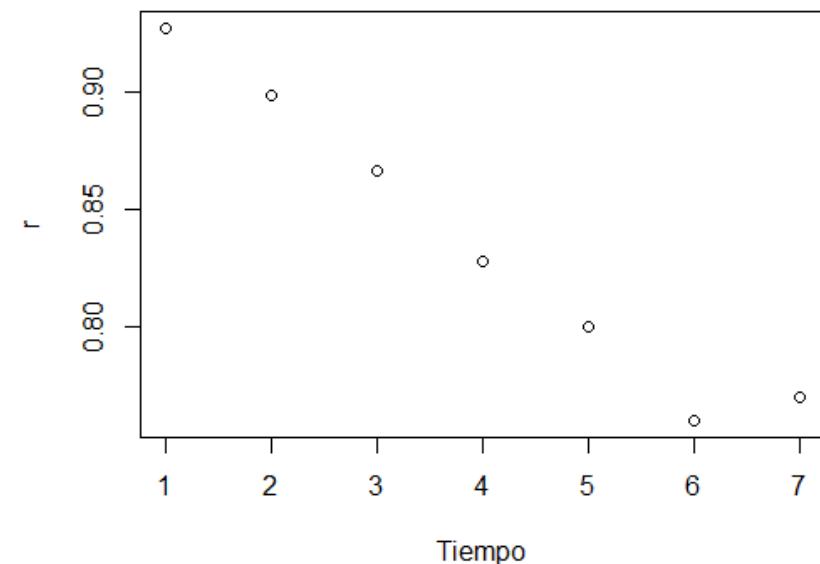
- Asume igual correlación entre cualquier par de MR

de parámetros?

Matriz de correlación (heatmap)



Coeficiente de correlación vs lapso entre tiempos



- La matriz de simetría compuesta puede ser poco realista en DMR: las observaciones adyacentes estarán más fuertemente asociadas que las más alejadas en el tiempo. Sin embargo puede funcionar para tiempos cortos.
- Es esperable en DBA, en anidados, en diseño de parcela dividida

Estructura autoregresiva de primer orden

17

- Supongamos que la correlación entre tiempos disminuye exponencialmente según la distancia entre tiempos Δt : $\rho_{t_i, t_{i+\Delta t}} = \rho^{\Delta t}$
- Supongamos que la correlación entre las observaciones de dos tiempos con la misma diferencia de tiempo es siempre la misma, ρ

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Matriz de correlación autoregresiva
de primer orden AR1
corAR1

- Para tiempos igualmente espaciados. Si no es el caso, usar Autoregresiva continua de orden 1 o Interdependencia de primer orden **corCAR1**
- Estos modelos suponen homocedasticidad (σ^2 común) pero pueden modelarse con heterocedasticidad

Matriz desestructurada

18

- No hay restricciones sobre los parámetros de la matriz
- Es la menos parsimoniosa, con menores restricciones (mayor cantidad de parámetros)

$$\begin{matrix} & \text{T1} & \text{T2} & \text{T3} & \text{T4} \\ \text{T1} & \sigma^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \text{T2} & \sigma_{12} & \sigma^2 & \sigma_{32} & \sigma_{42} \\ \text{T3} & \sigma_{13} & \sigma_{23} & \sigma^2 & \sigma_{43} \\ \text{T4} & \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma^2 \end{matrix}$$

Matriz de correlación
desestructurada
`corSymm`

$$\begin{matrix} & \text{T1} & \text{T2} & \text{T3} & \text{T4} \\ \text{T1} & \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \text{T2} & \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \text{T3} & \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} \\ \text{T4} & \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{matrix}$$

parámetros
Desestructurada homogénea: $t(t-1)/2$
Desestructurada heterogénea: $t(t+1)/2$

Modelos marginales

Modelamos la estructura de covarianza

19

- Ajusta un modelo general para la estructura promedio de la población de individuos
- No incluye VE de efectos aleatorios
- Se explicita una estructura para la matriz de covarianza de los errores

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \varepsilon_{ijk} \quad i = 1 \text{ a } 3, j = 1 \text{ a } 8, k = 1 \text{ a } 24$$

$$\varepsilon_{ijk} \approx N(0, \Sigma_j)$$

- ✓ donde Y_{ijk} es la respuesta de cada individuo a cada tiempo
- ✓ μ es la media general o media de la población
- ✓ α_i es el efecto fijo del tratamiento i
- ✓ β_j es el efecto fijo del tiempo j
- ✓ $\alpha\beta_{ij}$ es el efecto de la interacción fija tratamiento-tiempo
- ✓ γ_k es el valor basal
- ✓ ε_{ijk} es el error aleatorio

- En formato “regresión con v.indicadoras” (un plomo):

$$Y_i = \beta_0 + \beta_1 drogaA_i + \beta_2 drogaB_i + \beta_3 T_{2i} + \beta_4 T_{3i} + \dots + \beta_{10} T_{2i} + \beta_{11} T_{3i} + \dots + \varepsilon_i$$

Modelos marginales

```
library(nlme)  
gls
```

20

#Modelo 1: Simetría compuesta.

```
m1<-gls(vef ~droga*tiempo+basal, correlation = corCompSymm(form =  
~ 1 | paciente), bd)
```

#Modelo 2: Simetría compuesta. varianzas distintas

```
m2<-gls(vef ~droga*tiempo+basal, correlation = corCompSymm(form =  
~ 1 | paciente), bd, weights=varIdent(form= ~ 1|tiempo ))
```

#Modelo 3: AR1, varianzas iguales

```
m3<-gls(vef ~droga*tiempo+basal, correlation = corAR1(form = ~ 1  
| paciente), bd)
```

#Modelo 4: AR1, varianzas distintas

```
m4<-gls(vef ~droga*tiempo+basal, correlation = corAR1(form = ~ 1  
| paciente), bd, weights=varIdent(form= ~ 1|tiempo ))
```

#Modelo 5: matriz desestructurada

```
m5<-gls(vef ~droga*tiempo+basal, correlation = corSymm(form = ~ 1  
| paciente), bd)
```

Simetría compuesta

```
> summary(m1)
Generalized least squares fit by REML
Model: vef ~ droga * tiempo + basal
Data: bd2
      AIC    BIC   logLik 
401.2902 517.707 -173.6451
```

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

Correlation structure: Compound symmetry

Formula: ~1 | paciente
Parameter estimate(s):

Rho
0.7656626

```
> getVarCov(m1)
```

Marginal variance covariance matrix

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	0.26938	0.20626	0.20626	0.20626	0.20626	0.20626	0.20626	0.20626
[2,]	0.20626	0.26938	0.20626	0.20626	0.20626	0.20626	0.20626	0.20626
[3,]	0.20626	0.20626	0.26938	0.20626	0.20626	0.20626	0.20626	0.20626
[4,]	0.20626	0.20626	0.20626	0.26938	0.20626	0.20626	0.20626	0.20626
[5,]	0.20626	0.20626	0.20626	0.20626	0.26938	0.20626	0.20626	0.20626
[6,]	0.20626	0.20626	0.20626	0.20626	0.20626	0.26938	0.20626	0.20626
[7,]	0.20626	0.20626	0.20626	0.20626	0.20626	0.20626	0.26938	0.20626
[8,]	0.20626	0.20626	0.20626	0.20626	0.20626	0.20626	0.20626	0.26938

Standard Deviations: 0.51902 0.51902 0.51902 0.51902 0.51902 0.51902 0.51902 0.51902

$$\sigma_{Y_1 Y_2} = \rho_{Y_1 Y_2} \sigma_{Y_1} \sigma_{Y_2} \quad \rho = \frac{\sigma_{Y_1} \sigma_{Y_2}}{\sigma_{Y_1 Y_2}} = \frac{\sigma_{Y_1 Y_2}}{\sigma^2} \quad \hat{\rho} = \frac{0.20626}{0.26938} = 0.766$$

Simetría compuesta con varianzas distintas

```
> summary(m2)
Generalized least squares fit by REML
  Model: vef ~ droga * tiempo + basal
  Data: bd2
      AIC      BIC    logLik
 409.8817 556.4807 -170.9409
```

Correlation structure: Compound symmetry

Formula: ~1 | paciente

Parameter estimate(s):

Rho
0.7676804 $\hat{\rho}$

Variance function:

Structure: Different standard deviations per stratum

Formula: ~1 | tiempo

Parameter estimates:

1h	2h	3h	4h	5h	6h	7h	8h
1.000000	1.029784	1.001406	1.096472	1.071529	1.034628	1.128791	1.116516

Requiere tiempos
igualmente espaciados

```
> anova(m1,m2,m3,m4, m5)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
m1	1	27	401.2902	517.7070	-173.64508			
m2	2	34	409.8817	556.4807	-170.94086	1 vs 2	5.40844	0.6102
m3	3	27	329.0350	445.4519	-137.51752	2 vs 3	66.84668	<.0001
m4	4	34	324.5664	471.1653	-128.28318	3 vs 4	18.46869	0.0100
m5	5	54	261.4179	494.2516	-76.70896	4 vs 5	103.14842	<.0001

```
> summary(m5)
```

Generalized least squares fit by REML

Model: vef ~ droga * tiempo + basal

Data: bd2

AIC	BIC	logLik
261.4179	494.2516	-76.70896

Correlation structure: General

Parámetros:

- 1 varianza
- 28 covarianzas
- 24 (8x3) medias
- 1 Beta

```
> getVarCov(m5)
```

Marginal variance covariance matrix

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	0.26467	0.23843	0.23576	0.20718	0.18119	0.18029	0.13272	0.16837
[2,]	0.23843	0.26467	0.24099	0.22902	0.21071	0.18381	0.14981	0.18112
[3,]	0.23576	0.24099	0.26467	0.24009	0.21212	0.19635	0.16386	0.19186
[4,]	0.20718	0.22902	0.24009	0.26467	0.21395	0.19133	0.17070	0.19507
[5,]	0.18119	0.21071	0.21212	0.21395	0.26467	0.22473	0.18971	0.22059
[6,]	0.18029	0.18381	0.19635	0.19133	0.22473	0.26467	0.21254	0.23109
[7,]	0.13272	0.14981	0.16386	0.17070	0.18971	0.21254	0.26467	0.21365
[8,]	0.16837	0.18112	0.19186	0.19507	0.22059	0.23109	0.21365	0.26467

Standard Deviations: 0.51446 0.51446 0.51446 0.51446 0.51446 0.51446 0.51446 0.51446

Modelo condicional

24

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + B_{k(i)} + \varepsilon_{ijk} \quad \begin{matrix} i=1,2,3 \\ j=1,\dots,9 \\ k=1,\dots,24 \end{matrix}$$

- donde y_{ijk} es la respuesta de cada individuo a cada tiempo
- μ es la media general o media de la población
- α_i es el efecto fijo del tratamiento i
- β_j es el efecto del tiempo k
- $\alpha\beta_{ik}$ es el efecto de la interacción tratamiento-tiempo
- $B_{k(i)}$ es el efecto aleatorio del nivel j anidado en i (individuo anidado en tratamiento)
- ε_{ijk} es el error aleatorio

$$B_{k(i)} \approx NID(0, \sigma_{indiv}^2)$$

$$\varepsilon_{ijk} \approx NID(0, \sigma_e^2)$$

$B_{k(i)}$ y ε_{ijk} independientes entre sí

Modelo condicional

library(nlme)
nlme

25

No se modela la matriz de covarianza y se incluye el efecto aleatorio de individuo anidado en droga:

```
m6<-lme(vef ~droga*tiempo+basal, random = ~1|paciente, bd)
```

La matriz de covarianza que se induce implícitamente es la de **simetría compuesta**. Es decir que equivale a:

```
m1<-glm(vef ~droga*tiempo+basal, correlation = corCompSymm(form =  
~ 1 | paciente), bd)
```

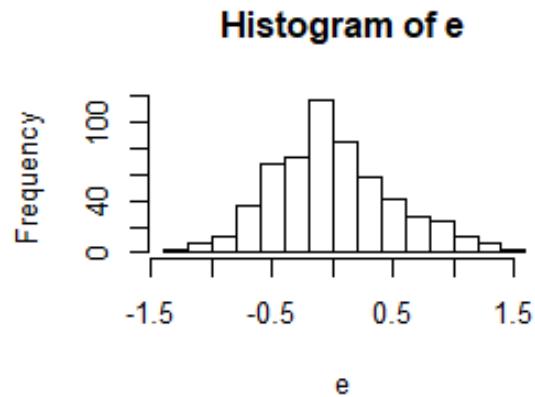
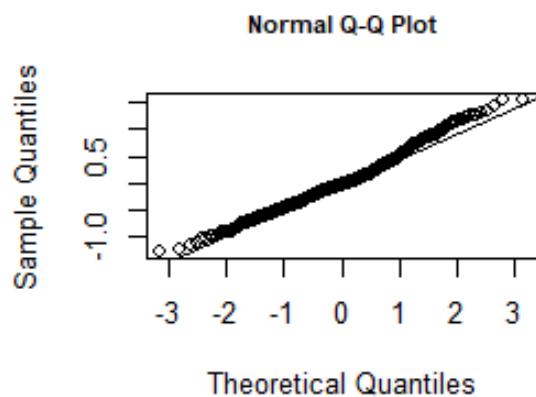
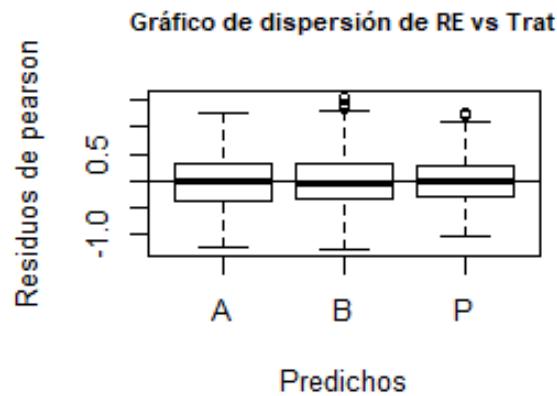
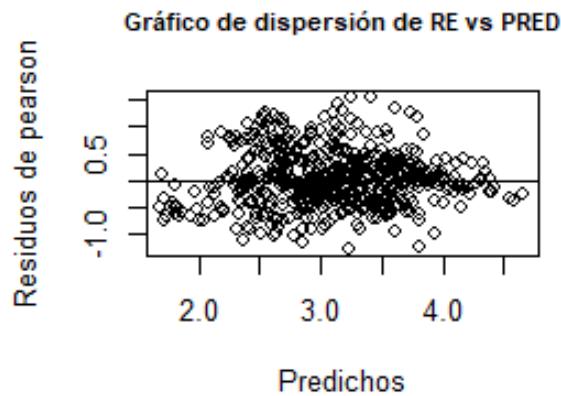
Los resultados para los efectos fijos son los mismos, pero estima componentes de varianza

No es muy aplicable a DMR

```
> summary(m6)  
Linear mixed-effects model fit by REML  
Data: bd2  
      AIC      BIC      logLik  
401.2902 517.707 -173.6451  
  
Random effects:  
Formula: ~1 | paciente  
             (Intercept) Residual  
StdDev:    0.4541552 0.2512505
```

Supuestos m5

26



```
> anova(m5)
```

Denom. DF: 551

	numDF	F-value	p-value
(Intercept)	1	3526.681	<.0001
droga	2	9.776	1e-04
tiempo	7	13.251	<.0001
basal	1	82.045	<.0001
droga:tiempo	14	3.963	<.0001

- ¿Son paralelos los perfiles de respuesta en los grupos?
- Si son paralelos, ¿difieren entre tratamientos?
- Si son paralelos, ¿son constantes en el tiempo?

```
> CLD(emmeans(m5, pairwise ~ droga | tiempo))
```

tiempo = 1h:

droga	emmean	SE	df	lower.CL	upper.CL	group
P	2.826998	0.1050225	551	2.620705	3.033292	1
A	3.471804	0.1050309	551	3.265494	3.678114	2
B	3.689948	0.1050156	551	3.483668	3.896228	2

tiempo = 2h:

droga	emmean	SE	df	lower.CL	upper(CL	group
P	2.892832	0.1050225	551	2.686538	3.099125	1
A	3.395137	0.1050309	551	3.188827	3.601447	2
B	3.625364	0.1050156	551	3.419084	3.831644	2

tiempo = 3h:

droga	emmean	SE	df	lower(CL	upper(CL	group
P	2.898248	0.1050225	551	2.691955	3.104542	1
A	3.182221	0.1050309	551	2.975911	3.388530	1
B	3.576198	0.1050156	551	3.369918	3.782478	2

tiempo = 4h:

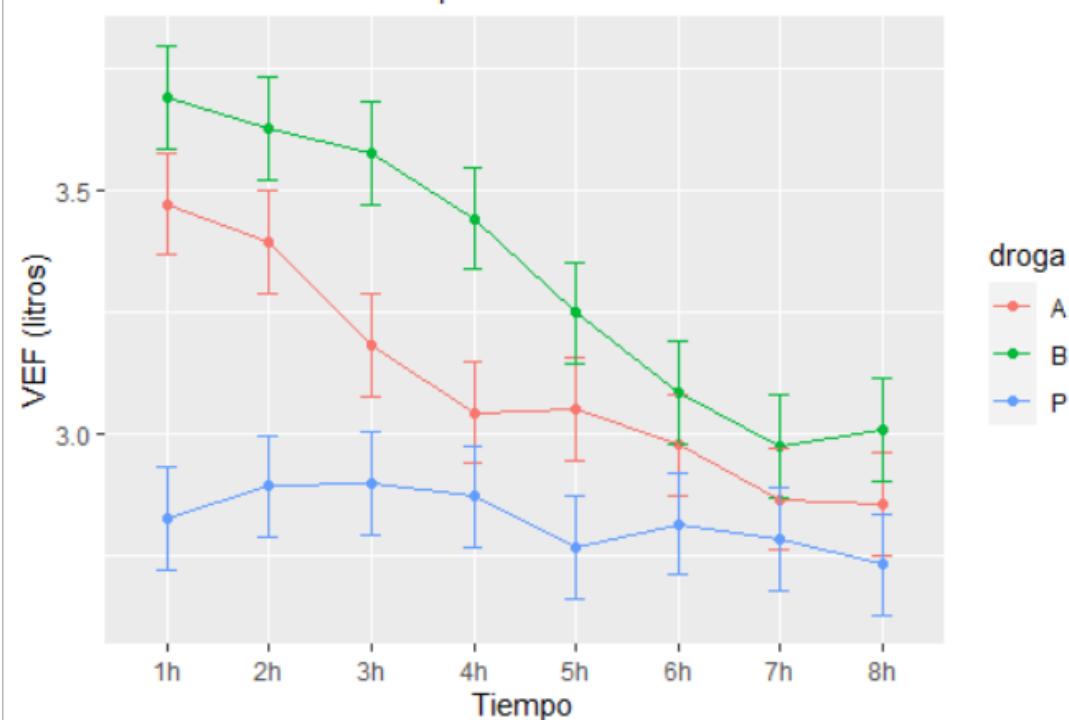
droga	emmean	SE	df	lower(CL	upper(CL	group
P	2.871582	0.1050225	551	2.665288	3.077875	1
A	3.044721	0.1050309	551	2.838411	3.251030	1
B	3.442448	0.1050156	551	3.236168	3.648728	2

tiempo = 5h:

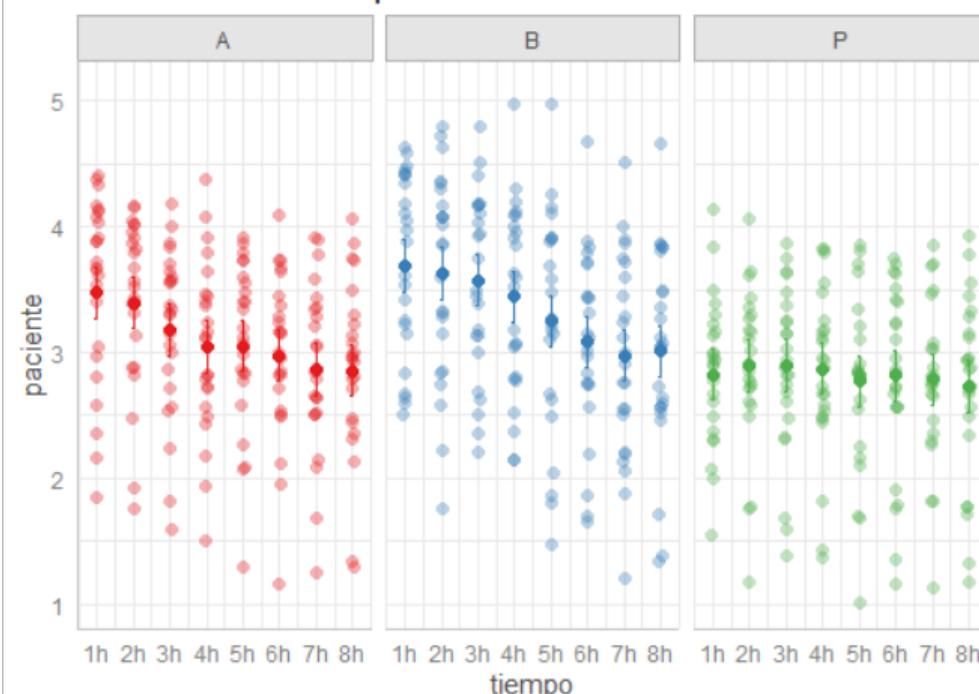
droga	emmean	SE	df	lower(CL	upper(CL	group
P	2.768665	0.1050225	551	2.562371	2.974959	1
A	3.051804	0.1050309	551	2.845494	3.258114	12
B	3.248281	0.1050156	551	3.042001	3.454561	2

Variación de VEF en función del tiempo según tratamiento

Media ± error estándar a partir del modelo



Predicted values of paciente



Otras opciones de análisis de MR

28

- Reducir el número de observaciones por individuo a 1, de manera de romper la dependencia entre observaciones:
 - Si se midieron sólo dos tiempos (inicial y final) calcular para cada individuo la diferencia en la respuesta (final-inicial por ejemplo) y aplicar un anova de un factor
 - Usar como VR el área bajo la curva para cada individuo
 - Modelar solo observaciones a un tiempo (final por ejemplo) 
- Usar el valor inicial como covariante. En experimentos, puede evitar que exista interacción significativa
- MANOVA (anova multivariado): supone una matriz de covarianzas desestructurada. Baja potencia si hay pocas réplicas en relación al DMR univariado

Diseño de parcela dividida (split plot)

- Se usa cuando se requieren UE mayores para un factor que para el otro
- Uno de los factores tratamiento (Factor A) se asigna a unidades experimentales de mayor tamaño (**parcela principal**) y dentro de cada parcela principal se identifican **subparcelas** o parcelas de menor tamaño sobre las cuales se asigna al azar el segundo factor tratamiento (Factor B)
- La aleatorización está **restringida** y es en dos etapas:
 - 1º los niveles del factor A en las parcelas principales
 - 2º los niveles del factor B en cada subparcela

A ₁	B ₁	B ₃	B ₄	B ₂
A ₃	B ₄	B ₂	B ₁	B ₃
A ₂	B ₄	B ₃	B ₂	B ₁

Parcela dividida

A ₁	T ₁	T ₂	T ₃	T ₄
A ₃	T ₁	T ₂	T ₃	T ₄
A ₂	T ₁	T ₂	T ₃	T ₄

Medidas repetidas

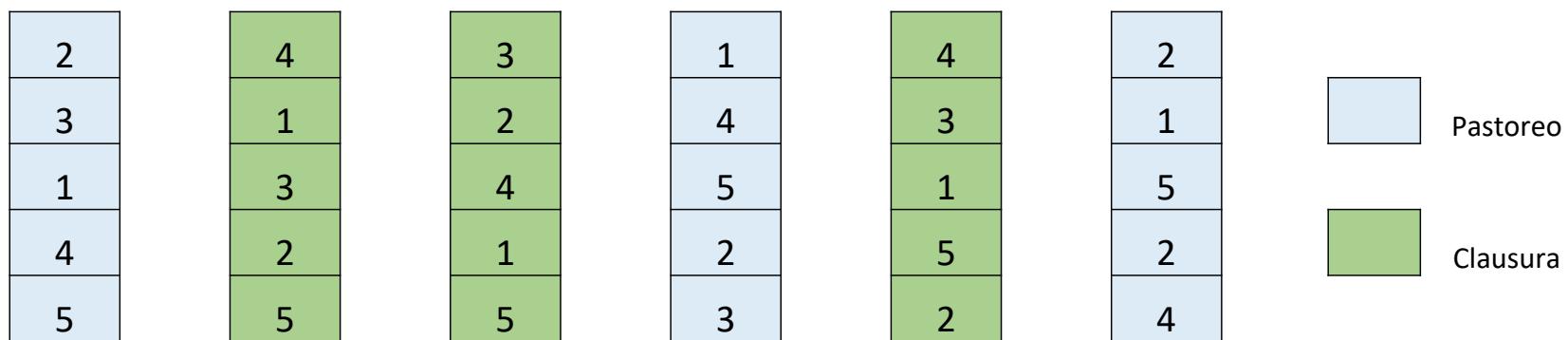
Experiencias de largo plazo para el manejo de una hierba invasora de pastizales: El caso de *Hieracium pilosella* L. en la Estepa Fueguina

Ecología Austral 24:135-144. Agosto 2014

P.A. CIPRIOTTI¹; M.B. COLLANTES²; C. ESCARTÍN³; S. CABEZA⁴; R.B. RAUBER⁵ & K. BRAUN²

El experimento a campo se estableció en la Estancia Cullen, ubicada al norte de la estepa Fueguina, en un sector utilizado para la implantación de pasturas hace 30 años y con una cobertura actual promedio de la especie invasora mayor al 15%. Se seleccionaron **seis áreas homogéneas** de aproximadamente 1 ha (parcelas) que contuvieran aproximadamente al menos 100 parches de *H. pilosella* de aproximadamente 2 m de diámetro. Tres de las seis parcelas delimitadas fueron seleccionados al azar y excluidas al ganado ovino mediante la construcción de un cerco permanente. De este modo, se consideraron a las tres parcelas abiertas al pastoreo doméstico pertenecientes al tratamiento “pastoreo” y a las tres parcelas excluidas al ganado ovino correspondientes al tratamiento “clausura”. Antes del establecimiento del experimento, las seis parcelas (con y sin pastoreo) tuvieron un manejo y carga animal similar durante las últimas décadas. Además, todas las parcelas tenían una posición topográfica, pendiente del terreno, cobertura vegetal y composición florística similar, incluyendo la cobertura de la maleza exótica *H. pilosella* (ca. 35%).

Dentro de cada parcela se establecieron **cinco sub-parcelas** (5×5 m) asegurando por lo menos la existencia de un parche de *H. pilosella* (diámetro aprox. 2 m), lo que resultó en un total de 30 subparcelas. Cuatro tratamientos de control y un testigo sin tratar fueron asignados al azar a las sub-parcelas: (1) Testigo sin tratar; (2) Inter-siembra de pastura + fertilización NP con fosfato diamónico; (3) Fertilización NP con fosfato diamónico; (4) Aplicación de herbicida selectivo de hoja ancha; y (5) Aplicación de herbicida no selectivo. Las variables de respuesta estudiadas al final de la séptima estación de crecimiento fueron la cobertura de la maleza, la cobertura de las formas de vida dominantes y la cobertura de suelo descubierto.



Análogo al DMR (parcela ppal ≈ individuos anidados en factor A; subparcela ≈ tiempos (pero con niveles del factor B aleatorizados)

An Integrated Shiny App for a Course on Repeated Measurements Analysis (completed)

March 18, 2016

By Dimitris Rizopoulos

Repeated Measurements Analysis

The screenshot shows a Shiny application running locally at 127.0.0.1:4904. The interface consists of two main sections: a sidebar on the left and a main content area on the right.

Left Sidebar:

- A title "Select chapter" above a dropdown menu set to "Chapter 0".
- A title "Select section:" above a dropdown menu.

Right Content Area:

- A top navigation bar with tabs: "Code", "Output", "Help", and "Slides" (which is currently selected).
- The main content area displays the following text:

Statistical Analysis of Repeated Measurements Data

Dimitris Rizopoulos
Department of Biostatistics, Erasmus University Medical Center
d.rizopoulos@erasmusmc.nl

March 27 – 31, 2017

A blue oval sticker on the left side contains the text "Ver script".

BIOMETRÍA II

CLASE 13

INTEGRACIÓN DE MODELOS



Estructura de los datos

2

Tipo de estudio: experimental u observacional	Determina las conclusiones (causalidad o asociación)
Tipo de VR: continua, discreta, binaria	Define la distribución de probabilidades y por lo tanto el tipo de modelo y el método de estimación
Tipo de VE: continua, discreta, binaria, cuali	Solo cuali: anova Cuanti y cuali: regresión con dummies (ancova) Cuanti: regresión
Relación entre VR y VE	Modelos lineales o no lineales (en los parámetros)
Estructura de agregación (independencia vs bloques, anidamiento, medidas repetidas, parcela dividida)	Modelos con VE de efectos fijos vs modelos marginales /condicionales
Declaración de dependencia entre las observaciones	Modelos marginales o modelos condicionales

Protocolo

3

Parte aleatoria

1. Basándose en el diseño experimental o en el método de muestreo empleado, incluir efectos aleatorios (modelos condicionales) o estructura de la matriz de covarianza (modelos marginales) si existe nivel de agrupamiento entre las observaciones

Parte fija

3. Si se trata de un experimento: incluir en el modelo todos los términos que quedaron definidos por el diseño experimental. El modelo no se debería simplificar (independientemente de la significación de cada término)
4. Si se trata de un estudio observacional: utilizar algún criterio de selección de modelos a fin de identificar las VE significativas o importantes, las restantes se eliminan del modelo a menos que interese incluirlas por razones teóricas

Modelos en R

Fun- ción	Modelo	Estima- ción
lm	Modelo lineal con errores normales y varianza constante; generalmente es usado para regresión con VE cuantitativas	MCO
gls	Modelo lineal con errores normales. Permite modelar heterocedasticidad y distintas estructuras de matriz de covarianzas; no admite VE de efectos aleatorios	MV
lme	Modelo lineal con errores normales y VE de efectos fijos y aleatorios. Permite modelar heterocedasticidad y distintas estructuras de matriz de covarianza	MV
lmer	Idem anterior, pero no modela estructuras de matriz de covarianza	MV

Funciones genéricas

5

Función	Modelo
summary	Proporciona estimaciones de los parámetros del modelo en forma de regresión o anova (summary.lm y summary.aov respectivamente)
plot	Gráficos de diagnóstico de supuestos del modelo
anova	Permite comparar modelos anidados
update	Modifica el último modelo ajustado
coef	Proporciona los estimadores de los parámetros del modelo
fitted	Valores predichos por el modelo lineal
predict	Idem anterior
resid	Diferencia entre el valor observado y el predicho por el modelo
AIC	Compara modelos por AIC

Modelos lineales sin VE de efectos aleatorios

Modelo	lineal general	lineal general	lineal generalizado
Método de estimación	cuadrados mínimos	máxima verosimilitud restringida	máxima verosimilitud
distribución de la VR	normal	normal	Familia exponencial
heterocedasticidad	sensible	modelable	modelable
desbalanceo	sensible	robusto	robusto
supuestos	Independencia, normalidad, homocedasticidad, linealidad	Independencia, normalidad, linealidad	Independencia, linealidad dispersión acorde a la distrib de la variable

Modelos lineales con VE de efectos aleatorios

Modelo	lineal general mixto	lineal general mixto	lineal general marginal
Método de estimación	cuadrados mínimos	máxima verosimilitud restringida	máxima verosimilitud
distribución de la VR	normal	normal	normal
Estructura de agregación	modelable pero arduo	Modelable mediante la inclusión de VE de efectos aleatorios	Modelable mediante la declaración de matriz de covarianza
Componentes de varianza	sí	sí	no
supuestos	normalidad, homocedasticidad, linealidad	Normalidad, linealidad; heterocedasticidad modelable	Normalidad, linealidad; heterocedasticidad modelable

Algunas pautas para una correcta experimentación

8

En el diseño:

- Aleatorización de los tratamientos
- Replicación en la escala adecuada. Atención a las seudorréplicas
- Determinar *a priori* el tamaño de muestra para una dada potencia y un efecto de tratamiento

Algunas pautas para un correcto análisis estadístico

9

- Recurrir a herramientas gráficas para tener idea del efecto de los tratamientos, variabilidad y posibles outliers
- Identificar tipo de VR, tipo de VE, si existen estructuras de agrupamiento de los datos
- Si se trata de un modelo con algún nivel de agregación de las observaciones, modelar primero la estructura aleatoria de los datos
- Centrar las VE si se desea darle sentido a la ordenada al origen
- Verificar los supuestos del modelo; patrones en los residuos pueden indicar la necesidad de incluir potencias o interacciones
- Evitar efectuar múltiples test estadísticos separados. Ajuste del nivel de significación
- Respetar el principio de marginalidad: siempre analizar primero la interacción de mayor orden; no excluir términos involucrados en interacciones
- No comparar p de distintas pruebas, ya que a menos que los GL sean iguales no son comparables

Algunas pautas para un correcto análisis estadístico

10

En las conclusiones:

- Alcance del estudio: aplicar la inferencia estadística sólo a la población de la cual se extrajo la muestra
- Diferenciar estudios experimentales de observacionales; atención con la causalidad
- Significación biológica vs estadística. Magnitud del efecto (diferencia de medias, pendiente), informar intervalos de confianza
- Presentar los resultados gráficamente utilizando las estimaciones del modelo

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3-14.

Presentación de resultados

Materiales & Métodos

11

- Diseño experimental
- Pruebas estadísticas
- Nivel de significación
- Software

Los resultados para (*variable respuesta*) fueron analizados mediante un modelo lineal general en un diseño (). Las variables explicatorias fueron (). Los supuestos () se estudiaron mediante (). El criterio utilizado para la selección de modelos fue () .

El nivel de significación empleado fue () / Se consideraron significativas aquellas pruebas con $p < ()$

Todos los análisis estadísticos fueron efectuados utilizando el programa estadístico (R, R Core Team 2021)

Plantas en cojín en la Puna

Análisis 1

12



- Las plantas en cojín son una de las formas de vida mejor adaptadas a las extremas condiciones de los ambientes de alta montaña
- Al proporcionar micrositios más adecuados para la adquisición de recursos, actuarían como nodrizas para el resto de las especies de la comunidad
- Se desea estudiar las características de los cojines de la especie *Laretia acaulis* que favorecen el establecimiento de plántulas de otras especies, a dos altitudes en la puna jujeña: baja (~2000 m.s.n.m) y alta (~3000 m.s.n.m)
- A cada altitud de la puna jujeña se seleccionaron al azar 65 cojines separados al menos 1 km. En cada uno se determinó la biomasa de otras especies vegetales (en gramos) y características de los cojines: diámetro (cm), altura máxima, temperatura del sustrato, potencial hídrico del suelo, distancia al cojín más cercano
- VR? VE? Tipo? Modelo? (asumiendo efectos aditivos) n? Base de datos

Modelando



13

1. Especule qué valores de VIF esperaría encontrar para las 5 VE cuantitativas
2. Si detecta colinealidad, explique qué consecuencias traería incluir a todas las variables en el modelo
3. Si detecta colinealidad, ¿se solucionaría centrando las variables?
4. ¿Cómo decidiría qué variables excluir?
5. Suponga que se retuvieron diámetro y altitud y que se sospecha que el efecto nodriza es mayor a mayor altitud. Plantee el modelo

Coefficients:

	Estimate	Std. Error	Pr(> t)	
(Intercept)	160.183	33.734	2.10e-04	***
diámetro	10.261	1.0835	2.23e-05	***
altitud.alta	-20.123	3.1471	2.18e-03	***
diámetro.				
altitud.alta	3.123	0.814	5.22e-03	***

6. ¿Interpretación de los coeficientes?
7. ¿Qué cambiaría si se centrase diámetro?
8. ¿Conclusiones biológicas?
9. ¿Y si se detecta que la variabilidad en la biomasa es mayor a menor altitud?

Análisis 2



15

- Los investigadores desean analizar las modificaciones microclimáticas que induce *L. acaulis* con respecto a los espacios abiertos.
 - Particularmente desean estudiar si existen cambios en el potencial hídrico matricial del suelo (PHMS).
 - Para ello, en cada uno de los cojines se midió el PHMS en el suelo en el cojín y en el suelo aledaño
1. Escriba el modelo condicional
 2. ¿Qué componentes de varianza pueden estimarse?
 3. ¿Qué diferencias tiene este modelo con el marginal?
 4. Si se midiese el PHMS a distintas distancias del cojín, ¿qué cambiaría?
 5. Si en cada sitio se tomasen observaciones a las 8, 12, 16 y 20 hs, ¿cómo se modificaría el modelo?

BIOMETRÍA II

CLASE 9

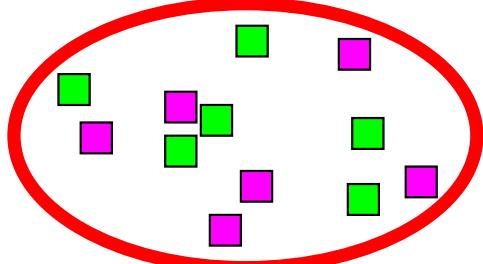
MODELOS LINEALES CON EFECTOS ALEATORIOS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

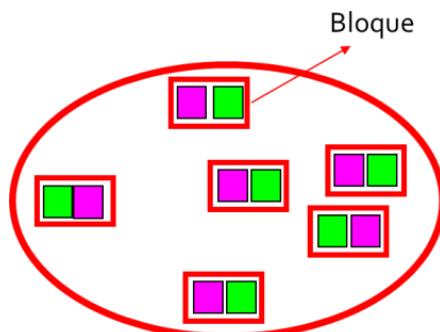
Efecto del la disponibilidad lumínica sobre la herbivoría en *Berberis buxifolia* (calafate)



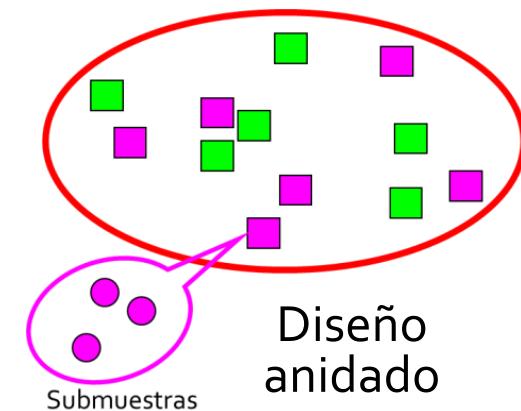
2



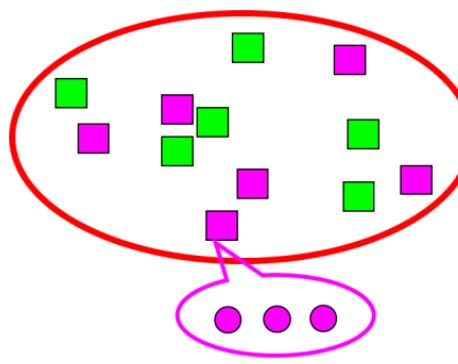
Diseño completamente aleatorizado



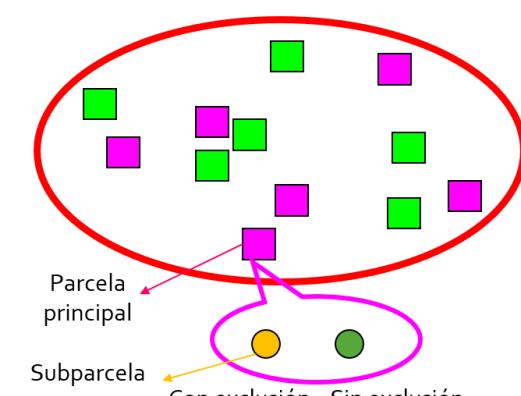
Diseño de Bloques al azar (DBA)



Diseño anidado



Diseño de medidas repetidas (DMR)



Diseño de parcela dividida (Split-plot)

Observaciones independientes

Observaciones no independientes (datos agrupados)

Datos agrupados

3

- Observaciones no independientes. Existe algún nivel de agregación de los datos, que induce una estructura de correlación entre las observaciones
- Ignorar esa falta de independencia afecta las estimaciones de los EE de los coeficientes y por lo tanto la inferencia no es correcta
- Se soluciona declarando en el modelo esa correlación entre las observaciones

¿Cómo modelamos datos correlacionados?

4

- En forma implícita, mediante la adición de VE de efecto aleatorio. Se induce la correlación entre observaciones. [Modelos condicionales o mixtos](#)
- En forma explícita, mediante la imposición de una estructura de correlación entre las observaciones (matriz de covarianzas). [Modelos marginales](#)
- Con errores con distribución normal ambas aproximaciones (formas explícita e implícita) son equivalentes
- Sin embargo, con datos no normales o cuando el modelo es no lineal, los parámetros bajo estas dos estrategias son intrínsecamente diferentes y se interpretan como:
 - ▣ “parámetros promedios poblacionales” (modelos marginales): ajustan un modelo general para la estructura promedio de la población de individuos
 - ▣ “parámetros sujeto-específicos” (modelos condicionales): proporcionan un modelo para cada individuo, pero donde la forma general del modelo es la misma para cada sujeto

VE de efectos fijos

5

- los niveles de la VE son elegidos deliberadamente por el investigador, porque le interesa estudiar esos efectos en particular
- La inferencia se efectúa sobre esos niveles
- Los tratamientos asignados por el investigador a las UE son de efectos fijos, aunque existen factores que no son asignables aleatoriamente pero son fijos (sexo, variedad, edad, precipitaciones, etc)
- $H_0: \mu_i = \mu \text{ o } \beta_i = 0$

VE de efectos aleatorios

- los niveles son elegidos al azar por el investigador de una población mayor de niveles de interés
- Las conclusiones se aplican sobre la población de niveles, no solo los estudiados
- El objetivo es modelar la falta de independencia y/o estudiar la variabilidad entre los niveles (espacial, temporal, genética, de la técnica experimental)
- Las VE de efectos aleatorios son cuali (bloques, individuos, días, líneas, etc); niveles no informativos
- Si se repite el experimento, no necesariamente se repiten los mismos niveles
- $H_0: \sigma^2_i = 0$

Volviendo a los ejemplos

6

VR: nivel de herbivoría

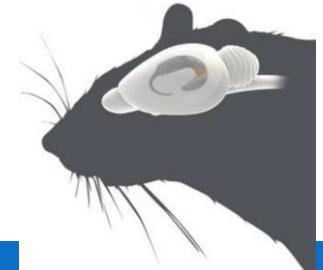
Diseño	VE de efectos fijos	VE de efectos aleatorios
Completamente aleatorizado		
bloques al azar		
anidado		
medidas repetidas		

Nivel de herbivoría:

1. biomasa consumida
2. Porcentaje de hojas con signos de herbivoría
3. Ramoneada/no ramoneada
4. Cantidad de herbívoros capturados en un día

Distribución de probabilidades?

Efecto de la exposición postnatal a etanol sobre el volumen del cerebro en ratones



7

- La exposición temprana al etanol causa alteraciones cognitivas y conductuales persistentes
- Se desea estudiar los efectos neuroestructurales asociados a esta exposición en ratones
- Para ello se seleccionaron 6 camadas de ratones de 7 días. De cada camada se eligieron 3 ratones al azar, que fueron asignados a uno de los siguientes tratamientos:
 - a) Solución salina, b) Etanol 1 g/kg, c) Etanol 2 g/kgA los 82 días se determinó el volumen cerebral por resonancia magnética

Experimento o estudio observacional?

VR:

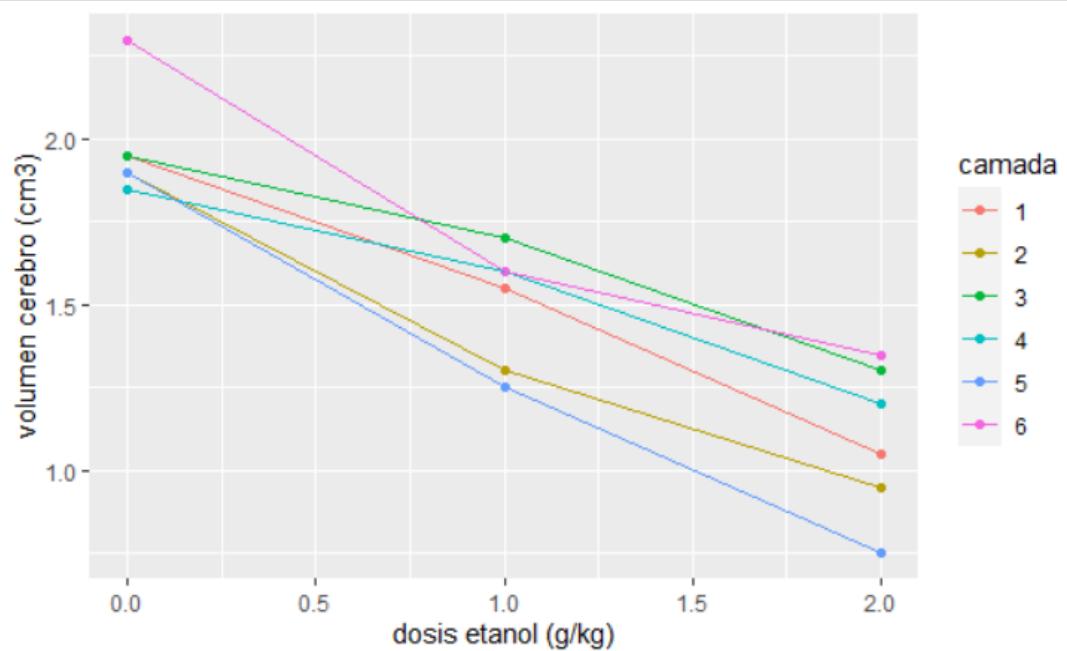
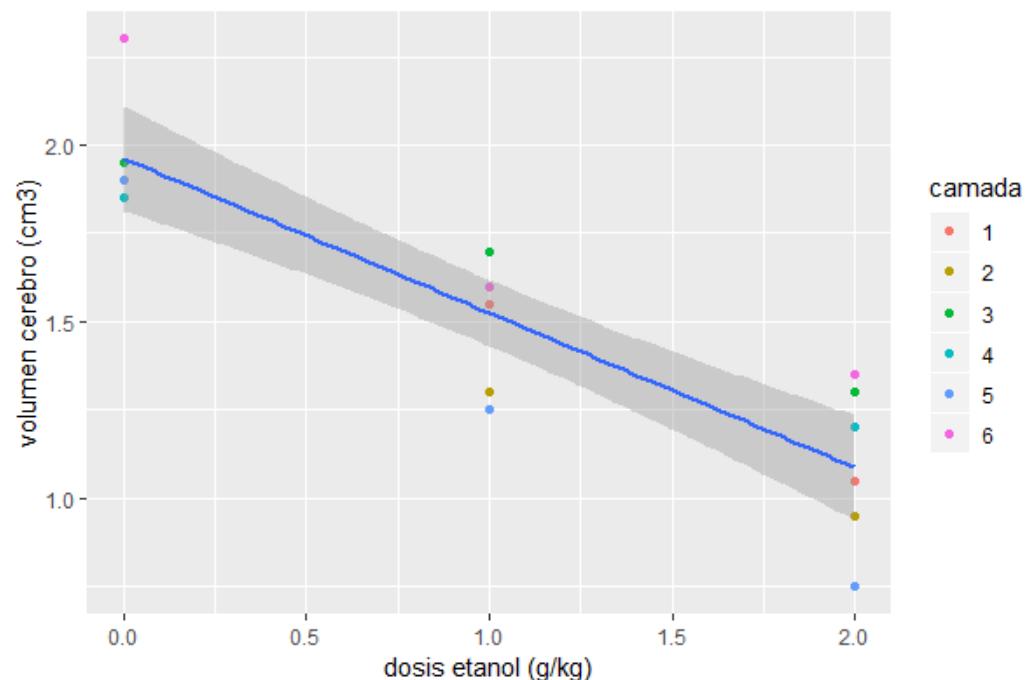
Tipo? Potencial distribución de probabilidades?

VE:

Tipo? De efectos fijos o aleatorios?

	camada	etanol	vol
1	1	0	1.95
2	2	0	1.90
3	3	0	1.95
4	4	0	1.85
5	5	0	1.90
6	6	0	2.30
7	1	1	1.55
8	2	1	1.30
9	3	1	1.70
10	4	1	1.60
11	5	1	1.25
12	6	1	1.60
13	1	2	1.05
14	2	2	0.95
15	3	2	1.30
16	4	2	1.20
17	5	2	0.75
18	6	2	1.35

Ratones_camada.txt



Sin considerar las camadas

9

$$Y_i = \beta_0 + \beta_1 etanol_i + \varepsilon_i \quad i = 1, \dots, 18$$

```
m1<-lm(vol~etanol, bd)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.96250	0.06999	28.038	4.96e-15 ***
etanol	-0.43750	0.05422	-8.069	4.96e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.1878 on 16 degrees of freedom
Multiple R-squared: 0.8028, Adjusted R-squared: 0.7904
F-statistic: 65.12 on 1 and 16 DF, p-value: 4.956e-07

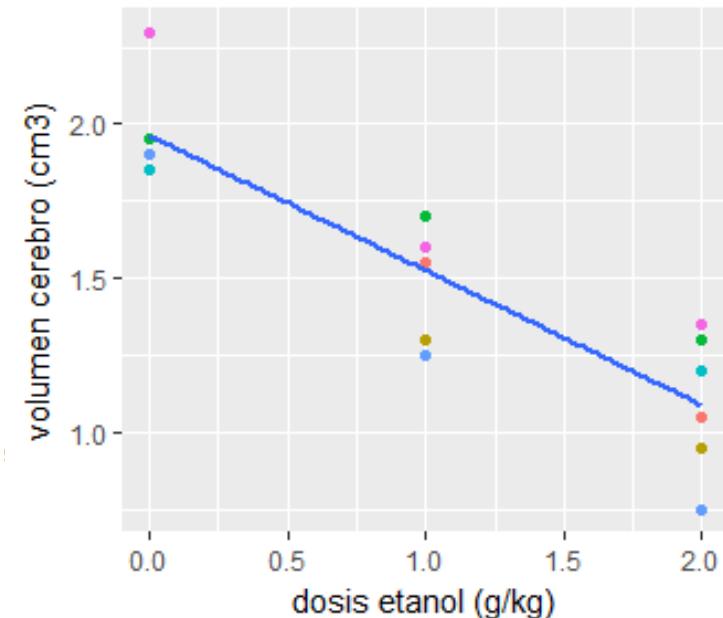
$$\sqrt{\hat{\sigma}^2}$$

```
> anova(m0)
```

Analysis of Variance Table

Response: vol

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
etanol	1	2.29688	2.29688	65.116	4.956e-07 ***
Residuals	16	0.56437	0.03527		$\hat{\sigma}^2$



La variabilidad explicada por las camadas va a parar al error!



Incorporando las camadas como VE de efectos fijos

10

BLUE: best linear unbiased estimate.
Efecto fijo de la camada

$$Y_i = \beta_0 + \beta_1 etanol_i + \beta_2 C_{2i} + \beta_3 C_{3i} + \dots + \varepsilon_i$$

```
m1<-lm(vol~etanol+factor(camada), ratones)
```

coefficients:

	Estimate	Std. Error	t value	Pr(> t)
etanol	-0.43750	0.03387	-12.92	5.44e-08 ***
camada1	1.95417	0.07574	25.80	3.43e-11 ***
camada2	1.82083	0.07574	24.04	7.37e-11 ***
camada3	2.08750	0.07574	27.56	1.68e-11 ***
camada4	1.98750	0.07574	26.24	2.85e-11 ***
camada5	1.73750	0.07574	22.94	1.22e-10 ***
camada6	2.18750	0.07574	28.88	1.01e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.1173 on 11 degrees of freedom

Multiple R-squared: 0.9966, Adjusted R-squared: 0.9945

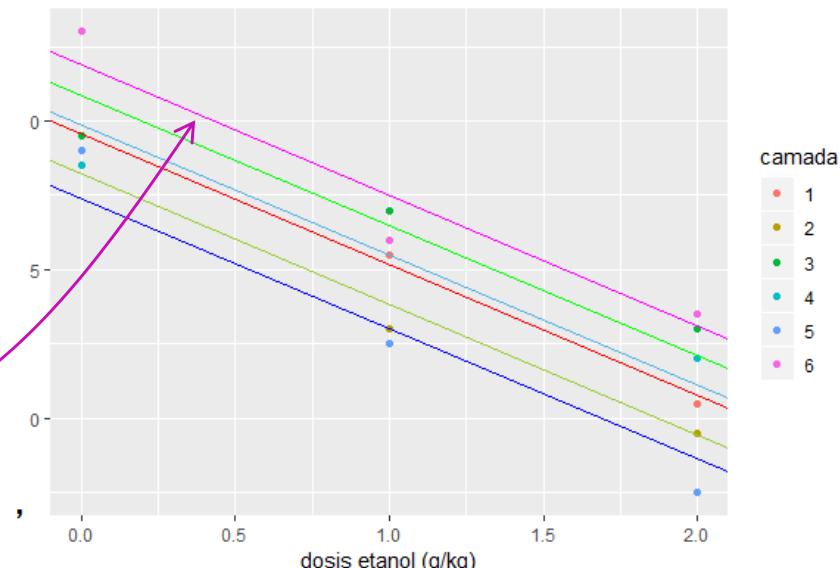
F-statistic: 462.4 on 7 and 11 DF, p-value: 1.085e-12

```
> anova(m2)
```

Analysis of Variance Table

Response: vol

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
etanol	1	2.29688	2.29688	166.8157	5.444e-08 ***
camada	5	0.41292	0.08258	5.9978	0.006458 **
Residuals	11	0.15146	0.01377		



- ✓ Mejora la precisión ✓
- ✓ Muchos parámetros ✗ para estimar
- ✓ Inferencia solo para ✗ los bloques medidos

Es la forma de evaluar si hay efecto de la camada

Modelo condicional: camadas como VE de efectos aleatorios

BLUP: best linear unbiased prediction.
Efecto aleatorio de la camada

11

Parte fija Parte aleatoria

$$Y_{ij} = \beta_0 + \beta_1 \text{etanol}_i + B_j + \varepsilon_{ij}$$

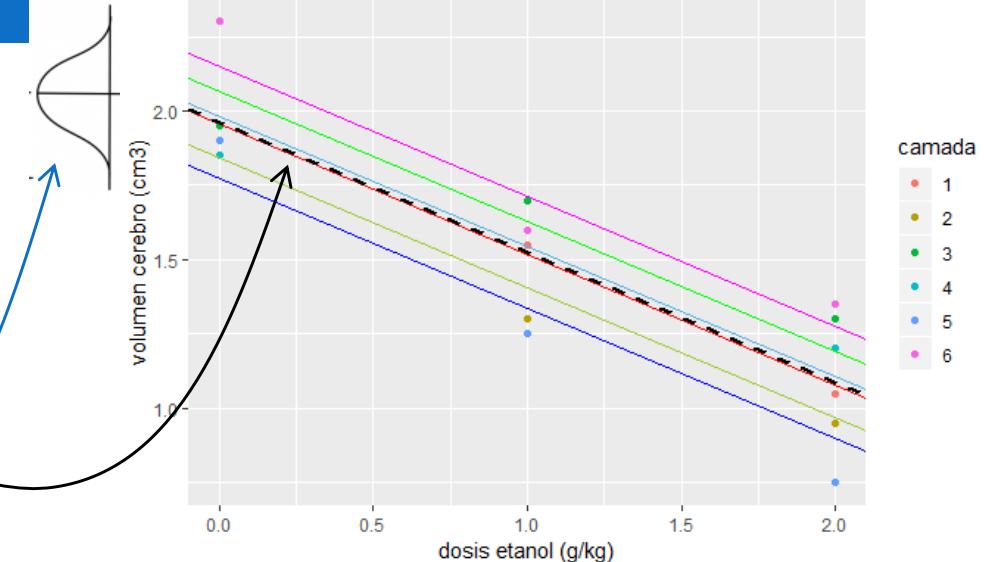
$i = 1 a 3, j = 1 a 6$

$$\varepsilon_{ij} \approx NID(0, \sigma^2)$$

$$B_j \approx NID(0, \sigma^2_{\text{camadas}})$$

ε_{ij}, B_j indep

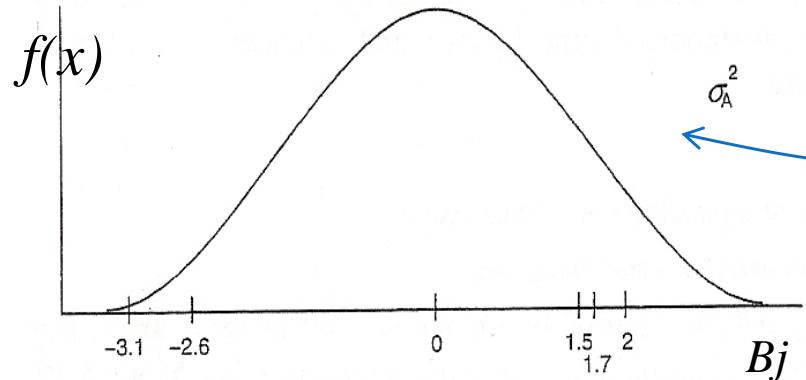
$$Var(Y) = \sigma^2_{\text{camadas}} + \sigma^2$$



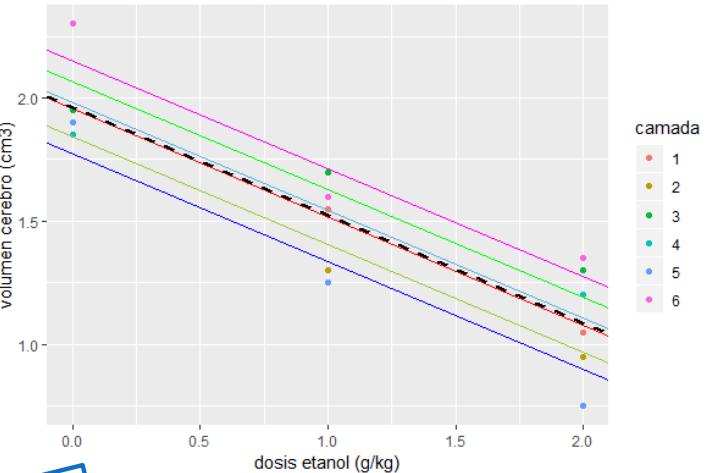
- **Modelo condicional o sujeto-específico:** proporciona un modelo para cada individuo, pero donde la forma general del modelo es la misma para cada sujeto. En este caso es un modelo de intercepto aleatorio, ya que la ordenada al origen es $\beta_0 + B_j$, que cambia aleatoriamente según B_j
- La inclusión de B_j (efecto aleatorio de la camada) induce una estructura de correlación entre las observaciones de la misma camada ✓
- La inferencia no está restringida a las 6 camadas estudiadas sino a toda la población de camadas ✓
- Un único parámetro $\sigma^2_{\text{camadas}}$ para modelar la variabilidad entre camadas ✓

B_j : Efecto aleatorio del nivel j de la VE Camada (BLUP: best linear unbiased prediction)

$$B_j \approx N(0, \sigma_{camada}^2)$$



Cuanto sube o baja la ordenada al origen de la camada j-ésima



- ✓ Suponemos que existe un gran número de niveles para el factor camada y por tanto una población de efectos B_j
- ✓ B_j es una variable aleatoria, que se considera normal e independientemente distribuida, con media 0 y varianza σ_{camada}^2
- ✓ Se seleccionan al azar niveles con el propósito de tratarlos como una representación de la población de efectos hacia la cual se pretende inferir.
- ✓ Los B_j se predicen I (no se estiman ya que no son parámetros sino VA)
- ✓ Se estima la variabilidad aportada por las camadas σ_{camada}^2

```
library(lme4)
m2 <- lmer(vol ~ etanol + (1 | camada), data = ratones)
```

- intercepto aleatorio por camada
- camada como factor

REML criterion at convergence: -8.8

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.2784	-0.8701	0.1124	0.6304	1.2784

Random effects:

Groups	Name	Variance	Std.Dev.
camada	(Intercept)	0.02294	0.1515
Residual		0.01377	0.1173

Number of obs: 18, groups: camada, 6

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)	
(Intercept)	1.96250	0.07573	7.59765	25.91	1.09e-08	*** $\hat{\beta}_0$
etanol	-0.43750	0.03387	11.00001	-12.92	5.44e-08	*** $\hat{\beta}_1$

Las estimaciones son por máxima verosimilitud restringida en lugar de máxima verosimilitud

de parámetros?

GL residuales?

library(lmerTest)
para obtener p-valores

Para ver significación de los coeficientes por test de Wald puede usarse la función Anova(modelo) del paquete car

```
#También puede utilizarse la librería nlme
library(nlme)
m2b <- lme(vol ~ etanol, random= ~1 | camada, data = ratones)
```

Linear mixed-effects model fit by REML

Data: ratones

AIC	BIC	logLik
-0.8271654	2.263189	4.413583

Random effects:

Formula: ~1 | camada

(Intercept)	Residual
-------------	----------

StdDev: 0.1514534 0.1173411

$$\hat{\sigma}_{camadas} \text{ y } \hat{\sigma}$$

Fixed effects: vol ~ etanol

	value	Std.Error	DF	t-value	p-value
(Intercept)	1.9625	0.07573227	11	25.91366	0
etanol	-0.4375	0.03387346	11	-12.91572	0

$$\hat{\beta}_0$$

$$\hat{\beta}_1$$

Correlation:

(Intr)

etanol -0.447

Standardized Within-Group Residuals:

Min	Q1	Med	Q3	Max
-1.2784414	-0.8700553	0.1124409	0.6303889	1.2784414

Number of Observations: 18

Number of Groups: 6

.

Según Pinheiro y Bates (2000) si los GL residuales son pequeños (<5?) los p no son confiables

Estimación de los parámetros en modelos lineales mixtos

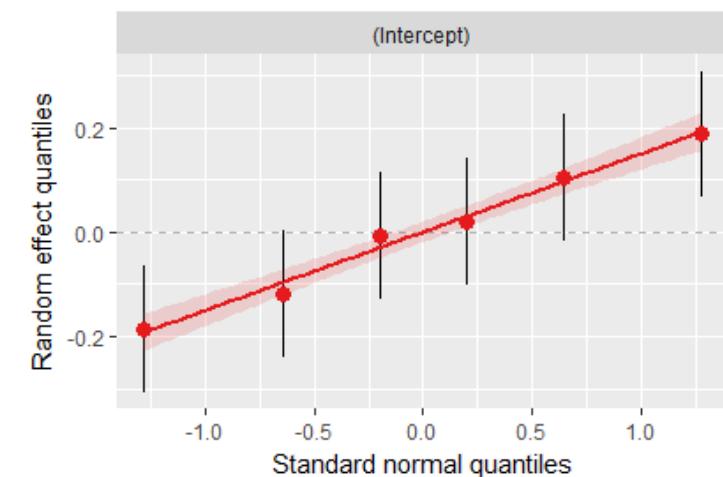
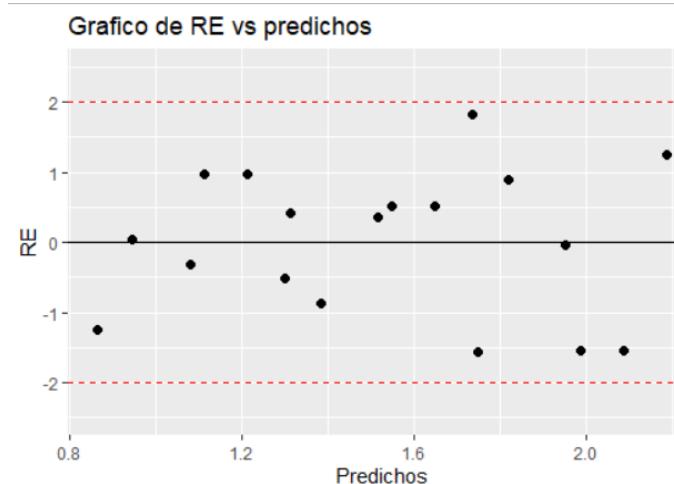
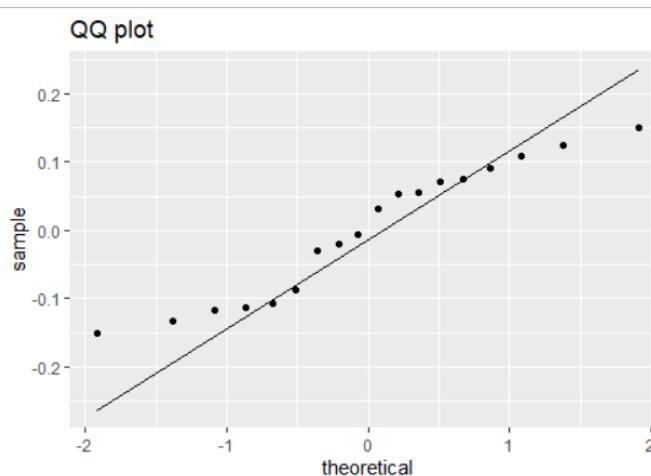
Los modelos con distribución normal y VE con efectos fijos y aleatorios se conocen como **modelos lineales generales y mixtos**. Dos métodos de estimación:

- **Mínimos cuadrados**: Si se verifica distribución normal de los errores, homocedasticidad y diseño sin desbalanceo grosero permiten modelar la falta de independencia. Pero debe indicarse cuál es el término de error para la construcción de las F, lo que lo torna un método tedioso en diseños complicados
- **Máxima verosimilitud**: Más versátil. Permite modelar distintas estructuras de correlación, heterocedasticidad, se bancan desbalanceo. Si n es grande, las estimaciones por MV proveen estimadores insesgados y consistentes. Sin embargo, las estimaciones de las varianzas están subestimadas (porque no corrigen por GL), por lo que en la práctica se aplica una corrección a los GL: **MV restringida o REML** (opción por defecto en Ime y Imer)

Supuestos

16

- Muestra aleatoria
- Linealidad para predictoras continuas
- $\varepsilon_{ij} \approx NID(0, \sigma^2)$; $B_j \approx NID(0, \sigma^2_{camadas})$; ε_{ij} y B_j independientes
- Además: ausencia de patrones en los residuos; chequeo de outliers



Shapiro-wilk normality test
data: e
W = 0.92747, p-value = 0.1752

Debería probarse también
normalidad con los B_j ,
aunque son pocos valores...

Parte fija

17

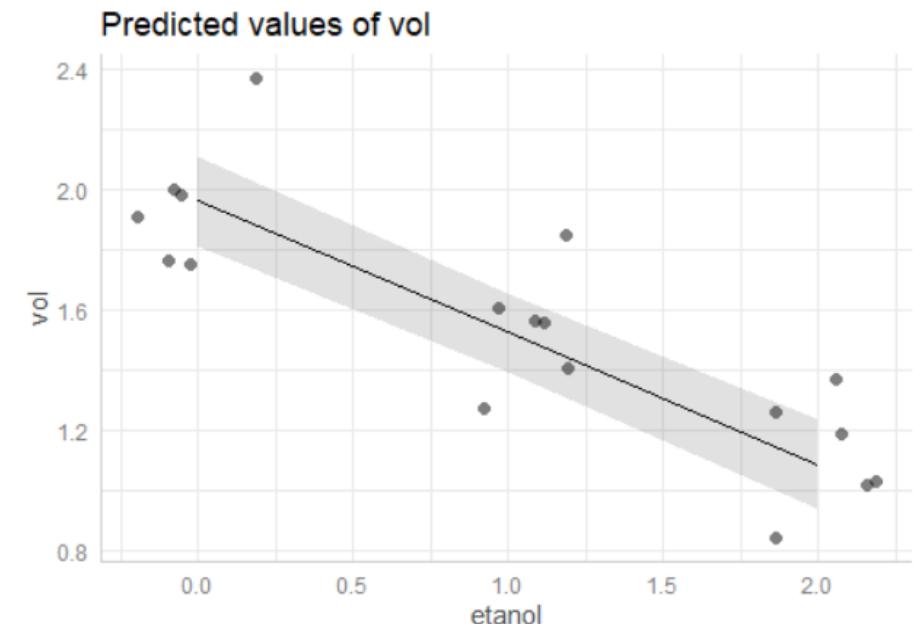
- Predicciones (sin incluir efectos aleatorios)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 etanol_i$$

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.96250	0.07573	25.91
etanol	-0.43750	0.03387	-12.92

- Interpretación de los coeficientes?
- Estimación de la respuesta media para las tres dosis de etanol?



```
> intervals(m2b)
Approximate 95% confidence intervals
```

Fixed effects:

	lower	est.	upper
(Intercept)	1.795814	1.9625	2.129186
etanol	-0.512055	-0.4375	-0.362945

Parte aleatoria

Efectos aleatorios

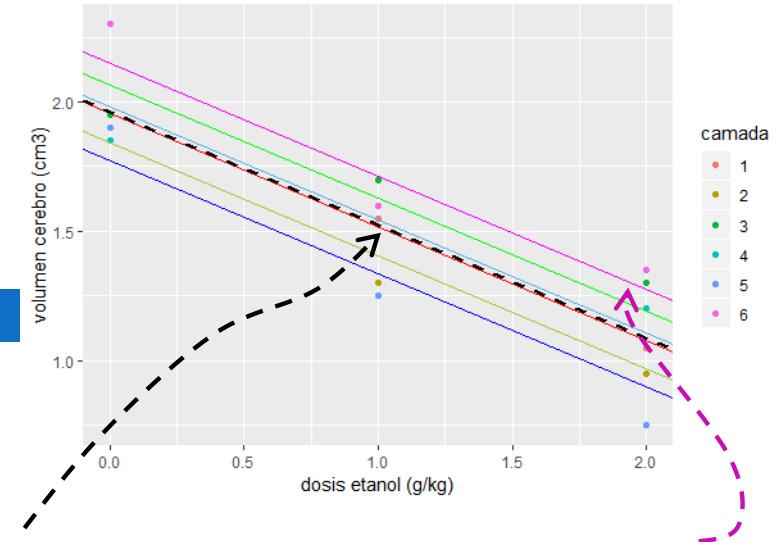
18

```
(Intercept) etanol
1 1.955556 -0.4375
2 1.844453 -0.4375
3 2.066659 -0.4375
4 1.983332 -0.4375
5 1.775014 -0.4375
6 2.149986 -0.4375
```

1.96250

```
ranef(m2) Bj
$camada
1 -0.0069
2 -0.1180
3 0.1042
4 0.0208
5 -0.1875
6 0.1875
```

	etanol	camada	vol	pred fija	efecto aleat	pred fija + aleat
1	0	1	1.95	1.9625	-0.0069	1.9556
2	0	2	1.90	1.9625	-0.1180	1.8445
3	0	3	1.95	1.9625	0.1042	2.0667
4	0	4	1.85	1.9625	0.0208	1.9833
5	0	5	1.90	1.9625	-0.1875	1.7750
6	0	6	2.30	1.9625	0.1875	2.1500
7	1	1	1.55	1.5250	-0.0069	1.5181
8	1	2	1.30	1.5250	-0.1180	1.4070
9	1	3	1.70	1.5250	0.1042	1.6292
10	1	4	1.60	1.5250	0.0208	1.5458
11	1	5	1.25	1.5250	-0.1875	1.3375
12	1	6	1.60	1.5250	0.1875	1.7125
13	2	1	1.05	1.0875	-0.0069	1.0806
14	2	2	0.95	1.0875	-0.1180	0.9695
15	2	3	1.30	1.0875	0.1042	1.1917
16	2	4	1.20	1.0875	0.0208	1.1083
17	2	5	0.75	1.0875	-0.1875	0.9000
18	2	6	1.35	1.0875	0.1875	1.2750



Predicciones
marginales

Predicciones
sujeto específicas

Parte aleatoria

Efectos aleatorios (BLUPs)

19

Fixed effects:

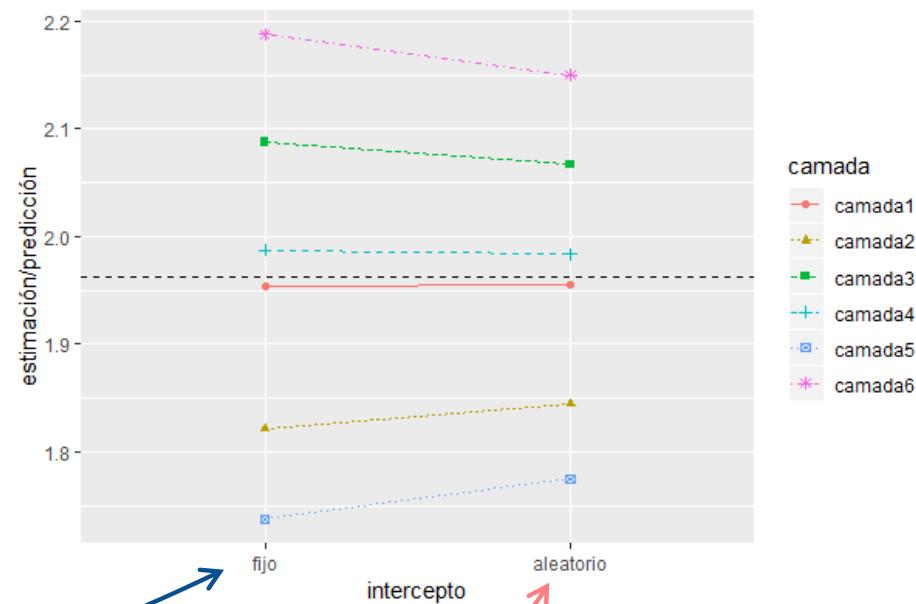
	Estimate	Std. Error	t value
(Intercept)	1.96250	0.07573	25.91
etanol	-0.43750	0.03387	-12.92

Los interceptos fijos y aleatorios difieren:

camada	interc_fijo	interc_aleat
camada1	1.954167	1.955556
camada2	1.820833	1.844453
camada3	2.087500	2.066659
camada4	1.987500	1.983332
camada5	1.737500	1.775014
camada6	2.187500	2.149986

y por lo tanto los efectos de las camadas:

camada	BLUE	BLUP
camada1	-0.008333333	-0.006943935
camada2	-0.141666667	-0.118046892
camada3	0.125000000	0.104159022
camada4	0.025000000	0.020831804
camada5	-0.225000000	-0.187486240
camada6	0.225000000	0.187486240



Los interceptos aleatorios están “encogidos” (shrinkage), es decir son más parecidos a la media. Es el costo de ampliar la inferencia

Parte aleatoria

Componentes de varianza

- Se denomina así a cada varianza que contribuye a la varianza aleatoria de Y y mide la variabilidad aportada por cada VE con efectos aleatorios- En este caso

$$\sigma_{Y_{ij}}^2 = \sigma_{camadas}^2 + \sigma^2$$

Random effects:

Groups	Name	Variance	Std.Dev.
camada	(Intercept)	0.02294	0.1515
Residual		0.01377	0.1173

$$\hat{\sigma}_{Y_{ij}}^2 = 0,02294 + 0,01377 = 0,03671$$

$$\hat{\sigma}_{Y_{ij}} = 0,19\text{cm}^3$$

> intervals(m2b)

Approximate 95% confidence intervals

Random Effects:

Level: camada

	lower	est.	upper
sd((Intercept))	0.07164942	0.1514534	0.3201442

Within-group standard error:

	lower	est.	upper
0.07726336	0.11734111	0.17820783	

- La estructura de correlación inducida por el factor aleatorio se denomina **coeficiente de correlación intraclasa**. Se calcula como

$$CCI = \frac{\sigma_{camada}^2}{\sigma_{camada}^2 + \sigma^2}$$

- 62% de la variación aleatoria (no impuesta por el tratamiento) en el volumen cerebral de los ratones está explicado por la variación entre camadas

Comparación de modelos

21

Prueba de devianza o cociente de verosimilitudes (LRT)

- Solo para modelos anidados, con distinta estructura de la parte fija o aleatoria

$$D = -2 \ln \frac{\mathcal{L}(\text{modelo 1})}{\mathcal{L}(\text{modelo 2})} \sim \chi^2_{p_2-p_1}$$

donde m_1 anidado en m_2
 p_1, p_2 = número de parámetros de cada modelo

- Para comparar modelos con distinta parte fija: `anova(modelo1, modelo2)`.
También puede usarse el comando `drop1(modelo2)`
- Para determinar la significancia del factor aleatorio `ranova(modelo2)`

```
#library(lmerTest)  
ranova(modelo2)
```

```
ANOVA-like table for random-effects: single term deletions  
  
Model:  
vol ~ etanol + (1 | camada)  
      npar logLik     AIC      LRT Df Pr(>Chisq)  
<none>       4 4.4136 -0.8272  
(1 | camada)  3 1.3663  3.2673 6.0945  1    0.01356 *  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparación por AIC/BIC

- No se recomienda comparar AIC calculados con distintos paquetes

Modelo marginal: estructura de correlación entre observaciones de una camada

22

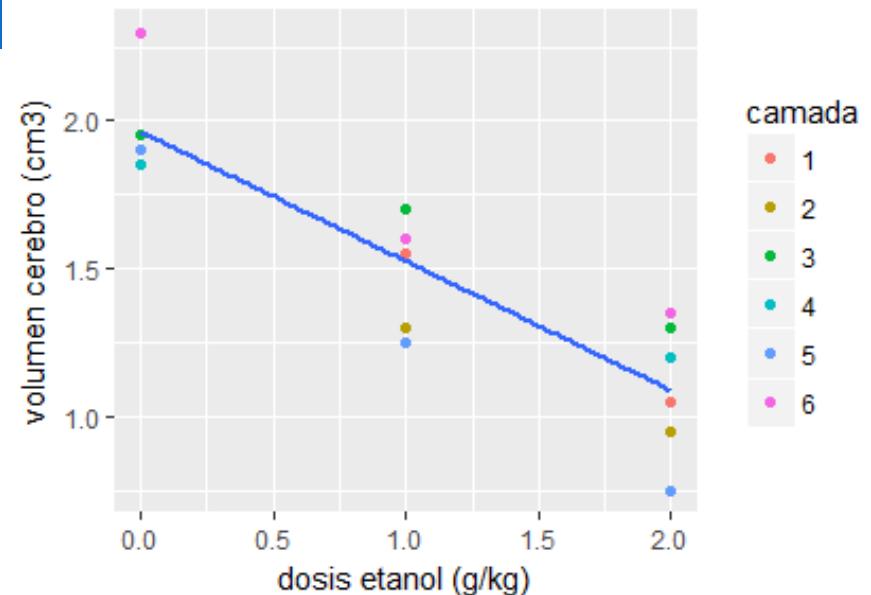
Parte fija

$$Y_{ij} = \boxed{\beta_0 + \beta_1 etanol_i} + \boxed{\varepsilon_{ij}}$$

$$i = 1 \text{ a } 3, j = 1 \text{ a } 6$$

$$\varepsilon_{ij} \approx NID(0, \boxed{\Sigma_j})$$

matriz de covarianza



- **Modelo marginal:** ajusta un modelo general para la estructura promedio de la población de individuos
- No incluye VE de efectos aleatorios
- La estructura de correlación entre las observaciones de la misma camada está explicitada en la matriz de covarianza

```
library(nlme)
m2c <- gls(vol ~ etanol, correlation=corCompSymm(form = ~ 1|camada),
data = ratones)
```

Estructura de correlación entre observaciones de una camada

```
Generalized least squares fit by REML
Model: vol ~ etanol
Data: ratones
      AIC      BIC    logLik
-0.8271654 2.263189 4.413583
```

Las estimaciones son por máxima verosimilitud restringida en lugar de máxima verosimilitud

```
Correlation Structure: Compound symmetry
Formula: ~1 | camada
Parameter estimate(s):
  Rho
0.6248966
```

Coefficients:

	value	Std.Error	t-value	p-value
(Intercept)	1.9625	0.07573224	25.91367	0
etanol	-0.4375	0.03387347	-12.91571	0

Residual standard error: 0.1915909
Degrees of freedom: 18 total; 16 residual

No descompone la varianza entre camadas y dentro de camadas

Modelo condicional vs modelo marginal

24

- Las estimaciones de la parte fija del modelo son las mismas!
- Con errores con distribución normal ambas aproximaciones son equivalentes. Es decir que postular un modelo mixto con factor aleatorio para la ordenada al origen implica ("es equivalente a") postular un modelo marginal con cierta estructura de correlación (simetría compuesta)
- Pero los modelos marginales no incluyen efectos aleatorios y por lo tanto no estiman componentes de varianza
- La interpretación no es la misma:
 - **Modelo marginal:** las predicciones obtenidas son para individuos promedio dadas características fijas (sexo, edad, etc)
 - **Modelo condicional:** las predicciones obtenidas son para individuos promedio dadas características fijas **y aleatorias** (sujeto-específica)

¿Cuándo tratar a una VE como de efectos aleatorios?

25

- Si no me interesa poner a prueba hipótesis entre los niveles de la variable
- Si no me interesa estimar magnitud del efecto
- Si quiero cuantificar la variabilidad entre los niveles de la variable
- Si es razonable suponer que los niveles analizados de la variable fueron aleatoriamente muestreados o son representativos de una población de niveles
- Si hay suficientes niveles del factor como para efectuar una estimación de confiable de la varianza de la población de los efectos (al menos 5)
- Si son los niveles del factor sólo etiquetas numéricas (no informativos)

Comentarios

26

- Es clave cómo se identifica a los niveles del factor aleatorio para determinar la estructura de agrupamiento. La recomendación es darle una identificación única a cada nivel
- Recordar que siempre son factores (VE categóricas)
- Es importante distinguir entre efectos aleatorios como control de la heterogeneidad del material experimental (como los bloques, incorporados al diseño experimental) y efectos aleatorios de interés (como en estudios genéticos evolutivos o en estudios ecológicos enfocados en la heterogeneidad). En el primer caso se deberían dejar en el modelo, independientemente de la cantidad de niveles y la significación. En el segundo caso podemos aplicar selección de modelos para decir si son relevantes o no
- Otra forma de lidiar con la heterocedasticidad es agregar como VE de afectos aleatorios al individuo (por más que existe una única observación)

Experimentos multiambientales

27

- Muchas veces un experimento se conduce en varios ambientes, donde los ambientes elegidos intentan representar una población relativamente mayor de ambientes. Dentro de cada ambiente se evalúan generalmente dos o más tratamientos bajo un cierto diseño experimental con o sin repeticiones. Los modelos posibles son:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (1)$$

$$Y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ij} \quad (2)$$

- los efectos de ambiente (β_j) podrían ser considerados como fijos o aleatorios según los supuestos que se hagan respecto a los ambientes incorporados en el experimento
- El modelo (1) corresponde a un DBA sin interacción
- El modelo (2) corresponde a un modelo con interacción tratamiento x ambiente y es necesario tener réplicas para poder estimarlo. Si los efectos del ambiente son considerados aleatorios, la interacción también lo es
- Ej: interacción genotipo-ambiente