Biometría

Modelando la heterocedasticidad

Ejemplo: Los vertebrados terrestres expuestos a metales pesados pueden presentar bioacumulación de dichos contaminantes en diferentes tejidos.

Las ratas (genero *Rattus*) que viven en áreas urbanas están expuestas a los metales pesados. Además, como presentan áreas de actividad relativamente pequeñas, se ha sugerido que estos roedores pueden ser usados para la detección de contaminación ambiental por metales pesados y de los riesgos de salud humanos asociados.

Un grupo de investigadores registró el nivel de acumulación de plomo en huesos de 143 ratas (*Rattus norvegicus*) provenientes de distintos ambientes de la Ciudad de Buenos Aires. Se analizaron fémures los cuales fueron digeridos con ácido nítrico y posteriormente llevados al Laboratorio de Química Analítica del Centro Atómico de Ezeiza (CEA) de la Comisión Nacional de Energía Atómica (CNEA), para la determinación de la concentración del plomo. El objetivo del trabajo fue comparar el nivel de acumulación media de plomo en fémures de ratas capturadas en 4 ambientes: *Espacios verdes; Barrios residenciales; Barrios carenciados* y *Costa del Riachuelo*.

a) Identifique la variable respuesta, factores y niveles.
b) Escriba el modelo en parámetros y en términos del problema.
c) Analice el cumplimiento de los supuestos del modelo.
d) Concluya en relación a las diferencias en la acumulación de plomo en ratas de distintos ambientes.

Caso	Ambiente	Pb
1	Esp verdes	2,48
2	Esp verdes	2,94
3	Esp verdes	2,77
4	Esp verdes	2,40
5	Esp verdes	2,89
6	Esp verdes	2,40
7	Esp verdes	2,08
8	Esp verdes	2,89
9	Esp verdes	2,94
10	Esp verdes	2,64
11	Esp verdes	2,64
…		
143		

Variables

Dependiente o respuesta:

Independientes:

Modelo:

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

Supuestos de un ANOVA de 1 factor:

Esp. verdes	B carenciados	Residencial	Riachuelo
2,48	3,87	2,64	4,70
2,94	2,64	3,53	4,32
2,77	4,06	2,71	3,53
2,40	2,94	3,22	5,19
2,89	5,50	3,71	3,26
2,40	4,85	4,19	4,32
2,08	2,30	2,94	4,85
2,89	5,19	3,71	1,79
2,94	4,06	1,39	5,41
2,64	2,71	5,02	4,52
2,64	2,30		3,64
1,95	2,94		5,88
3,18	2,94		1,39
3,71	4,06		4,52
...	...		...

ε_{ij} *independientes*

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

Los ε_{ij} se estiman con los residuos e_{ij}

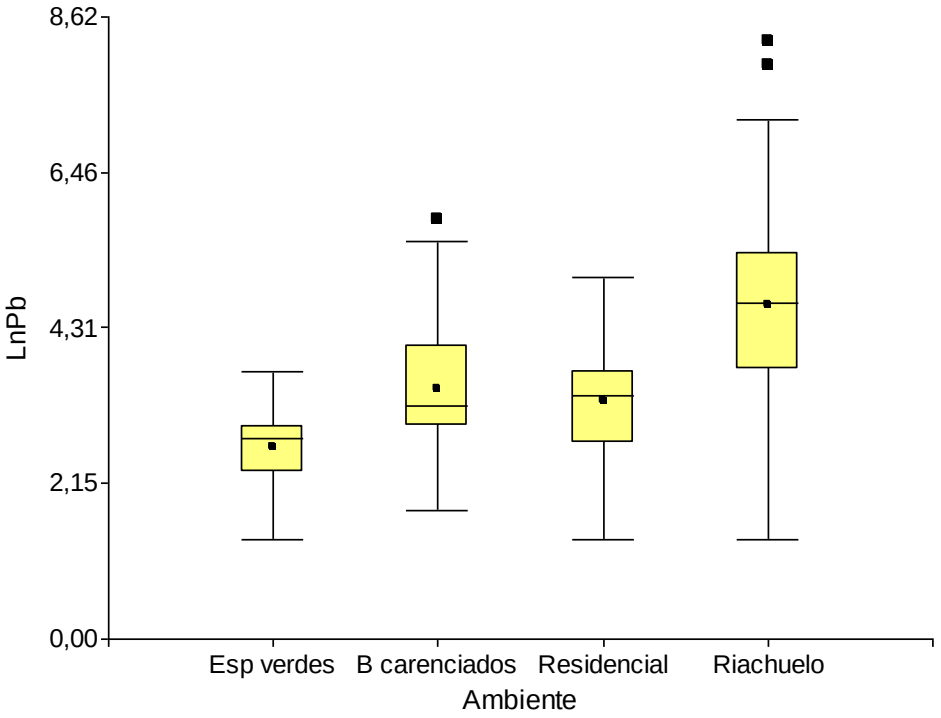
Estadística descriptiva

Variable respuesta: Pb

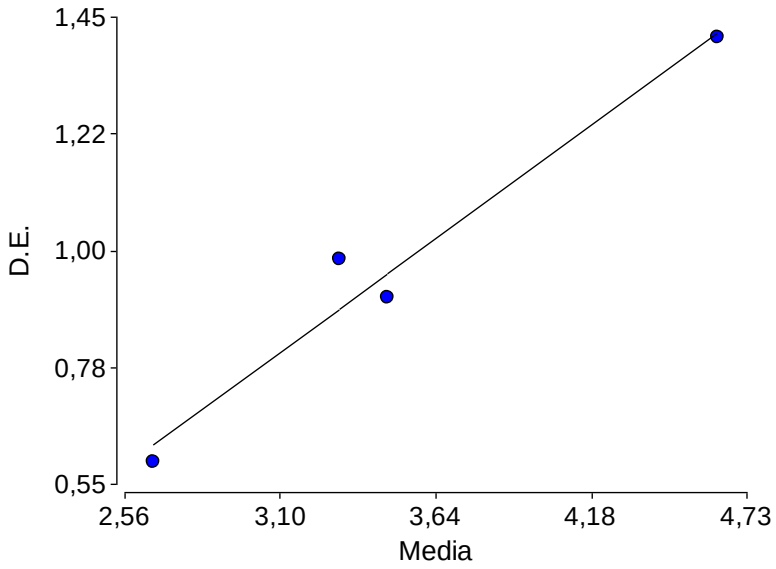
Medidas resumen

Ambiente	Variable	n	Media	D.E.	Mín	Máx
B carenciados	Pb	40	3,48	0,91	1,79	5,81
Esp verdes	Pb	37	2,65	0,59	1,39	3,71
Residencial	Pb	10	3,31	0,98	1,39	5,02
Riachuelo	Pb	56	4,63	1,41	1,39	8,27

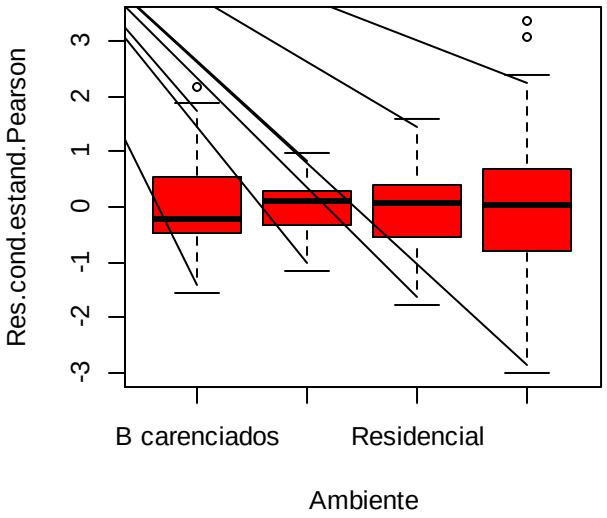
Gráfico de cajas (Box-plot)

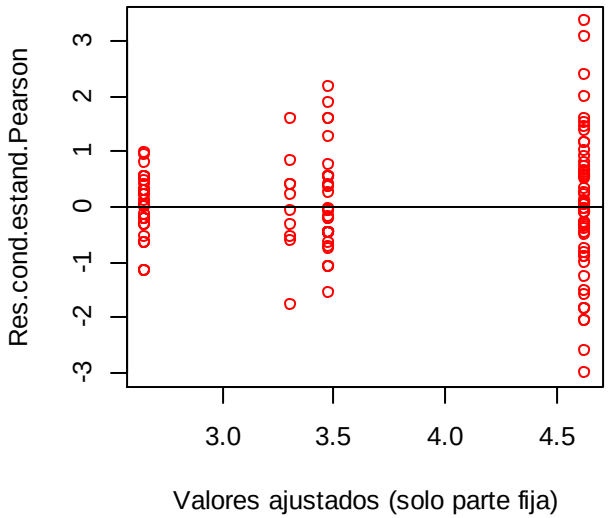


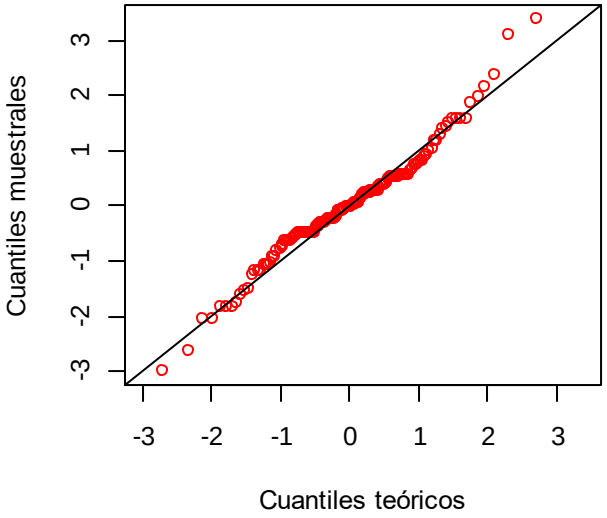
Relación Varianza-Media



Análisis de Residuos







Homocedasticidad

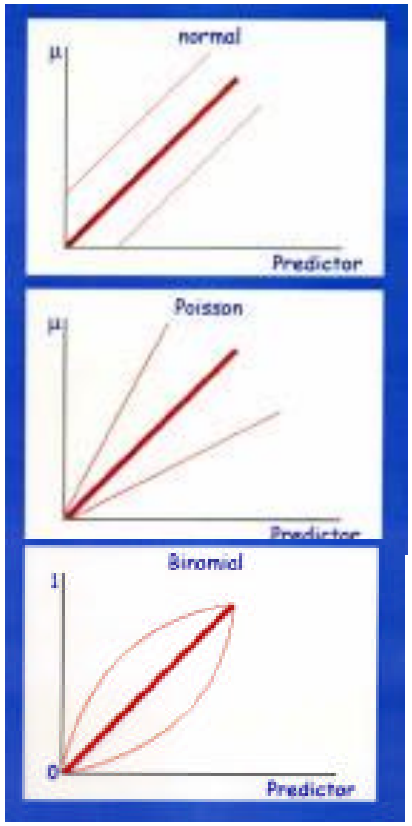
- Es uno de los supuestos más importantes
- Consiste en suponer que todos los tratamientos tienen la misma variabilidad σ^2 o alternativamente, que los errores tienen una variabilidad constante σ^2
- La violación al supuesto de igualdad de varianzas provoca:
  - Estimaciones erróneas de los EE de los tratamientos
  - Mayor probabilidad de cometer error tipo I
  - Las pruebas t o F no son válidas!
- Solución clásica: transformar. Pero la heterogeneidad puede implicar información biológica interesante!

Causas de heterocedasticidad

- Biológicas
- Presencia de outliers
- Distribución Poisson

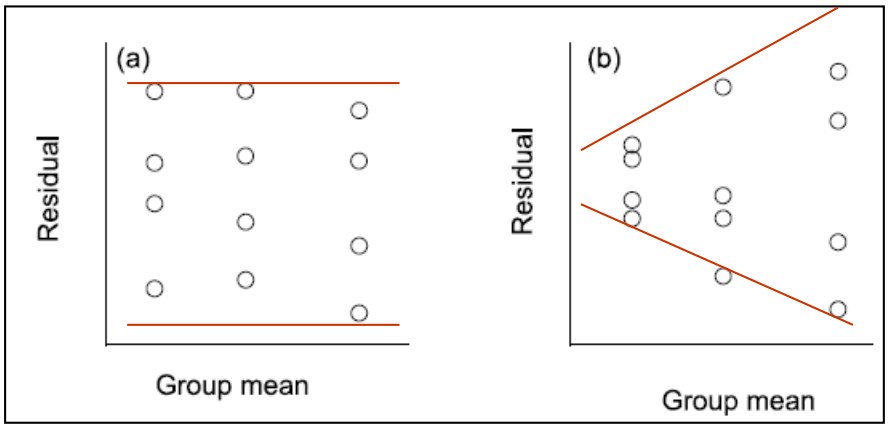
  (cantidad de eventos / continuo)
- Distribución binomial

  (proporción éxitos en una muestra de tamaño n)
- Otras distribuciones (gamma, log normal, etc)



Heterocedasticidad: ¿Cómo detectarla?

- Gráfico de residuos vs esperados o predichos. Se espera encontrar una distribución al azar y con variabilidad constante



- Pruebas analíticas: Prueba de Levene

Residuos

- Son fundamentales en el proceso de validación de los modelos
- Residuo e_{ij} es la diferencia entre el valor observado en *y* el valor esperado según el modelo

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

- Pero no son útiles cuando se modelan estructuras de varianza, ya que no cambian con las distintas estructuras
- Residuo estandarizado

$$e_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\sigma^2}}$$

O la función de varianza que corresponda

La buena noticia: podemos modelar la estructura de varianzas

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

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

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

Funciones de varianza

Identidad (varIdent) : una varianza distinta para cada grupo

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

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

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

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

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

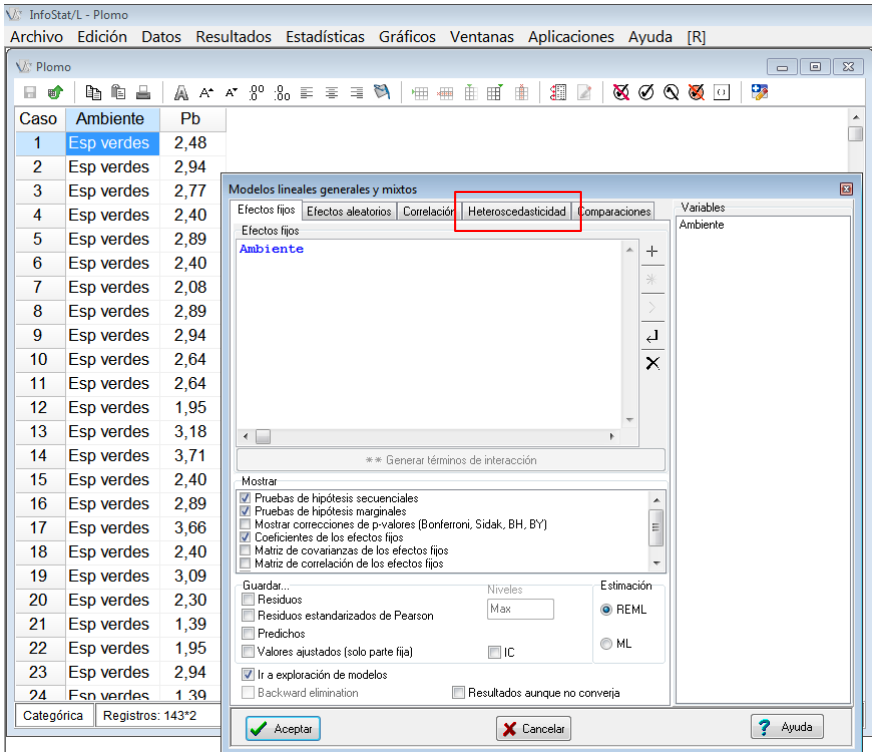
Potencia (varPower): la varianza como función de potencia de alguna covariable

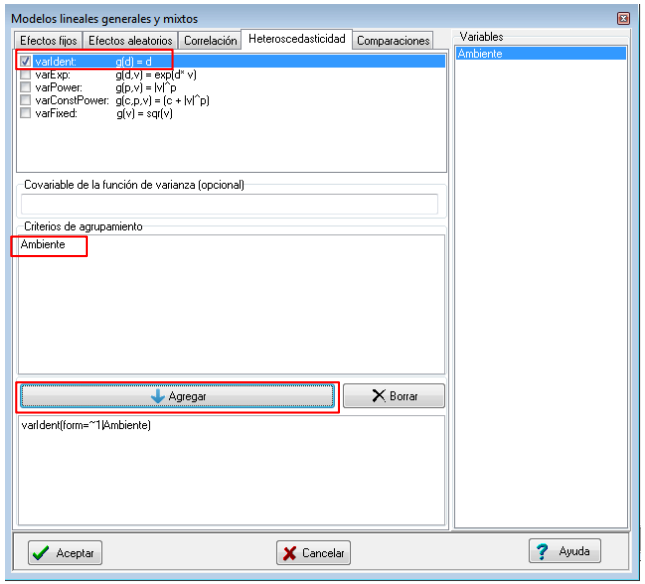
$$\varepsilon_i \sim N(0, \sigma^2 \cdot |X_i|^{2\delta})$$

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

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

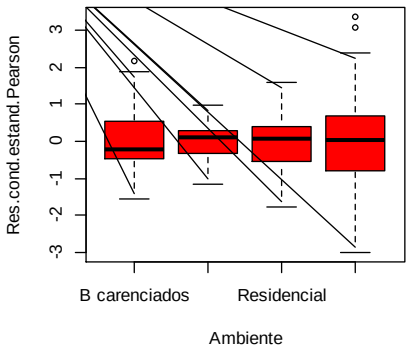
En Infostat

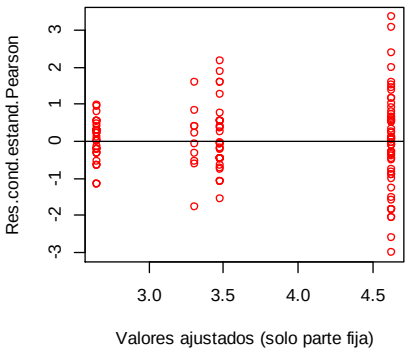

InfoStat/L - Plomo
Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Aplicaciones Ayuda [R]
Plomo
Caso Ambiente Pb
1 Esp verdes 2,48
2 Esp verdes 2,94
3 Esp verdes 2,77
4 Esp verdes 2,40
5 Esp verdes 2,89
6 Esp verdes 2,40
7 Esp verdes 2,08
8 Esp verdes 2,89
9 Esp verdes 2,94
10 Esp verdes 2,64
11 Esp verdes 2,64
12 Esp verdes 1,95
13 Esp verdes 3,18
14 Esp verdes 3,71
15 Esp verdes 2,40
16 Esp verdes 2,89
17 Esp verdes 3,66
18 Esp verdes 2,40
19 Esp verdes 3,09
20 Esp verdes 2,30
21 Esp verdes 1,39
22 Esp verdes 1,95
23 Esp verdes 2,94
24 Esp verdes 1,39
Categórica
Registros: 143*2
Modelos lineales generales y mixtos
Efectos fijos
Efectos aleatorios
Correlación
Heteroscedasticidad
Comparaciones
Variables
Ambiente
Efectos fijos
Ambiente
** Generar términos de interacción
Mostrar
Pruebas de hipótesis secuenciales
Pruebas de hipótesis marginales
Mostrar correcciones de p-valores (Bonferroni, Sidak, BH, BY)
Coeficientes de los efectos fijos
Matriz de covarianzas de los efectos fijos
Matriz de correlación de los efectos fijos
Guardar...
Residuos
Residuos estandarizados de Pearson
Predichos
Valores ajustados (solo parte fija)
Niveles
Max
IC
Estimación
REML
ML
Ir a exploración de modelos
Backward elimination
Resultados aunque no converja
Aceptar
Cancelar
Ayuda

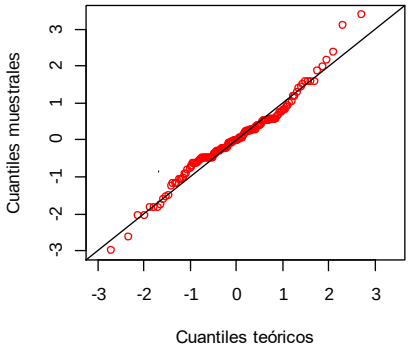

Modelos lineales generales y mixtos
Efectos fijos
Efectos aleatorios
Correlación
Heteroscedasticidad
Comparaciones
Variables
Ambiente
varIdent: g(d) = d
varExp: g(d,v) = exp(d* v)
varPower: g(p,v) = |v|^p
varConstPower: g(c,p,v) = (c + |v|^p)
varFixed: g(v) = sq(v)
Covariable de la función de varianza (opcional)
Criterios de agrupamiento
Ambiente
Agregar
Borrar
varIdent(form=~1|Ambiente)
Aceptar
Cancelar
Ayuda

Exploración de Modelos

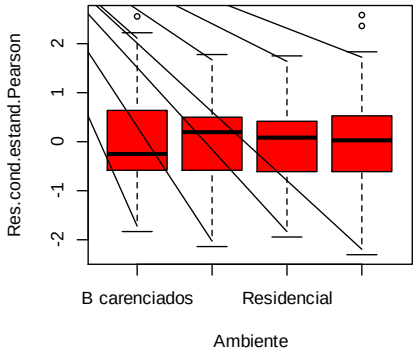
Modelo con σ constante

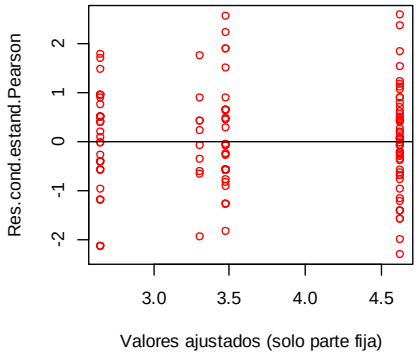


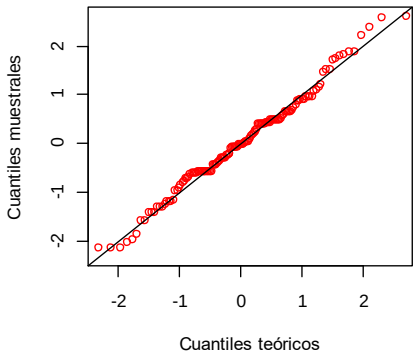




Modelo con σ_i (función varIdent)







Modelos lineales generales y mixtos

Modelo 2: función varIdent

Especificación del modelo en R

```
modelo.000_Pb_REML<-gls(Pb~1+Ambiente
,weights=varComb(varIdent(form=~1|Ambiente))
,method="REML"
,na.action=na.omit
,data=R.data00)
```

Resultados para el modelo: modelo.000_Pb_REML

Variable dependiente: Pb

$\hat{\sigma}_{Espacios\ verdes}$

Medidas de ajuste del modelo

N	AIC	BIC	logLik	Sigma	R2_0
143	416,41	439,88	-200,20	0,59	0,36

AIC y BIC menores implica mejor

Pruebas de hipótesis marginales (SC tipo III)

	numDF	F-value	p-value
(Intercept)	1	1217,97	<0,0001
Ambiente	3	30,95	<0,0001

Efectos fijos

	Value	Std.Error	t-value	p-value
(Intercept)	3,48	0,14	24,16	<0,0001
AmbienteEsp verdes	-0,82	0,17	-4,73	<0,0001
AmbienteResidencial	-0,17	0,34	-0,50	0,6195
AmbienteRiachuelo	1,15	0,24	4,87	<0,0001

Estructura de varianzas

Modelo de varianzas: varIdent
Formula: ~ 1 | Ambiente

Parámetros de la función de varianza

Parámetro	Estim
Esp verdes	1,00
B carenciados	1,53
Residencial	1,66
Riachuelo	2,37

"Espacios verdes" queda como referencia
Los estimadores están relativizados con respecto a "Esp verdes"

$$ej: \hat{\sigma}_{B\ carenciados} = 1{,}53 * \hat{\sigma}_{Esp\ verdes}$$

$$\hat{\sigma}_{B\ carenciados} = 1{,}53 * 0{,}59$$

Modelo 3: función varPower

Especificación del modelo en R

```
mlm.modelo.001_Pb_REML<-gls(Pb~1+Ambiente
,weights=varComb(varPower(form=~fitted(.)))
,method="REML"
,na.action=na.omit
,data=mlm.modeloR.data01)
```

Resultados para el modelo: mlm.modelo.001_Pb_REML

Variable dependiente: Pb

Medidas de ajuste del modelo

N	AIC	BIC	logLik	Sigma	R2_0
143	412.90	430.50	-200.45	0.14	0.36

AIC y BIC menores implica mejor

$\hat{\sigma}$

Pruebas de hipótesis tipo III - prueba

	Source	numDF	denDF	F-value	p-value
1	Ambiente	3	139	30.15	<0.0001

Estructura de varianzas

Modelo de varianzas: varPower

Formula: ~ fitted(.)

Parámetros de la función de varianza

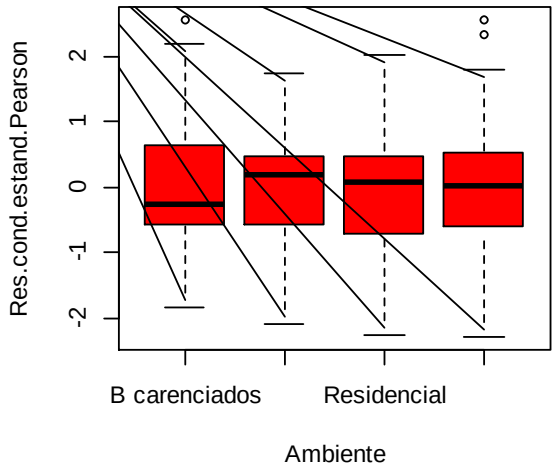
Parámetro	Estim
power	1.53

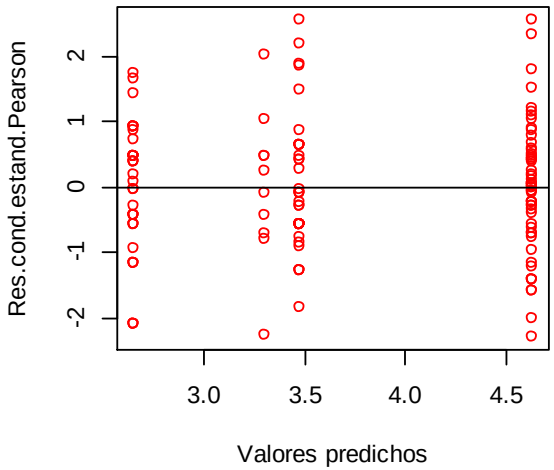
$\hat{\delta}$

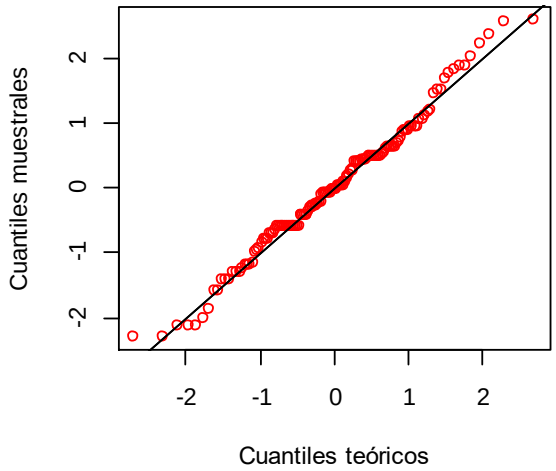
$$e_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{\sigma}^2 * |X_i|^{2*\hat{\delta}}}}$$

$$e_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{0{,}136^2 * |\hat{y}_i|^{2*1{,}53}}}$$

Modelo 3: función varPower







Modelo 4: función varExp

Especificación del modelo en R

```
mlm.modelo.000_Pb_REML<-gls(Pb~1+Ambiente
,weights=varComb(varExp(form=~fitted(.)))
,method="REML"
,na.action=na.omit
,data=mlm.modeloR.data00)
```

Resultados para el modelo: mlm.modelo.000_Pb_REML

Variable dependiente: Pb

Medidas de ajuste del modelo

N	AIC	BIC	logLik	Sigma	R2_0
143	413.37	430.98	-200.68	0.21	0.36

AIC y BIC menores implica mejor

$\hat{\sigma}$

Pruebas de hipótesis marginales (SC tipo III)

	numDF	F-value	p-value
(Intercept)	1	1465.11	<0.0001
Ambiente	3	29.15	<0.0001

Estructura de varianzas

Modelo de varianzas: varExp
Formula: ~ fitted(.)

Parámetros de la función de varianza

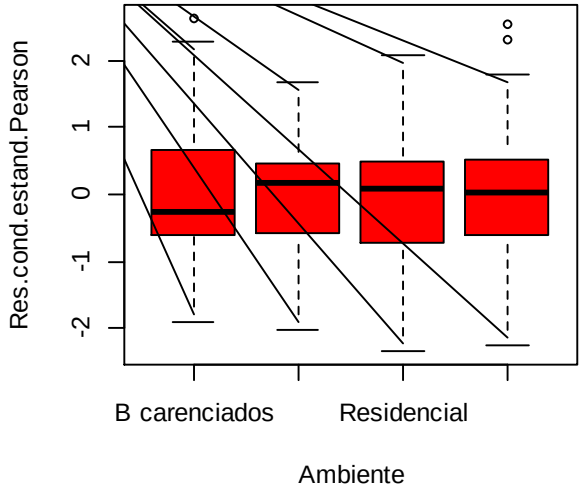
Parámetro	Estim
expon	0.42

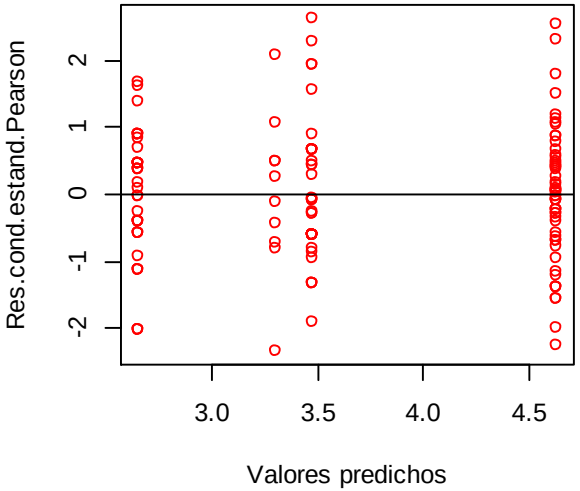
$\hat{\delta}$

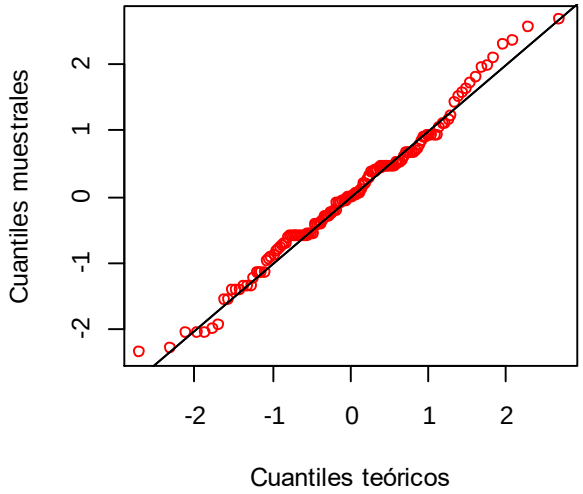
$$e_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{\sigma}^2 * e^{2*\hat{\delta}*X_i}}}$$

$$e_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{0.21^2 * e^{2*0.42*\hat{y}_i}}}$$

Modelo 4: función varExp



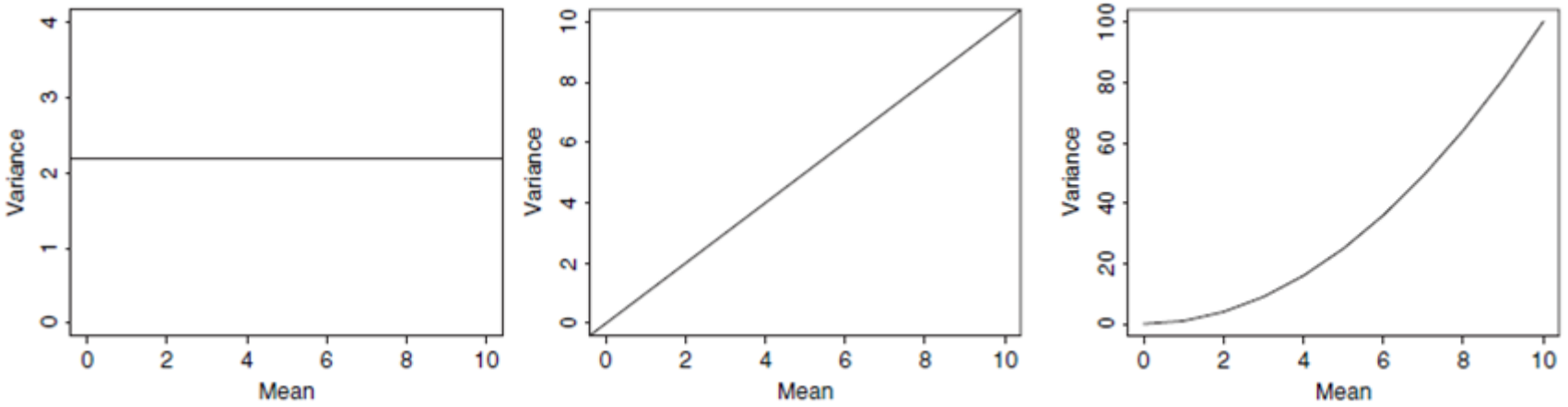




¿Cuál función utilizar?

- 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)
- 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)

Relación entre esperanza y varianza



Crawley 2007

Varios modelos posibles

	Residuos
Modelo 1: sin modelar varianzas. Se descarta por los residuos	**x**
Modelo 2: modelando varianzas por varIdent(ambiente)	**ok**
Modelo 3: modelando varianzas por varPower	**ok**
Modelo 4: modelando varianzas por varExp	**ok**

¿Cuál elegir?

Comparación de modelos

Criterios de información:

- Resumen la información de un modelo, teniendo en cuenta la función de verosimilitud $\mathcal{L}(\theta)$ (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

Criterios de información:

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

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

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

Medida del ajuste

Penalización por la complejidad del modelo

- 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

Varios modelos posibles

Modelo 1: sin modelar varianzas. Se descarta por los residuos
Modelo 2: modelando varianzas por varIdent(ambiente)
Modelo 3: modelando varianzas por varPower
Modelo 4: modelando varianzas por varExp

Comparación de modelos

	Model	df	AIC	BIC
~~mlm.modelo.000_Pb_REML~~	~~1~~	~~5~~	~~439.93~~	~~454.60~~
mlm.modelo.001_Pb_REML	2	8	416.41	439.88
mlm.modelo.002_Pb_REML	3	6	412.90	430.50
mlm.modelo.003_Pb_REML	4	6	413.37	430.98

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

Volvemos al ANOVA

Pruebas de hipótesis tipo III - prueba

	Source	numDF	denDF	F-value	p-value
1	Ambiente	3	139	30.15	<0.0001

P-valor< 0,05

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

Comparaciones

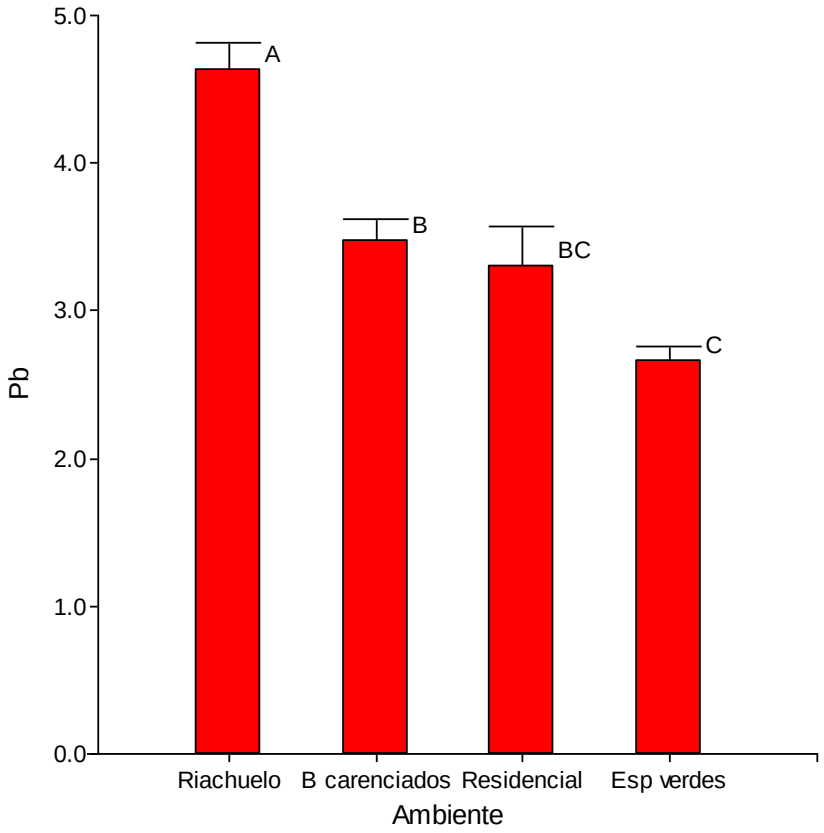
Pb - Medias ajustadas y errores estándares para Ambiente

LSD Fisher (Alfa=0.05)

Procedimiento de corrección de p-valores: Sidak

Ambiente	Medias	E.E.			
Riachuelo	4.63	0.19	A		
B carenciados	3.48	0.14		B	
Residencial	3.31	0.27		B	C
Esp verdes	2.65	0.10			C

Medias con una letra común no son significativamente diferente



Biometría

Anova como modelo lineal

Modelo lineal general

ANOVA vs **REGRESIÓN**

$$Y_i = \mu + \alpha_i + \varepsilon_i$$

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

ANOVA:

Las Var Exp o Indep. se denominan factores

Se las trata como cualitativas

Objetivo: Comparar Medias

REGRESIÓN:

Las Var Exp o Indep se denominan predictoras

Pueden ser cuantitativas o cualitativas (dummy)

Objetivo: Ajustar funciones, predecir la VR

ANOVA como GLM:

Variables auxiliares
o indicadoras o "Dummy"

Modelo de ANOVA

V Respuesta	V Exp	Xa	Xb	Xc
1.20	A	1	0	0
1.35	A	1	0	0
1.48	A	1	0	0
1.13	A	1	0	0
1.65	A	1	0	0
3.57	B	0	1	0
3.70	B	0	1	0
3.22	B	0	1	0
3.41	B	0	1	0
3.82	B	0	1	0
6.01	C	0	0	1
6.16	C	0	0	1
6.29	C	0	0	1
6.46	C	0	0	1
6.55	C	0	0	1

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1,2,3; \qquad j = 1,2,\ldots,15$$

$$Y_{ij} = \mu + \alpha_1 X_a + \alpha_2 X_b + \alpha_3 X_c + \varepsilon_{ij}$$

$$Y_{ij} = \underbrace{\mu + \alpha_1}_{} + \alpha_2 X_b + \alpha_3 X_c + \varepsilon_{ij}$$

$$Y_{ij} = \mu_1 + \acute{\alpha}_2 X_b + \acute{\alpha}_3 X_c + \varepsilon_{ij}$$

Var aleatoria cuantitativa

Var explicatoria cualitativa

Una de las variables auxiliares no aporta información novedosa ya que puede calcularse a partir de las otras dos

ANOVA como GLM:

Continuación…

$$Y_{ij} = \mu_1 + \alpha_2' X_b + \alpha_3' X_c + \varepsilon_{ij}$$

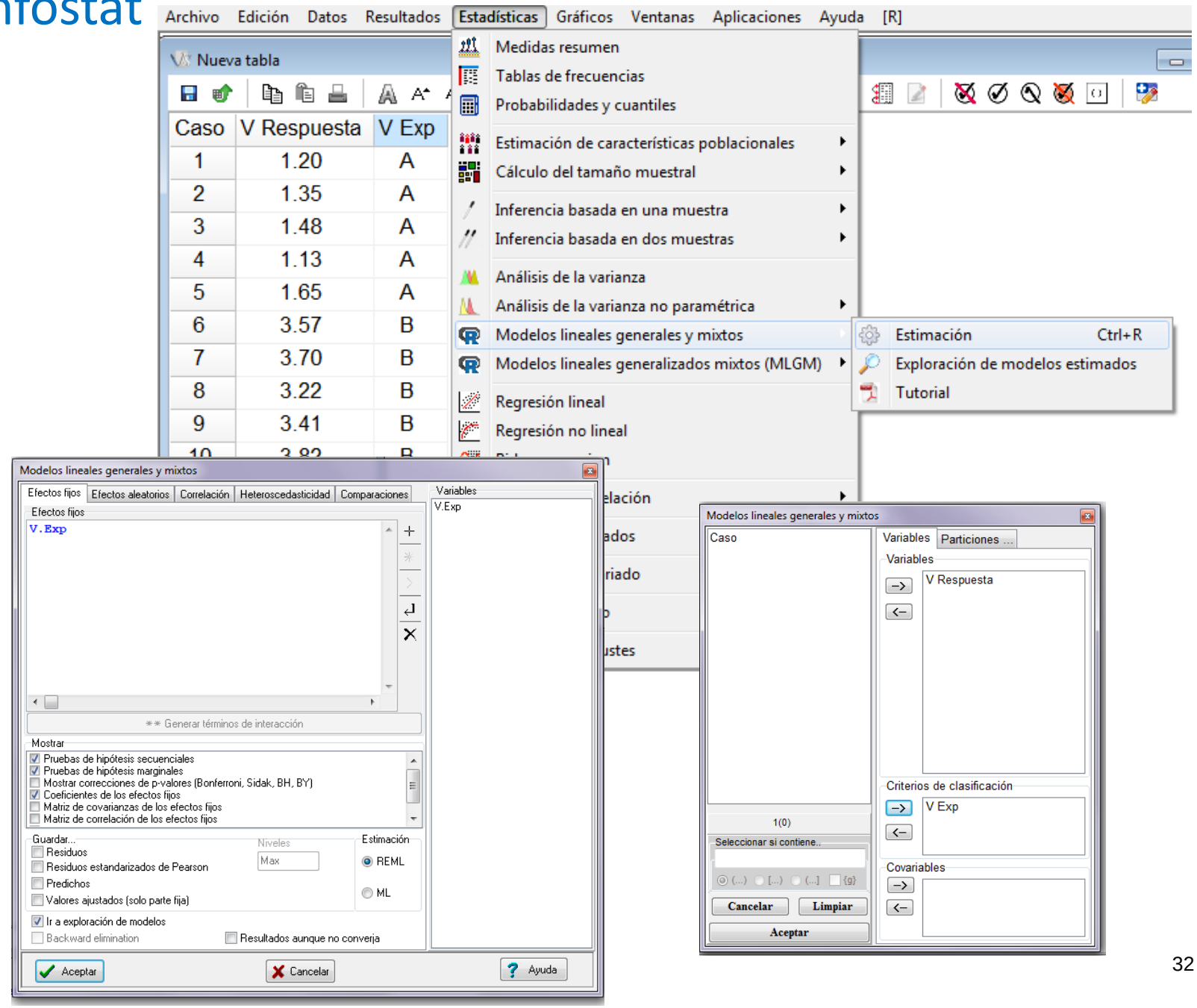
Ordenada al origen:
Valor Esperado de Y E(Y) cuando X_b y X_c valen cero. $\mu_1 = \mu_A$

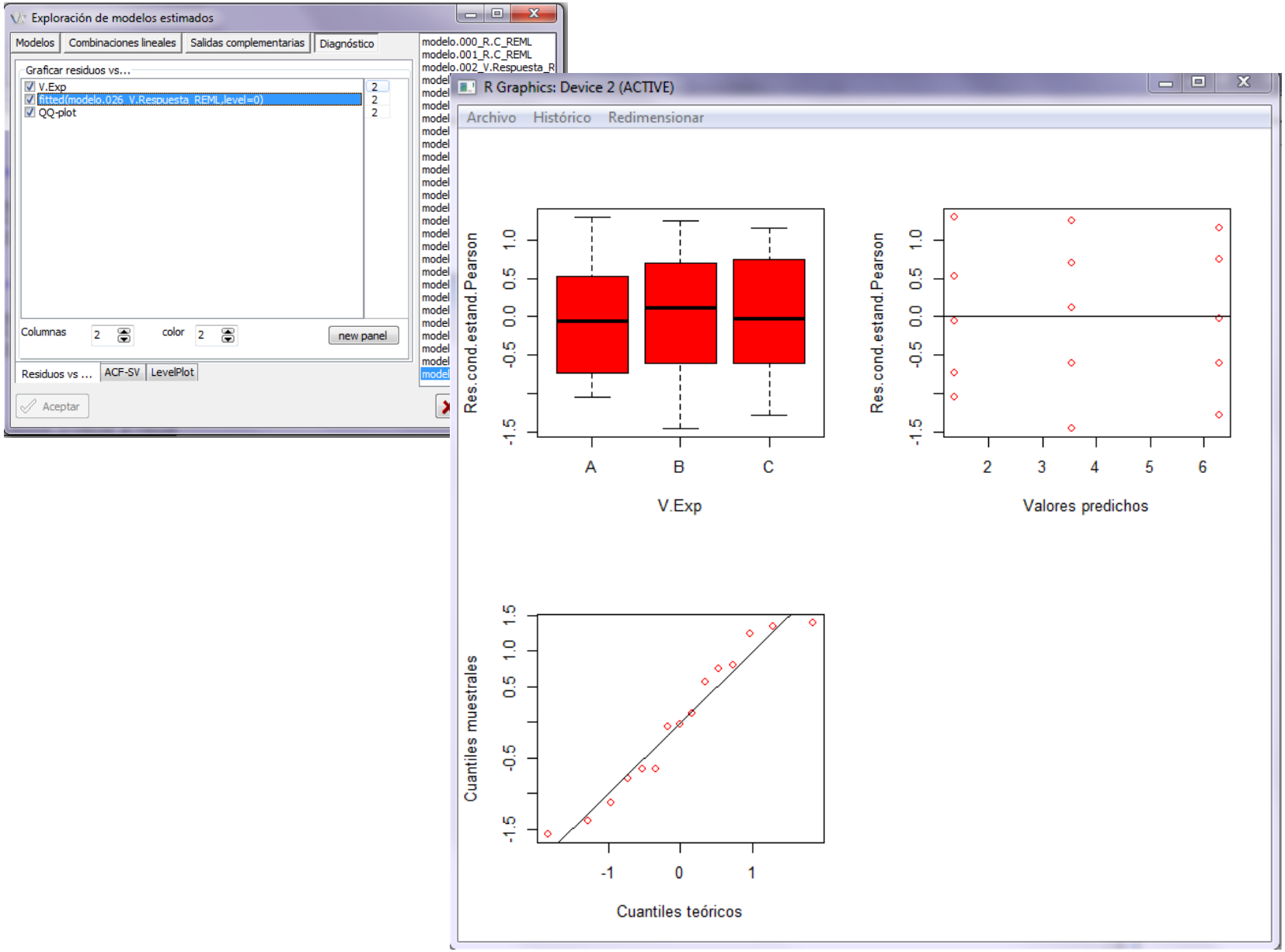
Efecto de los Tratamientos B y C con respecto al tratamiento de referencia (Tratamiento A)
$\alpha_2=\mu_2-\mu_1=\mu_B-\mu_A$ $\alpha_3=\mu_3-\mu_1=\mu_C-\mu_A$

$$Y_{ij} = \beta_0 + \beta_1 X_b + \beta_2 X_c + \varepsilon_{ij}$$

OJO: Si cambia el tratamiento de referencia, cambian los valores de los α

En Infostat


Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Aplicaciones Ayuda [R]
Nueva tabla
Caso V Respuesta V Exp
1 1.20 A
2 1.35 A
3 1.48 A
4 1.13 A
5 1.65 A
6 3.57 B
7 3.70 B
8 3.22 B
9 3.41 B
Medidas resumen
Tablas de frecuencias
Probabilidades y cuantiles
Estimación de características poblacionales
Cálculo del tamaño muestral
Inferencia basada en una muestra
Inferencia basada en dos muestras
Análisis de la varianza
Análisis de la varianza no paramétrica
Modelos lineales generales y mixtos
Modelos lineales generalizados mixtos (MLGM)
Regresión lineal
Regresión no lineal
Estimación Ctrl+R
Exploración de modelos estimados
Tutorial
Modelos lineales generales y mixtos
Efectos fijos Efectos aleatorios Correlación Heteroscedasticidad Comparaciones
Efectos fijos
V.Exp
** Generar términos de interacción
Mostrar
Pruebas de hipótesis secuenciales
Pruebas de hipótesis marginales
Mostrar correcciones de p-valores (Bonferroni, Sidak, BH, BY)
Coeficientes de los efectos fijos
Matriz de covarianzas de los efectos fijos
Matriz de correlación de los efectos fijos
Guardar...
Residuos
Residuos estandarizados de Pearson
Predichos
Valores ajustados (solo parte fija)
Niveles
Max
Estimación
REML
ML
Ir a exploración de modelos
Backward elimination
Resultados aunque no converja
Aceptar
Cancelar
Ayuda
Variables
V.Exp
Modelos lineales generales y mixtos
Caso
Variables Particiones ...
Variables
V Respuesta
Criterios de clasificación
V Exp
Covariables
1(0)
Seleccionar si contiene..
Cancelar
Limpiar
Aceptar


Exploración de modelos estimados
Modelos
Combinaciones lineales
Salidas complementarias
Diagnóstico
Graficar residuos vs...
V.Exp
fitted(modelo.026_V.Respuesta_REML,level=0)
QQ-plot
Columnas
color
new panel
Residuos vs ...
ACF-SV
LevelPlot
Aceptar
modelo.000_R.C_REML
modelo.001_R.C_REML
modelo.002_V.Respuesta_R
R Graphics: Device 2 (ACTIVE)
Archivo
Histórico
Redimensionar
Res.cond.estand.Pearson
V.Exp
Valores predichos
Cuantiles muestrales
Cuantiles teóricos

Modelos lineales generales y mixtos

Especificación del modelo en R

```
modelo.027_V.Respuesta_REML<-gls(V.Respuesta~1+V.Exp
,method="REML"
,na.action=na.omit
,data=R.data24)
```

Resultados para el modelo: modelo.027_V.Respuesta_REML

Variable dependiente: V.Respuesta

Medidas de ajuste del modelo

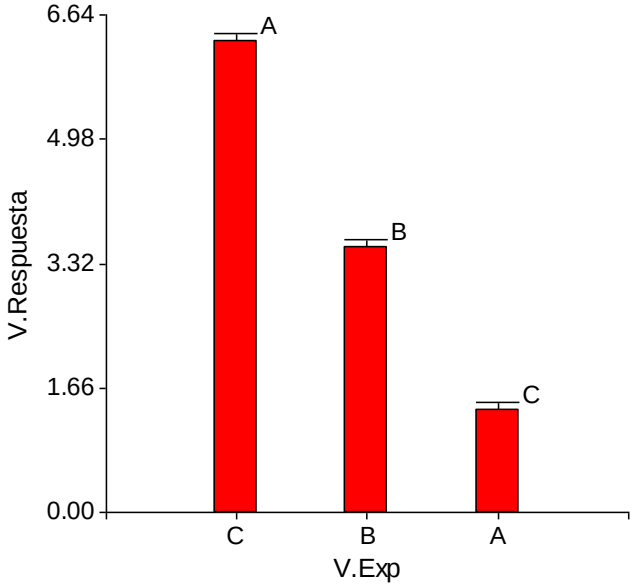
N	AIC	BIC	logLik	Sigma	R2_0
15	10.79	12.73	-1.40	0.22	0.99

AIC y BIC menores implica mejor

$\hat{\sigma}_{\text{error}}$

Pruebas de hipótesis marginales (SC tipo III)

	numDF	F-value	p-value
(Intercept)	1	4231.26	<0.0001
V.Exp	2	618.10	<0.0001



Efectos fijos

		Value	Std.Error	t-value	p-value
(Intercept)	$\hat{\beta}_0$	1.36	0.10	13.70	<0.0001
V.ExpB	$\hat{\beta}_1$	2.18	0.14	15.52	<0.0001
V.ExpC	$\hat{\beta}_2$	4.93	0.14	35.08	<0.0001

$$\hat{\beta}_0 = \hat{\mu}_A$$
$$\hat{\beta}_1 = \hat{\mu}_B - \hat{\mu}_A$$
$$\hat{\beta}_2 = \hat{\mu}_C - \hat{\mu}_A$$

V.Respuesta - Medias ajustadas y errores estándares para V.Exp

LSD Fisher (Alfa=0.05)

Procedimiento de corrección de p-valores: No

V.Exp	Medias	E.E.			
C	6.29	0.10	A		
B	3.54	0.10		B	
A	1.36	0.10			C

Medias con una letra común no son significativamente diferentes (p > 0.05)